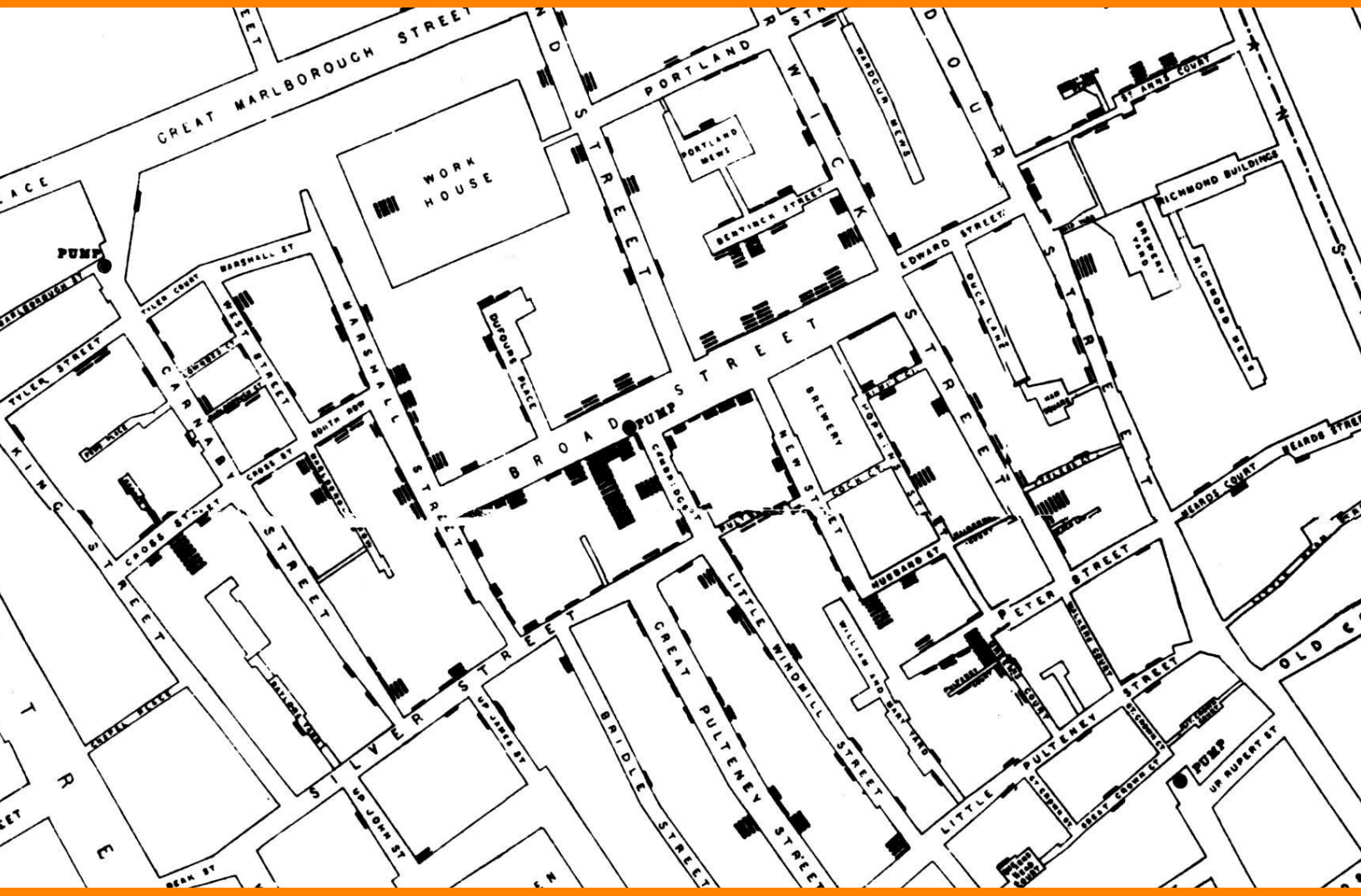
How Do We Know If a Program Made a Difference?

A Guide to Statistical Methods for Program Impact Evaluation



Peter M. Lance • David K. Guilkey • Aiko Hattori • Gustavo Angeles

How do we know if a program made a difference? A guide to statistical methods for program impact evaluation

Peter M. Lance

David K. Guilkey

Aiko Hattori

Gustavo Angeles

Suggested citation:
Lance, P., D. Guilkey, A. Hattori and G. Angeles. (2014). How do we know if a program made a difference? A guide to statistical methods for program impact evaluation. Chapel Hill, North Carolina: MEASURE Evaluation.

ISBN: 978-0-692-23861-5

Prepared by MEASURE Evaluation, University of North Carolina at Chapel Hill.

MEASURE Evaluation is funded through the U.S. Agency for International Development (USAID) under the terms of cooperative agreement GHA-A-00-08-00003-00, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, with Futures Group, ICF International, John Snow, Inc., Management Sciences for Health, and Tulane University. The views expressed in this publication do not necessarily reflect the views of USAID or the United States government. MS-14-87 (May 2014).

Asking for Forgiveness

The Universe is change. Life is opinion.
-Marcus Aurelius

There is a long-running, vigorous and evolving methodological debate about the appropriate or optimal way to evaluate the impact of a program. This manual strives not to convince readers of the merits of one particular alternative or another, but instead simply to present the various options in as approachable a fashion as possible and then let them decide for themselves where they stand. In other words, it strives to be impartial. However, that is probably an impossible standard. We cannot help but have our own views, which probably subtly intrude on the manual. To those who disagree with us we ask for their pardon but remind them that this shortcoming was probably inevitable.

Writing this manual has required review of a large number of contributions to the literature on impact evaluation methods, as well as many actual impact evaluations. Many of these are cited in this manual, which has now been through hundreds of revisions. In the course of those revisions, with the attendant re-arrangement, deletion and insertion, it is possible that some citations were accidentally dropped and others have become un-moored from the train of thought for which they were originally introduced. We offer unqualified apology for all such instances.

That said, updated versions of this manual will be released from time to time. We invite those with comments or criticism of the substance of the manual to submit their thoughts so that they can inform future versions. Errors and omissions in citations that are communicated to us will be corrected immediately.

Contents

Acknowledgements

No one ever truly does anything alone. The authors of this manuscript have benefitted from the support and insights of our colleagues and our families.

Wayne Hoover prepared many of the graphics in this manual. He took first draft graphics from the authors that probably looked like the efforts of a four year old and transformed them into something that looked polished and professional. This is no surprise: we have worked with Wayne for years and he is an unsung hero of many of our manuscripts.

We are grateful to Jim Thomas for having the vision to support this manual. An early draft stagnated for some years as the crises of the moment dominated our attention. Jim saw its potential and by his support created space for the enormous task of completing it. And he was very patient with the process: frankly speaking, we underestimated the challenge of closing out this manual and have (wildly) overrun the original time frame for completion that we provided him.

The manual has also benefitted from many conversations and exchanges (often not connected with the manual) over the years with colleagues, students and friends which gave us insights regarding how to present the ideas within in the most effective fashion. These experiences collectively built a kind of memory palace from which much of the presentation in the manual flows. A hardly complete list of those who so contributed includes Bates Buckner, Ilene Speizer, Chirayath Suchindran, Meghan Corroon, Siân Curtis, Heidi Reynolds, John Spencer, Lisa Calhoun, Tom Mroz, Kalee McFadden, Lauren Raymer-Heller, Dave Jones, Krishna Rao, Ruopeng An, Betty Tao, Albert Loh, William Kalsbeek, Mai Hubbard, Shankar Viswanathan, Philip Setel, Dan Gilligan, Tim Savage, Jason Dietrich, Sandy Darity, Kavita Singh, Minja Kim Choe, Ruth Bessinger, Paul Stupp, Stacey Gage, Jim Thomas, Tony Turner, Jin Shuigao, Cristina de la Torre, Krista Stewart, Will Dow, Jane Bertrand, John Akin, Abbas Bhuiya, Martha Skiles, Ronald Oertel, Hector Lamadrid-Figueroa, Quamrun Nahar, Karen Foreit, Erdal Tekin, Kanta Jamil, Phil Bardsley, Mike Foster, Syed Abbas, Elizabeth Ku, John Stewart, Emma Wang, Paul Hutchinson, Bernardo Hernández, Nahid Kemal, Mark Langworthy, Aimee Benson, Ahmed al-Sabir, Brad Schwartz, Scott Moreland, Chris Cronin, Edward Norton, Alex Cowell, Jessica Fehringer, Shams El Arifeen, Roland Sturm, Naoko Akashi-Ronquest, Peter Kim Streatfield, Alan de Brauw, Lewis Margolis, Jose Urquieta-Salomon, Ella Nkhoma, Janine Barden-O'Fallon, Sharon Weir, Lisa Mwaikambo, Ruchira Naved, Donna Gilleskie, Ken Hill, Dan Blanchette, Misha Lokshin, Nitai Chakraborty, Juan Carlos Negrette, Antonio Trujillo, Tim Frankenberger, Jipan Xie, Ozkan Zengin, Anne Swindale, William Sambisa and Oswaldo Urdapilleta.

Finally, we thank our families for putting up with us.

Prologue

Writing this manual has been a long journey, one that began in the Spring of 2005 while teaching a class for MEASURE Evaluation at the University of Hawaii at Manoa on the econometrics of program impact evaluation, with applications to health, nutrition and population. The students were international public health professionals drawn from around the world. Their backgrounds were not necessarily rooted in the statistics of causal modelling, which is really at the heart of the subject at hand.

They came to the class with rich and diverse technical backgrounds and a clear common need: to know more about the methods used to evaluate the impact of health, nutrition and population programs. They wanted, needed, to be better able to assess the evidence presented regarding the impact of health, nutrition and population programs in their own nations. They needed to be able to do this so that they could act meaningfully and effectively on such evidence, in the process crafting better programs in the hope of improving welfare in their societies.

Unfortunately, for them the body of literature describing and assessing the methods for program impact evaluation was largely inaccessible. This literature, concentrated in econometrics and statistics, featured dense, often inconsistent (across manuscripts and major players in the literature) mathematics typically linked with limited or no intuitive explanation.

The challenge for us as instructors was somehow to build a bridge for these students to the rich insights of the program impact evaluation literature. It is impossible to eschew mathematics altogether in explaining effectively program impact evaluation methodologies (more on this below) but we decided to adopt the simplest, most consistent mathematics (including at the mundane level of notation) possible. We tried to inform the math with careful intuitive explanations. Finally, we tried to strike a balance between developing nuanced understanding of particular topics (the trees, in some sense) and recognizing the relationships between the various program impact evaluation methodologies (the forest).

At the core of our approach was a simple belief: behind the mathematically dense discussion of the various methods for program impact evaluations was a far more approachable intuition. We tried to tease that intuition out from the math, and put it at the center of the discussion.

That said, you cannot have a discussion of program impact evaluation methods that avoids math while still delivering nuanced understanding. Above all, math allows one to be *precise* about what they mean in a fashion that words alone cannot. Math brings focus and specificity to the discussion in a fashion that words alone, no matter how carefully considered, cannot: there would always be some reasonable scope for misinterpretation with words alone. It was also necessary to use math because it empowers the students: a gentle, carefully guided introduction to the mathematics of program impact evaluation provides them with tools they need to continue their journey in the literature after the class has concluded.

This manual is designed to provide the sort of introduction to program impact evaluation that we attempted in that class, and for the same sort of audience. To be sure, this is a mathematically oriented manual. It *has to be* in order to convey certain subtle ideas precisely. At the same time,

math does not stand on its own in this manual. Rather, it is complemented by far richer intuitive discussion than is often the case in the literature.

Moreover, one element of the math is a relatively simple behavioral model through which nearly all estimators in the manual are explored. This will give the reader some sense of the kind of behavioral, and hence data generating, processes motivating these estimators and shaping the data structures they require.

Empirical examples run through much of the manual to provide illustrations of these behavioral models and impact evaluation estimators using simulated data that will hopefully strengthen the link for the reader between behavior, data structure, impact evaluation models and model performance. We chose to use simulated data for a number of reasons, but above all because simulated data allows us to establish explicitly links between behavior and observed data, to isolate the particular kind of statistical complications motivating each impact evaluation estimator and to examine model performance (which we can do because simulated data allows us to define true program impact, providing a clear yardstick against which to assess the various estimates generated by the program impact estimators).

A primary audience for this manual are those who motivated its writing in the first place: public health professionals at programs, government agencies and NGOs who are the consumers of the information generated by program impact evaluations. These are the professionals who often must commission impact evaluations. They serve as active stakeholders during the process of designing impact evaluations, providing key inputs about where the focus should lie given the advantages, disadvantages and costs of alternative impact evaluation approaches. Finally, they then must assess the information generated by the impact evaluation to decide how it will guide programmatic choices. They would be far better served if they could participate in these processes in an informed way: an informed consumer is an empowered consumer.

The audience that would find this manual useful should be much broader, however. To begin with, professionals serving the aforementioned role in any area of programming that influences human welfare could find this manual quite useful, and for the same reasons. It also provides a solid introduction to impact evaluation methods for anyone with an interest in the subject: graduate students, technical staff at evaluation projects, journalists looking for a more nuanced understanding of the steady stream of impact (and, more broadly, causal) studies on which they are asked to report, analysts at health analytics organizations and so on.

This manual is designed to be "stand alone". It is a more or less self-contained treatment of program impact evaluation methods. However, we feel that it is perhaps most powerful when offered as a text in the context of a class or training workshop designed to introduce program impact evaluation methods.

It is our hope that this manual opens the door for the reader to a truly powerful set of tools that allow for us to understand the implications of human welfare programs, in the process providing crucial information about what works and what does not in shaping health, education, fertility and a slew of other outcomes that ultimately shape the length and happiness of our lives. Program design informed by such tools has the power to change our world.

Chapter 1

Introduction

A dizzying array of programs seek to influence health, wealth, education, employment and other channels of human welfare.[1] An accurate understanding of what these programs actually achieve would allow society to focus scarce resources on those programs that most efficiently and effectively improve welfare. The aim of program impact evaluation is to learn whether and to what degree a program altered outcomes from what otherwise might have prevailed.

Measuring what might "otherwise have prevailed" is a challenging task. The phrase suggests an appeal to history's unrevealed alternatives. In the abstract, one might consider comparing the outcome of interest for an individual in circumstances under which they participate in a program or under which they do not.[2] Specifically, we might seek to measure differences in outcomes for an individual as their program participation, and only their program participation, varies. The use of the word *only* is important: if the only thing that varies is their participation in the program, then that must be the driving force behind differences in outcomes that we might observe when they participate and when they do not do so.[3]

Unfortunately, observation of an individual in two sets of circumstances that differ *only* in terms of their program participation is not intrinsically possible. Most obviously, an individual cannot be observed in two states at the same time. Varying individual program participation over time would violate the requirement that *only* program exposure varies: time (and hence any other factors or characteristics subject to evolution with time) would vary as well as program participation, making it questionable to ascribe wholly changes in outcomes of interest to variation in program participation alone.

What we have outlined in these paragraphs is what is often referred to as the "fundamental identification problem" of program impact evaluation. We would like to estimate program impact

[1] Examples include worker or employment training programs (designed to influence labor market experiences of participants), childhood development programs (which target physical, emotional or cognitive development), fertility control programs, health insurance delivery and health service provision programs (intended to influence health care utilization and, ultimately, health), legal or regulatory innovations to permit or stimulate institutional changes (for instance, charter school legislation to promote the development of alternative public educational institutions and, ultimately, to influence the quality of the learning experience provided by public education), alternative policing schemes (to reduce crime), agricultural extension programs, health care delivery schemes, etc.

[2] Throughout this manual, we rather freely interchange the terms program "participation" and program "exposure". Sometimes we refer to "treatment". Essentially, these mean the same thing. Instinctively, exposure or treatment probably imply less active engagement than participation. For instance, it seems more natural to say that one *participates* in a job training program but is *exposed* to a behavioral change communication program. In practice, however, the difference is mostly semantic and the statistical challenges are generally essentially the same.

[3] When we refer to *variation* in program participation, we typically mean the simple difference between participation and non-participation. However, at various points in this manual we will allude to the slightly different concept of a *dose-response* relationship, whereby program participation can vary in its intensity.

by measuring the difference in outcomes for a sample of individuals holding everything but their program participation constant. Unfortunately, this is not possible: we cannot observe an individual's outcomes while varying their participation status in a program and *only* their participation status in that program. Within the framework of varying exposure for the same individuals, we are therefore reduced to at most considering differences in outcomes over time for the same individual as their program participation varies.

A natural alternative would involve comparing (presumably at the same time) outcomes for *different* individuals who were participants and non-participants. However, under typical operational circumstances, program participation is determined in the messy laboratory of the human experience. It is generally not randomly determined, as would be the ideal under the laboratory and clinical traditions of experiments or randomized control trials, but instead the product of conscious decisions that balance the costs and benefits of participation. Since such costs and benefits likely vary with the characteristics and circumstances of individuals, we would expect different types of individuals to become participants and non-participants. However, this means that participants and non-participants would differ by more than just the experience of program participation, making it difficult to ascribe differences in their outcomes to the impact of the program.

The basic challenge is thus the same under this alternative: between those observed participating in a program and those who are not, there is always the possibility that more factors differ than just program participation. We thus cannot be certain whether program participation, and program participation alone, drove any observed differences in outcomes between participants and non-participants.

In recent decades various disciplines have offered a range of practical statistical tools or strategies for dealing with this essential challenge. This manual surveys these methods.

An important consideration in writing this manual has been to strike a balance between abstract presentation of methodological concepts and concrete application to actual data. A sound grasp of the conceptual issues is essential. It provides a meaningful framework for understanding the trade-offs between and limitations of alternative impact evaluation methods. However, the theoretical literature on the econometrics of program evaluation has been largely inaccessible to many of the applied researchers and analysts who might most benefit from its insights, since it relies on dense mathematics, often characterized by idiosyncratic notation and complemented by little intuitive explanation. In this manual we must rely on mathematics to present key ideas, but try to do so in the most straightforward, intuitive fashion possible. We always provide narrative discussion of the mathematics. Moreover, we attempt to apply a single consistent notational approach throughout.

Despite this desire for accessibility, an important and unavoidable complication in writing a manual such as this is the need for a certain degree of technical background required to understand impact evaluation methodologies in any sort of comprehensive, nuanced fashion. In short, while we attempt to make the mathematical statistics of this literature more accessible, doing so requires insuring that the reader has at their disposal some key tools from ... mathematical statistics. To do this, we have attempted to create relatively self-contained background briefings on these tools. The goal of these briefings is to focus on the most essential elements of a given tool from mathematical statistics for the purpose of providing the reader with the background information that they absolutely require for the discussion in the manual. These briefings strive to match the accessibility of the discussion in the main text.

Whenever the required discussion is fairly short these briefings are contained in Tool Boxes in the text. An example of a Tool Box is provided after this paragraph. These lend themselves best to tools that require only very short discussion, such as basic definitions or enumeration of a few properties that do not require further proof.

Tool Box: Tool Boxes

This is an example of a Tool Box. In these Tool Boxes, brief introductions are provided to key statistical concepts first invoked in the surrounding pages of the manual.

The purpose of the Tool Boxes and, where necessary, longer and more involved background explanations in the text is to make the manual as self-contained as possible. By doing this, we hope to avoid a hassle that we as researchers have frequently faced when reading manuscripts in the program evaluation literature: the need to constantly put down the manuscript to search for and read some other piece that more fully describes some concept, tool or technical device in order that we might fully understand what we are reading. We currently envision that updated versions of this manual will be periodically released as developments in the ever-changing impact evaluation literature warrant. We would be grateful for feedback regarding additional Tool Boxes or more lengthy background discussion that might improve the reader experience.

It is also important to move beyond concepts and provide a concrete sense of the implications of the abstract statistical phenomenon under discussion. Unfortunately, there is no single real world sample or data set that will allow us to demonstrate all of the data problems or estimation techniques discussed in this paper. This limitation illustrates one of the first considerations when deciding on an impact evaluation method: most come with very specific data requirements which will dictate to some degree the available options for the impact evaluation problem at hand.

At the same time, the very nature of the data problems that confound straightforward evaluation of human resource programs makes it difficult to illustrate them clearly with real world data for demonstration purposes. With real world data we often cannot rule out other inconvenient data features to focus on the particular data flaw motivating a given impact evaluation estimator: it is generally impossible to isolate with complete confidence that problem within a particular sample even if we have a high degree of confidence that it is somehow present.

Perhaps the most important problem with real world samples, however, is that we cannot ever know the true value of a parameter of interest (such as true program impact) for the population represented by a real world sample. This is a major consideration, since most of the techniques we consider in this manual are intended to generate estimates closer to such true population values than those offered by alternative estimation strategies. To know whether they have done so for expositional purposes in this manual, we must know what that true value is.

We will thus rely on two approaches to provide empirical illustrations of the various methods. First, wherever possible and efficient, we discuss the specific applications that applied researchers have used to explore particular estimators. This provides real world examples that can help to form a more concrete sense of how to approach the various constraints offered by particular data sets. These samples were often chosen by them *because* they met the data requirements of the method under consideration. By highlighting the features of these samples that made them so useful in those particular applications, we will shed light onto which data environments are particularly attractive for alternative impact evaluation methods.

However, we will also rely far more frequently on simulated data based on the idealized circumstances under which the various estimators are designed to operate. By "idealized" we mean a specific type of data failure or behavioral process, and a sample with the specific features that each estimator requires in order to be implemented. The simulations are described carefully in the text, and the full set of STATA .do files that implement them will be provided. The advantage of this approach is that, since we generate the sample, we know the true parameters (such as program impact) behind it.

When discussing STATA examples, we frequently place the relevant STATA results directly into the text as a "STATA Output". An example follows this paragraph. In the title to the example, the example name is given and, for the reader's convenience, the accompanying STATA .do file that produces the output is indicated. The STATA output itself is of a different font from the rest of the text, but matches that used in STATA itself.

STATA Output 1.1 (1.do)

```
. summarize
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          x1 |      10000   -.0005992    1.000659  -3.918931   3.641588
          x2 |      10000   -.0015025    .9946398  -3.379955   4.167702
          x3 |      10000   -.0071565    1.005172  -3.579602   3.915892
          x4 |      10000     .004938    .9915043  -3.794222   3.700645
          x5 |      10000    .0126444    1.010891  -3.536209   4.038878
-------------+---------------------------------------------------------
          x6 |      10000    .0058783    1.008324  -4.518918   3.598916
          x7 |      10000    .0200271    .9953145  -3.845707   4.448323
          x8 |      10000    .0186147     1.01171     -3.883   3.711688
          x9 |      10000       .00043    1.007457  -4.122494   3.578277
         x10 |      10000   -.0106295    1.000299  -3.659957   3.693912
.
. describe
Contains data
  obs:        10,000
 vars:            10
 size:       400,000
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
x1              float  %9.0g
x2              float  %9.0g
x3              float  %9.0g
x4              float  %9.0g
x5              float  %9.0g
x6              float  %9.0g
x7              float  %9.0g
x8              float  %9.0g
x9              float  %9.0g
x10             float  %9.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

As a final stylistic note, we confess to a certain degree of repetitiveness in this manual. We sometimes introduce and explain a concept repeatedly, albeit typically in slightly different contexts. While this might strike some readers as a touch tedious, we followed this course for two reasons. First, these are often important, foundational concepts. Thorough understanding of them, and how they manifest themselves in different settings, is crucial. Second, in our own intellectual lives, we have sometimes met with frustration reading complicated technical manuscripts on the occasions where a key concept, one that would repeatedly prove central to many of the discussions within the work, did not register clearly on its brief, isolated introduction.

Chapter 2

The Program Impact Evaluation Challenge

In this chapter, we introduce basic operational concepts in program impact evaluation. A thorough understanding of them is important since they provide the logical framework within which to understand and assess the various program impact estimators discussed in this manual. Though the concepts in this chapter may seem basic, many a program impact evaluation has foundered when the basic questions that these concepts speak to were not adequately addressed in the evaluation design: even the most sophisticated impact evaluation estimation methodology generally cannot compensate for an evaluation design that is not coherent in terms of the relatively basic respects discussed in this chapter.

2.1 Basic Concepts

An important step toward performing a program impact evaluation is to characterize the parameter one wishes to capture in order to decide what, if any, impact a program may have had. The definition of that parameter depends on what one wants to learn about the program. Typically, one wishes to know either the program's impact on a behavior the program seeks to influence or an ultimate human welfare outcome that it strives to shape.

At this stage, it is necessary to introduce some notation. Let Y represent an outcome of interest.[1] It is the behavior or outcome that we anticipate might be influenced by exposure to a program. It could be employment status, wages, income, health, health care utilization, performance on standardized tests, use of modern family planning, fertility, criminal behavior, farm output, or any of the myriad intermediate or ultimate human welfare outcomes that might be of interest.

Now suppose that we wish to consider the impact of some program on Y. Let Y^0 represent an individual's outcome if they do not participate in the program and Y^1 represent their outcome if they do participate in the program. In the simplest terms, for a given person, Y^0 is what happens when he or she doesn't participate in the program and Y^1 is what happens when he or she does participate in the program. The impact of a program on an individual is then the difference between what would have happened if they participated minus what would have happened if they did not participate, or $Y^1 - Y^0$.

Conceptually, we are supposing that each individual has two *potential* outcomes, one if they participate in the program and one if they do not do so. For example, in contemporary lower

[1] Here we abstract away from the terminology of program frameworks, which speak to concepts like outputs, outcomes, etc. Y is simply something the program may influence, and it is of interest what that influence is.

income societies policymakers employ a wide range of fertility control programs and it is common for household surveys to collect information regarding a woman's exposure to or participation in various fertility programs as well as, for instance, her use of family planning methods. At any given point in time, we observe her family planning behavior *under only one program participation regime*. We cannot observe her use of family planning while exposed to a program as well as in the absence of such exposure because at any given point in time she either is or is not exposed to the program. However, the woman would have experienced *some* level of fertility or utilization of family planning services either way. We can observe only one of her possible experiences.

The inability to observe both Y^0 and Y^1 for the same person at a point in time means that we cannot directly observe that person's program impact, $Y^1 - Y^0$. This is a fundamental obstacle in determining the impact of a program for each individual. Indeed, it is often referred to as the **fundamental identification problem of program impact evaluation**.

Let P be an indicator variable that equals 1 if an individual is exposed to or participates in a program and 0 otherwise. This is *observed* program participation. It tells us whether the person actually participated in the program or not. For instance, continuing the earlier example, at a given point in time we may *observe* some women to be exposed to or participating in a fertility control program (P =1) while others are not (P=0).

For each individual we can define their observed outcome Y as

$$Y = (1 - P) * Y^0 + P * Y^1$$

This simply says that the outcome we actually observe for each person depends on whether they actually participated in the program. If they did participate, we observe their outcome as a participant in the program (Y^1). If not, we observe their outcome in the absence of participation (Y^0).

This potential outcomes framework provides at the individual level our first analytical basis for approaching a basic concept in program impact evaluation: the **counterfactual**. In general, for the purpose of understanding the effect of any circumstance on an outcome, the counterfactual is what would have occurred in the *absence* of that circumstance. In the present context, we have defined potential outcomes Y^1 and Y^0 that tell us the outcome of interest that an individual would have experienced if, respectively, they participated in the program ($P = 1$) and if they did not do so ($P = 0$). Y^0 is the *counterfactual* to Y^1: it is the outcome that would have happened without program participation.

The goal of program impact evaluation is to answer the following question: did a program make a difference? It seems obvious that to do so we require at the very least information about outcomes for program participants. However, without some alternative against which to judge their outcomes, how are we to assess whether it made a difference? How can we know whether the outcomes that we observe among participants reflect changes brought on by the program or whether they would have occurred anyway? To answer these questions, we must observe the alternative state of the world, or the counterfactual.

Whenever one is designing an impact evaluation, it is quite useful to put serious thought into the meaningful counterfactual to whatever one wishes to assess in that particular application. Difficulty in defining the counterfactual should be taken as a sign that perhaps one has not yet developed a sufficiently focused sense of what it is that one wishes to assess: it is difficult to understand how an impact evaluation can provide much useful information if it does not begin with a well posed counterfactual question.

We now begin to consider some of the various types of program impact that might be of interest. Following conventional terminology, we liken exposure to a program to "treatment". First, we have

the **average treatment effect**

$$E\left(Y^1 - Y^0\right)$$

where $E(\cdot)$ is the expectations operator (see the Tool Box on the Expectations Operator). The average treatment effect is the average impact of the program across all of the individuals in the population of interest. It is defined, quite sensibly, as the average across the population of the program impact of each individual in that population.

Tool Box: The Expectations Operator

The expectations "operator" $E(\cdot)$ provides the expectation of a random variable. Suppose, for instance, that X is a discrete variable that takes on three values (x_1, x_2 and x_3) with a probability of each occurring. Its expected value is then

$$E(X) = x_1 \cdot Pr(X = x_1) + x_2 \cdot Pr(X = x_2) + x_3 \cdot Pr(X = x_3)$$

where $Pr(X = x_q)$ is the probability that X takes on the value x_q. Notice that this is of the basic form

$$E(X) = \sum_{q=1}^{Q} x_q \cdot f(x_q)$$

where $\sum_q^Q$ indicates summation over the various values that X takes on (x_1, x_2, ..., x_Q) and $f(x_q)$ is the probability of outcome x_q (i.e. $Pr(X = x_q) = f(x_q)$). If X was continuous, the analog to such summation is *integration*. For a continuous random variable with probability density function $f(x)$, the expectation is thus

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Thus, the expectation is basically the weighted average of a random variable, where the weights are the probabilities of the various values that that variable can assume.
Finally, expectations can depend on another variable through the conditional expectations operator $E(\cdot|\cdot)$. For instance, $E(Y|X)$ is the expectation of Y conditional on X (e.g. expected income Y given education X). The conditioning can be on more than one variable (e.g. $E(Y|X, Z)$, which, to fix ideas, could be the expectation of income Y conditional on education X and age Z and based on various restrictions on the conditioning variable (e.g. $E(Y|X > c)$ provides the expectation of Y conditional on X being greater than c, as in for instance, the expectation of income given that education is greater than 12 years).

Another common parameter of interest is the **average effect of treatment on the treated**:

$$E\left(Y^1 - Y^0|P = 1\right)$$

This is the impact of the program on those actually exposed to the program.

At first glance, it might seem to some that these two treatment parameters are effectively describing the same thing. Alternatively, others might wonder why the latter effect would be of interest. After all, don't we care primarily about how a program effects the *average* or *typical* person? However, the two are not necessarily the same and the latter may be more interesting in many instances. At the most basic level, the average treatment effect applies across the population, while the average effect of treatment on the treated captures impact for those who actually

participated in a program. As long as there are some non-participants and participation is not randomly determined, participants and nonparticipants are likely not populations identical in their characteristics.

Further, it is often not clear why the impact for the entire population would be a more useful parameter than that for those who experience an intervention. For example, a job training program focusing on basic work skills is likely to have little effect on the labor market outcomes of highly educated or skilled professionals with a great deal of work experience, who are unlikely to enroll in it in any case. On the other hand, its impact on those with lower skill levels and less labor market experience, who are other things being equal more likely to enroll in it than the aforementioned group, might be dramatic. In this case, the impact of the program on those who actually enroll might differ substantially from the impact on the average individual and is likely of greater interest.

Consider an income supplementation program designed to enhance the nutritional status of children. Eligibility for such programs often depends on some sort of means test (the program would most likely be targeted at the poor). It seems straightforward that, first, the impact of the program on the average child is likely to fall short of that for the children of poor families eligible for the program (because the latter is likely to be more nutritionally vulnerable, allowing greater scope for the program to have a discernible impact) and, second, that the impact among participants (who are drawn from the targeted group) is of greater interest for the purposes of policymaking than the impact on the randomly drawn child.[2]

Moreover, individuals often base their program enrollment decision on the anticipated gain that they will experience if they are exposed to the program. It is therefore entirely likely that in a setting of rational decision making those enrolled would be the type of individuals who anticipate larger gains from the program. The average program effect, taken as it is with respect to the randomly drawn person in society, might severely underestimate the gains experienced by actual participants.

2.2 The Estimation Challenge: Basic Ideas

In the program impact evaluation literature, it is not uncommon to see exposure to a program referred to as "treatment". The word treatment, with its connotation of a clinical intervention, suggests a convenient intellectual framework for examining the challenge of estimating the impact of programs with non-experimental data drawn from the messy laboratory of the human experience: evaluations of medical interventions often rely on data drawn from a setting that resembles the experimental ideal of randomization between treatment and control subjects. Randomized biomedical trials provide a simple platform for thinking about an ideal empirical circumstance for conducting impact evaluations, as well as consequences of the failures of randomization that frustrate simple impact evaluations.

When, for instance, the effect of a drug on the longevity of rats is evaluated in a laboratory setting, standard practice dictates establishing an experimental/treatment group of rats (which receives the drug) and a control group (which doesn't). Assignment of rats between the two groups is typically completely random.

Randomization insures that the rats are alike on average, so that idiosyncratic baseline differences between particular rats are statistically canceled out when comparing *average* outcomes between the two groups. Their statistical "sameness" is then reinforced by attempts to insure that their experiences in the course of the experiment vary only in terms of exposure to the drug. There

[2]Wooldridge (2001) also discusses the policy relevance of the two concepts. Heckman (1997) actually focuses on an example much along the lines of the job training example provided in the text.

are thus no differences between the groups in terms of food levels, veterinary attention, etc. during the experiment. To assess the average impact of the drug on longevity, it is then necessary only to compare mean (i.e. average) longevity in the control and treatment samples.

In this example, it is essential not only to have a sample of rats exposed to the program, but also a sample not exposed to it. The latter allows us to have some alternative state of the world against which to judge outcomes among those who received the drug. The "no drug" group provides information about the **counterfactual**.

We can think of this experimental setup in the terms introduced earlier. For every rat, there are two potential outcomes: its longevity with exposure to the drug (Y^1) and without exposure (Y^0). In the randomized setting, we still observe only one of the two outcomes for each rat (because, for a given experimental interval, rats either receive the drug or they don't). Nonetheless, randomization opens the door to powerful empirical possibilities: because of it we can be sure that the two groups of rats are alike *on average*. It is thus not necessary to observe both potential outcomes for each rat for the purpose of estimating the average treatment effect.[3]

Put a bit differently, because of randomization Y^0 and Y^1 are not related to actual treatment status: on average Y^0 and Y^1 are no different for the rats assigned to the treatment and control groups. Because of randomization, there are no differences between rats assigned to the two groups, and hence we would expect that there would be no average differences in the distribution of outcomes Y^0 and Y^1 between them.

Statistically, randomization insures that Y^0 and Y^1 are **independent** of actual assignment to the treatment or control groups, P. Another way of thinking about this is that assignment to the treatment or control group (i.e. the value of P) reveals no information about a rat's value of Y^0 or Y^1. We do not expect the rats in either group to be, on average, bigger, stronger, younger, smarter or fatter than those in the other group. Thus, we do not expect their outcomes on average in the face of treatment or the absence of it to be any different.

A related, but somewhat weaker condition is **mean independence**: $E(Y^0|P) = E(Y^0)$ and $E(Y^1|P) = E(Y^1)$. This means that the expected values of Y^0 and Y^1 should be the same whether the rat was observed to receive treatment or not. Independence always implies mean independence.[4]

Let's think about this in terms of attempting to estimate, for instance, the effect of treatment on the treated:

$$E(Y^1 - Y^0|P = 1)$$

By the properties of the expectations operator (see the Tool Box on Properties of the Expectations Operator)

$$E(Y^1 - Y^0|P = 1) = E(Y^1|P = 1) - E(Y^0|P = 1)$$

The challenge from the standpoint of recovering the impact of the program on the treated is to somehow estimate $E(Y^1|P = 1)$ and $E(Y^0|P = 1)$.

$E(Y^1|P = 1)$ can be estimated simply by averaging outcomes for the rats exposed to treatment (we are, after all, interested in the impact of treatment *on the treated*). But how does one recover the expected outcome for those actually receiving treatment *had they not received it*? Because

[3]There is actually one other condition that must be met: their experiences in the course of the treatment must differ only in terms of their exposure to the drug.

[4]It might seem intuitive that the opposite is also true (i.e. that mean independence implies independence). However, it is not. The distribution of Y^0 and Y^1 is about more than just the means of Y^0 and Y^1. For example, the distribution of Y^0 and Y^1 also involves measures of dispersion, such as variance. Mean independence simply guarantees that on average the values of Y^0 and Y^1 will be the same across treatment and control groups. It does not rule out other possibilities, such as the variance of Y^0 and Y^1 differing between treatment and control groups. Independence rules out *any* kind of statistical relationship between $\{Y^0, Y^1\}$ and P. Hence, it is the stronger condition.

of randomization, we know that Y^0 and Y^1 are independent of actual assignment to treatment (P). Thus, $E(Y^0|P = 1) = E(Y^0)$ and $E(Y^0|P = 0) = E(Y^0)$ and, therefore, $E(Y^0|P = 1) = E(Y^0|P = 0)$: we can use the average outcomes for rats that did not receive treatment as an estimate of the outcomes that those that did receive treatment would have experienced had they not received treatment (that is, the counterfactual). One important thing to note about this math is that although we established the independence of Y^0 and P at the outset, in practice we relied on the weaker condition of mean independence.

Tool Box: Properties of the Expectations Operator

The expectations operator has some important properties. First, if c is some constant (i.e. not a random variable, but a fixed number) then

$$E(c) = c$$

This leads to the next property, namely that

$$E(c \cdot X) = c \cdot E(X)$$

Finally, we have the additive property:

$$E(c_1 \cdot X_1 + c_2 \cdot X_2) = c_1 \cdot E(X_1) + c_2 \cdot E(X_2)$$

where the cs are constants. If $c_1 = c_2 = 1$, this reduces to

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

There are, however, limits to this flexibility. First,

$$E(X_1 \cdot X_2) \neq E(X_1) \cdot E(X_2)$$

unless X_1 and X_2 are independent. Furthermore,

$$E\left(\frac{X_1}{X_2}\right) \neq \frac{E(X_1)}{E(X_2)}$$

and

$$E[f(X)] \neq f(E[X])$$

A clear exception to the last inequality would be the case where $f(\cdot)$ is linear (e.g. $f(X) = a + b \cdot X$, where a and b are constants).

Let us now tweak this experimental design in a fashion that undermines randomization between the two groups. Suppose that instead of randomly assigning rats to the treatment and control groups, assignment is determined by a competition. For example, the researchers might place all of the rats in one of two adjacent boxes. The other box is partially filled with cheese. Researchers then allow rats to climb from one box to the other until they have achieved the desired numerical division of the available supply of rats between treatment and control groups.

It should be clear that the two groups of rats are no longer the same in terms of their average baseline characteristics. The rats that landed in the cheese box were probably quicker to the chase

because they smelled the cheese first and were better climbers. In short, they were keener, stronger and, most likely, younger. We now cannot be sure whether any differences between the two groups in terms of longevity are due to exposure to the drug or differences in the baseline characteristics as a result of sorting between the treatment and control groups.

In essence, treatment status P is serving two roles from an empirical standpoint: to signal the biomedical impact of the drug on longevity and to proxy for the differences in baseline characteristics between the two groups of rats. The empirical value of drug exposure as a signal of the drug's impact has been compromised by **self-selection** of the rats into the two groups.

Let us consider this in the terms presented above, supposing again that the goal is to estimate the average impact of treatment on the treated. Without loss of generality and for concreteness, assume that the rats in the box with cheese are the treatment group (the essential logic holds even if the reverse is true).

The goal is to provide estimates of $E(Y^1|P=1)$ and $E(Y^0|P=1)$. Once again, $E(Y^1|P=1)$ can be estimated by averaging outcomes (longevity) among those receiving treatment. However, recovering an estimate of $E(Y^0|P=1)$ is no longer straightforward. Since assignment to the treatment group is now non-random, we can no longer be sure that $E(Y^0|P=1) = E(Y^0)$ or, by extension, that $E(Y^0|P=1) = E(Y^0|P=0)$: outcomes are no longer independent, or even mean independent, of treatment.

In particular, we suspect that the rats in the treatment box are younger and stronger. These rats likely would have experienced greater longevity than those in the control group *even without the drug*. Mathematically, in this example $E(Y^0|P=1) > E(Y^0|P=0)$: the stronger, younger treatment rats would have lived longer even without treatment. This means, however, that the average outcomes for the control group can no longer serve as a valid estimate of the outcomes for the treatment group in the absence of treatment. Thus, we see the power of randomization by the consequences of its absence: one small complication has completely undermined our ability to assess the effect of treatment on the treated by straightforward means.

Although in this example the specific problem was $E(Y^0|P=1) > E(Y^0|P=0)$, in general the problem is simply that $E(Y^0|P=1) \neq E(Y^0|P=0)$. We had originally sought to estimate the average effect or impact of treatment on the treated

$$E\left(Y^1|P=1\right) - E\left(Y^0|P=1\right)$$

by taking the difference of two things we could observe (or at least estimate):

$$E\left(Y^1|P=1\right) - E\left(Y^0|P=0\right)$$

The problem is that now rats have selected themselves into boxes and, as a result,

$$E\left(Y^0|P=1\right) \neq E\left(Y^0|P=0\right)$$

and thus

$$\left\{E\left(Y^1|P=1\right) - E\left(Y^0|P=1\right)\right\} \neq \left\{E\left(Y^1|P=1\right) - E\left(Y^0|P=0\right)\right\}$$

The difference

$$\left\{E\left(Y^1|P=1\right) - E\left(Y^0|P=1\right)\right\} - \left\{E\left(Y^1|P=1\right) - E\left(Y^0|P=0\right)\right\}$$

$$= E\left(Y^0|P=0\right) - E\left(Y^0|P=1\right)$$

is called **self-selection bias** (or, in many works, just **selection bias**) because it is the bias generated by the self-selection of rats into boxes (in other words, by the breakdown of random assignment to the boxes).

This represents a huge statistical setback. Straightforward methods, such as simple comparison of estimated mean outcomes between participant and non-participant groups, are no longer effective for recovering program impact. Program effects can now be recovered only with more elaborate estimators, each of which carries a huge price tag in terms of the credibility of the inferences that can be made with them: additional assumptions. As the paper proceeds, the reader will gain a clearer understanding of exactly what is meant by this, but for present purposes it is important to understand that randomization had permitted the recovery of program effects with simple estimators with comparatively few strings (in the form of potentially unpalatable assumptions) attached.[5]

However, the problems in terms of straightforward evaluation of the drug's impact on longevity actually extend beyond these differences in baseline characteristics. One group of rats ended up in a box filled with cheese. All other things being equal, they were thus exposed to an additional sort of treatment: more cheese. And, for a variety of reasons, we might expect that exposure to cheese could influence longevity. The extra cheese will render them better fed, but also possibly overweight. Thus, in terms of the circumstances to which they were exposed, the two groups really differ along two lines that might influence longevity, rather than one.

To see this point more clearly, let us abstract away from the problem of self-selection and suppose instead that one box had been filled with cheese but that the researchers still randomly assigned the rats between the two boxes (rather than let them sort themselves). Then, the rats would be the same on average in terms of baseline characteristics, but their longevity would still differ simply because some had extra cheese by dint of what was in their box. Once again, exposure to the drug is serving two roles empirically: to signal the biomedical impact of the drug on longevity and to proxy for exposure (or non-exposure, as the case may be) to extra cheese. We refer to this as the problem of **confounding treatment**: a situation where assignment to the treatment under consideration is possibly related in some way to assignment to another treatment that also impacts the outcome of interest.

Turning once again to the challenge of estimating the impact of treatment on the treated and supposing again that the rats in the cheese box form the experimental/treatment group, we could once again estimate $E(Y^1|P = 1)$ by averaging longevity across those rats in the treatment group. However, because of the additional cheese, we have a problem: $E(Y^1|P = 1)$ now reflects the impact both of the drug treatment (which we hope to pick up) and of the cheese (which we don't). We thus can't use the outcomes for the rats actually exposed to the treatment to estimate the average outcome under treatment: because of the exposure to the cheese their experiences no longer statistically reflect just treatment. Note, however, that, because rats were randomized across groups and because under this scenario the rats not receiving treatment received no cheese in their box at the outset, we can still use $E(Y^0|P = 0)$ to estimate what the treatment group would have experienced in the absence of treatment.

One might wonder whether we could return to the salad days of simple estimation with randomized assignment to treatment simply by giving the rats in the control group the same amount of cheese. The answer depends on how one plans to interpret the results, highlighting a concern in some sense intrinsic to all impact evaluations. Since the impact of the drug may in some respect interact with cheese, it might be risky to generalize one's results after giving the control rats a compensating dose of cheese.

[5]This is not to say that randomized control trials do not involve complications in the real world, as we will see in the next chapter.

Cheese may dampen or amplify the drug's effect on longevity. For instance, it could be that cheese and the drug have little impact on longevity when administered separately, but together that may have a big impact. It thus might be inappropriate to generalize the evaluation results to more general circumstances where rats might not receive cheese with treatment.

This in some sense was an implicit problem even before the example of extra cheese was introduced: how certain could we as researchers be that the baseline circumstances and experiences in the course of the trial of *either* group of rats reflected the more typical circumstances under which the drug, if used more widely, might be expected to operate? More generally, program impact evaluations often rely on samples drawn from a distinct setting (institutional, programmatic, cultural, religious, economic, political, etc.). Thus, one must exercise great care when attempting to use results from that sample for the purposes of predicting program impacts in a very different setting. In this sense it is useful to make a distinction between **internal validity** and **external validity**. Estimates are internally valid if they correctly recover program impact within the sample at hand. They are externally valid if they provide a meaningful indication of the impact of the program under circumstances other than those from which the sample at hand emerged.

Before moving on, it might be useful to consider a graphical representation of the three situations that we have just discussed. As a preliminary to this, we offer a definition of the word *causal*, which will be bandied about in the discussion to follow. When we refer to a causal relationship between two variables, we mean a cause and effect relationship: variation in one of these variables *causes* variation in the other.[6] Figure 2.1 illustrates our graphical approach to modelling this: the solid arrow means that Z *causes* W (meaning, more precisely, that variation in Z causes variation in W). This is as opposed to mere correlation (represented by the dotted line in Figure 2.4), a situation where two variables are statistically correlated but variation in either of them does not actually cause variation in the other. It could be, for instance, that the correlation between the two is an artifact of the influence of a third variable, variation in which *does* cause them to vary.

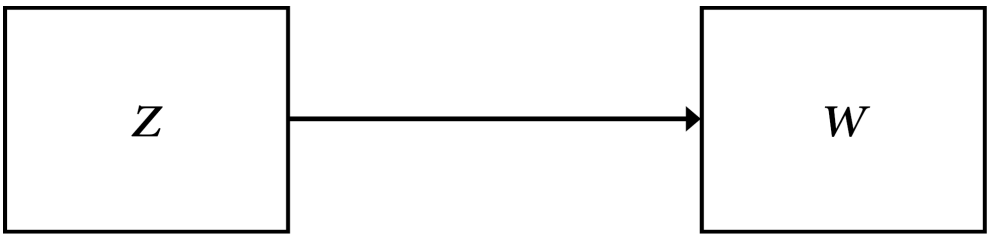


Figure 2.1: Z "causes" W

Let us begin with straightforward randomization, illustrated in figure 2.2. P and Y are defined in the same fashion: program exposure (in this case whether or not the rats received the drug) and the outcome of interest (their longevity), respectively. X are characteristics of the rats (such as their age, weight, etc.). Solid arrows indicate causal relationships. The figure shows that X and P both act to influence Y. Notice that there is no interrelationship between X and P: they are two separate boxes, and the only thing they have in common is that they both operate (independently) to affect Y. The lack of a relationship between X and P is a result of the randomization design, which insures that the two are independent. This allows us to consider the two pathways separately: we do not need to worry about the variation in characteristics X across rats when evaluating the effect of the program exposure P on outcome Y. The reason is that the variation in P has nothing to do with the variation in X, and hence the variation in Y from the two different sources can be assessed separately.

[6] We will explore alternative notions of causality later in this chapter.

Figure 2.3 illustrates self-selection. The individual characteristics of the rats X now *cause* exposure to the program P. From a purely statistical standpoint, we can no longer consider the effect of P on Y in isolation. Because of self-selection, certain types of rats (defined by different values of X) are more likely to be exposed to the program. For instance, if X is simply a measure of strength, stronger rats are more likely to be exposed to the drug and less strong rats are less likely to be exposed to the drug because stronger rats were more likely to reach first the cheese box that defines the treatment group. Thus, the average value of X will be different when $P = 0$ and $P = 1$. The point is that variation in P now goes hand in hand, statistically, with variation in X. If we consider the link between P and Y in isolation, it now serves to capture the link between P and Y *and* between X and Y (because the value of X now systematically varies according to the value of P). Exposure to the drug captures the effect we are interested in, that of the drug on longevity, but also serves as a proxy for X.

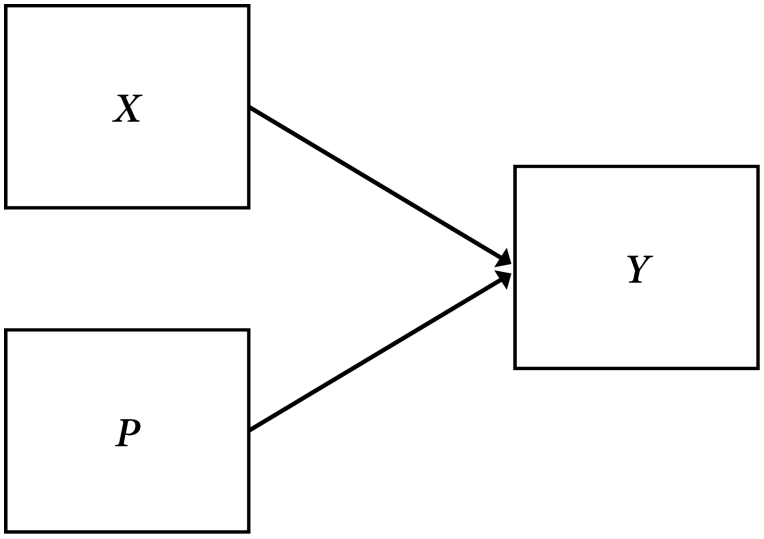


Figure 2.2: Randomization

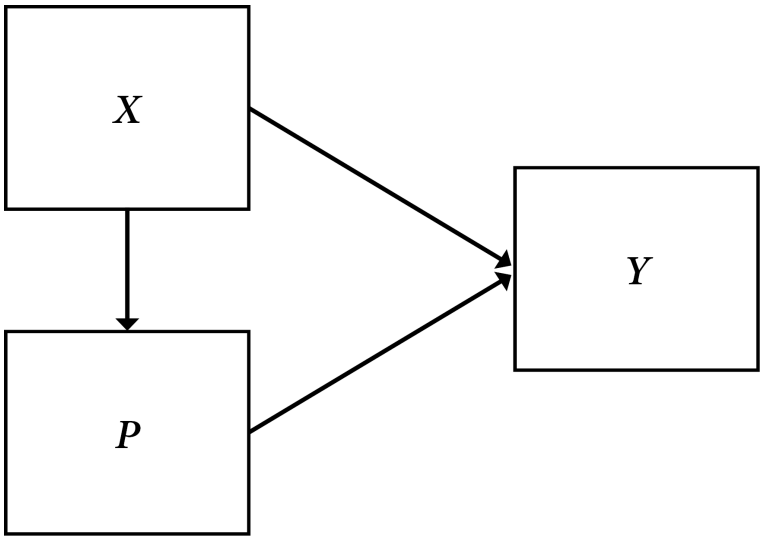


Figure 2.3: Self-selection

In our discussion of self-selection, we ignored the problem of confounding *treatment* created by the imbalance of cheese between the treatment and control groups. We now turn to that problem in Figure 2.4. The dashed line indicates a statistical relationship that is not causal in the strict behavioral sense: the fact that two variables are related does not mean that variation in one *caused* variation in the other.

Recall that we assumed for the purposes of discussing confounding treatment that rats were still randomly sorted between treatment and control groups, but that the treatment group received extra cheese along with the drug. In the figure, the extra cheese is represented by C. Both P and C are now once again independent of X because we re-established randomization of rats between the two groups. The difficulty in this case lies with the fact that the rats in the treatment group effectively received two treatments: the drug *and* the cheese. In general, we can only recover their combined effect by comparing the average values of Y when $P = 0$ and when $P = 1$. The effect that we recover by doing so reflects exposure not only to the drug (P) but also the cheese (C).

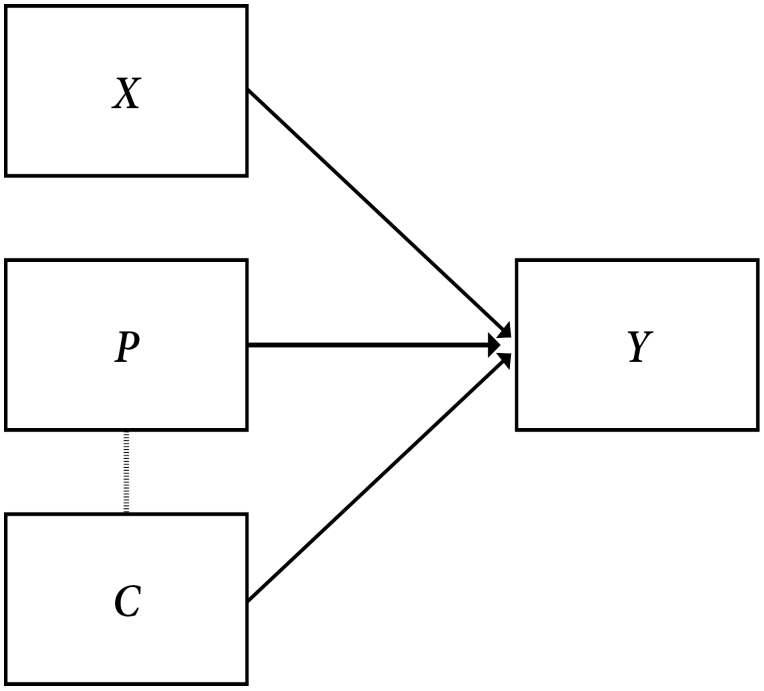


Figure 2.4: Confounding Treatment

These examples allow us easily to conceptualize, from the standpoint of the ideal random experiment, the two most common empirical challenges when evaluating social programs with non-experimental data drawn from the world at large. Participants often self-select (or are selected) into the program in a non-random fashion, and participation in the program may be associated with a series of other changes that might disguise the true impact of the program.

While we have focused to this point on rats in laboratories, the available data for evaluating human welfare outcomes, which usually provides information about program assignment and outcomes in the real world (which is, of course, not an experimental laboratory setting), is essentially contaminated by the same failures of randomization, preventing researchers from drawing conclusions about the impact of programs in a straightforward fashion (for instance, by simply comparing mean outcomes for treated and untreated individuals).

Consider, for example, job training programs. Researchers often want to identify the impact of these programs on income or wages. In practice, the information (that is, data) used to evaluate

the impact of the program on wages is gathered by observing the wages of a group of people who actually participated in the program and then comparing them with those of a group that did not. However, those who participated in job training programs often differ from those who did not by more than just exposure to the program. First, there are differences in baseline characteristics between participants and non-participants due to self-selection. For instance, it may be that, other things being equal, more motivated individuals enroll in job training programs. Moreover, these programs also often have qualification criteria that could in and of themselves be important in terms of future wages. Second, whether it is by the explicit desire of policymakers to link job training to other interventions or the fact that the qualification mechanisms just described are often similar for other programs as well, those enrolled in job training programs are often participating in other programs that might influence future wages through a variety of pathways. Alternatively, non-participants may in some cases be non-participants because they decided instead to enroll in some other, similar program. It could thus be impossible to evaluate the impact of the program on wages simply by comparing the wages of participants and non-participants: their wage differences might be capturing the impact of the program, other programs, or baseline differences in individual characteristics. Evaluation based on the simple difference in wages between participants and non-participants assumes that there are no important self-selection mechanisms or other confounding programs.

2.3 The Estimation Challenges: Some Common Estimators

Having described some basic program impact identification challenges in very general terms, we now focus on selection, which is the main concern in this manual.[7] We also become a bit more technical and begin to describe the estimation challenge in terms of the failure of two basic estimators: simple comparison of mean (or average) outcomes across control and treatment groups, and regression analysis designed to control for observed differences in observed characteristics across those exposed and not-exposed to a program.

2.3.1 Simple Comparison of Mean Outcomes Y Between Participants and Non-Participants

Let us consider estimation of program impact via comparison of mean outcomes between participants and non-participants. For concreteness, we focus on estimation of the average effect of treatment on the treated (similar conclusions can be reached for estimating the average treatment effect). Our discussion follows the reasoning of Heckman and his colleagues (Heckman et al. (1996), Heckman et al. (1997a, b), Smith and Todd (2001)).

Conditioning on ***observed*** characteristics $\vec{X}$ (age, sex, education, etc.), the average effect of treatment on the treated is given by

$$\Delta X = E(Y^1|\vec{X}, P = 1) - E(Y^0|\vec{X}, P = 1)$$

This setup allows the average effect of treatment on the treated to vary according to an individual's characteristics $\vec{X}$ (see the Tool Box on Vectors).

[7] We consider issues such as external validity mainly as a background interpretive issue for the purpose of understanding the implications of program impact estimates.

Tool Box: Vectors

Throughout this manual, we will sometimes present what appears to be a variable with an arrow running over it, as in $\vec{X}$. Such notation refers not to one variable, but to a set of them. $\vec{x}$ refers to a set of k variables:

$$\vec{x} = [x_1, x_2, ...x_k]$$

For instance, if $k = 3$, it could be that the three variables are age (captured in x_1), education (captured in x_2) and income (captured in x_3).

Estimation of $E(Y^1|P = 1)$, the expected outcome under participation across all participants, is fairly straightforward. For instance, one could take the average of Y^1 among those who actually participated in the program (i.e. $P = 1$) because it would be observed for them. Estimating $E(Y^1|\vec{X}, P = 1)$ would involve stratifying along the lines of the various values of $\vec{X}$ (i.e. stratifying the averaging of Y^1 by the various types of individuals found among participants), a trivial extension. We could thus, for instance, easily form estimates of $E(Y^1|P = 1)$ for the rich and the poor as long as we observe income as a part of $\vec{X}$: we would simply average Y^1 among rich and poor participants separately.

The challenge lies with forming an estimate of $E(Y^0|P = 1)$. We do not observe what participants would have experienced in the absence of participation. Therefore, we must form an estimate of $E(Y^0|P = 1)$ by less direct means than were applied in the case of estimation of $E(Y^1|P = 1)$.

One possibility would be simply to use the outcomes that non-participants experienced as an indication of what participants would have experienced in the absence of exposure to the program. In this context, Heckman and his colleagues define selection bias conditional on characteristics $\vec{X}$ as

$$S(\vec{X}) = E(Y^0|\vec{X}, P = 1) - E(Y^0|\vec{X}, P = 0)$$

In other words, selection bias is the difference between the outcomes participants with characteristics $\vec{X}$ experience in the absence of program participation and the outcomes non-participants with characteristics $\vec{X}$ experience in the absence of program participation.

Such selection bias reflects differences between participants and non-participants beyond just differences in their observed characteristics $\vec{X}$. For instance, some unobserved types (the more motivated, those with lower baseline unobserved health endowments, etc.) might appear more often among participants than non-participants. We focus on unobserved characteristics because the selection term $S(\vec{X})$ already involves explicit conditioning on observables $\vec{X}$. Thus, any selection bias must be driven by factors beyond $\vec{X}$, namely unobserved characteristics of participants and non-participants.

Notice that the term

$$S(\vec{X}) = E(Y^0|\vec{X}, P = 1) - E(Y^0|\vec{X}, P = 0)$$

is defined only for values of $\vec{X}$ occurring in both the participant and non-participant groups. However, the various values of $\vec{X}$ are defining observed "types" of participants and non-participants. For concreteness, if $\vec{X}$ includes just age and income, older poor people might represent one type. Thus, we could more intuitively say that $S(\vec{X})$ is defined only for types of individuals found in both the participant and non-participant groups. For instance if no older rich people participate in the program we cannot determine the selection bias for them. As you will see in a later discussion of

matching and propensity score estimation, this is sometimes referred to as a "failure of common support".[8]

One can isolate the selection bias not conditional on $\vec{X}$ (technically, we can "integrate out" $\vec{X}$ from $S(\vec{X})$), resulting in an expression of sample selection bias that does not depend on observed characteristics:

$$S = E(Y^0|P = 1) - E(Y^0|P = 0)$$

Think of this as overall selection bias across all types $\vec{X}$. Heckman and his colleagues show that S can be parsed into three terms which we (and generally they) refer to as S_1, S_2, and S_3:

$$S = S_1 + S_2 + S_3$$

There are thus three distinct factors driving selection bias.

The first term S_1 arises due to values of observed types of individuals in $\vec{X}$ that occur only among participants or non-participants, but not both. For instance, it could be that there were no high income people among participants (where we assume that we can observe income). Thus, one reason that selection bias can arise is that there are certain observed types of individuals occurring, for example, among participants for whom we can find no counterpart among non-participants. This is the aforementioned "failure of common support".

The second term S_2 reflects the possibility that some observed (i.e. characteristics captured *in* $\vec{X}$) types of individuals are found more frequently in one group than the other. Suppose, for instance, that we wish to evaluate the impact of a health program. Suppose as well that income level affects the health outcome the program targets and that the likelihood of program participation decreases with income. There would likely be more poor individuals in the participant group. The greater concentration of poor people among participants is one reason that the mean of Y^0 might differ between the two groups.

Finally, S_3 captures the possibility that there could be differences between the two groups in terms of their unobserved characteristics (i.e. characteristics not captured in $\vec{X}$). There might be some unobserved types that occur only among participants or non-participants and there may be some types that occur more often in one group than in the other. For example, if those with poorer initial health (many channels of which cannot be observed) are more likely to participate in a program, the mean of Y_0 might differ between the two groups because there are more sick people among participants.

To summarize, selection bias

$$S = E(Y^0|P = 1) - E(Y^0|P = 0)$$

can arise for three reasons:

1. There might be individuals with some observed characteristics who appear only among participants or non-participants;

2. There might be individuals with some observed characteristics who appear more often among participants or non-participants;

3. Either possibility can occur with unobserved characteristics.

[8]The term "failure of common support" offers a good opportunity for a lesson in the semantics of statistics. The "support" for a random variable are those values that have non-zero probability of occurring (i.e. there is some probability that they will occur). The issue with the failure of common support is the possibility of values for x that occur only when $P = 0$ or $P = 1$, but not both. Hence, there are values of x with non-zero probability when $P = 0$ or $P = 1$, but not both. In other words, the support for x is not completely common between the two values of P.

This is a simple but extremely useful framework for thinking about selection bias.

There are many important and interesting insights to be gained from this. For one thing, we have been able to understand conceptually the kinds of imbalances in covariates, observed and unobserved, that can contribute to baseline differences in average outcomes across groups. However, we have also learned a crucial lesson: in principle, we cannot necessarily be confident that we remove bias associated with differences between participants and non-participants *simply by controlling for their observed characteristics*. This is a point to which we will return in later chapters.

2.3.2 Regression

We now turn our attention to considering the challenge of estimating program impact through regression analysis. Let us begin, however, by considering the relatively easy case of *selection on observables*, under which program participation is guided exclusively by *observed* characteristics of the individual. For simplicity, assume that there is just one observed characteristic captured in the variable X.

If X plays a role in program participation (for instance, X might be income and it could be that poorer people are more likely to enroll in a program), then there is a statistical association between X and program participation status P. Intuitively, if different types in terms of X are more likely to participate in a program, then those types of individuals will be more common among participants than non-participants. For instance, if poor individuals are more likely to enroll in a program, then program participants will be more likely to be poor and program participation status P will be associated with income level.

This can be expressed mathematically in several ways. For one thing, there is now a correlation between P and X:

$$corr(X, P) \neq 0$$

Another way of looking at this is to note that it is now possible that the mean of X is no longer the same for the two groups:

$$E(X|P = 1) \neq E(X|P = 0)$$

For instance, expected income might differ between participants and non-participants.

We focus on the possibility that some types as defined by X might occur more often when $P = 0$ than $P = 1$. Thus we are more concerned with the second type of selection bias (S_2) from the preceding subsection. We are not in this subsection explicitly considering the first type of selection bias (S_1), wherein there are values of X that occur only when $P = 0$ or $P = 1$, but not both (i.e. the "failure of common support"). A statistical association between X and P might create problems if one plans to evaluate program impact simply by comparing $E(Y|P = 1)$ and $E(Y|P = 0)$ (perhaps by comparing sample mean outcomes for Y between participants and non-participants).

Consider the example of a health program (the outcome Y would thus be some channel of health). Suppose as well for simplicity that the one observed individual characteristic, X, is age. If age influences program participation, then program participants and non-participants should have different average ages. To fix ideas, suppose that older persons are more likely to participate.

This alone would not necessarily undermine impact evaluation by simple comparison of expected outcomes (presumably estimated as simple means for Y) between participants and non-participants

$$E\left(Y|P = 1\right) - E\left(Y|P = 0\right)$$

as an estimator of program impact: though the participant population would be older on average, that would not necessarily provide a source of difference between $E(Y|P = 1)$ and $E(Y|P = 0)$ outside of that driven by program participation P itself. This is important because it is our hope

in using this estimator that the differences between the two reflect the experience of program participation, and only that. If, however, age also influences Y, this simple estimator would no longer provide an unbiased estimate of program impact since any differences between $E(Y|P=1)$ and $E(Y|P=0)$ might reflect either program participation P, age X or both.

If one could control for the relationship between X and P even as they evaluated that between Y and P, they might be able to recover the causal relationship of interest (i.e. how variation in P *causes* variation in Y). **Regression analysis** allows us to control for relationships among the observed variables X and P that might influence the outcome of interest. Hence, selection on observables can generally be addressed through regression analysis (though, as we will see later, there are other possibilities, such as matching estimators and the closely related method of propensity score matching that can be applied in this setting).

What if we could **not** observe all of the variables that influence both program assignment and the outcome of interest? Continuing the example, suppose that we cannot observe X. Instead we observe Y and P for a sample of N individuals. That is, we observe the pair of realizations of Y and P $\{Y_i, P_i\}$ for each of $i = 1, ..., N$ individuals in the sample. With this data, the ordinary least squares estimators $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for the simple model

$$Y = \gamma_1 + \gamma_2 \cdot P + \epsilon$$

(i.e. the estimates that emerge when we **regress** Y on P) are given by

$$\hat{\gamma}_2 = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (Y_i - \overline{Y})}{\sum_{i=1}^{N}(P_i - \overline{P})^2} = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

and

$$\hat{\gamma}_1 = \overline{Y} - \hat{\gamma}_2 \overline{P}$$

where

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \text{ and } \overline{P} = \frac{\sum_{i=1}^{N} P_i}{N}$$

The key question before us is whether a causal interpretation can be attached to the estimate $\hat{\gamma}_2$. In other words, will the estimate of $\hat{\gamma}_2$ capture true program impact: the variation in Y *caused* by variation in P?

The first step in answering this question is to recognize that the estimator

$$\hat{\gamma}_2 = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

is a random variable. Clearly, we must have data to estimate any statistical model, including the parameters $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for the simple model

$$Y = \gamma_1 + \gamma_2 \cdot P + \epsilon$$

In other words, there must be N individuals for whom we observe P and Y for each. This data typically represents a sample of N individuals from some larger population of individuals.

However, in different samples we would expect some variation in the mix of values Y and P that the N individuals would experience. But such variation will inevitably lead to some degree of variation in the values of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ calculated for each sample. For instance, if we repeatedly selected samples of 100 individuals from the American adult population we would expect some variation between those samples in the mix of heights within them. We would then experience some variation across the samples in terms of the estimated average height obtained from each.

An immediate implication is that a particular estimate from a given sample is typically not going to equal the exact value of the parameter we are seeking to estimate. Rather, such estimates will typically vary from sample to sample. This variation is called ***sampling variation***. For instance, in the simple example of sampling 100 American adults and recording their height, the object might be to use that sample to form an estimate of the average height of American adults. However, the actual estimates of average height would vary across repeated samples of 100 American adults. Nonetheless, there is clearly some actual average height for the entire population of American adults. If nothing else, at any given point in time, there is an exact, finite number of American adults (even if we don't know what that number is) and each has a precise height. But the actual estimates based on samples from that population vary (hopefully) around that number.

Given such sampling variation, how can we ever know whether our method of estimation (or, in the language of statistics, our *estimator*) is any good? More precisely, how can we ever know whether it is estimating the thing we want it to estimate? There are two particularly commonly invoked indicators of this dimension of performance. They are:

> **Unbiasedness:** Because it is a random variable an estimator will have an expectation $E(\cdot)$. An estimator is unbiased if its expectation $E(\cdot)$ is in fact the true value of the parameter for which we wish to form an estimate. For instance, in the simple example of drawing a sample of 100 American adults and calculating the average height for that sample, the question is whether that sample average estimate is an unbiased estimator of the average height of American adults. Unbiasedness is unrelated to sample size. A somewhat crude but useful way of thinking about unbiasedness is that it asks whether an estimator will, on repeated selection of new samples and re-computation of estimates, produce a stream of estimates that are "right on average";
>
> **Consistency:** Because an estimator is a random variable, it has a probability distribution. In other words, it has probabilities attached to the possible values that it can take in a given sample. An estimator is consistent if, as sample size increases, the values that it can take on with positive probability become increasingly concentrated on the true value of the parameter one wishes to estimate. In the case of an estimator with continuous range, it is consistent if its probability density collapses around the parameter we wish to estimate as sample size grows. A somewhat crude but useful way of thinking about consistency is that it establishes that values for the estimator that deviate from that of the true parameter we wish to estimate become increasingly improbable as sample size increases.

An estimator can be biased but consistent. Roughly, it can be biased in smaller samples but nonetheless consistent as sample size grows.

For present purposes, we will focus on unbiasedness. To understand whether the expectation of $\hat{\gamma}_2$, $E(\hat{\gamma}_2)$, is unbiased (i.e. equals the true program impact), let us fix ideas by considering the implications of running the simple regression model (i.e. regressing Y on P) when the true data generating process is instead given by

$$Y = \beta_1 + \beta_2 \cdot P + \beta_3 \cdot x + e$$

where β_2 is the causal effect of the program (i.e. when one switches from being a non-participant ($P = 0$) to participant ($P = 1$) Y changes by β_2). We make the following assumptions about this model:

$$corr\,(P, e) = 0$$

$$corr\,(x, e) = 0$$

These two assumptions guarantee that the true regression error e is a truly independent, idiosyncratic error term unrelated to the regressors x and P.

The expected value of the ordinary least squares estimate (the key parameter of interest in this case) is given by

$$E(\hat{\gamma}_2) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (\beta_1 + \beta_2 \cdot P + \beta_3 \cdot x + e)}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

Notice that we insert the true data generating process $y = \beta_1 + \beta_2 \cdot P + \beta_3 \cdot x + e$ in order to take the expectation of $\hat{\gamma}_2$ from the incorrect (in the sense of not fully reflecting the true data generating process) model $y = \gamma_1 + \gamma_2 \cdot P + \epsilon$. We can then continue to take the expectation:

$$E(\hat{\gamma}_2) = E\left(\frac{\beta_1 \sum_{i=1}^{N}(P_i - \overline{P})}{\sum_{i=1}^{N}(P_i - \overline{P})^2} + \frac{\beta_2 \sum_{i=1}^{N}(P_i - \overline{P}) \cdot P_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2} + \frac{\beta_3 \sum_{i=1}^{N}(P_i - \overline{P}) \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2} + \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot e_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

$$= E\left(\beta_1 \cdot 0 + \beta_2 \cdot \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (P_i - \overline{P})}{\sum_{i=1}^{N}(P_i - \overline{P})^2} + \frac{\beta_3 \sum_{i=1}^{N}(P_i - \overline{P}) \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2} + \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot e_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

$$= \beta_2 + \beta_3 E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot e_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) = \beta_2 + \beta_3 \cdot E(\hat{\eta}_2)$$

The penultimate step exploits the independence of P and e (the independently distributed random error assumed earlier), so that

$$E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot e_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) = 0$$

We have the overall result that $E(\hat{\gamma}_2) = \beta_2 + \beta_3 \cdot E(\hat{\eta}_2)$.

Thus, when we estimate the parameters of the model $Y = \gamma_1 + \gamma_2 \cdot P + \epsilon$ when the true data generating process is in fact $Y = \beta_1 + \beta_2 \cdot P + \beta_3 \cdot x + e$ the estimate of γ_2, $\hat{\gamma}_2$, does not have an expected value equal to the true parameter of interest β_2. Instead, we get the rather odd looking hybrid result $E(\hat{\gamma}_2) = \beta_2 + \beta_3 \cdot E(\hat{\eta}_2)$. This means that the true parameter of interest β_2 is indeed buried in the estimate, but is obscured by the addition of the term $\beta_3 \cdot E(\hat{\eta}_2)$.

Notice that $\hat{\eta}_2$ is equal to

$$\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

But this is just the ordinary least squares estimate of η_2 from the model $x = \eta_1 + \eta_2 \cdot P + \zeta$ (where ζ is once again just a random error term independently distributed from the regressor P). Thus the extra term is simply the causal relationship between x and Y (i.e. β_3), weighted by the estimated relationship between x and the included variable P, or $E(\hat{\eta}_2)$. Note that if there is either no relationship between x and Y or no relationship between x and P, this troublesome term drops out and $E(\hat{\gamma}_2) = \beta_2$: the ordinary least squares estimator is indeed providing an unbiased estimate of β_2.

Behind all of this math is a simple intuitive point: when you fail to include an explanatory variable that is correlated both with the explanatory variable of interest *and* the outcome, you cannot cleanly recover an estimate of the causal effect of the included (in the regression) variable of interest on the outcome. The reason is that the explanatory variable of interest (in our case P) now serves two roles empirically: as a control for the variation you want to capture (that of the included

variable itself) and as a control for the excluded variable (with which it is correlated, and hence can serve as a proxy). In other words, we can think of our estimate of the effect of the variable of interest under these circumstances as now representing some linear combination of its own direct causal effect and the effect of the omitted variables for which it serves as a proxy weighted by the relationship between the included explanatory variable of interest and the excluded variable(s). For this reason the bias is often referred to as **omitted variable bias**.

The flip side of this bias is that it can, in the context of this basic example, be removed simply by adding x as a regressor and then regressing Y on P and x. In other words, it can be avoided simply by estimating the true model. In this example, we have focused on the possibility of a single omitted regressor x. In practice, however, an outcome of interest Y will probably have many determinants, some of which will be observed and some of which will not.[9] In general, regression analysis can recover an unbiased estimate of the causal effect of P on Y only if all other determinants of Y that are also correlated with P are also included in the regression. Once we move beyond the point of a single potentially omitted regressor, the direction of bias becomes much more complicated to determine. But, to be sure, there will be bias.

This is illustrated this in Figure 2.5. We consider three types of characteristics. $X1$ are those characteristics that directly influence the outcome Y but are unrelated to program participation P. $X2$ are characteristics that shape program participation P but have no direct role in determining the outcome Y. Finally, $X3$ are variables that influence both the program participation decision and the outcome. Of the three types of background characteristics, unbiased estimation of program impact through regression requires that $X3$ be included in the regression as controls. $X3$ are variables that influence program participation. Hence individuals sort into the participant and non-participant groups according to their values for $X3$, and the distribution of values for $X3$ would vary between the participant and non-participant groups. However, $X3$ also shapes the outcome Y. The participant and non-participant groups would thus differ by background characteristics that influence the outcome Y, making it difficult to ascribe differences in mean outcomes between the two groups to program participation alone. From a regression standpoint, P no longer signals just program participation, but also average background characteristics.

The same complication does not arise with either $X1$ or $X2$. $X1$ shapes the outcome but not program participation. Hence, we would expect the distribution of it to be the same between the participants and non-participants. $X2$ influences program participation, and hence its distribution should differ between the participant and non-participant groups. However, since it has no influence on the outcome this is of little consequence from the standpoint of evaluating program impact via either simple comparison of mean outcomes between participants and non-participants or regression of Y on P.

Finally, we consider omitted variable bias from a slightly different angle. We have considered the consequences of estimating the parameters γ_1 and γ_2 (with a special interest in γ_2) of the simple model

$$Y = \gamma_1 + \gamma_2 \cdot P + \epsilon$$

when the true model is

$$Y = \beta_1 + \beta_2 \cdot P + \beta_3 \cdot x + e$$

As we have seen, the ordinary least squares estimate $\hat{\gamma}_2$ will be biased when P and x are correlated and x influences Y (i.e. when $\beta_3 \neq 0$). Given that the latter is the true model and if $\beta_3 \neq 0$, in the simple model x is relegated to the error term ϵ of the simple model. If, as well, P and x are correlated

[9] Extending to this more complicated case would require either extraordinarily complicated equations or resorting to matrix algebra, either of which would represent a diversion at this point.

(the other condition required for bias), we are confronted with an important complication:

$$corr\,(P, \epsilon) \neq 0$$

Thus, the problem of bias in the simple model can be viewed as one of correlation between the observed regressor P and the error term in the estimated model. Omitted variable bias is often posed in this way: correlation between observed regressors and unobserved characteristics (we will often refer to these simply as *unobservables*) relegated to the error term.[10] This highlights the proxy role of the observed regressor(s) (in this case P) in the setting of omitted variable bias. The regressor serves not only as a control for itself but also for other regressors relegated to the error term with which the observed regressor(s) is correlated.[11]

We conclude with a numerical example (this numerical example is contained in the STATA do-file 2.1.do). This example considers the consequences of estimating the "wrong" model

$$Y = \gamma_1 + \gamma_2 \cdot x_1 + \epsilon$$

when the "true" model is

$$Y = \beta_1 + \beta_2 \cdot x_1 + \beta_3 \cdot x_2 + e$$

[10] If one were to estimate the simple model by ordinary least squares, in the estimation sample P would be uncorrelated with the estimated errors $\hat{\epsilon}$. This is true by construction and does not mean P is uncorrelated with the error ϵ in the underlying true regression model.

[11] It is important to remember at this point that the entire discussion of omitted variable bias in the regression has focused on the idea of a correlation between x and P reflected in a greater frequency of some values for x at $P = 1$ than $P = 0$. It has not introduced the explicit possibility that some values of x occur only when $P = 0$ or $P = 1$. This is the aforementioned *failure of common support*. This complication will be dealt more explicitly in a subsequent chapter considering matching estimators.

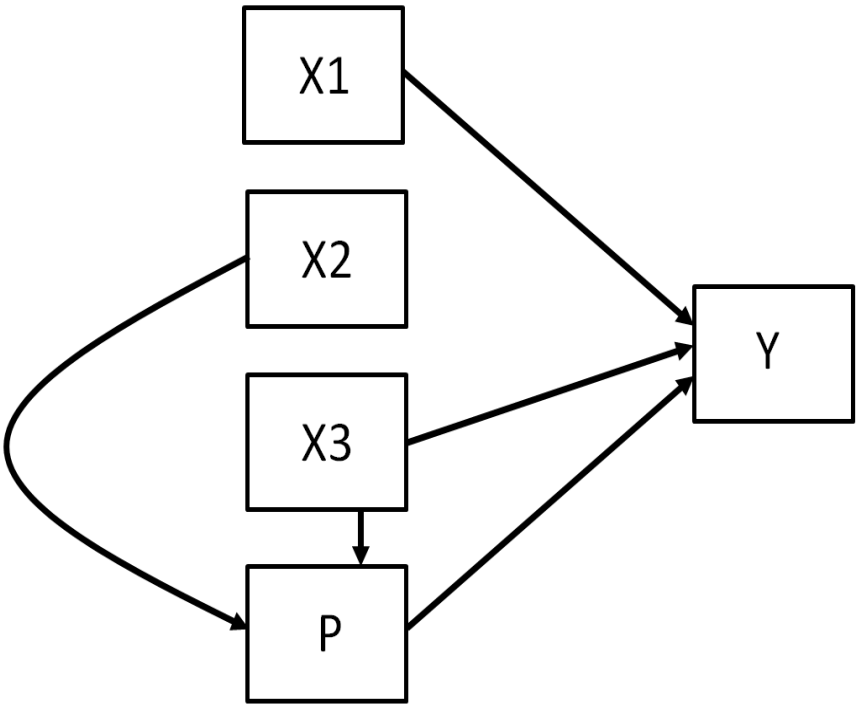


Figure 2.5: Controlling for Observables X

Specifically, suppose that we randomly generate 30,000 observations based on these two assumed equations:

$$x = 1 + P + e_x$$

$$Y = 1 - .5 \cdot P + 1.5 \cdot x + e$$

where $e_x, e \sim$ N(0,25).[12] P equals 1 if an N(0,4) random variable exceeds 0 and equals 0 otherwise. We have, with the second equation, generated data based on the true model. With the first equation we have established a correlation between the regressors P and x.

The resulting sample is considered in STATA Output 2.1. e and P have means of roughly zero at .0391012 and .5071333, respectively, and standard deviations of 5.017912 and .4999574. These are entirely sensible values. The mean and standard error for e are right in line with expectations for an N(0,25) random variable. The mean of P should be around .5: P equals 1 if a normally distributed random variable with a mean of 0 is greater than 0 (which we would expect half the time) and equals 0 if it does not. It is also a binary variable, with variance $\mu \cdot (1 - \mu)$ (where μ is the mean of P) or .5·(1-.5)=.25, leading to a standard error of .5 (i.e. the square root of .25). Notice as well that the mean of x is around 1.5, which it should be according by expectation:

$$E(x) = E(1 + P + e_x) = E(1) + E(P) + E(e_x) = 1 + .5 + 0 = 1.5$$

Finally, note P and x are correlated (at 0.0928 correlation), which we expect by construction.

Let us first regress Y on P and x. The results are presented in STATA Output 2.2. These results are right in line with where we would hope they would be, given the true data generating process behind them: the estimated coefficient on P is, at -.5152055, around -.5 while that on the coefficient on x is, at 1.501488, in the neighborhood of 1.5. In particular, note that the relationship between P and x has not complicated estimation since we control for both regressors.

Next we estimate the "wrong"regression model (i.e. omitting the variable x). The results are in STATA Output 2.3. We can now see that the omission of x has led to an estimate of the coefficient on P that deviates substantially from that underlying the true data generating process (.8838561, against a true value of -.5). This likely reflects omitted variable bias, and from this example it should be clear how bad it really can be.

We next demonstrate that the "bias" holds true to the earlier analytical result that

$$E(\hat{\gamma}_2) = \beta_2 + \beta_3 \cdot E(\hat{\eta}_2)$$

To do this, let us regress x on P to form an estimate of η_2. The estimates are provided in STATA Output 2.4. Combining these results, note that, following the results on omitted variable bias outlined above, the estimate of the coefficient on P (.8838561) from the "wrong" regression can be nearly recovered from the estimates from the correct model and the slope coefficient from the regression of x on P:

$$-.5152055 + 1.501488 \cdot .9317832 = .88385579$$

which is just about .8838561 (in the STATA .do file the fit is exact; the difference in the two calculations reflects rounding differences). Notice that we showed this exclusively with estimated parameters. The omitted variable bias likely evident in the "wrong" regression is substantial: the sign of the estimated effect of x actually became positive *and* significant!

[12] $a \sim$ N(b,c) means a is distributed as a normal random variable with mean 'b' and variance 'c'. In the context of this example we would thus be pseudo-randomly drawing 30,000 observations for a normally distributed variable a with mean 'b' and variance 'c'.

STATA Output 2.1 (2.1.do)

```
. * Summary statistics for the errors and regressors
. summarize e P x
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           e |      30000    .0391012    5.017912  -22.59459   20.19439
           P |      30000    .5071333    .4999574          0          1
           x |      30000    1.491703    5.022221  -18.59465   22.38237
.
. * Correlations between errors and regressors
. correlate e P x
(obs=30000)
             |        e        P        x
-------------+---------------------------
           e |   1.0000
           P |  -0.0014   1.0000
           x |   0.0013   0.0928   1.0000
```

In terms of the correlations, P is uncorrelated with the fitted errors from the regression of y on P. However, P's correlation with the true regression error for the "wrong" model (i.e. $1.5 \cdot x + e$, which is ϵ from $y = \gamma_1 + \gamma_2 \cdot P + \epsilon$) is .0764. Since the correlation between x_1 and the error term from the "true" regression (e) is essentially zero (-.00014) this means that the correlation between P and ϵ is driven by the presence of $1.5 \cdot x$ in ϵ.

STATA Output 2.2 (2.1.do)

```
. * Regressing y on P and x
. regress y P x
      Source |       SS       df       MS              Number of obs =   30000
-------------+------------------------------           F(  2, 29997) =33696.75
       Model |  1697036.77     2  848518.384            Prob > F      =  0.0000
    Residual |  755354.837 29997  25.1810127            R-squared     =  0.6920
-------------+------------------------------           Adj R-squared =  0.6920
       Total |  2452391.61 29999  81.7491118            Root MSE      =  5.0181

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |  -.5152055   .0582005    -8.85   0.000    -.6292809     -.40113
           x |   1.501488   .0057938   259.15   0.000     1.490132    1.512845
       _cons |   1.044592   .0416881    25.06   0.000     .9628816    1.126303
------------------------------------------------------------------------------
```

This bias will not disappear with increasing sample size because the ordinary least squares estimator of the effect of P in the second regression is *not consistent*. This was a simple example, but as such it serves as a powerful illustration of how easily omitted variable bias can arise, and how enormous its consequences can be: in this case, it would lead us to draw a completely wrong conclusion about the impact of P on y.

Precision matters when discussing the kind of topics at the center of this manual, and here we need to be precise. Technically, we have not directly demonstrated that estimated program impact from the "wrong" model is biased or inconsistent. For instance, biasedness means essentially that the estimate is wrong on average. To show biasedness we would then technically need to re-simulate many, many new samples in the fashion of this example, re-estimate program impact for

each sample with the "wrong" model, and then take the average of the program estimates across these samples. Only if that average differed substantially from -.5 could we be sure that the "wrong" model generates biased estimates of program impact.

STATA Output 2.3 (2.1.do)

```
. * Regressing y on x1 alone
. regress y P
```

Source	SS	df	MS		
Model	5857.81989	1	5857.81989	Number of obs =	30000
Residual	2446533.79	29998	81.5565633	F(1, 29998) =	71.83
				Prob > F =	0.0000
Total	2452391.61	29999	81.7491118	R-squared =	0.0024
				Adj R-squared =	0.0024
				Root MSE =	9.0309

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.8838561	.1042901	8.47	0.000	.6794431	1.088269
_cons	2.574856	.0742684	34.67	0.000	2.429287	2.720426

Thus, technically, we cannot say for certain that the estimate of impact from application of the "wrong model" to a single sample definitely reflects bias. Frankly, it is always possible that it reflects ordinary variation in estimate values from sample to sample. However, the degree to which the estimate matches expected bias per the omitted variable bias formula is powerful *prima facie* evidence for bias.

STATA Output 2.4 (2.1.do)

```
. * Regressing x on P
. regress x P
```

Source	SS	df	MS		
Model	6510.3235	1	6510.3235	Number of obs =	30000
Residual	750145.646	29998	25.006522	F(1, 29998) =	260.35
				Prob > F =	0.0000
Total	756655.969	29999	25.2227064	R-squared =	0.0086
				Adj R-squared =	0.0086
				Root MSE =	5.0007

x	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.9317832	.0577484	16.14	0.000	.8185937	1.044973
_cons	1.019165	.0411246	24.78	0.000	.9385589	1.099771

Nonetheless, the reader should always bear in mind that when estimation results that appear to match true program impact are presented in this manual it does not necessarily mean that the estimator is unbiased or consistent. Technically, it could be biased and/or inconsistent, and the closeness of the estimate presented to the truth could reflect a random fluke of sampling variation. By a similar token, a substantial difference between an estimate and true impact does not prove biasedness: it might simply be an artifact of sample by sample variation in estimates generated by an unbiased estimator.

2.3.3 Some Specific Examples

Let us consider a few more concrete examples. Suppose that we wish to understand the link between health (h) and income (I). To assess this relationship, we estimate the linear model

$$h = \beta_1 + \beta_2 \cdot I + e$$

The question is thus whether the ordinary least squares estimate $\hat{\beta}_2$ emerging from this exercise reveals the true causal relationship between income and health. The answer is *perhaps*, but only if there are not additional variables that might interfere with our ability to assess this causal link. For instance, education is thought to be a factor in the health production relationship. This is in and of itself not a problem as long as income and education are uncorrelated. However, the idea that income and education are uncorrelated is a difficult one to accept. Thus, unfortunately, it seems likely that $\hat{\beta}_2$ reflects not only the causal relationship between income and health but also the relationship between education and health. Put a bit differently, because education influences health and income *and* education and income are related, in a regression analysis of health on income the variable income serves not only to identify its own causal role but also serves as a proxy for education. Thus, $\hat{\beta}_2$ is biased as an estimator of the causal relationship between income and health.

Now let us shift focus toward a few examples more explicitly related to the problem of program evaluation. Suppose that we wish to understand the link between wages (w) and participation in a job training program (J, where J =1 if an individual participates and 0 if they do not). To assess this relationship, we might estimate the linear regression model

$$w = \beta_1 + \beta_2 \cdot J + e$$

by regressing w on J. Once again, the question is whether the ordinary least squares estimate $\hat{\beta}_2$ emerging from regression of w on J reveals the true causal relationship between wages and participation in the program (i.e. whether $E(\hat{\beta}_2) = \beta_2$). And, once again, the answer is *perhaps*, but only if there are not additional variables that might interfere with our ability to estimate this causal link. For instance, more motivated individuals may be more likely to enroll in the program. On the other hand, it could be that those with a poorer skill set, or less job market experience, enroll in the program. These possibilities are not a concern so long as these factors do not affect wages.

Yet, once again, in fact it seems likely that both motivation and skills or job market experience might influence wages. Thus, unfortunately, it seems likely that the estimate $\hat{\beta}_2$ emerging from simple regression of w on J reflects not only the causal relationship between enrollment in the program and wages but also the relationship between these other factors and wages. As this example makes clear, the pathways by which omitted variable bias can enter the picture might be varied and complex, and hence it can be difficult to anticipate the direction of the bias.

Finally, consider an example where the omitted variables are not necessarily individual-level explanatory variables. Suppose that we wish to understand the link between a woman's pregnancy status (F, which equals 1 if the woman is pregnant and 0 otherwise) and her community's exposure to a fertility control program (P). This really implies that we are interested in the underlying latent variable relationship between F^* (where F=1 if $F^* \geq 0$ and $F = 0$ if $F^* < 0$) and program participation P, and thus the latent variable model

$$F^* = \beta_1 + \beta_2 \cdot P + e$$

We would estimate this with a binary regression model such as the linear probability model (which is estimated by ordinary least squares). However, even though the program is assigned at the

community level (as opposed to instances where the design of the program is such that individuals must enroll themselves in the program), there is still reason to be concerned about omitted variable bias. The principal reason for concern is the possibility that the program is not randomly assigned across communities. If there are community-level factors that guide both program assignment *and* fertility, then estimates of β_2 from the simple regression of F on P will once again be biased (e.g. $E(\hat{\beta}_2) \neq \beta_2$) and inconsistent.

This is the **endogenous program placement** problem (Rosenzweig and Wolpin (1986), Pitt et al. (1993), Gertler and Molyneaux (1994), Angeles et al. (1998), etc.). Discussions of this problem often focus on the possibility of a program assigned by a central authority relying on an assignment rule that depends on community-level characteristics. However, this can become a problem even if the program is not centrally assigned, as in the case where community leaders might of their own initiative enroll their communities. For instance, communities where there is a greater, and typically unobserved, cultural commitment to reducing fertility (or perhaps less resistance to the idea of family planning) might be more likely to become enrolled in the program. However, these community-level factors are also likely to influence a woman's fertility, thus introducing an avenue for familiar omitted variable bias. In this case, the estimate $\hat{\beta}_2$ emerging from regression of F on P serves not only to capture the program effect, but also as a proxy for the community-level factors that guide program assignment (and fertility).

We have now dedicated considerable space to characterizing the problem of program evaluation, in part because a solid understanding of the various facets of the problem will make it much easier to understand the proposed solutions discussed in the coming chapters. The critical lessons from this discussion are:

- The ideal circumstance for evaluating a program is one in which assignment to the program across units of observation is random;
- Randomization of program assignment can easily break down. For the purpose of evaluating programs by comparing outcomes for groups exposed to and not exposed to the program, non-random assignment into the program (whether by means of observed or unobserved variables) introduces a potentially serious complication;
- However, methods that control for the role of observed variables in shaping program participation will not necessarily remove bias associated with unobserved variables that guide assignment. For instance, all omitted variables (i.e. those associated with program participation and the outcome of interest) must be included to remove the bias;
- The direction of the bias can be difficult to predict.

Before proceeding to our survey of the different program evaluation methodologies, we first briefly discuss a few other important issues surrounding program impact evaluation.

2.4 Other Considerations

Before shifting gears to review specific statistical methods for evaluating programs, there are a few issues that, while not necessarily a technically integral dimension of the various program impact evaluation methodologies per se, are important background considerations to bear in mind when designing, conducting and, in particular, drawing conclusions from evaluations. In some sense these issues transcend the particulars of the various estimators but at the same time are distinct from the basic challenge of program evaluation in the sense that they might speak to the usefulness of the

results generated by the particular methodology employed. Some of these issues have been touched on thus far, but it is still useful to round out the discussion.

2.4.1 Causality

In this manual, when we refer to a causal relationship between two variables, we mean a cause and effect relationship: variation in one of these variables *causes* variation in the other. Thus, as one variable is perturbed that perturbation will, all other things being equal, change the expected value of another variable. For simplicity, we also typically focus on one-way causality (i.e. we do not focus on the case where each variable shifts the other).

However, this is not the only conceptualization of causality that has been considered in statistics. Indeed, in statistics the word "causality" does not always refer to a cause and effect relationship. For instance, Granger (1969) focused on the case where one variable is simply empirically predictive of another. This is commonly referred to as **Granger Causality**. To fix ideas, let us consider this approach to causality in the context of estimation that Granger did: time series. First, let us assume that y and x are stationary time series. This means, essentially, that their probability distribution has a structure that is stable over time.[13] Then, consider the regression model

$$y_t = \beta_0 + \beta_1 \cdot y_{t-1} + \beta_2 \cdot y_{t-2} + ... + \beta_k \cdot y_{t-k} + \gamma_1 \cdot x_{t-1} + \gamma_2 \cdot x_{t-2} + ... + \gamma_s \cdot x_{t-s} + \upsilon_t$$

This is simply a regression of y at time t on lagged values of y as well as lagged values for another variable x thought to be a potential significant predictor of y. Roughly, if any of the γs are individually significant and those γ that are individually significant are also collectively significant (according to an F-test of their joint significance) then x "Granger causes" y.

Granger Causality does not necessarily indicate an actual cause and effect link between y and x, whereby variation in x actually causes variation in the expected value of y. For instance, it could be that the predictive power of x is due to an unobserved third variable (relegated to the error term υ in the above regression) that actually causes variation in both. In this sense it is not true behavioral causality, but instead a kind of predictive quality.

An approach to causality to which frequent appeal is made in epidemiology is the Bradford Hill criteria. Originally proposed by the British epidemiologist Sir Austin Bradford Hill (Hill 1965), the Bradford Hill criteria represent in essence a constellation of criteria for establishing a causal relationship. These criteria are:

1. **Strength-** The stronger the association between two variables, the more likely that a causal link exists;

2. **Consistency-** Essentially, the more consistently a link is found in different settings (places, samples, researchers) the more likely it is to be causal;

3. **Specificity-** The more specific the site and diseases, and hence the more readily alternative explanations can be ruled out, the more likely a statistical association is to reflect causality;

4. **Temporality-** Effects should come *after* their causes. In other words, one variable is more likely to cause another if variation in that variable precedes that in the variable whose variation it is thought to cause;

[13] Formally, a process is stationary if time itself does not shift its joint probability distribution. Consider Z_t to be a variable the value of which changes from one time period to the next. Let $F(z_k, z_{k+1}, ..., z_{k+d})$ be the joint distribution for Z_t during times $t = k, k+1, k+2, ..., k+d$. Then, Z_t is stationary if the joint distribution does not depend on time: $F(z_k, z_{k+1}, ..., z_{k+d}) = F(z_l, z_{l+1}, ..., z_{l+d})$ where l is some other base time period (i.e. $l \neq k$). In layman's terms, the underlying statistical relationships do not evolve as time passes.

5. **Biological gradient-** Essentially, a higher dosage or greater exposure should be more likely to provoke the effect in question. At the same time, there is some admission of possible threshold effects or effects if the hypothesized casual factor is simply present or occurs at all;

6. **Analogy-** Similar factors having a similar effect generally supports a causal interpretation;

7. **Coherence-** Essentially, that field (i.e. epidemiological, in the original context) and laboratory findings are consistent;

8. **Experiment-** The more plausibly experimental (in the sense of randomized variation of the independent variable in question) is the setting, the more likely an association is to be causal;

9. **Plausibility-** Whether there is a theoretically plausible mechanism for a casual and effect relationship.

Certainly some of these criteria are debatable. For instance, it is not clear that strength of association is really integral to a persuasive definition causality and, hence, why it should be a factor in determining it. Nonetheless, they remain a popular framework for approaching causality, particularly in epidemiology but also in other biomedical fields.

Given that the Bradford Hill criteria are a constellation of somewhat independent criteria and that many of them do not themselves imply a clear absolute standard (e.g. strength of association, plausibility, etc.), there is not necessarily a specific well-established threshold for judging whether, per the Bradford Hill criteria, a statistical association reflects a causal relationship. It would seem that in practice assessment of likely causality often rests on somewhat vague, subjective and inexact appeals to the extent to which a given statistical association generally meets these criteria. Indeed, in some sense there is an analogy to a circumstantial case in law in the sense that there is no lynchpin condition that establishes causality. Instead, causality is accepted per a preponderance of essentially circumstantial evidence, no element of which is unequivocally necessary and sufficient for establishing causality. Indeed, the criteria are arguably not even cumulatively sufficient for establishing it.

In this manual we frame causality in a simpler and more direct way. Put simply, the causal impact of a program is

$$Y^1 - Y^0$$

In other words, it is the difference between the outcome the individual experiences if they participate in the program and the outcome that they experience if they do not participate. Because the experience of participation is the only thing that differs for the individual between potential outcomes Y^1 and Y^0, any difference between these two potential outcomes must be caused by the program. The expected causal impact of the program would then be

$$E\left(Y^1 - Y^0\right)$$

The analog to this in any sample would be the average difference for that sample.

There can be variations on this that capture the causal impact of the program for different subgroups of the population of interest. For example, we might be particularly interested in the causal impact of the program for those who participate:

$$E\left(Y^1 - Y^0 | P = 1\right)$$

The foundation of such parameters is still, however, the difference in the potential outcomes Y^1 and Y^0.

The impact evaluation methods (that is, the estimators of program impact) that we consider in this manual generate estimates of the causal impact of a program so defined. The estimators typically involve assumptions, and provided the real world behavioral processes that generated the estimation sample conform to these assumptions, the resulting estimates of impact are causal per this simple but rather stringent definition. Therefore, these models do not offer a constellation of evidence, but instead rely on assumptions. If the assumptions are reasonable, the estimates are causal.

2.4.2 Representative Sampling

The program impact estimators discussed in this manual tend to rely on micro-level data (i.e. samples of individuals, households, firms, etc.). The collection of such data has been the subject of a vast technical literature and has utilized many different techniques, strategies, etc., to obtain data with a dizzying array of designs and features. The unifying theme is that, essentially, the process of data generation nearly always comes down to two steps: the selection of a sample of the units of observation (which is referred to as "sampling"), and the gathering of information from that sample. This subsection is concerned with the first step.[14]

The goal of the sampling process is typically to obtain a sample representative of the population from which that sample is selected. By representative we mean that the expected distribution of types of units of observation (for instance, in the case of individuals one can think of types in terms of age, wealth, education, etc.) in the resulting sample reflects that in the population as a whole from which it was drawn. In defining the term thusly, we abstract away from the possibility that, by random chance, a sample might differ (possibly substantially) in characteristics from the population from which it is drawn (for much the same reason that one really can flip a coin ten times with ten "heads" as the result). We also ignore the complication of probability sampling, whereby the sample itself is not strictly speaking representative but the sampling process yields information from which one can generate weights that might compensate for this.

Such a representative sample could, in theory, support unbiased estimation. As we have seen, an unbiased estimate is one that is "right on average" for the population from which the sample was selected. An implication of representative sampling is that, were one to draw many such samples from the population and form an estimate of the population parameter from each sample with an unbiased estimator, the average of those estimates would equal the true population value for that parameter.[15] For instance, suppose we drew many, many[16] representative samples of 10,000 Americans and interviewed them regarding their incomes. From each sample we could form an estimate of the average income of Americans. In principle, the average of these estimates would

[14]The second step is not explicitly considered very much in this manual beyond discussion of the structural requirements of data sources for various impact evaluation estimations (e.g. the estimator might require data that is repeated measures, contains certain types of variables in the sense of the operational role they might play in the implementation of that estimator, etc.). In short, the second step involves the specific information, or instruments, to be collected regarding the sample selected. We generally leave to the reader the task of deciding, for instance, which outcome Y is of interest in gauging program impact or which controls X are most appropriate in a given application (though, particularly in the case of the latter, we will often specify empirical properties that they must satisfy to insure internally valid program impact estimates). For further guidance on this subject, we refer the reader to the various compendiums of indicators for monitoring and evaluation that have been developed in various contexts. For instance, in the family planning setting Bertrand and Escudero (2002) is an excellent example.

[15]Per the discussion of internal and external validity below, the application of the unbiased estimator within each sample establishes that the estimate for that sample will be internally valid (i.e. in our context that it will capture a causal relationship). Representative sampling insures that the estimate will then be externally valid for the population from which representative sampling was done.

[16]In theory, by "many" we mean an infinite number of samples.

be the true average income of Americans. In our earlier discussion, the focus for considering unbiasedness was the estimator. Now it is the sampling process.

If a sample was not representative of the population from which it was drawn it is unclear how externally valid it would be (i.e how relevant it would be to that larger population). Suppose, for example, that the sample of 10,000 Americans was selected by telephoning families on land lines and then simply asking the respondent within the household who answers the phone about their income. This procedure will clearly under-sample (compared to their true presence in the population) those residing in large households (because the probability of being a respondent in such a household is, all other things being equal, lower) and those in households not served by traditional land lines.

An estimate of average income from such a sample will likely be skewed from the unbiased ideal (i.e. the true average income of Americans) because the types of individuals in the sample probably do not reflect the true distribution of individuals across the population. Moreover, this deviation is not due to random variation, but instead a systematic product of the sampling process. We would thus expect repeated samples drawn in this fashion to be skewed on average and thus estimates from them to be biased.

To fix ideas more directly on the program impact evaluation literature, suppose for the sake of discussion that we could in fact observe Y^1 and Y^0 for everyone sampled from some population. Suppose as well that we select a sample of size N from that population in some fashion that somehow leads to program participants (who likely had comparatively higher personal returns to program participation) having greater representation in the sample than their presence in the population might warrant. An estimate of average program impact (i.e. the average treatment effect) would be along the lines of

$$\frac{\sum_{i=1}^{N}(Y_i^1 - Y_i^0)}{N}$$

where $i = 1, ..., N$ indexes individuals in the sample. If the sample is dominated (compared to their true share in the population) by participants, who likely have a relatively high personal impact $Y^1 - Y^0$, we might expect that our estimate of the average treatment effect would exaggerate its true population value.

The same consideration might apply to the possibility of using a sample drawn (even in a representative manner) from one population to make inferences about the circumstances prevailing in another population. Consider our earlier example involving the income of Americans. Suppose that our sample was representative only of Americans between the ages of 18 and 50. Though such a sample might be the basis for an unbiased estimate of income for Americans in that age range, it might provide a misleading sense of average income for *all* Americans.

Consider as well a modification of the program impact evaluation example introduced earlier in this sub-section. Suppose now that the sample of N individuals was drawn from only program participants. It then becomes difficult to know how relevant the estimator

$$\frac{\sum_{i=1}^{N}(Y_i^1 - Y_i^0)}{N}$$

is to the program returns of non-participants, who are not represented in the sample used for program impact estimation.

Although it may seem to this point that the main concern in this discussion is *external* validity, failures of representativeness can be a tricky business. There are some failures of representativeness that can even threaten *internal* validity. In the extreme, it would not be possible to form an unbiased estimate even for the sample for which we do observe all of the information necessary to compute an estimate.

For now we conclude simply by noting that the representativeness of the sample used is an important determinant of the degree to which the resulting estimate of program impact can be assumed to hold in other settings, which might include the larger population from which the sample is drawn (here the concern would be that the sample was not selected from that population in a fashion that insured representativeness) or other populations of interest. Indeed, it can be an important determinant of whether the program impact estimate is even reasonable within the estimation sample (i.e. internally valid). Those conducting program impact evaluations should be aware that the sampling process by which their samples were selected are the foundation for their work, and that, as in architecture, a sound structure can rarely emerge from a flawed foundation.

2.4.3 Observer and Hawthorne Effects

It is sometimes the case that the individuals who are the subjects of observational studies are aware of the fact that they are being monitored or will be monitored. For instance, micro-level data may be drawn from a panel study of households and their members, in which case these individuals are typically aware of the prospect of their ongoing participation in a survey in which their future actions would be observed and recorded. More generally, a researcher will have available to them data on the behavior of and outcomes experienced by individuals only if they were monitored in some fashion or another. However, the fact that they knew that they were being monitored may have altered their behavior (for example, to avoid perceived shame or conform to social expectations).

In the social sciences, this idea is sometimes referred to as the **Heisenberg Uncertainty Principle** but perhaps more appropriately[17] defined as an **Observer effect**: the very fact that people know they are being watched might influence their decisions. This might limit our ability to draw broader conclusions with information drawn from observational data. The behavior of those observed might have been very different had they not been observed.

A related complication is often referred to as the **Hawthorne effect**, whereby participants in a study may alter their behavior as a result of their awareness of the uniqueness or distinctiveness of their status as participants. This possibility was first noted in the Hawthorne Plant of the Western Electric Company. At Hawthorne, Western Electric conducted a series of experiments designed to evaluate the effect of various innovations to working conditions on overall productivity. Many of the results seemed counterintuitive. In one perverse example, productivity increased both when the working environment was made brighter and when it was made darker.[18]

It has been widely hypothesized that the main factor driving the Hawthorne effect is the psychic effect of knowing that one has been selected for special attention, for instance as a participant in a study (perhaps that one's activity has been deemed sufficiently important to be studied).[19] For

[17] Werner Heisenberg (1901-76) was a physicist who introduced an uncertainty principle that is often popularly interpreted as stating that the very act of observing something might influence its behavior. Though this is a widespread and popular interpretation, it is not faithful to Heisenberg's original assertion. Strictly speaking, the Heisenberg Uncertainty Principle suggests that it is not possible to know simultaneously and with complete precision the values of certain pairs of variables (referred to in physics as "canonically conjugate variables"). A famous example is that one cannot simultaneously measure the precise position and momentum of a particle because there is a tradeoff with respect to the accuracy with which the two can be measured. Thus, strictly speaking, the original Heisenberg Uncertainty Principle is focused more on the limits of precise observation than the idea that observation would alter the behavior of the observed.

[18] Interestingly, Levitt and List (2011) has called into question the degree to which the "Hawthorne Effect" actually occurred during the original experiments at the Hawthorne plant itself.

[19] There seems to be a bit of disagreement about the basis for the Hawthorne effect. The Oxford English Dictionary describes it thusly: "an improvement in the performance of workers resulting from a change in their working conditions, and caused either by their response to innovation or by the feeling that they are being accorded some attention." In other instances, the focus seems to be more narrowly on involvement in a study or experiment.

instance, those that are selected to participate may feel important or even empowered, leading to effort levels that would not be forthcoming if they did not view their participation as such a mark of distinction (as in when a pilot program is eventually extended to all of society). Alternatively, it might be the case that participants in a more limited trial implementation of a program (or experiment) are aware of its importance, and thus modify their behavior according to their perception of the stakes of that experiment for society.

The implication of the Hawthorne effect is much the same as that of the Observer effect: it undermines confidence in the degree to which results can be generalized. For instance, it undercuts our ability to draw more general conclusions about the likely impact of a program based on the information gathered from a purposeful study, such as an intentional experiment.

At first glance, it might seem as if the two refer to the same thing (and indeed, on a definitional level some regard the two as identical). However, we believe that it is useful to make a distinction between them and view Observer effects as arising simply because of being observed (and not necessarily because of participating in a study, to which participants might be indifferent) and Hawthorne effects as reflecting the awareness of receiving special attention, as in *participation* in a study.[20]

The Observer effect can arise even with longstanding participants in a program already widely implemented throughout society who might modify their behavior once they perceive themselves to be under observation. For instance, women enrolled in a study of the efficacy of an established, ongoing child nutrition program may alter (or report altering) their behavior in front of observers in order to more closely conform with societal standards of appropriate parenting.

Hawthorne effects are more often ascribed to circumstances where participants somehow perceive themselves to be unique or important for their selection for the program or study. Conceptually, Hawthorne effects can emerge even in circumstances where the behavior of individual units of observation cannot be observed. For instance, it is conceivable that individual participants in an experiment (in program or control groups) may increase their effort levels as a result of their awareness of participation even if the end result of those intensified efforts is recorded only at the organizational level.[21] In any case, the critical concern with either Hawthorne or Observer effects is the degree to which the conclusions drawn from an impact evaluation can be generalized to settings where the program is implemented under more everyday (and thus less "special") circumstances when program participants and non-participants are not necessarily being observed or somehow receiving special attention in the context of a study.

2.4.4 Impact Evaluation Based on Pilot or Trial Programs

Often, when program impact is evaluated, the statistical model is estimated with data obtained from a setting where the program was operating on a limited pilot or trial basis. At that point, its operations were either sufficiently limited that it is reasonable to assume that it could not influence

[20] We introduce this distinction for present purposes because we believe that it is conceptually useful. It is not necessarily one that others draw or that is faithful to the details of the actual Hawthorne experiments (the interpretation of which remains the subject of debate).

[21] To cite one more example, police officers in a study of alternative policing strategies might alter their behavior because they fear that researchers will report sub-standard performance to superiors or simply because they wish to avoid appearing lazy or inefficient in front of anyone else. Alternatively, they may behave more cautiously to avoid endangering observers (as in when an officer cautiously awaits backup to avoid any collateral danger to a researcher "riding along"). Under our taxonomy, these possibilities would most appropriately be viewed as observer effects. They also might alter their behavior because their selection for inclusion in the study convinces them that their task within the police force is particularly vital or highly regarded. This would perhaps be most properly viewed as a Hawthorne effect. An interesting variation on the Hawthorne effect might arise if participating officers were aware of the implications of the experiment and modified their behavior in order to influence the conclusions drawn from it.

its larger operational basis or did so only to a miniscule degree reflective of its limited scale of operations. However, when a program expands beyond its trial or pilot phase, it might begin to influence aggregate variables that the researcher had been able to assume were constant (and thus ignorable) when conducting evaluations with data based on more limited implementations of that program.[22] The basic point is that, as a program expands, it might begin to have an impact on the overall environment within which it must operate. For instance, for the purposes of impact evaluation using data on the experience of participants and non-participants drawn from limited or pilot implementation of a job training program, the researcher can safely hold aggregates such as overall market wage rates constant. However, as the program expands, it may begin to have an aggregate, economy wide effect that will influence the value of various aggregate variables such as productivity and wages. This might serve to dampen or amplify the original training program impact observed for the more limited implementation (see, for example, Heckman et al. (1998)). This is a kind of external validity concern: estimates obtained from the setting of limited program deployment might be misleading indicators of program impact when the program is implemented on a much wider and more influential scale.

A distinct concern with pilot or trial data is that it is not always the case that more widespread implementation of the program will prove straightforward. In the extreme case, that which was practical on a limited basis will prove completely impractical on a wider one. While this extreme case this does not necessarily speak to the challenge of evaluating programs and the related concern of making inferences about the implications of the program when applied on a wider scale (since its infeasibility makes it a moot point), there are in the less extreme case more subtle complications that might arise.

When a program is implemented on a limited or trial basis, it might have features that are subtly altered when it is expanded. For instance, it is often the case that it is harder to maintain the motivation or training of program personnel when it is implemented on a large scale. It could also be the case that management constraints begin to assert themselves as the program expands. There are other intangible elements of a program, such as the degree to which it offers participants a friendly and personally attentive environment, that might erode as it expands. On the other hand, there could be positive qualities to a program present on a large scale but absent in more limited or trial implementations. Whatever the case may be, the essential point is the same: the qualities of a program when implemented on a limited basis that led to one set of conclusions might not remain after it has expanded. On a related note, one must also be mindful of the perils of using an impact estimate of a program developed under one set of circumstances for the purposes of extending it to another setting. For instance, even if a program could be replicated in all of its particular details in another country, its impact might not be the same: the entire institutional environment within which the program will operate will differ.

These in some sense appeal to the question of external validity analogous to that proposed in the preceding discussion of partial versus general equilibrium effects. In essence, in both cases the inferences concerning program impact obtained from a limited implementation of a program might not provide a good indication of impact when the program operates on a far wider scale. The two complications are distinct in that the partial versus general equilibrium tension reflects the possibility that a program will, when operating on a wide enough scale, begin to influence its operational environment, while the scaling up issue has more to do with the evolving internal

[22]The limited implementation impact is sometimes referred to as "partial equilibrium" while that when the program operates on an economy-wide scale can be referred to as "general equilibrium". The distinction is between circumstances where the program is too small to influence economy-wide equilibrium values for aggregates versus one where it is implemented on a large enough scale to do so, making the pathways influencing its impact more complicated.

dynamics and logistical and operational circumstances of a program as it expands.

2.4.5 Other programs

In some sense, the discussion in the preceding subsection speaks to the concept of external validity: the degree to which impact evaluation estimates can be generalized to settings beyond which they were drawn (even if they are internally valid in the sense of providing an unbiased or consistent estimate of program impact within the context of the sample used). Another complication is that, in the real world, for any given population there are typically a multitude of programs influencing various channels of human welfare. Any of these could influence the outcome of interest for a particular program and hence have implications for any impact evaluation of that program. For instance, one might wish to evaluate the impact of a health program in a given population. However, at the same time, an educational program might have been operating in that population that had some influence on outcomes that the health program also sought to influence.

This really has two potential implications. First, it could undermine internal validity to the extent that program participation rates for the educational program differ between participants and non-participants in the health program. This is no different from the imbalance of background characteristics problem that we have spent much of this chapter characterizing.

The other, more subtle, threat is to external validity. To fix ideas, suppose that participation rates in the education program were the same between health program participants and non-participants. Then, the presence of the education program does not present an obvious threat to internal validity. However, to the degree that the education program influences the efficacy of the health program, it could undermine external validity. For instance, it could be that the health program causes a far more pronounced improvement to health in combination with the education program. Impact evaluation based on a sample from the population receiving the health and education programs might then lead to an exaggerated sense of the benefit of the health program when applied in settings where the educational program is absent. Alternatively, it could be that the education program does not enhance the effectiveness of the health program, but does in some sense cushion non-participants in terms of the outcomes that they experience. If this were the case, internally valid estimates of program impact might actually understate the value of the program in populations where there was no educational program operating. Finally, there is always the potential for a type of reverse threat to external validity: applying program impact estimates drawn from a setting where the education program was not operating to circumstances were it would be.

There is not necessarily a statistical solution to this threat to external validity. However, it does suggest the need for interpretive caution when applying program impact estimates from one setting to another. We form program impact estimates with samples drawn from populations experiencing certain background circumstances, and hence those estimates are always conditional on those background circumstances. This requires caution when applying the lessons drawn from one set of circumstances to a new setting.

2.4.6 The Challenge of Defining the Counterfactual

Thus far in our discussion of characterizing the program impact challenge, we have generally assumed that we can observe a counterfactual. Specifically, we have assumed that there is a subpopulation not participating in or exposed to a program, and that we can at a minimum observe the experiences of a sample from that subpopulation. Suppose, however, that it is not evident that such a "counterfactual subpopulation" exists. This complication would most obviously arise

in the case of a program implemented throughout a society in a fashion that renders participation involuntary.

Consider, for instance, a health communication program including television and radio components in a small country within which the television and radio signals are received everywhere. Strictly speaking, there will be no subpopulations residing in areas not exposed to these components of the health communications program.

Sometimes this can be surmounted simply by refining the definition of program participation. For instance, Hutchinson et al. (2006) consider the impact of a health communications program in Bangladesh on various health and family planning outcomes. It proved essentially impossible to isolate an area not exposed to the television and radio components of the program, and so the authors redefined participation as recall of key messages from the campaign. This created a natural "counterfactual subpopulation" (those who could not recall these messages) despite the fact that the television and radio signals reached everywhere in the study area considered.

Another alternative, one which we will discuss briefly at the end of the manual, is to consider evaluating the impact of the program through deeply "structural" modelling. This approach seeks to estimate the parameters of a model that explicitly characterizes the decision making process for individuals. If one can successfully estimate such a model, then they could in principle use the fitted (i.e. with parameter estimates plugged in) model to simulate behavioral outcomes under various potential scenarios, including the presence and absence of a program. Note that this approach could in theory be used to obtain inferences about the likely impact of programs that have no yet been implemented (in which case there is in some sense no "factual subpopulation"). It is sometimes suggested that these models, by their purported capability of simulating the effects of policy variation not observed in the data from which they were estimated, overcome the Lucas Critique, although strictly speaking this is not exactly a correct application of that original critique.[23]

2.4.7 Levels of Implementation

Thus far we have usually, though not always, approached program participation as an individual level choice. However, this is not always a useful framework for considering program participation. For instance, some programs are implemented at essentially the community level. In wealthy countries, one can think of initiatives such as policing strategies, water fluoridation, etc. In lower income countries it is common to implement human welfare programs at the community level. For instance, health care programs are often implemented at the community level: an entire community might be considered participants if, for instance, a clinic is placed in that community. Just such a possibility has been considered in numerous works (e.g. Angeles et al. 1998).

In general, the impact evaluation methodologies discussed in this manual are presented through the framework of participation as an individual choice. However, the basic methods are typically easily adapted to other levels of participation choice.

[23]Robert Lucas is an American Nobel Laureate macroeconomist (i.e. his work focuses on the branch of economics concerned with understanding the movement of economy-wide aggregates such as prices, national incomes, employment, etc.). In Lucas (1976) he argued that the macroeconometric models of the time, estimated under historical data taken from eras when one or another policy regime held, were not useful for simulating the effects of new regimes. The reason is that the estimated parameters of those models were not structural (in these sense of being policy-invariant). This is referred to as the Lucas Critique. Though the original critique was quite a focused statement about a particular problem, the Lucas Critique is often invoked for the more general circumstances of using fitted models to simulate the effect of variation not observed in the estimating sample.

Chapter 3

Randomization

In this chapter, we discuss what some would regard as the most convincing response to the basic identification challenge of program impact evaluation: an experiment whereby program participation is somehow randomized. Indeed, experiments are often referred to as the "Gold Standard" in impact evaluation methodology. To others, they fall so short of their promise, in practice and even in some respects in principle, that they are something of a "Pyrite" standard.[1] In this chapter we will outline the basic requirements for a successful experiment, get a feel for the kind of experiment-based evaluations that have been done and then explore the debate over this approach.

3.1 Randomization: The Basics

We have now seen that the challenge in evaluating program impact with data drawn from uncontrolled real world circumstances is that program participation may not be randomly determined but instead a conscious, purposeful choice made by individuals. Their participation decision would most likely depend at least in part on their characteristics, both observed and unobserved. These characteristics would thus help to "sort" individuals into participants and non-participants. As a result, participants and non-participants would be different types of individuals.

The most straightforward estimators of program impact (e.g. comparison of average outcomes between participants and non-participants) may then fail to provide an unbiased estimate of the causal effect of the program on outcomes. The basic problem is very simple. Because of non-random assignment to the program, there may be systematic differences in the characteristics of participants and non-participants. In other words, these two groups differ by more than just the experience of program participation. We then cannot tell whether any differences in outcomes between them are driven by the experience of being in the program or by their other differences in characteristics.

An obvious potential solution to this identification challenge is an evaluation design under which program participation is somehow randomly determined. If agents were randomly assigned, there would be no basis for systematic differences between participants and non-participants beyond the experience of participation itself. In other words, while there may be idiosyncratic differences in characteristics between any particular participant and non-participant, these will tend to cancel out on average with the result that the mean characteristics (observed and unobserved) of the two groups should be the same. Any differences between the two groups can then be ascribed to their differences in terms of the experience of program participation. As we will see, this method lends itself particularly to estimating measures of average program impact.

[1] For the metallurgically disinclined, Pyrite, or FeS_2, is also known as "Fool's Gold".

We define this approach to program impact evaluation as an "experiment". The definition of an experiment can be a bit unclear. In many discussions of program impact evaluation, it refers to a purposefully crafted social experiment under which participation in the experiment and the randomly determined program participation assignment of participants in the experiment are per the design of an evaluator for the purpose of evaluating the impact of that program. Under this conceptualization, they are often referred to by the term **randomized control trials** (or **RCTs**, in short). This is the definition we had in mind in introducing this chapter of the manual and it represents the ideal around which the discussions in the chapter revolve.

We refer to the samples generated by *successful* RCTs as **experimental**. The cornerstone of our definition of an experimental sample is the complete randomization of program participation among the units of observation (typically individuals) within that sample. We refer to samples in which program participation is not completely randomized as non-experimental. In no sense are our definitions of experimental or non-experimental samples universally accepted or uncontestable. They do not need to be: they are simply convenient for sorting out issues in the context of the present discussion.

One can think of a somewhat broader and more inclusive definition of an experiment as an instance where there is some mechanism (purposefully introduced by evaluators or naturally arising by accident or happenstance and hence unrelated to the goal of conducting a program impact evaluation) that randomizes (wholly or partially) program participation and is completely external to the individuals for whom program impact is of interest. This broader definition includes what are often referred to as **natural experiments**. Natural experiments do not require purposeful design by evaluators, but can instead arise by happenstance from the natural course of events. They are experiments offered by the messy laboratory of the human experience: instances where the real world provides some mechanism beyond the individual's control that may have introduced a random element to their program participation decision.[2]

Though this chapter was motivated by the more restrictive definition of experiment reflected particularly by RCTs, we do discuss a few natural experiments within it. Natural experiments do have in common with RCTs that they involve some external (to the individual) source of random variation in program participation. There is thus value in assessing the credibility of a natural experiment in terms of the criteria by which RCTs are judged. Impact evaluations using data from natural experiments tend to utilize quasi-experimental estimators. These will be defined more precisely later in the chapter, but for present purposes they are estimators of program impact that seek some sort of random channel of program participation in data that is non-experimental (per our definition of experimental and non-experimental samples). The quasi-experimental estimator of program impact typically applied to data from natural experiments does however require a random channel of variation in program participation that is external to the individual. Natural experiments can thus be thought of as defining the borderlands between RCTs (for which the identification of causal program impact relies intrinsically on the complete randomization of program participation in a fashion external to the individual and hence reflects the ideal of biomedical trials as applied to program impact evaluation) and other quasi-experimental methods which do not rely on random variation in program participation external to the individual.

We generally do not describe the results of a successful natural experiment as an experimental sample, because natural experiments typically yield partial randomization of program participation.

[2]That said, the term natural experiment has been utilized in different contexts, leading to alternative operational meanings. Sometimes the term simply suggests the presence of a naturally arising control group of non-participants, but not necessarily randomization of participation. Conniffe (2000) adds the requirement of observing the non-randomly assigned program participants and non-participant controls both before and after the initiation of the program. Others have adopted still more inclusive or exclusive definitions.

In such cases we would classify their fruits as non-experimental samples. That said, the sample from a natural experiment that actually generated complete randomization of program participation (a rare bird but not unheard of) would properly be considered experimental. By this logic, an RCT that is unsuccessful, particularly in the sense of achieving less than full randomization of program participation, would yield a non-experimental sample.

We tend to retain the language of the last chapter and refer to those randomly selected to participate in the program as program participants and those randomly selected not to participate as program non-participants. In some works, these two groups are referred to as the treatment and control groups (an appeal to the metaphor of the clinical trial) or the experimental and control groups (which hints at the laboratory tradition of randomized biomedical experiments). We stick to the generally adhered to convention of Chapter 2 for the sake of consistency.

As Manski (1996) points out, the RCT approach to evaluation has a legacy that extends back at least to Fisher (1935). To quote Manski's summary of Fisher's framework:

> "Let random samples of persons be drawn and formed into treatment groups. Let all members of a treatment group be assigned the same treatment and suppose that each subject complies with the assigned treatment. Then the distribution of outcomes experienced by the members of a treatment group should be the same (up to a random sampling error) as would be observed under a program in which the treatment in question is received by all members of the population"(p. 710)

Despite their rather long pedigree, RCTs are an approach that seems to come in and out of fashion in various disciplines, and often when its popularity is on the rise it is discussed in terms almost suggesting a new innovation. Notice as well the use of the term treatment, which in this context is essentially synonymous with participation.

Manski (1996) goes on to summarize the assumptions that lie behind the RCT approach to program evaluation (paraphrasing Manski 1996, p. 711 closely):

1. Participants in the experiment are randomly selected from the population of interest and randomly assigned to their program participation status;

2. All participants in the trial comply with the program participation status to which they are assigned;

3. The experiment lasts long enough to replicate the program under consideration and to influence outcomes;

4. There are no "social interactions that may make a full-scale program inherently different from a smaller scale intervention" (p. 711).

Notice that assumption 4 in certain respects speaks to the partial versus general equilibrium effects discussed in chapter 2. It could also be interpreted in light of the discussion of the complications of scaling-up as well as that concerned with other, concurrently operating programs.

While the idea behind an RCT may seem straightforward, it can be difficult to grasp how they are conducted in practice. We therefore begin by discussing a few studies that appealed in one fashion or another to randomization designs. We take this approach because there is no real "method" to RCTs beyond simply achieving the criteria in Manski (1996) in the sense that there are well developed estimation procedures for the quasi-experimental estimators discussed in subsequent chapters. The review will hopefully leave readers with a richer sense of the body of work involving randomization.

3.2 Experimental Evaluations: Some Specific Examples

3.2.1 John Snow and the Causes of Cholera

Despite the fact that RCTs motivate this chapter, we begin by discussing a study that arguably relied on a natural experiment. We do so because it is the first and still one of the most important, oft-cited attempts to assess a causal relationship by appealing to an experiment (broadly defined). Indeed, it is often invoked by advocates of RCTs in building their case for their preferred approach to program impact evaluation!

John Snow (1813-1858) is widely regarded as a pioneer of modern epidemiology. In that context (he also had a broad record of achievement in areas such as anesthesia and medical hygiene), he is perhaps best known for his work identifying the causes of cholera.[3] Before Snow's work, it was widely believed that cholera was caused by a miasma. The term may seem silly at first glance to modern readers, but in effect it often amounted to the notion that the disease was somehow airborne (miasma theories typically appealed to the idea of "bad air"), which is not on its face implausible. Today we know that many diseases spread by air, for instance in aerosol form. The focus of the miasma theories was usually some sort of putrefaction of the air from rotting organic material. In the context of its time, this theory was not without some evidentiary support. For instance, it was noted that cholera was less common in elevated areas of London where the air was believed to be better (this had been found by William Farr, the Statistical Superintendent of the General Register Office, member of the Committee of Scientific Inquiries and, at least initially, general supporter of the miasma theory of cholera infection).[4]

Snow believed that cholera infection was waterborne. Today, we know that cholera is indeed spread via water infected with the bacterium *V. cholerae*. However, in Snow's time this was a very controversial idea and it had expensive implications: massive spending, both in terms of public works and investment by private water companies, would be required to improve the water supply. For instance, the cost of compliance had led to widespread non-compliance with London's Metropolitan Water Act of 1852, which aimed to improve water quality.

In an effort to gather evidence for his theory, Snow had carefully studied several cholera epidemics in mid-19$^{\text{th}}$ century London. One incident that he studied was an outbreak in a Northwest Soho neighborhood that appeared centered roughly on a water pump at the corner of Broad and Cambridge Streets. Snow produced a map (Figure 3.1) showing the cases at each point on each street around the pump. The density of cases at each location is indicated by a dark bar pointing in off the street from the location in question; among other contributions this is essentially an early and still particularly clever histogram. This clearly demonstrates a concentration of cases around the pump, but does not establish the pump as the cause of the outbreak. For instance, there were other establishments and features of Broad Street that might have arguably been the cause (perhaps by supplying a miasma). Some other feature of the area might have induced miasma. Indeed, the most cynical view would be that the map only unequivocally demonstrated what was already well-known: the location of the outbreak.

[3] See Deaton (1997), which helped to frame our thinking regarding Snow's work. All quotes are Snow (1855), by way of Deaton (1997). An extremely impressive resource regarding the life and work of Snow can be found at a site created by Professor Ralph R. Frerichs and hosted by the U.C.L.A. Department of Epidemiology (URL: http://www.ph.ucla.edu/epi/snow.html). We are grateful for the really nice clarifications of several key points provided by Prof. Frerichs' site.

[4] Farr did concede, and early on, that there was a statistical association between cholera infection rates and water source. However, he pointed out that water source was essentially confounded with elevation, and that the magnitude of the estimated associations lent more weight to the elevation based explanation. Recent work (e.g. Bingham et al. 2004) calls Farr's statistical conclusions in this regard into question.

Realizing that a more convincing test of his theory was required, Snow exploited a fascinating natural experiment. In that era, much of the water supplied to London households and establishments was provided by private water companies. In some areas of the city a given company had a near monopoly but in others their water coverage areas overlapped. This is illustrated in Figure 3.2, which shows the coverage areas for the Lambeth and Southwark-Vauxhall water companies. In the areas where their service delivery overlapped, the water lines of the two companies were completely intertwined and Snow believed there to be more or less happenstance assignment of households to Lambeth or Southwark-Vauxhall water. He argued essentially that a household's water company was the product of a gradual, idiosyncratic set of arrangements that had evolved over a long period of time, typically extending much farther into the past than the residents had lived in that home, and was hence for all intents and purposes independent of their characteristics. The overlap area contained a diverse population and, as can be seen in Figure 3.2, was considerable in size.

London sewage was at the time dumped straight into the Thames river at London.[5] Until the mid-1840s, both companies drew their water from the Thames River downstream from London and hence, crucially, downstream from the sewage flow from London into the river.

Then, in 1847, the Lambeth Company decided to move its intake upstream of London. The move was completed in 1852. Thus by the early 1850s one company still provided water potentially

[5]Today London sewage is treated to the point where it is nearly potable and then returned to the Thames below the Thames Barrier and hence well downstream of London.

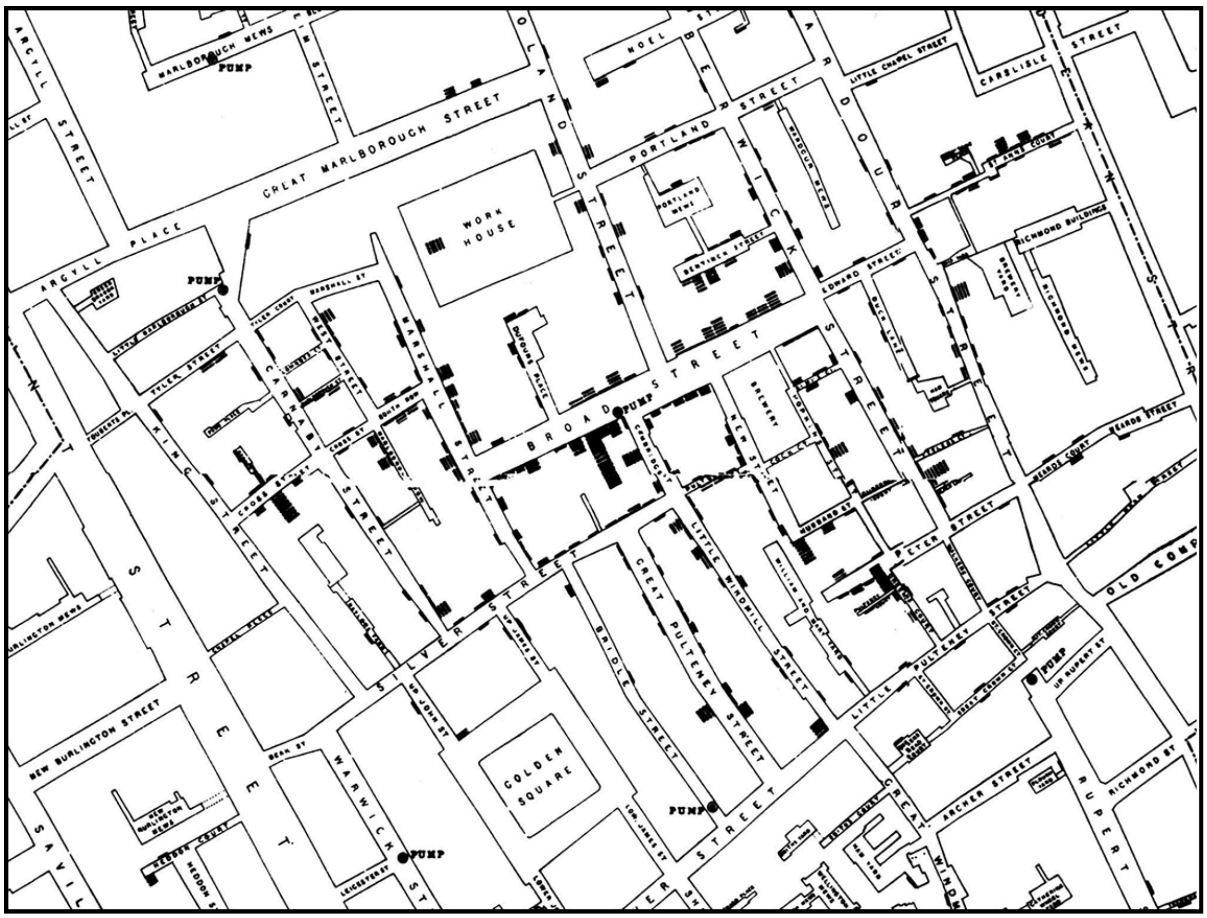


Figure 3.1: John Snow's Broad Street Map, image courtesy of Wikimedia Commons

contaminated by London sewage while the other did not. When a cholera epidemic struck once again in 1853-4, Snow used this unusual institutional circumstance to test his theory. Using household-level data, he found that those households served by Southwark-Vauxhall had experienced 8.5 times the deaths per thousand from cholera as those served by Lambeth. Applying the language and notation of chapter 2, provision of safe water to the household (as captured by receiving water from Lambeth Company) was the program P and the outcome was cholera mortality Y.

Figure 3.3 illustrates the basic mechanism for the experiment (using Baker Street, to be the home of a great detective four decades later, as a reference). Although London residents could (and often did) drink water outside of their household, their households were important water sources for them. The divergent intake points for the two companies sorted residents of the areas where services overlapped into two categories of exposure to contaminated water. As Deaton (1997) points out, the key to this experiment is that cholera is not directly caused by water intake but

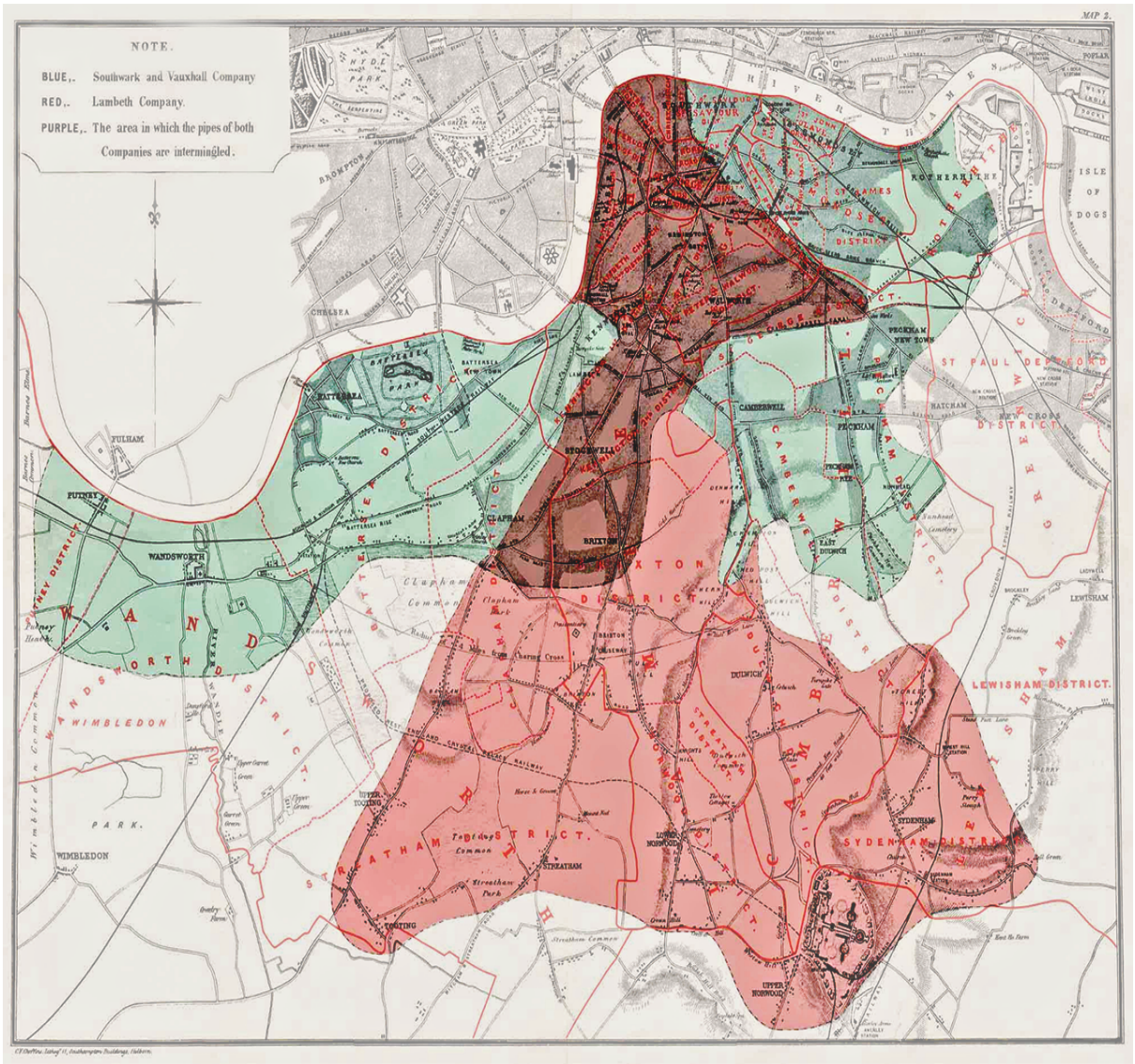


Figure 3.2: John Snow's Map of Lambeth and Southwark-Vauxhall Water Supply, retrieved from the U.C.L.A. School of Public Health John Snow Web site at http://www.ph.ucla.edu/epi/snow.html

rather by water contamination. The randomness of intake insured a source of random variation in exposure to contaminated water. As he notes, simply examining exposure to impure water would have been far less convincing: in practice drinking poor water would likely have been associated with poverty and other forms of environmental contamination. The key here is that individuals can get impure drinking water from all sorts of sources (outside of the home, impurities introduced within the home, etc.) and the tendency to drink impure water, itself a behavioral choice, is likely to be associated with other behavioral choices (such as other types of unsanitary conditions) that might also influence the outcome of interest, cholera mortality. But variation in the type of water company allowed Snow to identify one element of the variation in exposure to impure drinking water that was random.

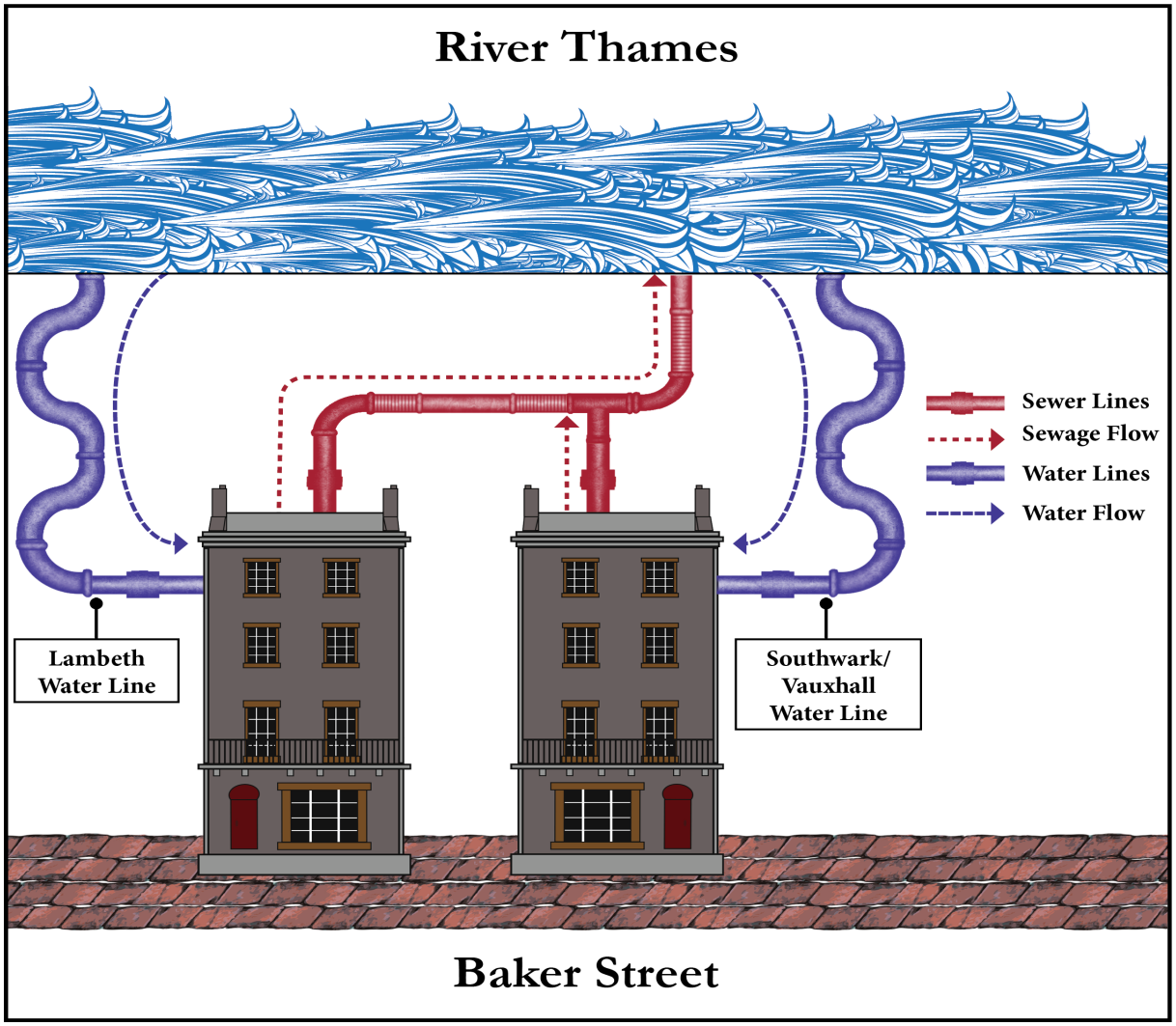


Figure 3.3: The Grand Experiment: Identification Mechanism

Snow's findings did not persuade everyone, but they were sufficient to prompt greater enforcement of the Metropolitan Water Act of 1852 (Southwark-Vauxhall was forced under the act to move their water intake upstream of London in 1855). The other water companies were gradually forced to follow Lambeth's lead. While many of the critics of Snow's investigation relied on somewhat specious reasoning, there were actually potentially serious flaws in this "experiment".

First, and most importantly, it is not clear that assignment of households to the two companies was completely random. Technically, households could elect to change sources, and there is some anecdotal evidence that there might have been reasons to do so (some reports suggest that Lambeth water unsurprisingly tasted somewhat better). Second, Snow differentiated household water source with a test of salinity, but there have been questions about the reliability of this test. This might lead to mis-assignment of households in Snow's data. Finally, Snow was unable to focus on the areas where services overlapped, but instead had to examine the entire coverage areas of the two companies. As we will see, some of these fundamental problems (e.g. potential non-compliance with original experimental assignment, in this instance by switching companies) are not unique to Snow's experience.

At the time, Snow's findings made somewhat more popular a multi-factorial (as opposed to merely miasma-based) approach to cholera infection. However, there is an interesting postscript to this story. One person that Snow's work did eventually convince was William Farr, whose earlier research had seemed to offer some support for the miasma theory. Farr would garner stronger support for Snow's theory with a similar study of an 1866 outbreak, where he found greatly elevated mortality for the customers of the East London Waterworks Company which, it turns out, had been cheating by continuing to draw water from the Old Ford Reservoir (which was contaminated by sewage).

Notice that in critiquing Snow's work we have appealed to some of the conditions for successful randomization laid out by Manski (1996) (for instance by questioning whether households in the study area really complied with their random "assignment" if some may have purposefully switched water companies). It might seem somewhat pedantic to focus on those rigid conditions when the randomization arises naturally. After all, Manski's conditions might at first glance assumed to speak most directly to purposeful social experimentation where randomization is explicitly introduced for impact evaluation purposes (i.e. "if you want to *perform* a successful randomized trial, such and such must happen"). However, the other way of looking at this is that Snow was effectively arguing that water supply status was as if it had been randomly assigned. In other words, he was arguing that his data were equivalent to what a purposeful experiment would have yielded. Viewed from that standpoint, it is altogether reasonable that natural experiments be scrutinized within the same framework as purposeful social experiments.

3.2.2 The RAND and Oregon Health Insurance Experiments

John Snow exploited an institutional circumstance beyond his control or design; it was a classic "natural experiment". There was no real possibility for a purposeful randomized design in his case. Partly this reflected his resources (e.g. purposefully randomizing water sources in London would have been prohibitively expensive) and the practice of his time and partly it reflected intrinsic ethical constraints: a purposeful experiment where exposure to a potentially serious risk factor was varied by design would not be ethical, in our age or his.[6] However, in some circumstances explicit and purposeful randomization (in other words, an RCT) is reasonable and logistically practical.

The largest and most elaborate (by a number of measures) explicitly randomized health policy experiment of which we are aware was the RAND Health Insurance Study (exhaustively described by Newhouse et al. (1993) and citations therein). The basic motivation of the study was the desire to learn more about the price-responsiveness of health care demand (though the study had many facets and yielded interesting insights far beyond this basic question). Learning more about the price-responsiveness of health care demand could serve a number of purposes. First, it could

[6] This is not to suggest that unethical randomization has never been attempted; unfortunately there is a long and dreadful history of such practices.

help to assess one of the most popular hypotheses regarding the dramatic increases in health care expenditures in America and other wealthy, industrialized nations in the post-WWII era, that these increases have been fueled by the expansion of health insurance coverage to broad segments of the populations of these nations. Health insurance coverage lowers the net price of care confronting consumers, and the theory suggests that they respond to this effective drop in the price of formal care by seeking more of it. However, the need to understand the price responsiveness of health care demand extended (and still extends) beyond explaining an historical curiosity: the financial implications and feasibility of various social insurance schemes designed to extend health insurance coverage (such as Medicare, Medicaid, the Medicare Drug Benefit, the Affordable Care Act, etc.) depend critically on the manner in which consumers will respond to the changes in the price of care that such programs induce.

To assess the price sensitivity of health care demand, researchers had until the RAND Health Insurance Experiment (and often since) to rely on non-randomized observational data containing information regarding the insurance status and health care utilization (measured in a variety of ways) of randomly drawn members of the population. The problem with such data is that health insurance demand is not randomly assigned: individuals generally *choose* their insurance status. For instance, in the United States they may have accepted employment involving some health insurance benefits. They may also have taken steps (e.g. drawing down their net assets) that rendered them eligible for some public insurance scheme. Finally, they may have been motivated to seek insurance due to poor health or the expectation of health problems. The insured and uninsured in such samples differ by all of these other factors and their determinants, making it difficult to ascribe differences in health or health related behaviors between the two groups to insurance alone. This makes it challenging to assess the causal impact of the price changes induced by insurance on health care demand and health. For instance, if those with coverage in real world data are less healthy than the typical individual, insured individuals would have been likely to have high health care expenditures in any case (because they are less healthy), possibly leading to an overestimate of the price sensitivity of health care demand.

The RAND experiment, which ran roughly and to varying degrees from 1971 to 1982, sought to recover a clearer understanding of the effect of insurance on health care demand (or, in other words, a clearer understanding of the price sensitivity of health care demand). Essentially, it involved randomly assigning 2,000 families at several sites to various health insurance plans. These plans were somewhat complex, but essentially varied mainly along two dimensions. First, their co-insurance rate (the percentage of medical bills paid by the family) was set to 0 (free care), 25, 50 or 95. Second, the maximum dollar expenditure for a family in a 12 month period varied between plans (5, 10 or 15 percent of family income). The various permutations led to roughly a dozen plans.

An incredible amount of thought and effort went into insuring the randomization of insurance status. For instance, the principal investigators calculated side payments for each family that insured that they would be left no worse off (financially) by the experiment (otherwise a potential confounder and disincentive to cooperate with the experiment). Nonetheless, the refusal rates across plans still varied substantially (see table 3.1). This of course threatens to introduce exactly the sort of biases that characterize non-experimental data. For instance, what if inherently less healthy individuals (a status that researchers cannot perfectly observe) were more likely to refuse more onerous plan assignments? Plan features would then become correlated with inherent health, making it more difficult to estimate the causal effect of insurance features on health care utilization or health.

Table 3.1: Refusal Rates by Plan	
Plan	**Refusal Rate (%)**
Free	8
25 and 50% coinsurance	11
95 % coinsurance	25
Source: Newhouse et al. (1993)	

Although Newhouse et al. (1993) argue forcefully (and, a reasonable person could conclude, fairly convincingly) that the variable refusal rates do not contribute substantially to biased estimates of the causal effect of health insurance status on health and health care demand, few (including, we suspect, Newhouse et al.) would argue that constant refusal rates across plans would not have been preferable. Thus, even with a carefully defined and conscientiously executed purposeful randomized trial such as the RAND effort,[7] one still sees that the basic challenge of defiance of random assignment can present itself (in much the same manner that it was often the case that London households could in fact change water companies).

Sidestepping this issue, it might seem that analysis could proceed simply by comparing mean outcomes and behaviors across plans. However, there can be a benefit to more sophisticated modelling. For instance, in some cases it can yield more precise estimates or a better fit to the data. In practice, the RAND work became the motivation for fertile discussions[8] about alternative modelling strategies. The main focus of the debate was the appropriate method for modelling health care expenditures, which typically exhibit a large mass at zero (reflecting the fact that in many intervals most households might have zero expenditures) in any given sample.

We conclude with mention of a more recent health insurance experiment (involving some of the same researchers as the RAND effort) in Oregon. In 2008, the Oregon Health Authority had funding to insure an additional 10,000 individuals through the Oregon Health Plan Standard, a Medicaid program for adults who are low-income, uninsured, "able-bodied" and not eligible for other public insurance in Oregon (see http://www.nber.org/oregon/documents.html for a full description of the Oregon Health Plan Standard and the design of the experiment). Unfortunately, nearly 90,000 expressed a desire to apply to enroll. The state decided to hold a lottery to determine who could apply for enrollment across a list of the 70,000 or so who remained after pre-lottery screening for likely eligibility. Interestingly, selection of an individual in the lottery meant that everyone in that person's *entire household* on the list was selected. An immediate implication of this is that winning individuals were more likely to come from larger households since those households had more individuals to sign up for the lottery. In the end, roughly 30,000 individuals were selected to apply, less than a third of whom would eventually be enrolled. The low enrollment rate was a result of low questionnaire return (applications were returned by around 60 percent of lottery winners) and the eventually determined ineligibility of around half of submitted applications (most often for income above the eligibility threshold). Data was then gathered from a variety of sources.

The "Oregon Experiment" experiment in some sense lies somewhere between John Snow's investigation of cholera in the Lambeth and Southwark-Vauxhall catchment areas and the RAND health insurance experiment. In the Oregon case, there was an institutional reality beyond the

[7] Newhouse et al. 1993 is worth reading for anyone planning social experiments regardless of the focus area if for no other reason than to see how much thought and effort is required for the design and execution of a seemingly straightforward experiment.

[8] It was quite a lively debate. A few of the major contributions include Duan et al. (1983), Hay and Olsen (1984), Duan et al. (1984) Maddala (1985), Duan et al. (1985) and, more recently, Dow and Norton (2003).

study designers' control (the limited number of slots in the Oregon Health Plan Standard) but they responded with a design that purposefully and explicitly generated experimental variation from that institutional opportunity.

Analyses using the Oregon such as Baicker et al. (2013) recognize that it was an imperfect experiment. Not all lottery winners became enrollees, partly out of behavioral choice: some never returned the enrollment questionnaire after winning the lottery. Thus, the Oregon Experiment could be viewed as a failed RCT, though this description seems rather harsh in light of the realities of what the researchers evaluating the program could and could not control. Within the taxonomy of samples adopted earlier, the Oregon sample would be non-experimental.

3.2.3 PROGRESA

One of the most well-known RCTs in a lower- or middle-income nation, at least in recent years, was the PROGRESA (Programa de Educacion, Salud y Alimentacion, subsequently renamed Oportunidades) program in Mexico. The random assignment of PROGRESA participation serves as a good example of the old adage that necessity is the mother of invention. PROGRESA is essentially a program offering cash grants to women on the condition that their children attend school and receive certain types of preventive health care (health education programs, nutritional supplements and formal care visits). The idea of the program is to break the inter-generational transfer of poverty in rural Mexico driven by the use of children as laborers at household enterprises (such as farms), in the process curtailing their formal education.

At the time the program was launched (in 1998), it was not financially possible to extend it to all of the villages (50,000) that would ideally receive it. Instead, policymakers selected 506 pilot communities (in the states of Guerrero, Hidalgo, Michoacan, Puebla, Queretaro, San Luis Potosi and Veracruz) and randomly assigned 320 to receive PROGRESA (see Duflo (2003), Duflo and Kremer (2003), Gertler and Boyce (2001), Buddelmeyer and Skoufias (2003) and many other authors for descriptions of PROGRESA's implementation). Baseline and follow-up data were collected for random samples of women and their children in 320 communities that participated in PROGRESA and the 186 randomly selected not to participate. This data became the basis of numerous evaluation studies that exploited the random assignment of communities to PROGRESA.

According to Duflo (2003) and Duflo and Kremer (2003), part of the reason for going ahead with the limited pilot implementation was to insure the program's political survival. Presumably, it is harder to cancel a successful program. PROGRESA's architects may have felt that the sort of randomization that they imposed would yield evidence of success (provided the program proved successful) that would have a fairly high degree of credibility. If that was the case, the strategy would seem to have paid off: the program has been widely implemented in rural Mexico, and an urban version has been introduced.

PROGRESA has achieved a tremendous degree of visibility, and attracted a great deal of research attention. Obviously, researchers have exploited the randomization design to assess the impact of the program itself (see, for example, Gertler and Boyce (2001), Skoufias and McClafferty (2001), Behrman et al. (2000, 2001), Schultz (2000a, b, c, 2001), Gertler (2000), Behrman and Hoddinott (2000), Hoddinott et al. (2000), Parker and Skoufias (2000), Teruel and Davis (2000), etc.). The evidence has generally shown that PROGRESA had an impressive effect on many of the human welfare outcomes it was designed to target. Interestingly, some researchers (e.g. Todd and Wolpin (2006), Buddelmeyer and Skoufias (2003)) have even cleverly exploited the randomization of PROGRESA to gather evidence in favor of certain non-experimental estimators.

Designs such as PROGRESA are not entirely immune from the possibility of experimental non-compliance. For instance, while communities per se cannot defy their assignment, in principle their

members could, simply by moving. This is referred to as **endogenous migration**.

3.2.4 The Work and Iron Status Evaluation (WISE)

Some RCTs in lower income societies have rather ambitiously involved purposeful randomization of participation at the individual level. In doing so they have sought not only to uncover important potential program impacts, but also to unravel some important behavioral pathways between human welfare outcomes. For instance, researchers have long sought to understand the link between health and economic prosperity. Many studies note an apparent correlation between the two. Unfortunately, the true causal effect of health on economic prosperity can be very hard to assess using real world non-experimental observational data. To begin with, there is a certain obvious potential for circularity to the relationship. It is clear that there are reasons that healthier people might be wealthier. Better health makes people more productive workers who lose less work time to illness. There are, however, equally plausible behavioral channels by which one might imagine that economic well-being influences health. For instance, it seems straightforward that wealthier people might, other things being equal, be more likely and able to make certain types of human capital investments (health care visits, better nutrition, etc.) that might improve health. Teasing out the part of the correlation between health and wealth that reflects the causal effect of health on wealth has proven to be a difficult empirical challenge.

The Work and Iron Status Evaluation (WISE) seeks to assess the effect of health on economic well-being using a randomization design. The particular channel of health on which it focuses is iron deficiency. Iron deficiency is a common health challenge in contemporary lower income societies. Additionally, there is "a very large literature in health, nutrition and biochemistry which provides a solid scientific foundation for understanding the biomedical consequences of iron deficiency"(Thomas et al. (2003), who go on to suggest Haas and Brownie (2001) for a thorough review of this literature). Thomas et al. (2003) point out that iron plays a central role in oxidative energy. This means that iron deficiency can translate into greater vulnerability to disease and fatigue. It can also impair childhood development, raise the likelihood of infant and child mortality, and reduce maximum aerobic capacity (Thomas et al. (2003)).

Iron deficiency is thus an ideal condition on which to focus because it is widespread (and thus has broad relevance) and also has substantial and well documented implications for health. WISE was an evaluation in the Purworejo district of Java, Indonesia. It involved 4,000 households, each of which were interviewed every four months for three years, beginning in 2002. The first two waves of interviews constituted baseline observations. Following the baseline interviews, households were randomly assigned to two groups: program participants and non-participants (where the "program" was receiving iron supplementation). All household members in the participant group received iron supplements on a weekly basis for a year. (The designers of the study anticipated that all members of the participant group would not be iron deficient by the end of the year, with most attaining that status in far less time.) The members of the non-participant group were not provided with iron supplementation. The surveys collected information about wealth and consumption, work, participation in community activities, health and cognitive status. Preliminary evidence (see Thomas et al. (2003)) suggests that those receiving the iron supplements were healthier and economically better off, though these initial effects seem to have been stronger for men than women.

3.2.5 Poverty Action Lab Studies

Beyond individual, stand-alone RCTs in lower income societies such as WISE and the PROGRESA work, MIT's Abdul Lateef Jameel Poverty Action Lab (http://www.povertyactionlab.org/) has

become a very active center for research involving randomized trials with a focus on development questions. The contributions of this institution are truly too numerous to do any sort of adequate justice to them (indeed, a thorough review of the Lab's achievements could be the subject of several manuals).[9] For the most part, the Lab's studies have involved purposeful RCTs, as opposed to natural experiments.

We list a few randomly selected early and recent contributions, simply to provide the roughest sort of feel for the work of the Lab:

- Miguel and Kremer (2004) consider the impact of a de-worming program by randomizing the order in which schools received it;
- Banerjee et al. (2007) examine the effect of a remedial education program that involved introducing young women from the community to provide remedial education to children in grades 2-4 who had not reached grade 1 levels of achievement. They did so by randomly introducing the program to government schools;
- Duflo et al. (2012) randomize school curriculums in Kenya to examine their implications for sexually transmitted disease infection and early fertility;
- Mobarak and Rosenzweig (2013) consider the effect of informal rainfall insurance on the demand for formal rainfall insurance and risk taking in India by randomizing villages to receive a marketing visit for a new rainfall insurance scheme;
- Ashraf et al. (2013) examine the effect of incentives (financial and non-financial) on the performance of those recruited to promote HIV prevention and sell condoms.

While these examples center on educational and health interventions, the Lab's studies have considered a wide variety of interventions and other social phenomenon. Further, it has not restricted its attention to policy questions in lower income societies. To cite one example, Bertrand and Mullainathan (2004) examine potential racism in hiring by randomizing names on resumes. The Lab has also been a focal point for methodological work exploring the possibilities of and limitations to randomized trials.

A general observation that should be made about the Lab's RCTs is that they tend to evaluate isolated theories of change (e.g. does school curriculum affect sexually transmitted disease and fertility?) or perhaps the interaction of a few, well-defined and distinguished theories of change. By contrast, much of the programming by developing country governments, as well as bi- (e.g. USAID, DFID, AUSAID, etc.) and multi- (e.g. the World Bank, Asian Development Bank, etc.) lateral donors is integrated in the sense that the overall program involve many separate areas of programming, each of which can be motivated by a distinct theory of change, delivered as a package in the hope that they will collectively influence outcomes, often per a (rather elaborate) results framework.

In principle an integrated program could be subject to an RCT through the deliberate randomization of package components across experimental subjects. In theory this could allow for identification of the average impact of different components, as well as exploration of how those impacts might depend on the specific package of other components. In practice, this has proven hard to do. For instance, an integrated program involving a suite of 4 components could potentially lead to 15 possible combinations of those components across which participation would need to be randomized, leading to 16 subsamples in the evaluation sample (one must be added to allow for a

[9] The Lab's studies are given so little real estate in this chapter because it has been so prolific that anything other than a brief overview would take over the chapter.

pure control with no exposure to program components whatsoever). As the number of components grow, evaluations could quickly grow very expensive. Nonetheless, donors could still learn much of value by evaluating the impact of the various components separately.

3.2.6 Methodological Benchmark Studies

Next we consider two important experiments in wealthy countries that have served as platforms for much methodological discussion of RCTs and impact evaluation. The National Supported Work Demonstration (NSW) is an important randomized trial, in part because data from it has been frequently utilized by researchers in the various methodological debates surrounding program impact evaluation. The NSW[10] provided temporary employment in order to give more marginal workers work experience and counseling. A sample of qualified applicants was admitted into the NSW randomly (and then received a job for 9 to 18 months), while the rest were left to provide insight on outcomes under non-participation. The program was run in the mid-1970s by the Manpower Demonstration Research Corporation at 10 sites. Qualified applicants included "AFDC women, ex-drug addicts, ex-criminal offenders, and high school dropouts"(LaLonde (1986)). The sample used by many researchers consists of 6,616 individuals (including treatment and control groups) and includes data on earnings and demographic characteristics at the baseline and every nine months thereafter.

Although there was by design some sample attrition, there was also unintended attrition that might serve to undermine impact evaluation. For instance, if more motivated people were more likely to participate in follow-up interviews, then the sample might become skewed toward those types of people (as opposed to the distribution of motivation among the overall population eligible for the program). Suppose that more motivated people are more likely to have positive labor market experiences, regardless of program participation. It could be that this would lead to a narrowing of the estimated program impact (because the sample becomes dominated by people who might have done comparatively well in either case), thus causing one to underestimate the true impact of the program when applied to a wider population. Of course, selective attrition might introduce other sorts of complications. The point is that this might lead to different inferences than might have been made had there been no selective attrition, in that sense undermining the randomization somewhat.

As a final note, consider what sort of treatment effect (of those introduced earlier in the manuscript) might be most readily estimated with the sample generated by this design. The first effect that we introduced was the average treatment effect:

$$E(Y^1 - Y^0)$$

The average treatment effect is the impact of the program on the typical or average person in the general population. In other words, it tells us the effect of the program on the randomly drawn person from a fairly broadly defined population (such as all Americans). The second was the average effect of treatment on the treated:

$$E(Y^1 - Y^0|P = 1)$$

This is the impact of the program for those actually exposed to it.

The NSW design yielded information on qualified applicants (some of whom were selected to participate in the program and a group refused the opportunity to enroll). It does not contain information about other types of individuals, such as those that might not have qualified for the program. It would thus seem reasonable to suggest that this data is most amenable to estimation

[10]This description draws heavily on LaLonde (1986).

of something akin to the average effect of treatment on the treated. Neither those enrolled in the program or the qualified applicants randomly refused enrollment into the program form a legitimate control for those enrolled. But neither group would seem particularly representative of the wider populace. Rather, they appear to be reasonably representative of those who would most likely receive such training (we offer the qualifier "reasonably" because some applicants might not have ultimately enrolled, even if their application was successful).

The National Job Training Partnership Act (JTPA) study (NJS) was designed to evaluate job training programs operating under the aegis of the JTPA. The JTPA is the most significant federal job training program in the U.S. It resembles both earlier federal training efforts and other job training programs throughout the world (Heckman et al. (1996)[11]). In the NJS, a group of accepted applicants were randomly assigned to participant and non-participant groups (with the later excluded from JTPA programs for 18 months). The study's designers also collected a sample of eligible individuals who elected not to enroll (on their own accord) as a non-experimental control. This setup thus implies randomization of participation among those who sought to participate. It should thus be clear that this is another reasonably good setup for examining the average effect of treatment on the treated (because the randomization is across those who wished to enroll in the program, as opposed to the general population), which is the program effect that Heckman and his colleagues did in fact pursue in their own work. The NJS sample was heavily utilized by Heckman and his colleagues (Heckman et al. (1996), Heckman et al. (1997a, b), Smith and Todd (2001)) in their work assessing propensity score matching estimators (which will be discussed in subsequent chapters).

The NSW and JTPA studies may have been subject to another potential complication for randomized trials. In both cases, some qualified applicants were randomly selected not to participate. What is to prevent them from trying to enroll in some other related program designed to achieve the same outcome? This is not just an academic possibility. Within any given society at any point in time, there are often many programs operating concurrently and independently to target the same human welfare outcomes for the same population subgroups. For instance, what if the program non-participants in either of these trials decided to enroll in some other training program, such as a state program? In that case we would not be able to characterize the control group as a whole as receiving no treatment. This could, for instance, lead researchers to under-estimate the effect of the program, since the outcomes experienced by some program non-participants may have been improved by enrolling in alternative programs.

Simply controlling for this possibility in a multiple regression framework might not yield the same estimates that true randomization would have provided. The problem is that non-random assignment to alternative programs within the non-participant group makes it impossible to isolate a subsample of truly randomized controls that did not experience any program intervention of any kind. Put differently, since enrollment by program non-participants in alternative programs is likely non-random, the non-participant group becomes sorted systematically by observed and unobserved characteristics into those enrolled in alternative programs and those not enrolled in any program. This non-random sorting means that the remaining group not enrolled in any program no longer has the same average characteristics as the original randomly determined program non-participant group. For exactly this reason, simply dropping those enrolled in an alternative program also would not restore the circumstances of a randomized trial.

[11] See as well Orr et al. (1995) for a thorough description of the program.

3.2.7 Negative Income Tax Experiments

The negative income tax is an idea first introduced by the American economist Milton Friedman. It can be regarded as a way of guaranteeing a minimum income. Negative income tax proposals typically involve some combination of a flat tax and an income transfer from the government. For instance, the government could impose a 30% flat tax on income, but offer a $30,000 transfer payment. An individual making nothing by their own efforts would then have an annual income of $0-.3*$0+$30,000=$30,000. An individual earning $50,000 would, on the other hand, end up with a net income of $50,000 - .3*$50,000 +30,000=$65,000. Finally, an individual making $150,000 per year would have a net after tax income of $135,000. It thus effectively amounts to a progressive tax, though in a very different manner than the income tax schemes generally in place in the contemporary wealthy, industrialized world.[12] The negative income tax is thought by proponents to have the advantage of eliminating the welfare trap. Critics claim that it creates a social safety valve that might allow employers in lower-skilled industries to pay workers far less. More generally, labor economists and others often feared that this approach might create disincentives to work.

To evaluate the negative income tax and its implications for behavior and public revenues, a number of field experiments were conducted. For instance, in 1968 the Office of Economic Opportunity, a central policy analysis body for the Johnson Administration, selected several sites in New Jersey to conduct such experiments. The experiment was overseen by the University of Wisconsin's Institute for Research on Poverty, with Mathematica Inc. handling actual field operations and data collection. Participants were randomly assigned to various permutations of the negative income tax (based on different tax rates and defined transfers). The control group was not assigned any transfer payment. Data was collected on the characteristics and behavior (e.g. labor market participation) of each. As an important footnote to history, this was to our knowledge the first social policy experiment in the U.S.

Unfortunately, the implementers of this experiment ran into a number of practical problems. Chief among them was that realistic combinations of transfers and flat tax rates made it very difficult to attract welfare recipients to the experiment, since they might be left financially worse off by their participation in the experiment. Any combination that might have removed this problem would likely have implied program parameters that, were they implemented on a broader scale, would make the program essentially infeasible.

We thus see a common pitfall of randomized trials: non-random refusal to participate in the experiment. The difficulties that the implementers of the New Jersey experiment experienced highlight another point: a relatively conceptually straightforward policy might prove very difficult to actually implement in the context of a actual randomized trial.

3.2.8 Unintended Random Experiments

We end our tour of experimental designs where we began: with the possibility that, on occasion, the messy laboratory of the human experience can provide randomization of program participation. While the potential range of sources of randomization for natural experiments may be limited only by the imagination (or the persuasive powers of the researcher attempting to argue that natural randomization design has occurred!), we discuss two examples (beyond John Snow's experience) to provide some notion of how such designs might arise naturally.

[12] Current income tax systems usually try to deliver progressivity by raising marginal tax rates with income.

Lotteries

Sometimes randomization can emerge due to lotteries that are employed to ration scarce resources or more equitably divide social burdens. A lottery determined eligibility for conscription into the U.S. military during World War II and parts of the Vietnam War. The exact details of the lottery evolved over time, but the essential scheme often involved prioritizing young men for conscription based on their birth date, making some far more likely to be inducted than others. For instance, in the 1969 Vietnam War-era lottery, each birth date was written on a slip of paper which was then placed in a plastic capsule which was in turn deposited in a bowl. The balls were then randomly withdrawn (Figure 3.4 shows Representative Alexander Pirnie drawing the first number in the 1969 draft), with the first birthdate date drawn assigned the number '1', the second '2' and so on until the 365^{th} and final date was drawn and assigned '365'. Table 3.2 provides a sampling of results from the 1969 and 1970 lotteries. The columns represent birth months and days and their draws from the 1969 and 1970 lotteries. For instance, the birth date March 7 drew 122 and 141 in the 1969 and 1970 lotteries, respectively, while the numbers for June 4 were 20 and 42.[13] Those with birth dates assigned a lower number were drafted first (in practice those drafted based on the 1969 lottery had birth dates assigned numbers of 195 or less).

Figure 3.4: The 1969 Draft Lottery Draw, courtesy of Wikimedia Commons

Ideally, a draft lottery would completely randomize participation in the military. This would imply that veterans and non-veterans would be the same on average, opening the door to the possible evaluation of the effect of military service (at least in that era) on many human welfare outcomes. Under this conceptualization, military service is the program participation experience to be evaluated.

Unfortunately, the draft lottery did not accomplish this. First, the lottery itself may have been slightly botched. For instance, it was quickly noted that November and December birth dates had overwhelmingly been assigned lower numbers in the 1969 lottery (only five birth dates in December were assigned numbers above 195, the effective cut-off number for conscription). It has been suggested that this may have resulted from the capsules not being sufficiently mixed in the bowl from which they were drawn. This could potentially introduce some complications for impact evaluation purposes since some human welfare outcomes are seasonal (to cite just one example that

[13]The draft numbers were retrieved from http://www.math.uah.edu/stat/data/Draft.html

the authors have recently encountered, schizophrenia appears to be more common among those born in colder months).

However, there were other complications. In practice there were a number of possible ways to avoid conscription, both before the lottery and even in the face of an adverse draw from the lottery. Some young men were likely in a better position to exploit those loopholes than others. These differences likely undermined the balance of characteristics between the drafted and un-drafted that one would hope true randomization might provide.

Even more importantly, however, military service status did not depend simply on the draft. Some (in fact, many) young men with a "better"lottery draw nonetheless enlisted voluntarily (even at the height of the unpopularity of the Vietnam war). During the Vietnam war the majority of the troops serving in country, and the majority of the casualties, were volunteers.

Table 3.2: Draft Lottery Numbers

M	D	1969	1970	M	D	1969	1970
3	7	122	141	9	11	158	288
8	22	339	250	11	1	19	243
4	18	90	138	6	4	20	42
7	12	15	257	7	13	42	349
5	9	197	357	12	30	3	192

What should an empirical researcher do with the volunteers? Simply ignoring them leaves out an important subgroup for gaining a comprehensive sense of the average impact of military service. However, if one includes them, what cohort could serve as a legitimate control? This example illustrates the fact that randomization mechanisms that seem straightforward at first glance can become far less appealing on closer examination of the institutional and social circumstances within which they were implemented.

Notice that there is somewhat of an analogy to the circumstances of the Oregon Health Insurance Experiment (which could just as easily have fit into the present discussion of lotteries). A lottery determined who would be allowed to apply for insurance, but behavioral choices then influenced the actual application decision among those offered the opportunity to apply. As we will see in the chapter on instrumental variables, it is possible to exploit the random element of the participation process for evaluation purposes, even if non-random factors also drove participation.

There are other interesting naturally occurring (as opposed to occurring by the explicit design of researchers) lotteries. For instance, in the U.S. magnet and other school sometimes accept students by lottery. Cullen et al. (2006) take advantage of just such a mechanism, as did Angrist et al. (2002) (though the latter considered a school voucher program in Columbia, where vouchers were distributed by lottery). Whenever one appeals to such sources of randomization, it is very important to carefully establish in a convincing fashion that the lottery did in fact represent an allocation mechanism that yields a truly random sample. Duflo (2003) discusses lotteries as a source of randomization in general, in more detail.

Happenstance Randomization

Policy implementation is certainly an imperfect process, more so in some cases than others. In certain instances, it may be reasonable to assume that individuals or other units of observation (families, communities, etc.) participated or did not participate in a program in a more or less random fashion due to accidents, incompetence or a haphazard or inconsistent adherence to program

assignment rules.

A novel application that more or less appeals to randomization by happenstance is Miller (2010). Miller's goal was to evaluate the impact of PROFAMILIA, one of the world's largest and most established family planning programs, on fertility and other human welfare outcomes in Colombia. Miller notes the difficulty in parsing out the true effect of family planning programs on fertility and other outcomes given non-random program placement. His approach to this problem is particularly notable for a novel appeal to an unusual source of exogenous program variation: arbitrary program assignment by PROFAMILIA administrators. For instance, Dr. Gonzalo Echeverry, one of the PROFAMILIA administrators during its main period of expansion (the 1960s and 1970s) related how he often decided to introduce the program into an urban community based on a chance meeting with a fellow physician "at cafes or bus stops"(Miller 2010 p. 714).

One must be very careful when assuming that such circumstances actually result in random assignment to a program. For instance, suppose that, in some fictional country X, we learned that there had been a health program that was to be assigned at the community level based on a systematic allocation rule that reflected the characteristics of the communities in that country. This might happen if policymakers have only limited resources and thus had to prioritize in terms of the communities receiving the program. To assess the impact of the program with observational data, we as researchers would normally need to observe all of the factors that guided the allocation rule and the health outcome. Otherwise the communities exposed to the program would systematically differ from those not exposed by more than just the experience of the program (namely, they would differ in terms of their average values for the characteristics that shaped the allocation rule) and we would not be able to ascribe with any degree of confidence differences in health outcomes between them to the program.[14]

Now imagine that we as researchers learned that the program's administrators had in practice ignored this rule and simply extended the program to communities as local leaders requested it, at least until they ran out of resources to extend the program any more widely. Could we necessarily conclude that program assignment was therefore random? To answer this question, one must ask whether this haphazard assignment mechanism is likely to lead to systematic (i.e. average) differences in the communities that receive the program. It is already obvious that they differ in terms of one crucial feature: whether their leaders asked for the program. It seems straightforward that leaders who asked for the program are more likely to be sensitive to the health needs of their constituents and/or operate in an institutional setting (cultural, political, etc.) that is more sensitive to health needs, demands a leadership that is more sensitive to the health needs of the populace, etc. However, these traits are likely to influence health in other ways (beyond their impact through the program). These largely unobserved community characteristics would likely contaminate any estimates of the programs impact.

Miller's (2010) argument regarding the arbitrary nature of the expansion of PROFAMILIA is compelling because he provides comprehensive evidence that the program was indeed more or less randomly assigned. For instance, he finds no statistical association between the program and the fertility or socioeconomic status of women just beyond their reproductive years when the program was introduced. If the introduction of the program did in fact vary systematically with community-level observed and unobserved characteristics that influence fertility, program exposure should then serve as a proxy for these factors. We would thus anticipate a significant (though spurious) statistical relationship between program exposure and the socioeconomic status

[14]In a multiple regression context, we could in principle regress health outcomes on a variable indicating program exposure and a set of controls for community characteristics. However, unless that set of controls included all of the variables that influence health and program allocation, our estimate of the effect of the program would be tainted by classic omitted variable bias.

and fertility of women just beyond reproductive years when it was introduced because the program variable would be picking up other community-level factors that influence these outcomes for these women. Randomization due to haphazard implementation should not be taken at face value. In that sense Miller (2010) sets a high and appropriate standard of proof when appealing to such a source of randomization.

Other Studies

In no sense has this been intended to serve as an exhaustive review of the randomization designs to date. Indeed, it misses entire institutions that have made major contributions (e.g. the efforts under the aegis of the International Initiative for Impact Evaluation[15]). Nor should the studies included be regarded as anything other than what they are: a rather random selection of important randomization driven evaluations. Nonetheless, this list should give the reader some sense of the range of studies performed, as well as common pitfalls encountered in the pursuit of randomization designs.

3.3 The Case for Randomization

3.3.1 Quasi-Experimental Estimators

Before proceeding, we must define **quasi-experimental estimators**. Quasi-experimental estimators obtain program impact estimates from non-experimental samples, for which program participation is *not* fully randomized. Broadly speaking, quasi-experimental estimators can be conveniently divided into two classes. First, there are quasi-experimental estimators that rely on non-experimental data that includes some mechanism generating random variation in program participation that is incomplete but external to the individuals for whom program impact is of interest. The most obvious example of this vein of quasi-experimental estimation is what is called instrumental variables. The other class requires random variation in program participation not external to the individual or their program participation decision. Examples might include within and matching estimators. All of these will be the subject of subsequent chapters.

They are called 'quasi'-experimental because they rely on less than complete randomization of program participation. Appealing to a quasi-experimental estimator generally requires assumptions. The logic is roughly as follows:

> *If such and such is true of the real world processes that gave rise to our non-experimental sample, then the estimates of program impact generated by this quasi-experimental estimator provide the causal impact of program participation on outcomes of interest.*

Along with assumptions, quasi-experimental estimators typically involve data requirements (i.e. certain information regarding the sample) that vary in stringency across quasi-experimental estimators.

Quasi-experimental estimators are thus the most obvious alternative to the full randomization present in experimental samples for identifying program impact. In some sense quasi-experimental estimators can be thought of as the competing approach to designs that generate experimental samples (per our earlier definition). Since most experimental samples emerge from RCTs, this is for all intents and purposes essentially the same as saying that quasi-experimental estimators are the competing approach to RCTs.

[15] See http://www.3ieimpact.org/ .

Experimental samples can support **experimental estimates** of program impact. To capture the causal impact of program participation on an outcome, experimental estimates rely on the complete randomization of program participation.

3.3.2 LaLonde's Critique

We focus on just one particular contribution to the case for randomization, LaLonde (1986), largely because it was an important catalyst for the modern advocacy of RCTs. LaLonde (1986) explores the shortcomings of quasi-experimental estimators. This seemed an appropriate route to take since the affirmative theoretical case for RCTs appeared so straightforward, at least for the purpose of estimating average program impact. The argument contained within LaLonde (1986) is sometimes referred to as **"LaLonde's critique of non-experimental estimators"** or, more simply, LaLonde's critique.

LaLonde (1986) used experimental data on the experiences of participants and non-participants in the NSW (described above) to test the effectiveness of estimators often used to evaluate such training programs with non-experimental data. Basically, those randomly assigned to participate in the NSW labor market program and to not do so provided benchmark results. That is, the RCT experimental sample provided benchmark results regarding program impact. LaLonde then explored the performance of a range of quasi-experimental estimators in terms of their ability to recover these benchmark results. Specifically, he used the sample of participants in the program (i.e. the subsample randomly selected to participate), and, following the practice of many quasi-experimental studies, used a non-randomized control group constructed from other nationally representative surveys. In other words, he crafted a non-experimental sample. The basic idea was to see how much is lost by using quasi-experimental estimators to evaluate the effect of the program with such samples.

The particulars of the quasi-experimental estimators LaLonde (1986) examined are not important at this stage (many will be discussed in detail in chapters to come). The important point is that these estimators are all designed to generate estimates of program effects with non-experimental data. LaLonde found that these quasi-experimental estimators generally did not deliver reliable estimates of the impact of the program. Perhaps more disturbingly, his results hint at the possibility that conventional specification tests that researchers at the time might have used to select an appropriate specification for the quasi-experimental estimators generally did a poor job in terms of recommending that specification which yielded results closest to those provided by the fully randomized sample. Put simply, LaLonde's work suggests that quasi-experimental estimators are not reliable.

Not surprisingly, Lalonde's critique has not gone unchallenged. Using the JTPA data to expand on LaLonde's (1986) basic exercise, Heckman and Smith (1995) point out that much better performance can be had from quasi-experimental estimators with sufficiently rich data on non-participants from the same labor markets; LaLonde's results thus partly reflect the "crudity of his data"(Heckman and Smith 1995 p. 91). They also note that the NSW sample that LaLonde used allowed him to assess only a limited range of quasi-experimental estimators and question his failure to pursue a broader range of tests to guide model selection (see also Heckman and Hotz (1989)).

That said, LaLonde's (1986) basic criticism remains persuasive to many. As time has gone by a more general argument for RCTs has emerged. It essentially focuses on the assumptions required for quasi-experimental estimates produced from non-experimental samples to be viewed as reflecting causal program impact. Estimates generated with experimental samples do not, by contrast, entail such assumptions. Aside from fundamental questions about the validity of those assumptions, a certain degree of skepticism about quasi-experimental methods has also arisen in

the face of a near endless methodological discussion about the intricacies of their performance. In short, the attractiveness of RCTs and the experimental samples they generate has in part been driven by ongoing skepticism about the reliability of the quasi-experimental methods that serve as the most obvious alternative.

3.4 Randomization and Its Discontents

Of course, skepticism can be a two-way street. RCTs have also proven controversial. The attractiveness of RCTs hinges on meeting in practice Manski's (1996) conditions for a successful experiment as well as their ability to answer a broad enough range of questions to inform policymaking usefully. Both have been called into question.

It is *generally* accepted that an experiment satisfying all of the conditions mentioned by Manski (1996) can in fact yield valid estimates of average program impact. That being understood, it should already be clear to the reader that there are a number of ways in which RCTs can be easily undone by events, in the process falling short of Manski's conditions. To begin with, generally speaking individuals cannot be compelled to fulfill their randomly assigned role in an RCT. This has several manifestations.

First, individuals cannot be forced to participate in an RCT. For instance, some refused from the outset to participate in the RAND health insurance experiment. The Negative Income Tax Experiments struggled to attract participants. Since refusal to participate is a behavioral choice, those who refuse to participate likely differ systematically from those who do not. Because of this, experimental estimates of program impact can fail to be externally valid, even when they are internally valid. In this context, internal validity refers to whether the estimates of program impact are unbiased for that population represented by those who actually take part in the RCT, while external validity essentially speaks to the usefulness of the experimental estimates of program impact for predicting behavior in the population of interest for the RCT. Somewhat distinct from this, many (e.g. Heckman (1992)) have pointed out that, since the expected return to participating in an RCT is not the same as the expected return from program participation under normal operational circumstances, those willing to participate in one might be very different from those who would enroll in a program in the everyday setting in which the program would normally operate. This is **randomization bias**.

In this vein, Heckman and Smith (1995) highlight the fact that the practical challenges of randomization sometimes necessitate a change in the types of individuals participating in an RCT. For instance, JTPA job training sites feared the "effects of randomization on the quality of the applicant pool ... to form an experimental control group, centers had to expand the set of persons deemed acceptable for the program"(p. 100). The basic idea here is that, in order to generate randomized control samples of sufficient size, sites would be forced to admit lower quality applicants (to be faithful to randomization, they could not recruit individuals explicitly for the control sample: they had to recruit them for the trial and *then* randomly assign them), thus undermining the representativeness of the sample of participants (in terms of the characteristics of participants under normal operational circumstances). If this was the case with an already large and established program, one should be even more concerned about this possibility (this can be thought of as a manifestation of randomization bias) in the setting of the limited pilot studies that are the place where RCT advocates argue most forcefully for applying their approach (e.g. Duflo 2003).

Second, individuals cannot be forced to accept their random assignment within the RCT. This can manifest itself in three ways. They can simply drop out in the face of what they perceive to be an unfavorable assignment. In the case of the RAND health insurance experiment, the

high refusal rates for the more onerous plans likely reflected this. The arguments of Newhouse et al. (1993) that this was effectively random in the RAND case notwithstanding, in general those who accept their assignment probably differ systematically from those who do not, potentially generating a correlation between participation status and the characteristics of the individual. Another manifestation is that, within the context of their selection for participation in the RCT, they can defy their random assignment. The London citizens in Snow's Grand Experiment could (in theory) and probably did sometimes decide to switch water companies. Lastly, they could attempt to participate in an RCT even if they were not selected to do so and then influence their assignment within it. Families could elect to move to communities selected for treatment under PROGRESA.

Third, those assigned to the program non-participant group might seek some alternative intervention. This is called **substitution bias** (Heckman and Smith 1995). For instance, the children randomized not to receive a de-worming treatment in the context of an RCT could seek one from some other source outside of the RCT. This may serve to bias downwards experimental estimates of the true effect of the program under evaluation.

Finally, they do not have to adhere to their random assignment or remain participants in the RCT. This is especially problematic since many RCTs involve an initial assignment to a program participation status and then observation of individuals' outcomes over time. Even those who agree initially to participate and accept their assignment can decide to drop out with the passage of time. When it is non-random (which is probably nearly always the case) it is called **selective attrition**. Individuals who accepted their experimental assignment might also defy it with time. For instance, some of the control subjects in the WISE experiment might conceivably have decided to take iron supplementation on their own.

The defiance of experimental assignment has led some to focus on alternative evaluation parameters (see Tool Box on Alternative Evaluation Parameters). While this might seem natural alternatives, they involve a huge cost: losing our focus on the parameters that originally motivated the RCT. Moreover, it is really unclear how in any given evaluation what information, if any, these alternative parameters provide about average program impact.

Tool Box: Alternative Evaluation Parameters

Intention to Treat (ITT) analysis compares the outcomes of interest by assignment to the participant and non-participant group, rather than actual participation status. Intention to treat analysis yields the impact of *treatment assignment* (which is not the same as treatment because some have an actual treatment status that does not match their treatment assignment).

Per Protocol analysis generally involves restricting attention to those who adhere to their participation assignment. Because those who do adhere to their assignment are probably a non-random sub-sample of those selected to participate in an RCT, it is unclear how generalizable their experiences may be. Indeed, if refusal of randomly assigned status depends on assignment, their experiences may not even support internally valid estimates of impact. Clearly, intention to treat and per protocol analysis do not seek to estimate the same treatment impact parameters, and hence are not the same thing (a point about which there is sometimes some degree of confusion).

All of these potential complications of RCTs have led to a certain degree of skepticism regarding their true credibility as estimators of average program impacts for the population of interest. In

terms of whether their results are credible, reflecting on Hausman and Wise (1985) and Manski and Garfinkel (1992), Manski (1996) writes "these assumptions [ie those on which RCTs rest] may sometimes be appropriate but very often there is good reason to doubt their realism ... Many of the contributors to the volumes edited by Hausman and Wise (1985) and Manski and Garfinkel (1992) find that the classical argument for random assignment is not credible when applied to recent experimental evaluations of income maintenance, welfare, training, and other social programs"(p.711). Manski goes on to explore the inferences that can be made with purportedly experimental data when one or more of the classical assumptions justifying RCTs fails to obtain, and finds the picture significantly complicated. For our purposes, Manski (1996) makes several important points. First, treating data as experimental does in fact require making certain assumptions (listed earlier in this chapter). Second, these assumptions are, even in the case of RCTs, often not credible. Third, when they fail to hold, the inferences one can reasonably and easily make with the resulting sample fall considerably short of what was originally advertised.

In many applications there is probably no way to operationalize tests against all of the possible ways that randomization can fail and we are unaware of any omnibus test that covers all possibilities. Moreover, it isn't really clear exactly how RCTs could be improved in light of these possible failures. As Manski (1996) puts it:

> "An initially appealing but ultimately empty response to this situation is to say that we should strive to design and execute better experiments. This response is ultimately empty because it is logically impossible to observe outcomes under alternative programs. Each member of the population actually receives exactly one treatment in one program. All attempts to compare programs, whether based on experimental or non-experimental data, face the problem of counterfactual inference."(p. 730).

In this respect experimental estimates may not in practice always be so conceptually different from quasi-experimental estimates derived from non-experimental data: both approaches rely on assumptions that may or may not be credible or testable, depending on the empirical objective and the context (the sample, institutional environment, etc.).

Heckman and Smith (1995) offer a very useful list of cautionary points regarding the utility of RCTs in terms of generating sufficiently useful, relevant information to meaningfully inform programming. These include:

1. RCTs cannot estimate many important program impact parameters. For instance, while it is clear that one can in theory easily recover average program impact with experimental data, it is generally not so straightforward to estimate median or quantile effects, the proportion of participants experiencing a positive program impact, etc. Heckman and Smith offer one rather dramatic instance of an experimental sample with which one cannot, given the mean program effect, distinguish between a situation where the program had a beneficial impact for many and a detrimental impact for many others, and a situation where the program serves and harms only a few.

   Suppose that we want to estimate the average treatment effect $E(Y^1 - Y^0)$. By the properties of expectations, $E(Y^1 - Y^0) = E(Y^1) - E(Y^0)$. Experimental data insures that program participants and non-participants are alike on average (with the exception of their program participation experience). Then, the average of Y^0 for those not participating provides a valid estimate of $E(Y^0)$ for all, while the average of Y^1 for those participating provides a valid estimate of $E(Y^1)$ for all. Experimental data thus allows us to make inferences about the average program impact without building an empirical model of individual outcomes (as some quasi-experimental estimators do at the cost of assumptions).

But such quasi-experimental models can provide some payoff for those additional assumptions. Certain such models provide the joint distribution of Y^0 *and* Y^1. In other words, assuming that the model is correct (i.e. the assumptions are justified), they provide us with a way of obtaining a valid estimate of Y^0 and Y^1 for *each* observation in our sample. This is exactly the information one needs to estimate a median program effect. Letting *med* indicate "median."The basic challenge for forming an experimental estimate of median program effects is that

$$med(Y^1 - Y^0) \neq med(Y^1) - med(Y^0)$$

Thus, even with experimental samples there is no easy way to estimate median program effects without resorting to exactly the sort of quasi-experimental assumptions which the randomized trial approach had sought to avoid;

2. RCTs carried out in the context of a limited pilot implementation of a program raise several concerns, perhaps the most importance of which is that the experimental estimates must be regarded as estimating strictly partial equilibrium effects (see, for example, Heckman et al. (1998)). This means that the evaluation of the pilot program takes as given all sorts of broader environmental factors that may themselves be changed as the program expands and begins to effect, in equilibrium, aggregate circumstances. However, these changes to the environment within which the program operates may serve to limit the degree to which results obtained from the pilot study can be generalized. To be sure this is potentially a concern with non-experimental samples as well. However, evaluations with such non-experimental samples tend to involve more often active programs already instituted on a fairly wide scale, thus reducing this concern;

3. In many instances, there are real institutional constraints on the implementation of purposeful RCTs. It is not straightforward *how* to randomize meaningfully participation in many programs. Heckman and Smith discuss the stages involved in program enrollment, and point out that the questions that RCTs can answer hinge on the stage at which the randomization occurs (see as well Angrist and Imbens (1991) and Heckman and Smith (1993)).

All of these points speak to the possible limits to the usefulness of the information generated from experimental samples, even if internally valid estimates of some average treatment effects are possible with them.

Deaton (2010) has recently tossed a few more logs on the fire regarding the usefulness of the information RCTs yield. He seems to feel that RCTs have been somewhat oversold. He offers several distinct critiques. Some of his points effectively echo the concerns raised by Heckman and Smith (1995). However, he emphasizes a number of other reasons for concern regarding RCTs, both practical and philosophical.

Perhaps most importantly, Deaton (2010) emphasizes that RCTs cannot shed any light onto the process behind program impact. This is in some sense a natural extension of Heckman and Smith's (1995) concern about the limited information that randomization can yield. Deaton's (2010) criticism could apply to most quasi-experimental approaches as well. Estimation models for quasi-experimental estimators, like experimental estimates, tend not to be directly and explicitly motivated by theoretical or behavioral models within which program participation and outcomes of interest are linked via carefully outlined pathways by which the former influences the latter. They simply strive to capture some statistical average impact measure by teasing out some random channel of otherwise non-randomly determined program participation. Randomization is in that sense the ultimate example of a purely statistical, atheoretical evaluation approach that seeks simply

to recover average impacts. However, without explicitly considering, for instance, the pathways by which programs influence outcomes, we cannot know how they work (in the sense of having a positive average impact) if they do or why they failed (in the sense of having zero or even negative average impact) if they did not work. Similarly, without a model that captures the heterogeneity of the processes across the population, one cannot know for whom they worked within the population or whether they worked for different reasons for different types of members of the population. Sometimes this may not make a difference (e.g. the salutary adjustments to water quality made based on John Snow's work essentially preceded the modern germ theory of disease). But for many other sorts of programs this information would be critical for achieving more effective program delivery or more effective targeting of the program within the population.

Tool Box: Data Designs

A **pre-post** design is one in which observations are made before and after the start or application of treatment/participation.
A **post-only** design involves observation only after the initiation of treatment/participation.
A design in which both participants and non-participants (alternatively, the treated and un-treated) are observed is called a **treatment-control** design.
Data for which participants and non-participants are observed before and after the initiation of program exposure for participants is often referred to as a **treatment-control pre-post** design.

3.5 Estimation Methods

It might seem intuitive that experimental samples can yield estimates of average treatment effects[16] simply by averaging Y^1 across the program participant sample and then subtracting from that the average of Y^0 across the program non-participant sample. Indeed, this is sometimes more or less the practice.

In practice, however, average program impacts are frequently estimated with more elaborate modelling. Indeed, somewhat ironically, average program impact is often estimated from experimental samples with quasi-experimental models. Sometimes, modelling can simply generate more precise estimates of program impact. For this purpose multiple regression has been a popular choice, though recent work has suggested better alternatives. Alternatively, the desire to model can reflect some unusual feature to the outcome distribution of $\{Y^1, Y^0\}$. We have already seen an example with the RAND Health Insurance experiment, where one goal was to estimate the impact of insurance status on health care expenditures, which in their distribution often exhibit pronounced mass at zero, reflecting the fact that in any given observational interval many households and individuals might have no health care expenditures. The RAND team's preferred solution to this was something called the two-part model. Sometimes the modelling strategy has the attractive feature of providing something of a test of randomization. A ready example is the frequent application of the difference-in-differences model to treatment-control pre-post randomized samples (see Tool Box on Data Designs). In no sense does this undermine the estimates as experimental per se. The goal is usually the better fit and precision these more elaborate quasi-experimental models might offer. In this context, the estimates are not regarded as relying on the non-experimental

[16] As we have seen, depending on the design of the experiment they might more properly be thought of as the average effect of treatment on the treated.

data assumptions of these quasi-experimental estimators. Rather, complete randomization is still viewed as the source of identification of average program effects. This may be less true in instances where the quasi-experimental methods are applied to obtain impact parameters beyond average program impact. In such instances the estimates should generally not be viewed as resting exclusively on an assumption of full randomization (since that might not be sufficient to identify the impact parameter in question).

Quasi-experimental estimators can offer some recourse when randomization of program participation is less than complete. For instance, we have seen cases with the Oregon Health Insurance Experiment and the draft lottery where there is a genuinely random element to assignment, but final assignment is still to some degree a behavioral choice. In these instances appeal has frequently been made to a modelling strategy called instrumental variables.

We will discuss all of these methods in subsequent chapters. We did not do so in this chapter because a full understanding of a quasi-experimental model's implications for analyzing experimental samples requires a complete technical discussion of that model, which would have been a digression in the context of this chapter.

3.6 Some Closing Thoughts

Purposeful social experiments (a.k.a. randomized control trials) are currently experiencing a "Golden Age" of popularity as an impact evaluation tool in international development. Randomized control trials provide perhaps the most theoretically compelling solution to the challenge of program impact evaluation: insure that the participants and non-participants differ by only the experience of program participation by eliminating their capability to select themselves into the participant and non-participant groups (indeed, by randomly assigning them to the participant and non-participant groups) and, by doing so, introduce ways that they could differ other than just the experience of participation. As a result of random assignment to the participant and non-participant groups, any average difference in the outcome of interest between the two can be ascribed to the experience of program participation. In other words, program participation *caused* any observed differences in outcomes between the participant and non-participant groups. The promise of the randomized control trial approach is thus immense, opening the way for the possibility of clear cut evidence for guiding policymaking where all too often there have been muddled or confusing messages and, frankly, in many instances nothing but great story telling.

At the same time, this approach has generated some degree of reasonable skepticism. It seems to the authors that in many arenas of human welfare policymaking randomization seems to follow something of a common arc as it grows increasingly popular for its promise of uncomplicated and uncontroversial inference regarding program impact, followed by something of a period of skepticism as many of the practical limitations become apparent. These limitations are worth reviewing.

First, many parameters of interest simply cannot be estimated via the randomization approach. We cited the specific and simple example of median treatment effects, but the list extends to virtually any parameter where the calculations of impact for the participant and non-participant groups cannot be additively separated. Perhaps as importantly, many sorts of interventions simply cannot be evaluated via this method for ethical or practical reasons. For instance, it becomes immediately obvious that randomization is a tough ethical sell if it seems persuasive *ex ante* that either participants or non-participants would likely be harmed by their assignment. For instance, a relatively straightforward way to resolve the debate over whether vaccines cause autism would be to withhold vaccinations for a randomly chosen group of children, but one does so with the near certainty that some of those children would become very ill or even die without the benefit of the

randomly withheld vaccinations.

More subtly, many of the programs that program implementing stakeholders (for instance, development agencies such as the World Bank, USAID, the Gates Foundation, etc.) need evaluated are not amenable to randomization. First, they might be established programs. Second, they often involve integrated packages of interventions that would be impractical to fully randomize. Finally, the realities of the funding cycles of donors often render impractical the notion of a preliminary randomized control trial before full roll out of the program. It is telling that many of the persuasive randomized evaluations of, for instance, the Poverty Action Lab have involved relatively simple interventions involving little or no integration of distinct program components and were being examined outside of the context of the normal formal programming process of donors or other stake holders (such as national governments). Indeed, one could argue that they are often evaluating not so much programs per say as "theories of change". The researchers at the Lab would probably argue that this is not a reason not to pursue randomization as an ultimate standard of program impact evaluation (and they would have a point about that) but that leaves open the question of how to conduct evaluations under the more challenging realities of today's programming environment. In short, the "randomization or nothing" tack (which, to be sure, the Lab has not advocated) would simply leave us with no information about the implications of many if not most of the interventions potentially shaping (for better or worse) human welfare in the world today.

Second, and on a somewhat more discouraging level, randomized designs can be very hard to execute. We highlighted the shortcomings of, for instance, the RAND Health Insurance experiment sample. A hasty, but in the end rather unpersuasive, answer to the uneven drop out rates that the RAND study team encountered would be to try harder to avoid these complications when executing future randomized control trials. In fact, the RAND study team put an enormous amount of careful planning and effort into protecting the integrity of that experiment. It is not obvious what, exactly, they could have done to avoid the uneven participation rates encountered. As long as individuals have a say in their own affairs (including the experiments in which they do and do not participate) it is frankly impossible to insure absolutely the integrity of randomization.

Randomized control trials are an extremely powerful tool for program impact evaluation that hold the potential to offer persuasive evidence regarding the efficacy of many sorts of programs and interventions where none has existed to this point. But it is already clear that randomization cannot be taken for granted, and this method cannot answer all of the questions that urgently need answering as we attempt to allocate scarce resources across alternative programs.

Chapter 4

Selection on Observables

We have now reached a major turning point in our discussion of program impact evaluation methods. For the remainder of this manual the focus will be on quasi-experimental methods. We will discuss these quasi-experimental models in the context for which they were originally developed: the estimation of causal program impact with non-experimental samples. We defined non-experimental samples in Chapter 3 as ones within which program participation is not completely randomized. As such, they involve some channel of non-random determination of program participation, even if there were random elements influencing participation as well.

Program impact is the change in some outcome of interest (which could be a behavior, actual outcome, etc.) *caused* by program participation. Quasi-experimental estimators recover valid estimates of program impact from non-experimental samples by assuming the presence of some channel of random variation in program participation out of the total variation in participation, which is on the whole not fully random. Sometimes, this is equivalent to assuming some hypothesized mechanism for the failure of fully randomized program participation in the observed sample. The credibility of a particular quasi-experimental estimator in the context of a given impact evaluation using non-experimental data essentially reflects the plausibility (in that context) of the assumptions on which that estimator is premised.

Basic and arguably reasonable hypothetical guidelines for comparing quasi-experimental estimates (assuming that available data allows pursuit of more than one method) might be as follows:

1. Given equally credible assumptions across models, results yielded by methods requiring fewer assumptions are likely more credible;

2. Given models with an equal number of required assumptions, results yielded by models featuring particularly strong assumptions are less credible.

Of course, this does not always suggest a clearly preferred quasi-experimental estimate since each rule is based on a *ceteris paribus* (other things being equal) condition unlikely to be met in practice. The first condition requires a careful understanding of each quasi-experimental method. This will allow practitioners to enumerate clearly and comprehensively the assumptions behind a given method. The second condition should, wherever possible, be informed by any available statistical test that might allow for assessment of the assumption (as opposed to reliance on arbitrary opinion). In any case, best practice requires transparency in terms of the assumptions behind a given quasi-experimental estimate. This is crucial for assessing the credibility of that estimate.

Despite this focus on the application of quasi-experimental methods to non-experimental data to recover program impact, quasi-experimental methods are sometimes applied to the task of generating average program impact estimates using experimental samples. In the last chapter we

defined experimental samples as featuring the complete randomization of program participation, most naturally through a purposeful social experiment where randomized program participation is explicitly engineered (i.e. an RCT). Where appropriate such applications of quasi-experimental methods are discussed after the given quasi-experimental method has first been fully discussed in the non-experimental data setting.

In this chapter, we discuss methods that rely critically on a "selection on observables" assumption to identify program impact with non-experimental samples. This assumption means essentially that we can observe in our data all of the non-random factors that guide the outcome of interest and the program participation decision. If we were able to observe all of these factors we could, by a variety of methods, control for their role in shaping the outcome of interest. The remaining variation in that outcome would then be that driven by program participation and purely random variation. Since the latter would presumably be the same on average between participants and non-participants (and hence any representative samples of the two) the average difference in outcomes between the two groups would thus reflect the impact of program participation.

Broadly (and, frankly, crudely) speaking, "selection on observables" methods fall into three broad categories: multiple regression modelling; matching; various combinations of the two. We begin by discussing regression, since it is probably the most widely used (both historically and, taking the broad view across social science disciplines, even presently) strategy in the class of methods relying on the "selection on observables" assumption. We then move on to matching methods, and discuss the link between the two approaches.

This chapter will not offer an exhaustive, detailed coverage of all of the various specific impact estimation approaches relying on the "selection on observables" assumption. First, although the above categorization of the models relying on this assumption is arguably not unreasonable, it is very crude in the sense that there have been an incredible number of specific variations on each, in recent years particularly where matching is concerned. Second, the methodological literature is ever-evolving, offering seemingly unending new variations on these approaches to modelling under this assumption.

Instead, the focus in this chapter will be the broad properties of this class of impact evaluation estimators. These properties generally apply, in one fashion or another, to most of the specific twists that have been developed in this vein of the impact evaluation literature. Indeed, a major theme of the chapter is that the empirical differences between the various methods relying on a "selection on observables" assumption are likely of secondary importance to their common foundation in that assumption.

4.1 Regression

The first major estimation method relying on a "selection on observables" assumption that we consider is multiple regression. This is a technique that we have already discussed in Chapter 2 in the context of exploring the basic challenge of program impact evaluation. In that earlier discussion, we learned that regression can be used as a technique to control for factors (in the earlier discussion we focused on just one factor) that might be related both to program participation and the outcome of interest. Moreover, it was demonstrated that the failure to include that factor that influenced program participation and the outcome of interest in the regression resulted in a "biased" estimate of program impact. We will now review and elaborate a bit on that simple case before moving on to the more general case of multiple regression, wherein we can control for numerous factors influencing both program participation and the outcome of interest, before concluding with a caveat about adding controls, as well as a discussion of estimating the precisions of estimates and

the application of the method to experimental data. The regression discussion is one of the longer and more detailed of the manual. Though the technical discussion (and, to be sure, math) can at times seem tedious, a thorough understanding of the mechanics of regression provides a solid foundation for understanding the rest of the topics discussed in the manual.

4.1.1 Regression Basics: The Simple Case

In the introduction we suggested that the credibility of an estimate of program impact provided by a quasi-experimental estimator of program impact applied to non-experimental data hinges on the assumptions behind that estimator. This subsection explores the precise assumptions required to interpret the estimate of program impact emerging from multiple regression analysis as actually reflecting the change in the outcome of interest *caused* by program participation (i.e. true program impact).

We begin by developing a very simple model of program impact. To do so, we appeal to the basic potential outcomes framework. We define Y^1 as the outcome of interest an individual experiences when he or she participates in a program and Y^0 as the outcome of interest he or she experiences when they do not do so. That individual's program impact is then

$$Y^1 - Y^0$$

In other words, the individual's program impact is simply the difference in outcomes he or she would experience were they to participate in the program and were he or she not to do so. Suppose that the potential outcomes of each individual were determined by

$$Y^0 = \beta_0 + \epsilon$$

$$Y^1 = \beta_0 + \beta_1 + \epsilon$$

The βs are population parameters that we cannot observe but might be able to *estimate*. ϵ is an unobservable, purely random component of Y^0 and Y^1 that we will assume has an expected value of zero (i.e. $E(\epsilon) = 0$).

For the purposes of this subsection, we assume that we are interested in learning the average program impact for the general population.[1] The program impact for that population is the average treatment effect across it:

$$\begin{aligned} E\left(Y^1 - Y^0\right) &= E\left(Y^1\right) - E\left(Y^0\right) \\ &= E(\beta_0 + \beta_1 + \epsilon) - E(\beta_0 + \epsilon) \\ &= \beta_0 + \beta_1 + E(\epsilon) - \beta_0 - E(\epsilon) \\ &= \beta_0 + \beta_1 + 0 - \beta_0 - 0 \\ &= \beta_1 \end{aligned}$$

β_1 is thus the population program impact that we wish to learn about through program impact estimation.

The observed outcome is

$$Y = P \cdot Y^1 + (1 - P) \cdot Y^0$$

where P equals 1 if an individual participates in the program and 0 otherwise. Inserting the formulas for Y^1 and Y^0 we have

$$Y = P \cdot Y^1 + (1 - P) \cdot Y^0$$

[1] As opposed to some specific subpopulation, particularly program participants.

$$= P \cdot (\beta_0 + \beta_1 + \epsilon) + (1 - P) \cdot (\beta_0 + \epsilon)$$
$$= P \cdot \beta_0 + \beta_0 - P \cdot \beta_0 + P \cdot \beta_1 + P \cdot \epsilon + \epsilon - P \cdot \epsilon$$
$$= \beta_0 + P \cdot \beta_1 + \epsilon$$

This is a model by which the observed outcome Y is determined by program participation status P, unobserved (but hopefully estimable) population parameters β_0 and β_1 and an individual-specific purely random factor ϵ. Note that at this point we have said nothing about how P is determined, particularly whether the program participation decision is in any way related to the random element ϵ.

Regression is, simply put, a method for estimating the parameters of relationships between variables. We have just derived a simple model

$$Y = \beta_0 + \beta_1 \cdot P + \epsilon$$

Through the regression approach to program impact evaluation we could obtain estimates of β_0 and β_1. Clearly, the hope is that the estimate of β_1 would indeed reflect what that parameter represents: program impact. It would thus hopefully reflect *only* the causal effect of program participation P on Y: the change in Y that occurs when the participation decision "switches" an individual from Y^0 to Y^1 (i.e. changing from $P = 0$ to $P = 1$). Perhaps the most important issue before us then are the conditions we must assume hold for the regression estimate of β_1 to be interpreted as providing the causal impact of program participation, P, on Y.

Suppose that we have a sample of N units of observation. We use this formal term, but usually this means a sample of individuals, since readers of this manual are probably typically concerned with the impact of human welfare programs on individuals.[2] For expositional simplicity, for the rest of this manual we will assume simply that the units of observation are individuals. For each individual in our sample we observe Y and P. Specifically, we observe $\{Y_i, P_i\}$ for each of the $i = 1, ..., N$ individuals in our sample.

We can then apply the regression approach to this sample to form estimates of β_0 and β_1. These estimates are represented by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. There are a variety of ways of forming the regression estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Three popular alternatives are ordinary least squares estimation, method of moments estimation, and maximum likelihood estimation. The differences between these approaches are not of great importance for the discussion to follow. We will touch on maximum likelihood estimation a bit later in this chapter, and discuss the method of moments in a subsequent chapter the focus of which is instrumental variables methods for program impact evaluation. For now, we concentrate on the (relatively) simple approach of ordinary least squares.

To begin with, the estimate of the random component ϵ for the i^{th} individual in the sample per the regression estimates is

$$\hat{\epsilon_i} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i$$

Although we have to this point described ϵ as the random component of $\{Y^0, Y^1\}$ (and hence Y), in the language of regression this is often referred to as the **regression residual** or **regression error term**. These terms are motivated by the idea that ϵ is some factor driving Y beyond the variation generated systematically by P_i given the parameters $\{\beta_0, \beta_1\}$. It is the residual remaining beyond the variation that can be explained by P, and can also be considered as an error term that is the difference between actual Y and what the value of P predicts Y should be given $\{\beta_0, \beta_1\}$.

[2] To be sure, the units of observations could be other entities. For example, we might be interested in the impact of a program on firm or organization behavior.

Ordinary least squares regression involves finding the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of estimated squared residuals across the sample. Thus, given the data described above, ordinary least squares finds the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{N} \hat{\epsilon_i}^2 = \sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i \right)^2$$

One way of thinking about this is that we seek to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively, that minimize the estimated role of the random component ϵ in explaining variation in Y. In other words, we seek the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize the variation in Y explained by P, and in so doing capture fully the systematic role of P in shaping Y.

It turns out that the solutions to this minimization problem are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (Y_i - \overline{Y})}{\sum_{i=1}^{N}(P_i - \overline{P})^2} = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{P}$$

where

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

and

$$\overline{P} = \frac{\sum_{i=1}^{N} P_i}{N}$$

Because of these specific formulas by which they are computed given a specific sample (containing observations of $\{Y_i, P_i\}$ for $i = 1, ..., N$), we refer to $\hat{\beta}_0$ and $\hat{\beta}_1$ not just as estimates but also **estimators** of β_0 and β_1, respectively. (By comparison, β_0 and β_1, the parameters we wish to estimate, are the **estimands**.[3]) When we refer to them as *estimators*, we mean the formula for estimating β_0 and β_1. When we refer to them as *estimates*, we mean the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that these formulas yield given a specific sample (containing observations $\{Y_i, P_i\}$ for $i = 1, ..., N$).

In chapter 2, we saw two methods of assessing whether $\hat{\beta}_1$ is capturing the causal effect of P on Y: we can examine whether it is either an unbiased or consistent estimator of the effect of program participation on the outcome of interest. To begin with, we must recognize that $\hat{\beta}_1$ is a random variable, if for no other reason than that different samples will have, by random variation, somewhat different mixes of Y_i and P_i and, perhaps, different specific sizes N. Thus, the value of $\hat{\beta}_1$ will vary somewhat from sample to sample. Given that $\hat{\beta}_1$ is then a random variable, in Chapter 2 we learned the subtly different meanings of consistency and unbiasedness:

1. **Unbiasedness:** Because it is a random variable the estimator $\hat{\beta}_1$ will have an expectation $E(\hat{\beta}_1)$. In general, an estimator is unbiased if its expectation $E(\cdot)$ is in fact the true value of the parameter for which we wish to form an estimate. Per the fancy terminology above, an estimator is unbiased if its expected value is indeed the estimand. In this case, we wish to form an estimate of the impact of P on Y, which in our simple regression model is β_1. Therefore, the estimator $\hat{\beta}_1$ is unbiased if $E(\hat{\beta}_1) = \beta_1$. What does this mean in plain terms? An easy way to think about this is to conceptualize drawing many, many samples of size N and forming the estimate $\hat{\beta}_1$ for each sample. If the average of these estimates across these many samples is β_1, then $\hat{\beta}_1$ is an unbiased estimator of β_1. For this reason, unbiasedness is sometimes referred to as being *right on average*;

[3]The term estimand is often used a bit loosely in practice. Sometimes it would seem to mean what an estimator is *actually* estimating, at others it would seem to indicate what one *wishes* to estimate. We adopt the latter definition.

2. **Consistency:** Because an estimator is a random variable, it has a probability distribution. In other words, it has probabilities attached to the possible values that it can take in a given sample. An estimator is consistent if, as sample size increases, the values that it can take on with positive probability become increasingly concentrated on the true value of the parameter one wishes to estimate. In the case of an estimator with continuous range, it is consistent if its probability density collapses around the parameter we wish to estimate as sample size grows. A somewhat crude but useful way of thinking about consistency is that it establishes that values for the estimator that deviate from that of the true parameter we wish to estimate become increasingly improbable as sample size increases.

Both concepts speak to the idea of a central tendency of the estimator but involve different ways of approaching the concept.

Unbiasedness involves the average value of the estimates generated by an estimator across many samples of the same size. It is a property that therefore holds regardless of sample size. Consistency, on the other hand, is concerned with the probability distribution of the estimator as the sample size becomes increasingly large. Figure 4.1 provides an illustration of the basic dynamics of consistency for an estimator $\hat{\mu}$ of some population parameter μ. At a lower sample size N, the probability density of the estimator $\hat{\mu}$ is concentrated around a value other than the true population parameter (and hence presumably biased). However, as the sample size N grows, the density begins to concentrate or collapse around the true population value for the parameter. Consistency does not necessarily imply unbiasedness nor does unbiasedness necessarily imply consistency.[4]

We will start by focusing on the conditions under which $\hat{\beta}_1$ is an unbiased estimator of the impact of program participation P on Y. (This is an entirely reasonable departure point since we are rarely blessed with samples of infinite size.) We start this process simply enough, by writing the expectation down:

$$E\left(\hat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

The next step is to insert the true model outlining the causal link between P and Y by substituting out Y_i:

$$E\left(\hat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (\beta_0 + \beta_1 \cdot P_i + \epsilon_i)}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

Multiplying out, we have

$$E\left(\hat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \beta_0}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \beta_1 \cdot P_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

β_0 and β_1 are population parameters from the regression model and as such basically mathematically operate as constants:

$$E\left(\hat{\beta}_1\right) = \beta_0 \cdot E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P})}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + \beta_1 \cdot E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot P_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

[4] To provide some examples of these possibilities, suppose you have a sample of N observations of a variable Z. Then

$$\frac{\sum_{i=1}^{N} Z_i}{N} + \frac{1}{N}$$

is a consistent but biased estimator of the population mean of Z. On the other hand, Z_1 (i.e. the first observation on Z in a given sample) is an unbiased but inconsistent estimator of the population mean of Z. For the technically inclined, any estimator that is unbiased and also converges (convergence is a topic we will discuss below) to a constant is consistent.

We next note that $\sum_{i=1}^{N}(P_i - \overline{P}) = 0$ and, therefore,

$$E\left(\hat{\beta}_1\right) = \beta_0 \cdot E\left(\frac{0}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + \beta_1 \cdot E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot P_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

$$= \beta_1 \cdot E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot P_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

Finally, we note that $\sum_{i=1}^{N}(P_i - \overline{P}) \cdot P_i = \sum_{i=1}^{N}(P_i - \overline{P})^2$. Therefore

$$E\left(\hat{\beta}_1\right) = \beta_1 \cdot E\left(1\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

$$= \beta_1 + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

We have thus established that whether $E(\hat{\beta}_1) = \beta_1$ hinges on whether

$$E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) = 0$$

What remains is to explore what this condition means.

Let us begin by recalling a key property of expectations:

$$E\left(Z_1 \cdot Z_2\right) = E\left(Z_1\right) \cdot E\left(Z_2\right)$$

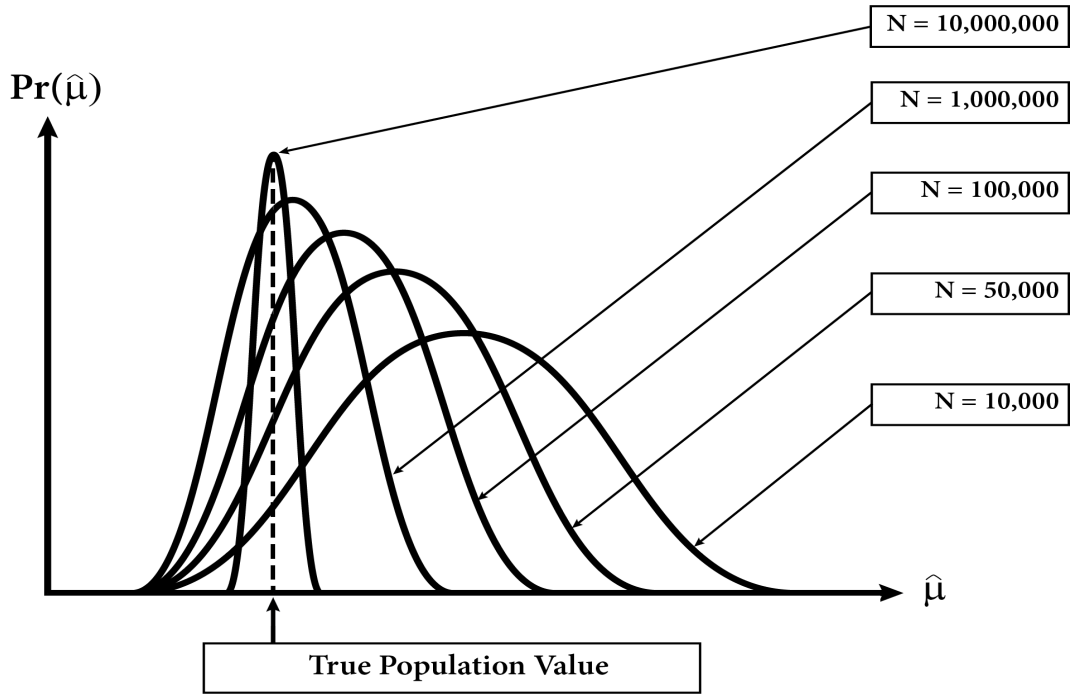


Figure 4.1: Consistency

only if Z_1 and Z_2 are independent. If we assume P and ϵ are independent, we could write

$$E\left(\frac{\sum_{i=1}^{N}(P_i-\overline{P})\cdot\epsilon_i}{\sum_{i=1}^{N}(P_i-\overline{P})^2}\right)=\sum_{i=1}^{N}\left(E\left(\frac{\sum_{i=1}^{N}(P_i-\overline{P})}{\sum_{i=1}^{N}(P_i-\overline{P})^2}\right)\cdot E\left(\epsilon_i\right)\right)$$

Since $E(\epsilon)=0$, independence of P and ϵ insures that

$$E\left(\frac{\sum_{i=1}^{N}(P_i-\overline{P})\cdot\epsilon_i}{\sum_{i=1}^{N}(P_i-\overline{P})^2}\right)=0$$

In other words, independence of program participation P and ϵ insures that $E(\hat{\beta}_1)=\beta_1$ (in other words, that the least squares estimator provides an estimate of true, causal program impact). We should be clear, however, what full independence means: it means that the regressor P is completely independent of the error term ϵ. In other words, it means that all of the Ps ($\{P_1,P_2,...,P_N\}$) are independent, in every sense, of all of the ϵs ($\{\epsilon_1,\epsilon_2,...\epsilon_N\}$).

Unbiasedness can actually be established with the slightly weaker condition of **mean independence** of P and ϵ. Specifically, let

$$\vec{P}=\begin{bmatrix}P_1\\P_2\\\cdot\\\cdot\\\cdot\\P_N\end{bmatrix}$$

In other words, $\vec{P}$ is just the set of values for P_i that occur across all of the $i=1,...,N$ observations in our sample. Unbiasedness requires that ϵ_i is mean independent of $\vec{P}$, or

$$E\left(\epsilon_i|\vec{P}\right)=E\left(\epsilon_i\right)=0$$

for all $i=1,...,N$ observations in the sample.[5] In other words, none of $\{P_1,P_2,...,P_N\}$ reveals any information about the expected value of ϵ_i.[6]

To show that this condition is sufficient to establish the unbiasedness of $\hat{\beta}_1$ requires appeal to something called the **Law of Iterated Expectations**, which states that, for two variables Z and W,

$$E\left(W\right)=E_Z\left(E\left(W|Z\right)\right)$$

where $E_Z\left(\cdot\right)$ is the expectation over the distribution of Z. This basically tells us that the unconditional expectation of W, $E\left(W\right)$, is the expectation across the distribution of Z of the conditional expectation $E\left(W|Z\right)$. It might be easiest to think of this in terms of Z being a discrete variable. Then, $E\left(W\right)$ is the weighted sum of the terms $E\left(W|Z=z_k\right)\cdot Pr\left(Z=z_k\right)$ across the $k=1,...,K$ values of Z with positive probability of occurring.

We now return to the expression

$$E\left(\hat{\beta}_1\right)=\beta_1+E\left(\frac{\sum_{i=1}^{N}(P_i-\overline{P})\cdot\epsilon_i}{\sum_{i=1}^{N}(P_i-\overline{P})^2}\right)$$

[5]Mean independence is weaker than independence because independence implies no relationship whatsoever between P and ϵ. Under mean independence, there could still be, for instance, some relationship between the variance of ϵ and P.

[6]The condition is sometimes written as either $E(\epsilon_i|\vec{P})=0$ or $E(\epsilon_i)=0$ (for all $i=1,...,N$ in both cases). Basically, either of these conditions imply $E(\epsilon_i|\vec{P})=E(\epsilon_i)$ since they leave $\vec{P}$ no role in shaping the expectation of ϵ_i.

Conditioning on the observed $\vec{P}$ (i.e. proceeding as if $\vec{P}$ was not random but assumed fixed at whatever values for $\{P_1, P_2, ..., P_N\}$ we happen to observe in our sample of size N), we have

$$E\left(\hat{\beta}_1|\vec{P}\right) = \beta_1 + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}|\vec{P}\right)$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot E\left(\epsilon_i|\vec{P}\right)}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

Since we are making the mean independence assumption $E\left(\epsilon_i|\vec{P}\right) = 0$,

$$E\left(\hat{\beta}_1|\vec{P}\right) = \beta_1$$

But the Law of Iterated Expectations tells us that

$$E\left(\hat{\beta}_1\right) = E_{\vec{P}}\left(E\left(\hat{\beta}_1|\vec{P}\right)\right) = E_{\vec{P}}\left(\beta_1\right) = \beta_1$$

Thus, the mean independence condition $E(\epsilon_i|\vec{P}) = E(\epsilon_i)$ is sufficient to insure the unbiasedness of $\hat{\beta}_1$ (i.e. that $\hat{\beta}_1$ is an unbiased estimator of true program impact β_1).

Independence and mean independence have implications for the assumed process governing program participation. Independence means that program participation is completely unrelated to the ϵs. Mean independence means that P reveals no information about the value of ϵ. This means, among other things, that the Ps and ϵ are completely uncorrelated. Thus, under independence one assumes that the ϵs are completely unrelated to (including uncorrelated with) the program participation decision. Under mean independence, one assumes that the ϵs are uncorrelated with the program participation decision. In other words, in either case, one assumes that program participation is completely uncorrelated with the unobserved determinants of the outcome of interest.

If the condition $E(\epsilon_i|\vec{P}) = E(\epsilon_i)$ is not met, then in general $\hat{\beta}_1$ is not an unbiased estimator of β_1. In particular, there is the possibility that even if P_i is uncorrelated with ϵ_i (i.e. there is no *pairwise* correlation: $corr\left(P_i, \epsilon_i\right) = 0$) it may be correlated with some ϵ_j where $j \neq i$. Thus, ϵ_i is not fully mean independent of $\vec{P}$.[7] A classic example of how this could arise would be given by the following hypothetical model:

$$Y_i = \varphi_0 + \varphi_1 \cdot Y_{i-1} + \epsilon_i$$

In this case the regressor Y_{i-1} is clearly correlated with ϵ_{i-1}.[8] In such an instance the least squares estimator is biased. However, it *is* still consistent as long as there is no pairwise correlation (i.e. as long as $corr\left(\epsilon_i, P_i\right) = 0$).

Establishing this requires the introduction of something called a **probability limit**. First, let $\hat{Z_N}$ represent a sequence of estimates for different sample sizes N: $\{\hat{Z_1}, \hat{Z_2}, ..., \hat{Z_{100}}, ...\}$. In other words, this is the sequence of estimates obtained as sample size grows ever larger. This sequence **converges in probability** to a value Z if

$$lim_{N\to\infty} Pr\left(|\hat{Z_N} - Z| \geq \nu\right) = 0$$

[7] If, for instance, P_j and ϵ_i (where $j \neq i$) are correlated, then the value of P_j does reveal something about the expected value of ϵ_i.

[8] To be consistent, we use only the subscript i, which indexes individuals, in the main text. However, this model most often arises in the time series context in which the observations are from individual i at time t:

$$Y_{i,t} = \varphi_0 + \varphi_1 \cdot Y_{i,t-1} + \epsilon_{it}$$

In this context this is often referred to as the **lagged dependent variable model**.

Although this seems like a complicated expression, it says something quite simple. As sample size N grows infinitely large, the probability that the estimate $\hat{Z_N}$ is outside of the interval $[Z-\nu, Z+\nu]$ becomes 0. If this is true, the sequence (across successively larger sample sizes N) of estimates $\hat{Z_N}$ **converges in probability** to Z. This is also captured by saying Z is the probability limit (or "plim") of the sequence $\hat{Z_N}$:

$$plim\left(\hat{Z}_n\right) = Z$$

This expression simply means that as the sample size becomes infinite, the probability that Z_n differs from Z by more than some tiny amount eventually becomes zero.

Probability limits, with their appeal to infinite sample sizes, are the natural tool for considering the consistency of estimators. Probability limits have some very handy properties that expectations do not. In particular,

$$E\left(Z_1 \cdot Z_2\right) \neq E\left(Z_1\right) \cdot E\left(Z_2\right)$$

unless Z_1 and Z_2 are independent. Furthermore,

$$E\left(\frac{Z_1}{Z_2}\right) \neq \frac{E\left(Z_1\right)}{E\left(Z_2\right)}$$

However, for probability limits,

$$plim\left(Z_1 \cdot Z_2\right) = plim\left(Z_1\right) \cdot plim\left(Z_2\right)$$

(regardless of whether Z_1 and Z_2 are independent) and

$$plim\left(\frac{Z_1}{Z_2}\right) = \frac{plim\left(Z_1\right)}{plim\left(Z_2\right)}$$

Otherwise, probability limits share all of the basic properties of expectations.[9] Hence probability limits are in some sense much more flexible than expectations.

To examine the probability limit of the estimate $\hat{\beta}_1$ we simply repeat the process of the examination of the expectation $E(\hat{\beta}_1)$ but substitute the probability limit operator, $plim(\cdot)$, for the expectations $E(\cdot)$ operator[10] with the ultimate result being

$$plim\left(\hat{\beta}_1\right) = \beta_1 + plim\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

However, probability limits allow us to go further:

$$\begin{aligned} plim\left(\hat{\beta}_1\right) &= \beta_1 + plim\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) \\ &= \beta_1 + \frac{plim\left(\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i\right)}{plim\left(\sum_{i=1}^{N}(P_i - \overline{P})^2\right)} \end{aligned}$$

[9] With one other exception. In general,

$$E\left(g\left(Z\right)\right) \neq g\left(E\left(Z\right)\right)$$

where $g(\cdot)$ is a continuous function. However,

$$plim\left(g\left(Z\right)\right) \neq g\left(plim\left(Z\right)\right)$$

This is referred to as **Slutsky's Theorem**.

[10] By "*something* operator", we mean in effect the statistical function that yields *something*. Hence, for instance, $plim(\cdot)$ is the probability limit operator because it yields the probability limit of the variable within the parentheses.

Substituting in $1/N$ in the numerator and denominator, we have

$$plim\left(\hat{\beta}_1\right) = \beta_1 + \frac{plim\left(\frac{\sum_{i=1}^{N}(P_i-\overline{P})\cdot\epsilon_i}{N}\right)}{plim\left(\frac{\sum_{i=1}^{N}(P_i-\overline{P})^2}{N}\right)}$$

Notice that the numerator is the covariance of P_i and ϵ_i. Hence, if the covariance of P_i and ϵ_i is zero (i.e. the pairwise correlation $corr(\epsilon_i, P_i) = 0$), $\hat{\beta}_1$ is a consistent estimator of β_1. If P_i and ϵ_i have a non-zero covariance (i.e. they are correlated) then $\hat{\beta}_1$ is <u>not</u> a consistent (or unbiased) estimator of program impact β_1.

P and ϵ being pairwise uncorrelated (i.e. $corr(\epsilon_i, P_i) = 0$) implies that the program participation decision of each individual is not correlated with the random component of their outcome. This is technically weaker than the conditions that would yield unbiasedness, but still speaks to the idea that program participation is unrelated to the unobserved determinant of the outcome of interest, in this case for each individual. We thus see that consistency essentially relies on a (behaviorally) very similar assumption to that required to view the least squares estimator as an unbiased estimator of program impact.

We have thus established the following:

- If ϵ_i is independent of $\vec{P}$ or even just mean independent of $\vec{P}$, $\hat{\beta}_1$ is an unbiased estimator of β_1;
- If ϵ_i and P_i are uncorrelated (i.e. the pairwise correlation condition $corr(\epsilon_i, P_i) = 0$ is met), $\hat{\beta}_1$ is a consistent estimator of β_1;
- If ϵ_i and P_i are correlated (i.e. $corr(\epsilon_i, P_i) \neq 0$), $\hat{\beta}_1$ is a biased and inconsistent estimator of β_1.

In everyday practice in the program impact evaluation literature, there can be some imprecision about the conditions under which least squares provides a biased or consistent estimate of program impact $\hat{\beta}_1$. One often hears the basic assertion "$\hat{\beta}_1$ is unbiased if ϵ and P are uncorrelated". While we have seen that this is not exactly true (unbiasedness requires more) this is still a useful way of thinking about things in the following sense: clearly if $corr(\epsilon_i, P_i) \neq 0$ then $\hat{\beta}_1$ will be a biased and inconsistent estimator of program impact β_1.[11]

We have now uncovered the conditions under which the ordinary least squares estimator $\hat{\beta}_1$ is an unbiased or consistent estimator of true program impact β_1. Unbiasedness and consistency both require that $corr\left(\epsilon_i, P_i\right) = 0$ (though unbiasedness requires more than just this). Our next step is to explore a little bit more thoroughly what is happening, particularly from a behavioral standpoint, when the regression estimator is biased or inconsistent because $corr\left(\epsilon_i, P_i\right) \neq 0$.[12] We do this because the credibility of the assumptions required to view regression as yielding an unbiased or consistent estimate $\hat{\beta}_1$ of program impact β_1 is best considered through the lens of their implications for behavior.

[11] It is worth noting that this is an instance where unbiasedness implies consistency but not the reverse. Either the independence or mean independence assumption required for unbiasedness implies no pairwise correlation between P_i and ϵ_i. However, the absence of pairwise correlation does not imply either mean independence or independence.

[12] Once again, to be sure unbiasedness requires more than just the pairwise condition $corr(\epsilon_i, P_i) = 0$. However, in program impact evaluation we are most often concerned with pairwise correlation, and so we focus our attention on that case. This is likely, in the program impact evaluation context, to be a feature of any violation of either independence or mean independence of ϵ_i from $\vec{P}$.

To do this we introduce a slightly more elaborate model of program impact and hence the determination of observed Y. Now let us suppose that the true model determining observed Y is given by[13]

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

We make the following key assumption: $E\left(\epsilon|\vec{P}\right) = 0$ (i.e. that the random component ϵ for each individual is independent of the program participation decision of themselves or any other member of their population[14]). This assumption means that the random component of $\{Y^0, Y^1\}$ (or the errors from the true population regression model) are mean independent of program participation. We will also assume that $E(\epsilon|\vec{x}) = 0$ (where $\vec{x}$ is defined analogously to $\vec{P}$) but refrain from discussing the importance of this assumption until the subsection concerned with "bad controls".

Suppose that, once again, we have a sample of N individuals for whom we observe Y and P, but may or may not observe x. If we perform ordinary least squares estimation by regressing Y on P with that sample, our estimate of β_1, $\hat{\beta}_1$, is, once again,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

The question, of course, is what exactly we would be estimating in this instance.

Once again, we will examine the expectation of $\hat{\beta}_1$:

$$E\left(\hat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

Next, we substitute in the true model,

$$y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \epsilon_i$$

for Y_i. This yields

$$E\left(\hat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \epsilon_i)}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

$$= E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \beta_0}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \beta_1 \cdot P_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \beta_2 \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$
$$+E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

[13] This is easy to motivate with a simple extension of our model of program impact. We now assume

$$Y^0 = \beta_0 + \beta_2 \cdot x + \epsilon$$

$$Y^1 = \beta_0 + \beta_1 + \beta_2 \cdot x + \epsilon$$

with the rest of the development of the model proceeding in the same fashion but with these two formulas.

[14] Here we admittedly engage in a bit of hand waving: we earlier defined $\vec{P}$ as the vector containing program participation $\{P_1, P_2, ..., P_N\}$ for all N members of some sample from a population. Because we have not yet described the estimation sample for this example, we for the time being think of $\vec{P}$ as containing the program participation status for all members of the population. However, if $E(\epsilon|\vec{P}) = 0$ when we think of $\vec{P}$ as containing the program participation of all members of the population, then $E(\epsilon|\vec{P}) = 0$ when we think of $\vec{P}$ as containing the program participation of any random sample of size N from that population.

Using the same math as above, we can eliminate the first term and simplify the second:

$$E\left(\hat{\beta}_1\right) = \beta_1 + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \beta_2 \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) + E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

Since we assume that P and ϵ are uncorrelated, we have

$$E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right) = 0$$

Then,

$$E\left(\hat{\beta}_1\right) = \beta_1 + \beta_2 \cdot E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

(where we also engage in some re-arrangement since β_2 is a parameter and hence can be treated as a constant).

Notice that

$$E\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot x_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

is the expectation of the least squares estimator $\hat{\gamma_1}$ from the model

$$x = \gamma_0 + \gamma_1 \cdot P + \upsilon$$

In other words, it is reflective of the relationship between P and x. Our result is thus that

$$E\left(\hat{\beta}_1\right) = \beta_1 + \beta_2 \cdot E\left(\hat{\gamma_1}\right)$$

The situation is now coming into focus. When we regress Y on P when the true model is

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

the expectation of our estimate of β_1, $E(\hat{\beta}_1)$, does not necessarily actually equal the true population program impact β_1. Instead, it equals program impact β_1 (i.e. the estimand) plus the term

$$\beta_2 \cdot E\left(\hat{\gamma_1}\right)$$

This term is the effect of x on Y (i.e. β_2) weighted by the effect of P on x.

This term reflects the fact that P is potentially forced to play two roles when we regress Y on P: a control for itself and a proxy for x. The value of the proxy role depends on the strength of the relationship between x and Y, weighted by the strength of the relationship between x and P. In other words, P serves as a proxy for x, but only to the degree that P and x are themselves related.

The bias to $\hat{\beta}_1$ when P is also forced to serve as a proxy for x is known as **omitted variable bias**. It is so known because this bias is an artifact of the omission of x as a regressor. Although we focus on one omitted regressor, in practice there may be many (a point to which we will return in the discussion of multiple regression).

If x has no effect on Y (i.e. $\beta_2 = 0$) or x and P are unrelated (i.e. $\gamma_1 = 0$) then the proxy term disappears and

$$E\left(\hat{\beta}_1\right) = \beta_1$$

In other words, $\hat{\beta}_1$ would be an unbiased estimator of β_1 if $\beta_2 = 0$ or $\gamma_1 = 0$.

Thus, when we employ the regression approach to impact evaluation, we are effectively assuming that there are no omitted factors from our model that are related to both program participation and the outcome of interest. This is the key assumption of the regression (and, as we will see, matching) approach to program impact evaluation: that we observe and control for all of the factors associated statistically (for whatever reason) with both the outcome of interest and program participation. In other words, we are essentially assuming that *we have measured all that matters* from the standpoint of recovering unbiased estimates of program impact. In many applications, this is likely a fairly strong assumption.

We now consider a numerical example (this numerical example is contained in the STATA do-file 4.1.do, but is essentially the same in structure as that in 2.1.do). This example considers the consequences of regressing Y on P when the "true" model is

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

Specifically, suppose that we randomly generate 30,000 observations based on these two assumed equations:

$$Y = 1 + .5 \cdot P + 1.5 \cdot x + e$$

where $e \sim$ N(0,25) and $x \sim$ N(0,4).[15] Note, crucially, that true program impact is .5. P equals 1 if

$$.5 \cdot x + e_p > 0$$

where $e_p \sim$ N(0,36) and 0 otherwise. x is thus a determinant of Y and P, the exact circumstance to which models relying on a "selection on observables" assumption appeal.

STATA Output 4.1 (4.1.do)

```
. * Summary statistics for the errors and regressors
. summarize e eP x P
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           e |      30000    .0024446    4.991197   -19.5649   21.35932
          eP |      30000    .0105079    5.948241  -24.66446   23.28767
           x |      30000   -.0044134    1.996313  -8.141914   8.420812
           P |      30000        .498    .5000043          0          1
.
. * Correlations between errors and regressors
. correlate e eP x P
(obs=30000)
             |        e       eP        x        P
-------------+--------------------------------------
           e |   1.0000
          eP |  -0.0058   1.0000
           x |   0.0020   0.0022   1.0000
           P |  -0.0016   0.7871   0.1290   1.0000
```

The resulting sample is considered in STATA Output 4.1. The error terms e and e_p have means of roughly 0 and standard deviations roughly in line with what we might expect (e.g., since $e \sim$ N(0,25), leaving us to expect an empirical standard deviation estimate in the neighborhood of 5). Notice that the true regression error e is essentially uncorrelated with P and x (with correlations

[15] Once again, $a \sim$ N(b,c) means a is distributed as a normal random variable with mean 'b' and variance 'c'. In the context of this example we would thus be pseudo-randomly drawing 30,000 observations for a normally distributed variable a with mean 'b' and variance 'c'.

of -0.0016 and 0.0020, respectively). Importantly, P and x are correlated, as we might expect since x played a role in determining P.

Let us first regress Y on P and x. The results are presented in STATA Output 4.2. These results are right in line with where we would hope they would be, given the true data generating process behind them: the estimated coefficient on P (or, another words, estimated program impact) is, at .4808799, around .5, while that on the coefficient on x is, at 1.505581, in the neighborhood of 1.5.

Next we estimate the "wrong"regression model (i.e. omitting the variable x). The results are in STATA Output 4.3. We can now see that the omission of x has led to an estimate of program impact that deviates substantially from that underlying the true data generating process (1.256391, against the true program impact of .5). This likely reflects omitted variable bias, and from this example it should be clear how bad it really can be: it has led us to over-estimate the impact of the program by more than one hundred percent!

In terms of the correlations, P is uncorrelated with the fitted errors from the regression of Y on P (this is by construction but will become clearer when we discuss the method of moments approach to regression estimation). However, P's correlation with the true error per the "wrong" model (i.e. $1.5 \cdot x + e$) is 0.0649. Since the correlation between P and the error term from the "true" regression (e) is essentially zero (at -.0016) this means that the correlation between P and the error per the true model is driven by the term $1.5 \cdot x$.

Finally, we consider the nature of the omitted variable bias. In short, when we regressed Y on P alone, our estimate of the coefficient on P was 1.256391, more than double the true value of .5. To relate this to our earlier omitted variable bias formula, we first regress x on P. The results are shown in Output 4.4. We see from that Output that the coefficient on P is .5150909. The omitted variable bias equation tells us that the estimated coefficient on P when we regress Y on P alone should be

$$.4808799 + 1.505581 \cdot .5150909$$

which equals 1.256931, exactly as the formula predicted.

STATA Output 4.2 (4.1.do)

```
. * Regressing y on P and x
. regress Y P x
```

Source	SS	df	MS		
Model	278329.976	2	139164.988	Number of obs	= 30000
Residual	747330.934	29997	24.9135225	F(2, 29997)	= 5585.92
				Prob > F	= 0.0000
				R-squared	= 0.2714
				Adj R-squared	= 0.2713
Total	1025660.91	29999	34.1898367	Root MSE	= 4.9913

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.4808799	.0581213	8.27	0.000	.3669597	.5948001
x	1.505581	.0145573	103.42	0.000	1.477048	1.534114
_cons	1.011991	.0408499	24.77	0.000	.9319236	1.092059

Before proceeding we remind the reader of a qualification offered in Chapter 2: program impact estimates from particular samples can never prove or exclude the possibility of unbiasedness. Unbiasedness means that the estimator is "right on average". The way to establish this is to simulate many samples, estimate program impact for each and then average the impact estimates from those

many samples. Whether the average does or does not match the true population expected program impact is the true indicator of unbiasedness.

Particular impact estimates that differ substantially from or adhere closely to the true population expected impact might always simply reflect ordinary sample-by-sample variation in unbiased or biased estimators, respectively. By themselves they do not prove anything. As Tony Soprano once observed, even a broken clock is right twice a day.

Nonetheless, the fact that the discrepancy between true population estimated impact and the impact estimate essentially reflected the prediction of the omitted variable bias model is highly indicative of likely bias. Moreover, the reader knows that within the context of the model the examples are engineered by the authors to show what we claim they show, and for the reasons we suggest. Nonetheless, one would do well always to remember what *cannot* be learned from an estimate from just one particular sample.

STATA Output 4.3 (4.1.do)

```
. * Regressing y on P alone
. regress Y P
```

Source	SS	df	MS		Number of obs =	30000
					F(1, 29998) =	350.30
Model	11838.6955	1	11838.6955		Prob > F =	0.0000
Residual	1013822.21	29998	33.7963269		R-squared =	0.0115
					Adj R-squared =	0.0115
Total	1025660.91	29999	34.1898367		Root MSE =	5.8135

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	1.256391	.0671286	18.72	0.000	1.124816	1.387966
_cons	.6191419	.0473721	13.07	0.000	.5262906	.7119931

We have just seen the essence of omitted variable bias. We regressed Y on P alone when the true model was

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

In other words, we omitted x, one of the true systematic determinants of Y. By doing so, we made it impossible to obtain an unbiased estimate of the casual effect of P on Y, β_1: the omission of x biased our estimate of β_1. The extent of the bias would depend on the strength of the relationship between Y and x (β_2) and the strength of the relationship between x and P. The intuition should be familiar at this point: when we omit x, P must serve not only as a control for itself but also as a proxy for x.

Suppose, for example, that Y is health, P is participation in a health program and x is motivation. Suppose as well that more motivated individuals are more likely to participate in the program (establishing a relationship between participation P and motivation x) and that motivation x does indeed influence health Y. Then, when we regress health Y on program participation P, our estimate of the effect of P on Y reflects not only the casual effect of participation on health (which we want) but also picks up to an extent (i.e. the extent to which program participation P and motivation x are related) the effect of motivation x on health Y. In other words, the omission of x from the estimation has muddied the waters in terms of obtaining an unbiased estimate of the impact of program participation P on health Y.

Of course, the bias to $\hat{\beta}_1$ could, in principle, be resolved easily: simply add x as a regressor (in other words, regress Y on P *and* x). Assuming that this is a viable strategy (i.e. assuming that x

is observed along with Y and P for each individual in some sample), the bias to the estimate of β_1 would disappear. The mathematics of such a regression, called a **multiple regression**, are more complex (typically involving matrix algebra, which is beyond the reasonable scope of this manual). However, it can be shown easily that the bias to the estimate of β_1 disappears with the inclusion of x in the estimation (it can also be shown easily that the estimate of β_2 will be unbiased).

STATA Output 4.4 (4.1.do)

```
. * Regressing x on P
. regress x P
      Source |       SS       df       MS              Number of obs =   30000
-------------+------------------------------           F(  1, 29998) =  507.74
       Model |  1989.85791     1  1989.85791            Prob > F      =  0.0000
    Residual |  117564.144 29998  3.91906605            R-squared     =  0.0166
-------------+------------------------------           Adj R-squared =  0.0166
       Total |  119554.001 29999  3.98526622           Root MSE      =  1.9797

------------------------------------------------------------------------------
           x |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .5150909   .0228594    22.53   0.000     .4702856    .5598962
       _cons |  -.2609287   .0161317   -16.17   0.000    -.2925474     -.22931
------------------------------------------------------------------------------
```

Omitted variable bias can also be explained in terms of the potential outcomes framework introduced in Chapter 2. We defined Y^1 as the outcome an individual experiences when they participate in a program, and Y^0 is the outcome that individual would experience in the absence of participation in that program. $Y^1 - Y^0$ is that individual's program impact. Their observed outcome is $Y = P \cdot Y^1 + (1-P) \cdot Y^0$. Suppose that x influences program participation P. This means that some types as defined by x are more likely to participate. For instance, if x is motivation (with larger values associated with more motivation) and more motivated individuals are more likely to participate, then the average value of x should be higher among participants than non-participants.

To fix ideas, suppose that we have data on Y and P for N individuals. N_1 of those N individuals are participants ($P = 1$), while the remaining N_0 of the individuals are non-participants ($P = 0$). We observe Y^1 (but not Y^0) for the N_1 participants, and observe Y^0 (but not Y^1) for the N_0 non-participants. In principle, we could estimate program impact by considering the difference in average Y between the participant and non-participant subsamples:

$$\frac{\sum_{i=1}^{N_1} Y_i^1}{N_1} - \frac{\sum_{j=1}^{N_0} Y_j^0}{N_0} = \frac{\sum_{i=1}^{N_1} Y_i}{N_1} - \frac{\sum_{j=1}^{N_0} Y_j}{N_0}$$

This estimator of program impact is motivated by the idea that the differences in average outcomes between the two groups is program impact. As such, it seeks to estimate the same thing as is captured by the parameter β_1.

The problem is that we know that these two subsamples differ by more than just their program participation experience P. Those in the subsample of N_1 participants are more motivated (i.e. have a higher value for x). If x influences Y, it is unclear whether the difference in average Y between the two subsamples reflects program participation P or motivation x. In essence, the estimate of β_1 obtained by regressing Y on P is plagued by the same problem: P is forced to serve as a proxy for x because x has a different average value at $P = 1$ than $P = 0$.

It is worth re-posing the problem at this point for the purpose of getting at the core assumption

of the "selection on observables" models. We are interested in obtaining an estimate of β_1 from

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

that actually reflects the population parameter β_1, the impact of program participation P on Y. There is another systematic variable (x) that appears in the regression model. It might potentially influence both participation P and the outcome of interest Y. Suppose we are willing to assume that x is the only such variable. This is tantamount to assuming that the correlation between P and ϵ is zero: $corr(P, \epsilon) = 0$. Then, by regressing Y on P and x we would obtain a valid estimate of the causal effect of P on Y, β_1. The key assumption is that x is the only variable that shapes both Y and P. Multiple regression allows for us to control for the effect of x on Y, and hence recover the causal program impact because the impact of program participation P is the only systematic source of variation in Y left after doing so.

Because motivation also influences participation, we would expect that its average value among participants ($P = 1$) would be different from that among non-participants ($P = 0$). This alone can introduce a difference in the average values of Y^1 among participants and Y^0 among non-participants (even if participation had no effect and $E(Y^1 - Y^0) = 0$ across everyone). But another way of saying this is that an individual's participation status reveals something about their values of Y^1 and Y^0 because participation is acting as something of a proxy for x.

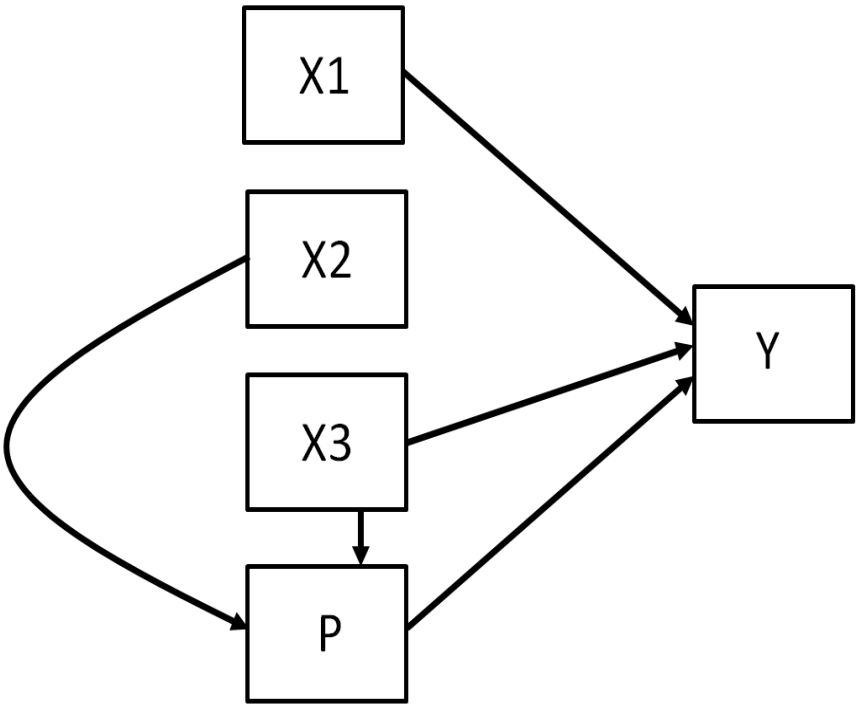


Figure 4.2: Controlling for Observables X

Multiple regression of Y on P and x controls for the effect of x on P, so that remaining systematic (as opposed to random) variation in Y must reflect the impact of program participation P. In other words, since $Y = P \cdot Y^1 + (1 - P) \cdot Y^0$, once the effect of x on Y is controlled for, the remaining variation in Y comes from *switching* between Y^1 and Y^0 depending on each individual's value of P. In other words, P offers no information about $\{Y^1, Y^0\}$. It acts merely as a switching mechanism between Y^1 and Y^0.

This is in essence the key assumption of "selection on observables" methods: once we control for the observables (in this case x), P reveals no information about the values of $\{Y^1, Y^0\}$. This is often written as follows:

$$E\left(Y^1|P,x\right) = E\left(Y^1|x\right)$$
$$E\left(Y^0|P,x\right) = E\left(Y^0|x\right)$$

In other words, once you condition on x, the expected values of Y^1 and Y^0 do not depend on P. In other words, re-visiting a term from Chapter 2, once we condition on x, Y^1 and Y^0 are *mean independent* of P. Sometimes this is written

$$\left\{Y^1, Y^0\right\} \perp\!\!\!\perp P|x$$

where $\perp\!\!\!\perp$ means "independent". This line simply states "Y^1 and Y^0 are independent of P conditional on x".

In essence, what is being said in this assumption is that we can observe and hence control for all of the factors that systematically shape both P and Y. This is the crucial assumption of the regression approach to program impact evaluation, as well as matching (the other major "selection on observables" strategy). If we can measure all of the factors that systematically shape both P and Y, and hence would allow for a reason why Y might differ between participants ($P = 1$) and non-participants ($P = 0$) aside from the impact of participation, we can control for them.

It is worth revisiting an earlier figure (now presented as Figure 4.2) to clarify what must be observed under the "selection on observables" assumption. In Figure 4.2, it would be necessary under this assumption to observe and control for every factor that behaved like 'X3', because such factors influence Y *and* P. It would not be necessary to control for 'X1' because there is no mechanism by which such factors can differ by participation status P (and hence no way that P could serve as a proxy for them). The factors in 'X2' may influence P, but they do not independently influence Y. The factors in 'X3' are associated with program participation P and hence will differ on average between participants and non-participants. At the same time, because they also influence Y, one cannot be sure whether differences in average outcomes between participants and non-participants reflect participation or differences in 'X3' between participants and non-participants. In some disciplines it is most common to refer to such variables as **confounders**.

This assumption would not hold if there were some factors that influence both Y and P (i.e. elements of 'X3' from Figure 4.2) and which could not be observed. In that case, regression of Y on P and the factors that we can observe would not be sufficient to recover an unbiased estimate of program impact.

At this point we round off the interpretive phase of the discussion by explicitly defining **endogenous** and **exogenous** variables. These two terms are frequently used in discussions of statistical modelling for program impact evaluation (and, more generally, causal analysis).[16] In empirical program impact evaluation they have come to take on several slightly different meanings, but their

[16] Classically, in mathematics an endogenous variable is simply one determined within a specified system while an exogenous variable is one determined outside of that system. Take, for instance, the following system:

$$x_1 = x_2 - 2 \cdot x_3$$
$$x_3 = x_1 + x_4$$

In this system, the endogenous variables are x_1 and x_3, while the exogenous variables are x_2 and x_4. Note that, as is often the case in such systems, the endogenous variables can be solved for in terms of the exogenous variables:

$$x_1 = \frac{x_2 - x_4}{3}$$

statistical implication is consistent (at least for the purposes of regression modelling). Often, in the context endogenous means a variable subject to or influenced by individual choice. Behaviors and outcomes that reflect or are influenced by personal choices are, of course, driven by myriad characteristics of the individual and the larger environment within which they make decisions, many of which influence other choices but only some of which can be observed and hence controlled for in a regression. For instance, our level of motivation probably influences many behavioral choices (e.g. does an individual make good nutritional choices for their child?) and outcomes (e.g. their child's health). On the other hand, some variables probably are not subject to individual choice. For instance, in many lower income societies whether a child is a twin is probably less subject to the behavioral influences of their parents.[17]

Let us consider what this might mean in a regression context. Suppose that we are interested in understanding the determinants of height-for-age in a poor society, which we will denote, as a variable and for brevity sake, as HFA. Clearly HFA reflects many health investments, shocks, etc., each of which are determined at least in part by the observed and unobserved characteristics of a child and their parents. It also might be determined in part by child nutrition (which we denote as a variable by NTR) and twinning (which we denote as a variable by TWN and which could conceivably influence HFA in a number of ways). Consider two separate regression specifications:

$$HFA = \gamma_0 + \gamma_1 \cdot NTR + \epsilon$$

$$HFA = \delta_0 + \delta_1 \cdot TWN + \upsilon$$

Clearly, ϵ and υ contain all sorts of determinants of HFA that are, for the purposes of these regressions, unobserved. Aside from the obvious omission in each regression[18], these two error terms contain all sorts of characteristics of the child and their family that have direct and indirect influence on HFA.

We have suggested that NTR is a behavioral choice (i.e. it fits our working definition of endogeneity to this point). In other words, it likely depends on observed and unobserved characteristics of a child and their family. But we have also just suggested that these sorts of variables are in ϵ. This means that NTR and ϵ are likely correlated:

$$corr\left(NTR, \epsilon\right) \neq 0$$

This is a familiar condition from the preceding discussion: it means that a regression of HFA on NTR will not yield an estimate of γ_1 that reflects the causal effect of NTR on HFA. It is a natural consequence of NTR being a behavioral choice.

The condition

$$corr\left(NTR, \epsilon\right) \neq 0$$

also serves as the foundation for the more general definition of endogeneity: an endogenous variable is one correlated with unobserved determinants of some outcome of interest. This is more general

$$x_3 = \frac{x_2 + 2 \cdot x_4}{3}$$

In other words, the endogenous variables are ultimately just a function of the exogenous variables. The last two equations (in which we solved for the endogenous variables in terms of the exogenous variables) are called the **reduced forms** for x_1 and x_2.

[17] In wealthier societies the variation in twinning across the population is likely influenced in part by fertility treatment and other purposeful health inputs, which are behavioral choices associated with all sorts of observed and unobserved family characteristics. Admittedly, however, fertility treatments are becoming more common among the middle and upper classes even in some quite poor societies.

[18] To the extent that NTR and TWN actually influence HFA, TWN is relegated to ϵ and $NUTR$ is in υ.

than the definition that focuses on the variable in question being subject to individual choice: a variable can be correlated with the unobserved determinants of an outcome for lots of reasons. For instance, even if it were not an individual choice child nutrition might be correlated with unobserved factors (such as a community level social norms for protecting child welfare) beyond the control of the individual that might also influence HFA.

As we have approached it, TWN is not a choice variable. By the logic initially applied to NTR, one could argue that

$$corr\left(TWN, \upsilon\right) = 0$$

This is also the more general definition of exogeneity: for a given outcome of interest, an exogenous variable is one that is not correlated with unobserved variables that influence that outcome. Whether a given variable is truly exogenous is of course never certain.[19] For instance, it is possible that TWN is partly the result of unobserved community-level environmental factors that also shape HFA.[20]

A variable that is endogenous in one regression context can potentially be considered exogenous in a more richly specified model. Earlier, we considered a situation where the true population regression determining an outcome Y was given by

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

Crucially, we assumed that ϵ contained purely idiosyncratic determinants of Y in the sense that

$$corr\left(P, \epsilon\right) = 0$$

Regression of Y on P alone did not yield an estimate of program impact that was causal to the extent that

1. x had an impact on Y (i.e. $\beta_2 \neq 0$);

2. x and P were correlated.

Hence, in a regression of Y on P alone P might be an endogenous variable because x, with which it is correlated, is in the error term for that regression. Hence, for the purpose simply of regressing Y on P alone, P would qualify as an endogenous regressor. However, it could be rendered exogenous simply by including x as a regressor.[21] A regressor of interest is thus exogenous if the "selection on observables" assumption is satisfied. Endogeneity and exogeneity are thus natural concepts for the consideration of this assumption, which brings us back to the original assumption of this class of impact evaluation models.

How reasonable is the "selection on observables" assumption? This question has been debated fiercely. Some sweeping patterns have been suggested (e.g. Heckman, Ichimura and Todd (1997a) find that the assumption is more plausible when there are strong reasons to believe that participants and non-participants have similar background circumstances). It does seem reasonable to suggest that this is a strong assumption. Part of the reason for suspecting this is that many of the factors that probably systematically influence program participation and outcomes of interest are intrinsically unobservable. For instance, thus far we have sometimes focused on "motivation" as a

[19] As we will see in subsequent chapters, there are statistical tests of exogeneity, but they are just that: statistical tests. They do not prove exogeneity in the pure sense.

[20] And, as we have argued, in some societies it is likely influenced by behavioral choices like fertility treatments.

[21] As we will see in the discussion of "bad" controls below, we will see that P being truly endogenous, and hence the estimated coefficient of P representing true program impact, also depends on the exogeneity of x. However, for the moment we set this complication aside.

particular confounder. However, motivation cannot really be observed. There are probably many analogous emotional and psychological characteristics and values of individuals that shape both their program participation decisions and their outcomes.

Perhaps it might be reasonable to link the credibility of the "selection on observables" assumption to the degree to which program enrollment variation within the estimation sample is driven by some sort of criteria beyond the control of the individual, as opposed to the discretion of the individuals themselves. For instance, Angrist (1998) considered the role of voluntary military service on later life earnings. In that application, estimation was restricted to a sample of voluntary military applicants. What separated those who actually served from those who did not within this sample was then clear (from an empirical standpoint) enlistment criteria established by the military mostly rooted in individual background characteristics and readily observable by the researcher. If, however, eventual enlistment among applicants was ultimately more of an individual decision, it would likely have been driven by factors (like motivation, courage, etc.) unlikely to be intrinsically observable, making the "selection on observables" assumption less plausible.

We now conclude by briefly extending the omitted variable bias discussion to the limited dependent variable setting. The limited dependent variable case is, as one might expect from the term, one in which the variation in Y is limited in nature. We consider the basic case where Y takes on only two values, 1 and 0. These can be thought of as corresponding to binary outcome possibilities. For instance, many of the behaviors and outcomes recorded in health interview surveys are binary in nature: an individual either did or did not visit a clinic, get their children immunized, experience a symptom, use contraception, get pregnant, etc. Nor is interest in binary behaviors and outcomes limited to health: individuals do or do not graduate school, work, receive fringe benefits, commit crimes, etc.[22] Since a binary outcome captures simply whether something happens or not, the typical regression modelling goal is to capture the role that P and x play in determining the *probability* that the binary outcome event occurs.

Confronted with a binary outcome, one could still resort to least squares regression of the limited dependent variable Y on P and x. While this approach, called the **linear probability model**, has its drawbacks, it also has a certain appeal. The drawbacks include faulty standard error estimates and the possibility of predicted probabilities above 1 or below 0 when the probability of an event occurring can by definition never be outside of those bounds. On the other hand, the linear probability model is a relatively straightforward approach that involves few assumptions and can be readily integrated into other impact evaluation methods, such as the within estimators or linear instrumental variables approaches that will be considered in subsequent chapters.

Perhaps its biggest advantage for the present discussion is that the estimated coefficients are readily interpretable in probability terms. For instance, suppose that Y can take on values 0 or 1. A simple linear probability model analogous to the first regression model considered in this section would involve least squares regression of Y on P. This estimates the parameters of the model

$$Y = \beta_0 + \beta_1 \cdot P + \epsilon$$

In other words, the model is exactly the same as that at the outset of this section. All that has changed is that the variation in Y is constrained so that Y can only take on values 0 or 1. The estimate $\hat{\beta}_1$ arising from least squares regression of Y on P has a direct interpretation in probability terms: it is the increase in the probability that $Y = 1$ that would occur when P switches from 0 to 1. This is the interpretation of program impact when dealing with an outcome constrained to be

[22]Though our focus is on the binary outcome, the basic insights extend to other limited dependent variable outcomes. For example, multinomial outcomes involve choice between mutually exclusive and exhaustive categories. As another example, count outcomes capture the number of integer times some event occurs.

either 0 or 1: the degree to which participation influences the probability of an outcome of interest occurring. In terms of unbiasedness or consistency, all of the conclusions reached about the least squares estimator in the more general case (where Y is a continuous variable) carry over to the limited dependent case where Y is binary.

We will thus use results from the linear probability model as a reference for what program impact should be. However, we will use this only for comparison purposes as our main focus will be the popular **logit** binary regression model. We focus on the logit model because it allows us to introduce an example of the application of the regression approach to a non-linear model.

As with the linear probability model, we consider a binary outcome of interest Y that can take on values 1 or 0. Underlying the observed binary outcomes are "indirect utility functions" with continuous outcomes:

$$V^{1*} = \gamma_0 + \gamma_1 \cdot P + \gamma_2 \cdot x + \epsilon_1$$
$$V^{0*} = \phi_0 + \phi_1 \cdot P + \phi_2 \cdot x + \epsilon_0$$

V^{1*} is the indirect utility or welfare that is experienced when the observed outcome $Y = 1$, given P, x and a random element to utility ϵ_1. A similar interpretation obtains for V^{0*} when $Y = 0$. To begin with, $Y = 1$ if the indirect utility from making that choice is greater than or equal to the indirect utility from not doing so. Thus, we observe $Y = 1$ if:

$$V^{1*} \geq V^{0*}$$

or

$$\gamma_0 + \gamma_1 \cdot P + \gamma_2 \cdot x + \epsilon_1 \geq \phi_0 + \phi_1 \cdot P + \phi_2 \cdot x + \epsilon_0$$

or, re-arranging and collecting terms,

$$\epsilon_1 - \epsilon_0 \geq (\phi_0 - \gamma_0) + (\phi_1 - \gamma_1) \cdot P + (\phi_2 - \gamma_2) \cdot x$$

If we multiply both sides by -1, we have

$$\epsilon_0 - \epsilon_1 \leq (\gamma_0 - \phi_0) + (\gamma_1 - \phi_1) \cdot P + (\gamma_2 - \phi_2) \cdot x$$

or, simplifying further,

$$\epsilon_0 - \epsilon_1 \leq \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x$$

where $\beta_k = (\gamma_k - \phi_k)$ for $k = 0, 1, 2$ just collects the ϕ and γ terms.

The random elements to the indirect utility functions are key to casting the regression in terms of the probability that the outcome $Y = 1$ would occur. Specifically, we have

$$Pr\left(Y = 1|P, x\right) = Pr\left(\epsilon_0 - \epsilon_1 \leq \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x\right)$$

Notice that this means that the probability that $Y = 1$ is just the cumulative probability of the random variable $\epsilon_0 - \epsilon_1$ at $\beta_0 + \beta_1 \cdot P + \beta_2 \cdot x$. **Parametric**[23] binary choice regression models like

[23]Parametric is one of those terms that can mean different things in different contexts in statistical research, but usually somehow appeals to the idea that a particular functional form is assumed. Thus, in this context we apply the term parametric to binary choice models that assume a specific distributional assumption for $\epsilon_0 - \epsilon_1$. Logit is not the inevitable outcome of the parametric approach to binary regression modelling. For instance, assuming $\epsilon_0 - \epsilon_1$ is normally distributed gives rise to the **probit** model. Nor is the parametric approach inevitable. Non-parametric modelling is an exciting emergent area of limited dependent variable regression modelling. To some extent, the linear probability model can be seen as an example of this. A discussion of more sophisticated non-parametric modelling would represent an un-necessary digression from the manual.

logit operationalize an actual regression form by making some assumption about the probability distribution of $\epsilon_0 - \epsilon_1$.

The logit model assumes that the ϵs are distributed as Type-I Extreme Value.[24] Though this may sound like a very exotic distribution, it is shaped roughly like the Normal distribution but with higher kurtosis (i.e. fatter tails) and, depending on the values of the parameters of the distribution, some degree of skewness. The reason this distribution has attracted so much attention in this context is that it turns out that the distribution of the difference of two Type-I Extreme Value random variables has a cumulative density given by the logistic function. Thus, if Z and W are Type-I Extreme Value Distributed random variables,

$$Pr\left(Z - W \leq c\right) = \frac{exp\left(c\right)}{1 + exp\left(c\right)}$$

In the era when binary choice models were under development computer resources were dear, and such a computationally simple form had great appeal.

For present purposes, assuming the ϵs are Type-I Extreme Value Distributed leads to the choice probability

$$Pr\left(Y = 1|P, x\right) = \frac{exp\left(\beta_0 + \beta_1 \cdot P + \beta_2 \cdot x\right)}{1 + exp\left(\beta_0 + \beta_1 \cdot P + \beta_2 \cdot x\right)}$$

Since Y can assume values of only 1 and 0, $Pr(Y = 1) + Pr(Y = 0) = 1$ and, therefore, $Pr(Y = 0) = 1 - Pr(Y = 1)$. From this simple truth, we can obtain the other choice probability

$$Pr\left(Y = 0|P, x\right) = 1 - Pr\left(Y = 1|P, x\right) = 1 - \frac{exp\left(\beta_0 + \beta_1 \cdot P + \beta_2 \cdot x\right)}{1 + exp\left(\beta_0 + \beta_1 \cdot P + \beta_2 \cdot x\right)}$$

$$= \frac{1}{1 + exp\left(\beta_0 + \beta_1 \cdot P + \beta_2 \cdot x\right)}$$

We now turn to estimation of this model.

The data requirements are essentially the same as in the earlier ordinary least squares estimation case. Suppose that we have a sample of N individuals. For each (indexed by i) we observe $\{Y_i, P_i, x_i\}$. Since the regression outcome is cast in terms of the probabilities that certain events occur (namely that $Y_i = 1$ or $Y_i = 0$), the most popular method of estimation is by **maximum likelihood**. Basically, maximum likelihood estimation involves choosing the values for β_0, β_1 and β_2 that maximize the probability of the observed sequence of values (0 or 1) for $Y_1, Y_2, ..., Y_N$.

Formally, individual i's probability for their outcome Y_i is

$$Pr_i\left(Y_i|P_i, x_i, \beta_0, \beta_1, \beta_2\right) = Pr_i\left(Y_i = 1|P_i, x_i, \beta_0, \beta_1, \beta_2\right)^{Y_i} \cdot Pr_i\left(Y_i = 0|P_i, x_i, \beta_0, \beta_1, \beta_2\right)^{(1-Y_i)}$$

$$= \left(\frac{exp\left(\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i\right)}{1 + exp\left(\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i\right)}\right)^{Y_i} \cdot \left(\frac{1}{1 + exp\left(\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i\right)}\right)^{(1-Y_i)}$$

Notice that their choice probability will depend on their choice Y_i in the sense that it determines whether their contribution to the likelihood takes the form of $Pr(Y_i = 1|\cdot)$ or $Pr(Y_i = 0|\cdot)$.

As the probability of their observed outcome Y_i given their program participation P_i, characteristic x_i and the parameters β_0, β_1 and β_2, $Pr_i\left(Y_i|P_i, x_i, \beta_0, \beta_1, \beta_2\right)$ is the individual i's contribution

[24] Sometimes in developing the logit model reference is made to the Gumbel or log-Weibull distribution. They refer to the same distribution. The "Extreme Value" name is derived from the fact that these distributions are often used to model the probability distribution of the maximum value of series of random variables.

to the **likelihood function**. The overall likelihood $L(\beta_0, \beta_1, \beta_2)$ is simply the product of such probabilities for all individuals in the sample of N individuals:

$$L(\beta_0, \beta_1, \beta_2) = Pr_1(Y_1|P_1, x_1, \beta_0, \beta_1, \beta_2) \cdot Pr_2(Y_2|P_2, x_2, \beta_0, \beta_1, \beta_2) \cdot ... \cdot Pr_N(Y_N|P_N, x_N, \beta_0, \beta_1, \beta_2)$$

$$\prod_{i=1}^{N} Pr_i(Y_i|P_i, x_i, \beta_0, \beta_1, \beta_2)$$

Maximum likelihood estimation then involves finding the values of β_0, β_1 and β_2 that maximize the value of the likelihood function. Intuitively, maximum likelihood estimation seeks to find values for β_0, β_1 and β_2 that make the observed sample sequence of outcomes $[Y_1, Y_2, ..., Y_N]$ most likely to have occurred given the corresponding observed sample sequences $[P_1, P_2, ..., P_N]$ and $[x_1, x_2, ..., x_N]$. Simply put, it seeks parameter estimates under which what you observe in your sample is most likely to have occurred. Looked at from this standpoint, the phrase "maximum likelihood" is intuitively quite sensible.

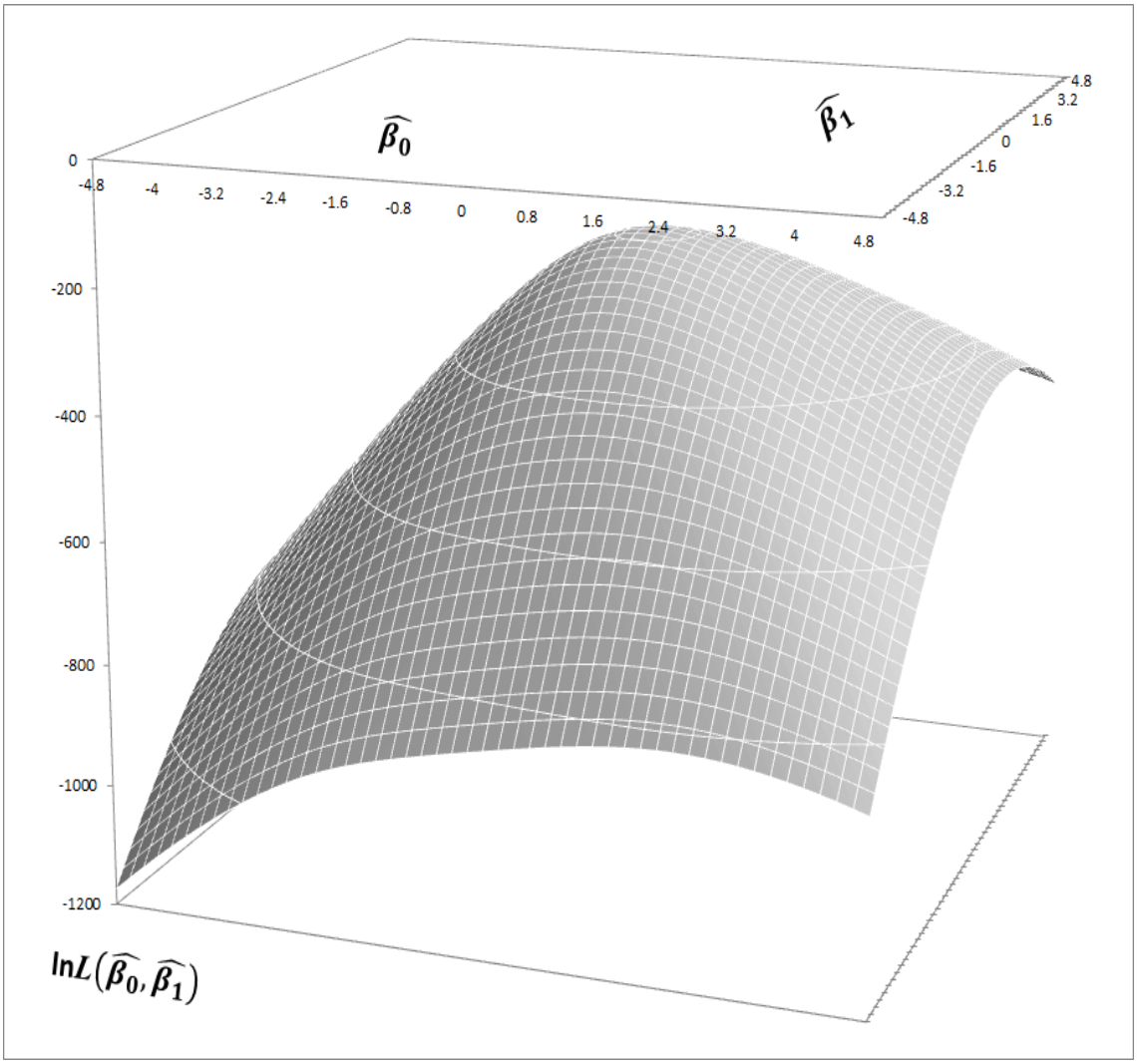


Figure 4.3: Logit Log Likelihood

In practice, the **log-likelihood** (usually denoted by $\mathcal{L}(\beta_0, \beta_1, \beta_2)$ produced simply by taking

the natural logarithm of the likelihood function) is maximized:

$$\mathcal{L}(\beta_0, \beta_1, \beta_2) = ln\{L(\beta_0, \beta_1, \beta_2)\}$$

$$= \sum_{i=1}^{N} lnPr_i(Y_i|P_i, x_i, \beta_0, \beta_1, \beta_2)$$

There are several reasons for focusing on the log-likelihood, but the main one is computational simplicity. For instance, if there were many individuals in the sample (i.e. a large N), then

$$L(\beta_0, \beta_1, \beta_2) = \prod_{i=1}^{N} Pr_i(Y_i|P_i, x_i, \beta_0, \beta_1, \beta_2)$$

is the product of a very large number of small numbers (since probabilities must be no smaller than 0 and no larger than 1) and hence a very small number itself. This risks an under-flow error[25] on many computers.

To get some visual sense of this, in Figure 4.3 we illustrate the log-likelihood "surface" for the case where we seek estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively. The easiest way to motivate this hypothetical case is simply to assume that $\beta_3 = 0$ and that x_i thus is not included as a regressor in the model. We cut the problem down to just two parameters, β_0 and β_1, simply because we cannot illustrate the log-likelihood surface in *four* dimensions, which would be what we were facing if we assumed that we were estimating β_3 as well. The focus on estimating just two parameters is a temporary device which we employ for discussing Figure 4.3.

In Figure 4.3 we have a likelihood surface. Notice that it has a single defined "summit" or "peak". Log-likelihood functions with just one "peak" such as this are known as **globally concave**. This is a fancy term that basically means that there is only one set of estimate values $\{\hat{\beta}_0, \hat{\beta}_1\}$ for which the log-likelihood function has a "peak". Put slightly differently, there is just one "global" peak, and not many "local" peaks. Maximum likelihood estimation procedures usually involve an iterative approach to estimation whereby in each iteration the values of the parameter estimates $\{\hat{\beta}_0, \hat{\beta}_1\}$ are updated by increasing or decreasing their value. The adjustment to their values is whatever adjustments increase the value of the log-likelihood (remember, the goal here is to maximize the log-likelihood). The estimation process ends when there is no change to the values of $\{\hat{\beta}_0, \hat{\beta}_1\}$ that could improve the log-likelihood value. Clearly, the iterative approach would thus require adopting some initial values for the parameter estimates $\{\hat{\beta}_0, \hat{\beta}_1\}$. When there is a global maximum as in Figure 4.3, the starting point does not matter: whatever the starting point, incremental changes in the parameter estimates $\{\hat{\beta}_0, \hat{\beta}_1\}$ in one direction will ultimately yield estimate values at the global maximum for the log-likelihood. Matters become a bit more complicated when there are many peaks (i.e. local maximums): then, the starting point for the parameter estimates $\{\hat{\beta}_0, \hat{\beta}_1\}$ might be quite consequential, since different starting points can lead to different peaks.[26] Fortunately, the simple logit model is globally concave, and so this is not concern in the context of our discussion of binary regression.

The maximum likelihood estimator has a number of properties that will be relevant to later discussions in this manual. For instance, one important one is that maximum likelihood estimates

[25] A situation where one attempts to work with a number smaller than the computer is capable of handling.

[26] This becomes a more likely possibility with more complicated models, for which global concavity might not hold. Dealing with this problem has led to the introduction of techniques such as **simulated annealing** which is simply a comparatively efficient method of identifying a global maximum. Of course, a less efficient but in many problems probably just as effective (in terms of final product) method it to re-estimate with many different candidate starting values for the parameter estimates.

(in this application $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$) are asymptotically normal, a very handy property that we will shortly exploit. However, it also has pitfalls. It typically requires assumptions (in the sense that some specific probability distribution must be adopted to operationalize the probabilities that are the foundation of the likelihood function) that some regard as strong. Depending on the context and particular assumption, inferences might not be very robust to violations of those assumptions. (Translation: the assumptions are really important and if they are wrong the estimates may be very misleading.) It can also have some undesirable numerical properties, such as the "incidental parameters problem" that we will discuss in the next chapter.

We conclude this discussion with a numerical example of a logit regression and the appropriate manner for calculating program impact from it. This example is captured in STATA do-file 4.2.do. In this example, we carry over much of the example from STATA do-file 4.1.do. Once again, this example considers the consequences of regressing Y on P when the "true" model is

$$Y^* = \beta_1 + \beta_2 \cdot P + \beta_3 \cdot x + \epsilon$$

Specifically, suppose that we randomly generate 30,000 observations based on these two assumed equations:

$$Y^* = 1 - .5 \cdot P + 1.5 \cdot x + e$$

where $e \sim$ N(0,25) and $x \sim$ N(0,4). P equals 1 if

$$.5 \cdot x + e_p > 0$$

where $e_p \sim$ N(0,36). x is thus a determinant of Y and P, the exact circumstance to which models relying on a selection on observables assumption appeal.

The main innovation over STATA do-file 4.1.do is that the observed Y is now binary. Specifically, for each observation i, we set Y_i as follows:

$$Y_i = \begin{cases} 1 & if\ Y_i^* > 0 \\ 0 & otherwise \end{cases}$$

We display the result in Output 4.5. Y takes on the value 0 for 41.82 percent of the sample, while it is 1 for the remaining 58.18 percent.

STATA Output 4.5 (4.2.do)

```
. * The binary outcome
. tab Y
```

Y	Freq.	Percent	Cum.
0	12,545	41.82	41.82
1	17,455	58.18	100.00
Total	30,000	100.00	

We first estimate the linear probability model. This is presented in Output 4.6. This involves estimation of the "correct" model (i.e. regressing Y on P and x) by ordinary least squares. The estimate of the coefficient of P (.0396309) is thus an unbiased and consistent estimate of program impact. In other words, a correct estimate of program impact should suggest that the probability that Y equals 1 should increase around 3.96 percentage points.

Next estimate the full model by regressing Y on P and x via logit regression using maximum likelihood regression. The results are shown in Output 4.7. The estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are, respectively, .3061813, .1967581 and .499398. It cannot be emphasized strongly enough, however, that none of these are really directly interpretable since none are in the metric of interest in a logit regression, the probability that the outcome of interest Y equals 1 given program participation P and characteristic x (i.e. $Pr(Y = 1|P, X)$). In particular for this context, $\hat{\beta}_1$ is <u>not</u> the impact of program participation on the probability that outcome Y occurs. Plugging in the estimates this probability is, as we have seen, simply

$$\hat{Pr}\left(Y = 1|P, X\right) = \frac{exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot P + \hat{\beta}_2 \cdot x\right)}{1 + exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot P + \hat{\beta}_2 \cdot x\right)}$$

or, substituting in the actual estimate values,

$$\hat{Pr}\left(Y = 1|P, X\right) = \frac{exp\left(.3061813 + .1967581 \cdot P + .499398 \cdot x\right)}{1 + exp\left(.3061813 + .1967581 \cdot P + .499398 \cdot x\right)}$$

Thus, the parameter estimates are not directly behaviorally interpretable but instead work through the logistic function.

STATA Output 4.6 (4.2.do)

```
. * Linear Regression y on P and x
. reg Y P x
```

Source	SS	df	MS		Number of obs = 30000
					F(2, 29997) = 3079.61
Model	1243.40331	2	621.701657		Prob > F = 0.0000
Residual	6055.69585	29997	.201876716		R-squared = 0.1704
					Adj R-squared = 0.1703
Total	7299.09917	29999	.243311416		Root MSE = .44931

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.0396309	.0052319	7.57	0.000	.0293762	.0498857
x	.1002254	.0013104	76.48	0.000	.0976569	.1027938
_cons	.5625395	.0036772	152.98	0.000	.555332	.5697469

The goal of the logit regression is to understand how the regressors P and X work to shape the probability that the outcome Y equals 1. Our main regressor of interest is P, which is itself binary. The impact of P on the probability that Y equals 1 is captured by comparing that probability when $P = 1$ with its value when $P = 0$:

$$\hat{Pr}\left(Y = 1|P = 1, X\right) - \hat{Pr}\left(Y = 1|P = 0, X\right)$$

This is the **marginal effect** of P on $\hat{Pr}(Y = 1|P, x)$ and it is the actual estimate of program impact. In practice, this is computed for each of the $i = 1, ..., N$ individuals in our sample at their various values for x_i (remember that $\hat{Pr}(Y = 1|P, x)$ depends on the value of x as well). The marginal effects for each of these $i = 1, ..., N$ individuals is then averaged across them to form an overall average marginal effect estimate for the sample.[27] When we perform this exercise with

[27] An alternative might simply be to calculate one marginal effect evaluated at the average of x, but this appears to be less common practice.

this example (see STATA do-file 4.2.do for the calculations), this leads to an average marginal effect estimate of .0397071. In other words, participation in the program appears to increase the probability that Y equals 1 by just under 4 percentage points. Reassuringly, this is right in the neighborhood of the benchmark estimate of program impact of 3.96 percentage points provided by the linear probability model.

STATA Output 4.7 (4.2.do)

```
. * Logit regressing y on P and x
. logit Y P x
Iteration 0:   log likelihood =   -20390.8
Iteration 1:   log likelihood = -17607.838
Iteration 2:   log likelihood = -17584.319
Iteration 3:   log likelihood = -17584.285
Iteration 4:   log likelihood = -17584.285
Logistic regression                               Number of obs   =      30000
                                                  LR chi2(2)      =    5613.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -17584.285                       Pseudo R2       =     0.1376
------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .1967581   .0258382     7.62   0.000     .1461163       .2474
           x |    .499398   .0077492    64.45   0.000     .4842098    .5145862
       _cons |   .3061813   .0181661    16.85   0.000     .2705765    .3417861
------------------------------------------------------------------------------
```

STATA Output 4.8 (4.2.do)

```
. * Logit regressing y on P alone
. logit Y P
Iteration 0:   log likelihood =   -20390.8
Iteration 1:   log likelihood = -20262.266
Iteration 2:   log likelihood = -20262.234
Iteration 3:   log likelihood = -20262.234
Logistic regression                               Number of obs   =      30000
                                                  LR chi2(1)      =     257.13
                                                  Prob > chi2     =     0.0000
Log likelihood = -20262.234                       Pseudo R2       =     0.0063
------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .3763136   .0235249    16.00   0.000     .3302056    .4224216
       _cons |   .1458089   .0163407     8.92   0.000     .1137817    .1778361
------------------------------------------------------------------------------
```

Finally, we attempt to regress Y on P alone by logit. There is reason to believe this might cause problems again, given that x influences both P and Y. However, this is not necessarily evident from the regression output (see Output 4.8). While it is true the coefficient on P is larger in magnitude than when Y was regressed on P and x (.3763136 versus .1967581) it is not clear what this means for the marginal effect via the highly non-linear logistic function. True understanding of the implication of excluding x requires re-calculation of the marginal effect of program participation P on the probability that Y equals 1 using this new fitted model. When we re-compute the marginal effect with this new model, the average value is now .0912561. In other words, excluding x has

nearly doubled our estimate of the impact of program participation.

Unlike the linear ordinary least squares case, for which the mathematics are comparatively easy, it is not straightforward to demonstrate analytically omitted variable bias in the context of marginal effects computation with the logistic function. Nonetheless, this example demonstrates that it clearly is a concern even in the limited dependent variable setting. The reason for the bias is much the same: the exclusion of x has forced P to serve the two roles of a control for itself (which we want it to do) and a proxy for x (which we do not want). The result is an estimate of the marginal effect of P on the probability that Y equals 1 that is muddied since it partly reflects the effect of x on Y.

Moreover, the logic of the regression approach to selection on observables obtains in the limited dependent variable setting: the omitted variable bias can be remedied simply by including x as a regressor. Whether this fully remedies the omitted variable bias problem depends crucially on whether x is the only variable that systematically influences both the outcome Y and program participation P.

4.1.2 Multiple Regression

We now briefly digress to consider multiple regression models with many controls. The preceding discussion considered the possibility of a single confounder (x). While that discussion technically revolved around multiple regression (since regression of Y on P and x does involve multiple regressors, namely P and x), this was a somewhat restrictive example in the sense that there were only two potential controls. In reality, there are probably many factors that shape both program participation and the outcome of interest.

We now relax the framework of the preceding discussion to consider many potential controls for factors that might influence both Y and P. For concreteness, we consider 5 such factors $(x_1, x_2, ..., x_5)$. The choice of 5 is arbitrary. We extend thusly for two reasons. First, we establish that partial control only for *some* of the elements of $x_1, x_2, ..., x_5$ is not sufficient to recover an unbiased estimate of program impact. This will highlight how strong an assumption "selection on observables" can be: in general, we must control for all factors that influence both Y and P to insure unbiased estimation of program impact. Second, we introduce a key property of multiple regression that can be very important for understanding the behavior of multiple regression models, the regression anatomy property.

The multiple regression framework with more than one potential confounder is mathematically more complicated to discuss than the case of the single confounder.[28] In general, analysis of the multiple regression case relies on linear algebra, a mathematical approach beyond the scope of the manual.

A simple point can still be made, however, with simple math. Suppose that the true population regression model governing Y is given by

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x_1 + \beta_3 \cdot x_2 + \beta_4 \cdot x_3 + \beta_5 \cdot x_4 + \beta_6 \cdot x_5 + \epsilon$$

where we assume ϵ is mean independent of all observed regressors (i.e. the Ps and xs). We now have program participation status (P) and 5 potential confounders $(x_1, ..., x_5)$. Suppose that we attempt to estimate the impact of the program P on the outcome Y (in other words, β_1) simply by regressing Y on P alone. To fix ideas, suppose that we have data in which we observe $\{Y_i, P_i, x_{1i}, x_{2i}, ..., x_{5i}\}$.

[28] Indeed, the mathematics grows more complex with more than one regressor, but our simple discussion of omitted variable bias in the preceding subsection used the single regressor case (i.e. regressing Y on P alone) to form the basis for the mathematical discussion of omitted variable bias.

Our estimate of β_1, $\hat{\beta}_1$, from regression of Y on P alone is once again given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot Y_i}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}$$

To examine the expectation of this estimate, $E(\hat{\beta}_1)$, we substitute in the true data generating process behind Y_i:

$$E\left(\hat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot Y_i}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

$$= E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot (\beta_0 + \beta_1\cdot P_i + \beta_2\cdot x_{1i} + \beta_3\cdot x_{2i} + \beta_4\cdot x_{3i} + \beta_5\cdot x_{4i} + \beta_6\cdot x_{5i} + \epsilon_i)}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

$$= E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_0}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_1\cdot P_i}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_2\cdot x_{1i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

$$+E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_3\cdot x_{2i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_4\cdot x_{3i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_5\cdot x_{4i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

$$+E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \beta_6\cdot x_{5i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \epsilon_i}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

Using the same math as in the last subsection, we have

$$E\left(\hat{\beta}_1\right) = \beta_0\cdot 0 + \beta_1\cdot 1 + \beta_2\cdot E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot x_{1i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

$$+\beta_3\cdot E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot x_{2i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + \beta_4\cdot E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot x_{3i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + \beta_5\cdot E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot x_{4i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right)$$

$$+\beta_6\cdot E\left(\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot x_{5i}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}\right) + 0$$

$$= \beta_1 + \beta_2\cdot E\left(\hat{\gamma_{11}}\right) + \beta_3\cdot E\left(\hat{\gamma_{12}}\right) + \beta_4\cdot E\left(\hat{\gamma_{13}}\right) + \beta_5\cdot E\left(\hat{\gamma_{14}}\right) + \beta_6\cdot E\left(\hat{\gamma_{15}}\right)$$

where $\hat{\gamma_{1j}}$ is an estimate of the slope coefficient γ_{1j} from $x_j = \gamma_{0j} + \gamma_{1j}\cdot P + \upsilon_j$, where $j = 1, 2, ..., 5$.

This is a generalization of the omitted variable bias result from the last subsection to the more general multiple omitted variable (i.e. potential confounder) setting. There are three major conclusions to draw from this exercise:

1. When one regresses Y on P alone, the estimate of β_1, $\hat{\beta}_1$, is still biased to the extent that any of the omitted regressors $x_1, ..., x_5$ are correlated with Y and P;

2. The direction of the bias is much harder to gauge *a priori*, since it depends on the net influence of the biases introduced by all of the omitted variables;

3. Assuming that all of $x_1, ..., x_5$ are related to Y and P, recovering an unbiased estimate of β_1 requires controlling for <u>all</u> of the regressors $x_1, ..., x_5$.

The result from the simple case of a single omitted variable thus generalizes, but the resulting bias is more complicated in terms of magnitude and direction since it depends on the net bias across all of the omitted regressors.

A numerical example is perhaps in order. The example is to be found in STATA do-file 4.3.do. In that .do file we draw 50,000 observations as follows. First, the overall data generating process is given by

$$Y_i = 1 + .5 \cdot P_i + 1.5 \cdot x_{1i} + 1.3 \cdot x_{2i} + .5 \cdot x_{3i} + 1 \cdot x_{4i} - .5 \cdot x_{5i} + e_i$$

where $e_i \sim$ N(0,25) is independent of $x_{1i}, x_{2i}, ..., x_{5i}$ and eP_i (described below), $x_{ji} \sim$ N(0,4), and program participation P_i equals 1 if

$$-.5 \cdot x_{1i} + x_{2i} - 1.5 \cdot x_{3i} + 2 \cdot x_{4i} + 4 \cdot x_{5i} + eP_i > 0$$

and 0 otherwise (and $eP_i \sim$ N(0,144)).[29] We hence have preserved and extended the basic setup from STATA do-file 4.1.do. There are now five potential controls $x_{1i}, x_{2i}, ..., x_{5i}$. They each shape both Y and P. The fact that e_i is independent of $x_{1i}, x_{2i}, ..., x_{5i}$ and eP_i means that it is also independent of P_i. Once again, true program impact is .5.

STATA Output 4.9 (4.3.do)

```
. * Regressing Y on P and x1-x5
.
. reg Y P x*
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  6, 49993) = 7192.22
       Model |  1083600.01     6  180600.002            Prob > F      =  0.0000
    Residual |  1255346.73 49993  25.1104501            R-squared     =  0.4633
-------------+------------------------------           Adj R-squared =  0.4632
       Total |  2338946.74 49999  46.7798705            Root MSE      =   5.011

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    .5188164    .051738    10.03   0.000     .4174093    .6202235
          x1 |    1.490441   .0112458   132.53   0.000     1.468399    1.512483
          x2 |    1.292802   .0113638   113.77   0.000     1.270529    1.315075
          x3 |    .5112961   .0113817    44.92   0.000     .4889879    .5336044
          x4 |    .9998604   .0115517    86.56   0.000     .9772189    1.022502
          x5 |   -.5005947   .0123521   -40.53   0.000    -.5248049   -.4763846
       _cons |    .9789824   .0341271    28.69   0.000     .9120928    1.045872
------------------------------------------------------------------------------
```

In Output 4.9 we provide results from regression of Y on P and $x_1, ..., x_5$. The estimate of the coefficient on program participation P is .5188164, which is right in the near neighborhood of the true value of .5. The xs are significant predictors of Y, as we would expect by construction.

[29] The specific values of the data generating process were, as always, rather arbitrarily chosen. As with all of the .do files, we encourage the reader to study results under alternative values.

Moreover, the estimates of their coefficients are roughly in line with what we would expect, given their assumed values used to generate the data.

STATA Output 4.10 (4.3.do)

```
. * Regressing Y on just P and x2, x4, x5
.
. reg Y P x2 x4 x5
```

Source	SS	df	MS		
Model	595869.311	4	148967.328		
Residual	1743077.43	49995	34.8650351		
Total	2338946.74	49999	46.7798705		

Number of obs	=	50000			
F(4, 49995)	=	4272.69			
Prob > F	=	0.0000			
R-squared	=	0.2548			
Adj R-squared	=	0.2547			
Root MSE	=	5.9047			

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	-.2856236	.0598212	-4.77	0.000	-.4028738	-.1683734
x2	1.324834	.0133844	98.98	0.000	1.2986	1.351068
x4	1.049367	.0135961	77.18	0.000	1.022718	1.076015
x5	-.41929	.0145117	-28.89	0.000	-.447733	-.390847
_cons	1.365239	.0397911	34.31	0.000	1.287248	1.44323

STATA Output 4.11 (4.3.do)

```
. * Regressing Y on just P and x1, x3, x5
.
. reg Y P x1 x3 x5
```

Source	SS	df	MS		
Model	585667.707	4	146416.927		
Residual	1753279.04	49995	35.0690876		
Total	2338946.74	49999	46.7798705		

Number of obs	=	50000			
F(4, 49995)	=	4175.10			
Prob > F	=	0.0000			
R-squared	=	0.2504			
Adj R-squared	=	0.2503			
Root MSE	=	5.9219			

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	2.245759	.059019	38.05	0.000	2.130081	2.361437
x1	1.522749	.013287	114.60	0.000	1.496707	1.548792
x3	.5878619	.0134327	43.76	0.000	.5615337	.6141901
x5	-.6770191	.0145045	-46.68	0.000	-.705448	-.6485902
_cons	.1219913	.0395227	3.09	0.002	.0445264	.1994563

In Output 4.10 and Output 4.11, we provide results when Y is regressed on P and a subset of the xs. Specifically, Output 4.10 provides the coefficient estimates when Y is regressed on just P and x_2, x_4 and x_5. The estimated coefficient on P is now -.2856236, wildly off from the true value of .5. Indeed, the program would now seem to have a negative effect on the outcome of interest Y! This demonstrates that correcting for some but not all of the factors x that influence both Y and P would not be sufficient to recover an unbiased estimate of program impact. In Output 4.11 we present estimation results for the regression of Y on P and x_1, x_3 and x_5. The estimated program impact is now 2.245759, more than four times the true value of .5. Comparing these results with those in Output 4.10 demonstrates that, not only will bias remain if one fails to correct for all of

the factors x that shape both program participation P and the outcome Y, but that the direction of the bias can depend on which particular confounders x are excluded from the regression.

We conclude this subsection with a brief digression to present the "regression anatomy formula".[30] This is a property of multiple regressions that the authors have found quite useful for understanding various phenomena involving them. The property itself is quite simple. Suppose that one regresses Y on P and a series of confounders $x_1, x_2, ..., x_5$. Call $\hat{\beta}_1$ the resulting estimated coefficient on P. The same estimate can be recovered by following these steps:

1. Regress P on all of the regressors $x_1, x_2, ..., x_5$. Thus, continuing our specific example, regress P on x_1, x_2, x_3, x_4 and x_5 to estimate the coefficients $\mu_0, ..., \mu_5$ of the model

$$P_i = \mu_0 + \mu_1 \cdot x_{1i} + \mu_2 \cdot x_{2i} + \mu_3 \cdot x_{3i} + \mu_4 \cdot x_{4i} + \mu_5 \cdot x_{5i} + \zeta_i$$

   This simply captures the variation in P driven by the observed confounders x and the remaining variation (ζ_i) not explained by the xs;

2. Compute predicted program participation $\hat{P}_i$ for each observation:

$$\hat{P}_i = \hat{\mu_0} + \hat{\mu_1} \cdot x_{1i} + \hat{\mu_2} \cdot x_{2i} + \hat{\mu_3} \cdot x_{3i} + \hat{\mu_4} \cdot x_{4i} + \hat{\mu_5} \cdot x_{5i}$$

   This is simply the value of P predicted by the observed confounders x and its variation across a sample of N individuals is the variation in P in that sample associated with variation in the confounders x;

3. Compute the predicted residuals $P_i^{\hat{RES}}$

$$P_i^{\hat{RES}} = P_i - \hat{P}_i$$

   This is simply the portion of P not associated with x. Its variation across the sample is the variation in P not associated with variation in x;

4. Regress Y on $P_i^{\hat{RES}}$.

The regression of Y on $P_i^{\hat{RES}}$ estimates the model

$$Y_i = \nu_0 + \nu_1 \cdot P_i^{\hat{RES}} + \vartheta_i$$

The ordinary least squares estimator for this model is

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i^{\hat{RES}} - \overline{P_i^{\hat{RES}}} \right) \cdot Y_i}{\sum_{i=1}^{N} \left(P_i^{\hat{RES}} - \overline{P_i^{\hat{RES}}} \right)^2}$$

$$= \frac{\sum_{i=1}^{N} P_i^{\hat{RES}} \cdot Y_i}{\sum_{i=1}^{N} \left(P_i^{\hat{RES}} \right)^2}$$

The last step exploits the fact that $P^{\hat{RES}}$ is an ordinary least squares regression residual, hence its mean

$$\overline{P_i^{\hat{RES}}}$$

[30] We are unsure of the origin of this term, but it does appear in Angrist and Pischke (2009).

is zero. This estimate yielded by this estimator actually equals the estimate $\hat{\beta}_1$ emerging from estimation of the regression model

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x_1 + \beta_3 \cdot x_2 + \beta_4 \cdot x_3 + \beta_5 \cdot x_4 + \beta_6 \cdot x_5 + \epsilon$$

by regression of Y on P and all of the xs.

STATA Output 4.12 (4.3.do)

```
. * Regressing P on x1-x5
.
. reg P x*
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  5, 49994) = 3324.55
       Model |  3119.0354     5   623.80708             Prob > F      =  0.0000
    Residual |  9380.69082 49994  .187636333            R-squared     =  0.2495
-------------+------------------------------           Adj R-squared =  0.2495
       Total |  12499.7262 49999  .249999524            Root MSE      =  .43317

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0124907   .0009705   -12.87   0.000     -.014393   -.0105885
          x2 |    .027836   .0009744    28.57   0.000     .0259262    .0297459
          x3 |  -.040647    .0009669   -42.04   0.000    -.0425422   -.0387518
          x4 |   .0518833   .0009712    53.42   0.000     .0499796    .0537869
          x5 |   .1014909   .0009665   105.01   0.000     .0995966    .1033851
       _cons |   .4974507   .0019373   256.78   0.000     .4936536    .5012478
------------------------------------------------------------------------------
```

STATA Output 4.13 (4.3.do)

```
. * Regressing Y on "Residual P"
.
. reg Y PRES
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  1, 49998) =   54.03
       Model |  2525.00481     1  2525.00481             Prob > F      =  0.0000
    Residual |  2336421.74 49998   46.730304            R-squared     =  0.0011
-------------+------------------------------           Adj R-squared =  0.0011
       Total |  2338946.74 49999  46.7798705            Root MSE      =   6.836

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        PRES |   .5188164     .07058     7.35   0.000     .3804787    .6571541
       _cons |   1.222157   .0305713    39.98   0.000     1.162237    1.282077
------------------------------------------------------------------------------
```

The intuition behind this goes to the heart of what multiple regression actually does. Multiple regression controls for the interrelationships between regressors to isolate the channel of variation in each regressor associated with that regressor alone, and not somehow reflective of its relationship with another regressor. In Output 4.12 and Output 4.13 we extend the example from STATA do-file 4.3.do to demonstrate the regression anatomy formula. To begin with, in Output 4.12 we show the estimates from regression P on x_1, x_2, x_3, x_4 and x_5. We used this fitted model to obtain predicted program participation and then predicted residual program participation. In Output 4.13

we show results from regressing Y on the predicted residual program participation (denoted P^{RES}, or $PRES$ in the STATA output). Notice that the coefficient on P^{RES} is .5188164. This is exactly the same coefficient estimate as that obtained for P when regressing Y on P and x_1, x_2, x_3, x_4 and x_5 (see Output 4.8).

4.1.3 "Bad" Controls

A key lesson of the past subsection is that failure to control for all factors that influence both the outcome of interest Y and program participation P will result in a biased estimate of program impact from the regression of Y on P and whatever potential confounding factors are included. A natural temptation would be to adopt an "everything and the kitchen sink" approach to covariate inclusion in program impact evaluation regressions. After all, it is tempting to assume that doing so is a good insurance strategy against excluding a confounder than influences both Y and P.

Unfortunately, some controls can actually *introduce* bias. In other words, suppose that you had a sample of N individuals for whom you observed $\{Y_i, P_i, x_i\}$. The preceding discussion might lead one to conclude that it would be safer to regress Y on P and x rather than on just P in case x influences both P and Y. However, if x has certain statistical properties (explained below) one might actually worsen the regression estimate of the impact of program participation P on Y from a bias standpoint by introducing x. We call such problematic variables "bad controls" (following Angrist and Pischke (2009), which provides one of the best discussions of this possibility that we have seen).

In most of the discussions of the problem of bad controls that the authors have seen (e.g. Angrist and Pischke 2009[31]) the problem is framed in terms of regressors that are other outcomes of the experiment the effect of which one is studying. In plain terms, program impact analysis seeks to estimate the impact of a program on some outcome of interest that that program sought to influence. The classic definition of a bad control focuses on a regressor that happens to be yet another outcome that depends on program participation.

Let us consider a simulated example (see do-file 4.4.do). To fix ideas, let us assume that program participation P is randomly determined. This rules out the possibility that any bias to the estimate of program impact from a regression of Y on P owing to some intrinsic unobservable associated with both Y and P. Intuitively, because P is randomly determined participants and non-participants differ only by their program participation; any differences between them then can be attributed to their program participation. P does not also serve as a proxy for some other determinant of the outcome of interest. Therefore, simply regressing Y on P should yield an unbiased estimate of program impact.

The data generating process involves simulating 50,000 observations. First, program participation is determined simply by whether a standard normal distributed random variable exceeds 0. We next draw two variables ey and ex from the bivariate normal distribution with correlation .3. The potential outcomes for the outcome of interest Y (Y^1 and Y^0, represented by `Y_1` and `Y_0`, respectively, in the do-file 4.4.do since STATA does not allow variable names with explicit superscripts) are generated as follows:

$$Y^0 = 1 + ey$$
$$Y^1 = 3 + ey$$

The true average treatment effect is then

$$E\left(Y^1 - Y^0\right) = E\left(Y^1\right) - E\left(Y^0\right) = E\left(3 + ey\right) - E\left(1 + ey\right) = 3 + 0 - 1 - 0 = 2$$

[31] Frankly, many of the discussions of bad controls that we have seen simply outline that of Angrist and Pischke (2009).

The observed outcome is then $Y = P \cdot Y^1 + (1 - P) \cdot Y^0$.

We also consider another outcome, X. X is an outcome in the sense that its value depends on program participation P. Specifically, the potential outcomes X^1 and X^0 are determined as follows:

$$X^0 = 1 + ex$$

$$X^1 = 4 + ex$$

(Notice that, by logic similar to that in the last paragraph, the true impact of P on X is 3.) Observed X is then $X = P \cdot X^1 + (1 - P) \cdot X^0$.

First we regress Y on P. The result is shown in Output 4.14. Unsurprisingly, the estimated program impact is, at 1.964227, right around the expected program impact of 2. This is a natural consequence of the randomization of program participation P: it cut off any avenue for omitted variable bias in the regression of Y on P. In particular, the correlation between P and the unobserved determinant of Y, ey, is essentially 0 at -0.0045 (see Output 4.15).

STATA Output 4.14 (4.4.do)

```
. reg Y P
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  1, 49998) = 3015.82
       Model |  48223.5896     1  48223.5896            Prob > F      =  0.0000
    Residual |  799478.746 49998  15.9902145            R-squared     =  0.0569
-------------+------------------------------           Adj R-squared =  0.0569
       Total |  847702.335 49999  16.9543858            Root MSE      =  3.9988

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    1.964227   .0357675    54.92   0.000     1.894122    2.034331
       _cons |    .9787719   .0251799    38.87   0.000      .929419    1.028125
------------------------------------------------------------------------------
```

STATA Output 4.15 (4.4.do)

```
. corr P ey
(obs=50000)
             |        P       ey
-------------+------------------
           P |   1.0000
          ey |  -0.0045   1.0000
```

We then regress Y on P and X, with the results in Output 4.16. The results are quite striking: the introduction of X as a regressor has resulted in a program impact estimate of 1.062924, half of the true population program impact[32] and half of the previous estimate. This reflects the fact that the introduction of the regressor X has created a bias to the estimate of program impact where none had been evident in the simpler regression of Y on P. In other words, the addition of the

[32]In simulations such as these, one can think of an infinite population of which the 50,000 simulated observations are akin to a sample and whose outcomes and program participation are determined by the indicated data generating process.

regressor X, a move the preceding discussion would suggest might move us closer to a bias free estimate of program impact, has completely backfired by actually *introducing* a bias to the estimate of program impact!

STATA Output 4.16 (4.4.do)

```
. reg Y P X
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  2, 49997) = 4091.19
       Model |  119221.309     2  59610.6543            Prob > F      =  0.0000
    Residual |  728481.027 49997  14.5704948            R-squared     =  0.1406
-------------+------------------------------           Adj R-squared =  0.1406
       Total |  847702.335 49999  16.9543858            Root MSE      =  3.8171

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   1.062924   .0365027    29.12   0.000     .9913784     1.13447
           X |   .2973571   .0042598    69.80   0.000     .2890078    .3057065
       _cons |    .691562   .0243857    28.36   0.000     .6437656    .7393583
------------------------------------------------------------------------------
```

Let us begin to unravel what has happened by remembering what it is that generates biased estimates in a regression context: correlation between a regressor and some unobserved determinant of the outcome.[33] By this logic, the general truth of what has happened should be clear. The introduction of the regressor X has somehow generated a correlation between P and an unobserved determinant of Y. Since the only unobserved determinant of Y in this example is ey, this is tantamount to suggesting that the introduction of X as a regressor has somehow generated a correlation between P (or, more precisely, the variation in residual P used in the attempt to identify program impact) and ey where none necessarily exists in the data.

STATA Output 4.17 (4.4.do)

```
. reg Y PRES
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  1, 49998) =  739.46
       Model |  12354.6364     1  12354.6364            Prob > F      =  0.0000
    Residual |  835347.699 49998  16.7076223            R-squared     =  0.0146
-------------+------------------------------           Adj R-squared =  0.0146
       Total |  847702.335 49999  16.9543858            Root MSE      =  4.0875

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        PRES |   1.062924   .0390881    27.19   0.000     .9863109    1.139537
       _cons |   1.952243   .0182798   106.80   0.000     1.916414    1.988071
------------------------------------------------------------------------------
```

To see how this is possible, we return to the regression anatomy formula. According to the regression anatomy formula, regressing Y on P and X is the same as regression Y on P^{RES}, the predicted residual from the regression of P on X. Given the regression model

$$P = \mu_0 + \mu_1 \cdot X + \zeta$$

[33] There is a slightly more complicated path to omitted variable bias in a regression context due to reliance for estimation on a non-random sample, but that is a subject for a subsequent chapter.

we have

$$P^{RES} = P - (\hat{\mu_0} + \hat{\mu_1} \cdot X)$$

where $\hat{\mu_0}$ and $\hat{\mu_1}$ are the estimates of μ_0 and μ_1 generated by least squares regression of P on X. Output 4.17 provides the results from the regression of Y on P^{RES}. At 1.062924, the estimate of program impact is what was obtained by regressing Y on P and X.

To understand the source of the omitted variable bias in the regression of Y on P^{RES} where none existed in the regression of Y on P, it is necessary to look more closely at P^{RES}. Least squares regression attempts to parse the variation in P into two parts: a systematic component associated with X and another, random component not associated with X. Thus, by construction, P^{RES} should be uncorrelated with X.

However, that only means that P^{RES} should be uncorrelated with X *as a whole*. It need not be uncorrelated with the *components* of X. In particular, P^{RES} could be correlated with ex. Indeed, the regression exercise creates a natural avenue by which P^{RES} depends on ex. To see this, let us substitute in for X:

$$\begin{aligned}
P^{RES} &= P - (\hat{\mu_0} + \hat{\mu_1} \cdot X) \\
&= P - \left(\hat{\mu_0} + \hat{\mu_1} \cdot \left(P \cdot X^1 + (1 - P) \cdot X^0\right)\right) \\
&= P - (\hat{\mu_0} + \hat{\mu_1} \cdot (P \cdot (4 + ex) + (1 - P) \cdot (1 + ex))) \\
&= P - \hat{\mu_0} - \hat{\mu_1} \cdot P \cdot 4 - \hat{\mu_1} \cdot P \cdot ex - \hat{\mu_1} \cdot 1 - \hat{\mu_1} \cdot ex + \hat{\mu_1} \cdot P \cdot 1 + \hat{\mu_1} \cdot P \cdot ex \\
&= P - \hat{\mu_0} - \hat{\mu_1} \cdot 3 \cdot P - \hat{\mu_1} - \hat{\mu_1} \cdot ex
\end{aligned}$$

Thus, ex is a determinant of P^{RES}.

Output 4.18 provides the correlations between P, ey, ex and X. To begin with, P is uncorrelated with the terms ey and ex, which is unsurprising given that P is independently, randomly determined. P is, however, correlated with X, a natural consequence of the fact that program participation is a determinant of X by operating as the switching mechanism that determines whether, for a given observation, X is equal to X^1 or X^0:

$$X = P \cdot X^1 + (1 - P) \cdot X^0$$

Output 4.19 provides the correlations between P^{RES}, ey, ex and X. The situation is somewhat different for P^{RES}. P^{RES} is completely uncorrelated with X, which it should be as the predicted residual from the regression of P on X. However, P^{RES} is highly correlated with ey and ex. The correlation with ex is -.3501. The reason for this is made clear by the regression anatomy formula: ex is a determinant of P^{RES}. (Remember that P^{RES} should be uncorrelated by construction with X as a whole, but would not necessarily be uncorrelated with the individual components that determine the values of X.) The correlation between P^{RES} is derivative of the correlation between P^{RES} and ex: because ex and ey are correlated, correlation between P^{RES} and ex generates correlation between P^{RES} and ey. Notice that the correlation between P^{RES} and ey is about a third of the value of the correlation between P^{RES} and ex: P^{RES} should be correlated with ey, but less so than ex by a factor that reflects the correlation between ey and ex.

We have thus seen the genesis of the bad control problem in this instance. Via the regression anatomy formula, we have seen how P^{RES} is correlated with ex where P was not. However, because ex and ey are correlated, the correlation between P^{RES} and ex generates a correlation between P^{RES} and ey. Thus, a classic omitted variable bias to the estimate of program impact exists in the regression of Y on P^{RES} (which is, for the purpose of estimating program impact, equivalent to regressing Y on P and X) where none was present in a regression of Y on P alone.

STATA Output 4.18 (4.4.do)

```
. corr P ey ex X
(obs=50000)
```

	P	ey	ex	X
P	1.0000			
ey	-0.0045	1.0000		
ex	0.0039	0.2980	1.0000	
X	0.3537	0.2772	0.9367	1.0000

STATA Output 4.19 (4.4.do)

```
. corr PRES ey ex X
(obs=50000)
```

	PRES	ey	ex	X
PRES	1.0000			
ey	-0.1096	1.0000		
ex	-0.3501	0.2980	1.0000	
X	0.0000	0.2772	0.9367	1.0000

Intuitively, X is an endogenous regressor. This means that X is correlated with unobservables that determine Y. The regression anatomy formula tells us that the variation in P used to identify program impact in a regression of Y on P and X is the variation in P^{RES}. However, P^{RES} is contaminated by the endogenous variation in X. Regression of Y on P and X thus creates an avenue for P to become contaminated by the endogenous variation in X.

The example we have covered focuses on the possibility of X being another outcome variable the value of which depends on program participation. However, the problem of bad control bias is much broader than this. A bad control is any control which:

1. Is correlated with the regressor of interest (even if it is not explicitly *caused* by it);
2. Is correlated with the unobservables in the regression of interest (i.e. is endogenous);

The broadness of these criteria suggest that the bad control problem is likely not a marginal concern but instead a significant consideration in selecting appropriate regression specifications (i.e. in this context, which controls to include). We were able to ignore this potential complication in earlier examples of omitted variable bias and multiple regression because we assumed the error or random term ϵ was mean independent of the regressors introduced to serve as controls for potential omitted variable bias (e.g. x or $x_1, x_2, ..., x_5$). We thus did not then need to be concerned with this possibility, a luxury typically not available with a real world sample.

To demonstrate, we briefly discuss another example (also capture in do-file 4.4.do). Once again we draw 50,000 observations from a contrived but illustrative data generating process. First, we draw two variables, μ_1 and μ_2 from the a bivariate normal distribution with correlation .5. Next, we separately draw two additional variables ex and ey from a bivariate normal distribution with correlation .5. We then determine P, Y^1, Y^0 and Y as follows:

$$P = \begin{cases} 1 & if\ \mu_1 > 0 \\ 0 & otherwise \end{cases}$$

$$Y^1 = 3 + 4 \cdot ey$$
$$Y^0 = 1 + 4 \cdot ey$$
$$Y = P \cdot Y^1 + (1 - P) \cdot Y^0$$

Per the logic applied earlier in this discussion, the true average treatment effect should be 2. Output 4.20 provides results from the regression of Y on P. The estimated program impact of 2.05918 is right in the neighborhood of the true average treatment effect of 2. This reflects the lack by construction of a correlation between program participation P and any unobservables that determine the outcome Y.

STATA Output 4.20 (4.4.do)

```
. reg y P
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  1, 49998) = 3320.49
       Model |  53002.7548     1  53002.7548            Prob > F      =  0.0000
    Residual |  798085.331 49998  15.9623451            R-squared     =  0.0623
-------------+------------------------------           Adj R-squared =  0.0623
       Total |  851088.086 49999  17.0221022            Root MSE      =  3.9953

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |     2.05918    .035735    57.62   0.000      1.98914    2.129221
       _cons |    .9908723   .0252563    39.23   0.000     .9413697    1.040375
------------------------------------------------------------------------------
```

We now introduce a complication. Specifically, we generate a variable X as follows:

$$X = \mu_2 + ex$$

Notice that x is correlated with P because μ_1 and μ_2 are correlated (indeed, the correlation between P and X is .2846). However, as this framework is presented, P does not really *cause* variation in X (as, for instance, by serving as a switching mechanism between some hypothesized potential outcomes X^1 and X^0). Rather, the framework is more consistent with the existence of some unobserved variables (the effect of which is captured by μ_1 and μ_2) that influence both X and P. X is also clearly correlated with the unobservable ey that shapes Y because ex and ey are correlated. Indeed, the correlations between ex and Y and ey are, respectively, .4116 and .3538. Thus, X is heavily correlated with the "unobservable" ey that determines Y.

Output 4.21 provides the results from a regression of Y on P *and* X. Plainly, the inclusion of the variable X as a regressor has had a huge effect on the estimate of program impact: it is now 1.1889, only around 60 percent of the true program impact of 2. The reason for this is much the same as in the first example: where the overall variation in P is uncorrelated with ey (with a sample correlation of essentially zero at 0.0074) the variation in "residual" P (P^{RES}, using the variable definitions from the first example) per the regression anatomy formula has a correlation with ey of -.0973. Once again, the endogeneity of X has contaminated the empirically utilized variation in P.

The difference in this example, of course, is that P does not cause variation in X. In other words, we have seen that, essentially, a bad control is an *endogenous* control also correlated with program participation. This is indeed a frightening notion, because it suggests that the bad control threat might be quite widespread.

Unfortunately, ready tests for bad controls, so broadly defined, are not obvious. A bad control must be correlated with the regressor of interest (a condition typically easily tested) and correlated

with unobservables that determine the outcome of interest (a condition typically impossible to test). In the end, selection of controls must be guided by judgment and informed appeals to theory to avoid (hopefully) the pitfalls of bad controls. A useful guiding principle might be to prefer controls less subject to the choices of the individuals whose outcomes are under consideration.

STATA Output 4.21 (4.4.do)

```
. reg Y P X
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  2, 49997) = 5806.89
       Model |  160432.176     2  80216.0881            Prob > F      =  0.0000
    Residual |  690655.909 49997   13.813947            R-squared     =  0.1885
-------------+------------------------------           Adj R-squared =  0.1885
       Total |  851088.086 49999  17.0221022            Root MSE      =  3.7167

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |      1.1889   .0346772    34.28   0.000     1.120932    1.256868
           X |    1.079302   .0122388    88.19   0.000     1.055314     1.10329
       _cons |    1.440349   .0240418    59.91   0.000     1.393227    1.487471
------------------------------------------------------------------------------
```

For instance, consider the evaluation of the impact of a health program on some health outcome. It may be that those more likely to participate in the program are also likely to make healthier choices, like smoking less or not at all. One might be tempted to control for the individual's smoking choices in a regression of the health outcome on program participation and potential controls for confounding factors. In light of this discussion, it might be safer to control for the price of cigarettes than the actual demand for them, for example, since the former is less under the control of the individual.

One could argue that the application of such logic to the selection of controls reduces the likelihood of the bad control problem. Variables less under the control of the individual probably do not vary so much with their individual level preferences and other unobserved characteristics that also work to shape outcomes of interest.

To be sure, however, such controls are not guaranteed to be exogenous: even if the individual does not determine specifically and purposefully determine the value of a variable, the variation in that variable might still be related to unobservables that influence the outcome of interest at the individual level. To return to the example, the variation in cigarette prices across individuals in a typical empirical sample reflects differences in the equilibrium price of cigarettes across the markets in which those individuals operate. However, equilibrium market prices reflect supply and demand determinants at the market level. For instance, it could be the cigarettes are, other things being equal, cheaper in markets where the community has stronger preferences for health and hence lower demand for cigarettes. However, the community preference for health might still shape individual-level health outcomes. For instance, there are often externalities to certain channels of health. Some individuals might, via social learning, be influenced by community-level norms.

4.1.4 The Program Participation Decision

We conclude our discussion of the unbiasedness and consistency of the regression estimator of program impact by confronting briefly an issue about which we have thus far been somewhat evasive: the determination of program participation. This serves three purposes. First, it will

make explicit from a behavioral standpoint how a correlation between program participation P and the regression errors might arise. Second, it is a good opportunity to introduce in simple terms a model of program participation to which we will appeal repeatedly in this manual. Finally, it represents our first stab at estimation of program impact in a setting where impact is not constant across individuals.

We begin by extending the simple model introduced at the beginning of this section. Suppose again that Y^1 is an individual's outcome when they participate in a program, and Y^0 is their outcome when they do not do so. Suppose as well that the individual's potential outcomes can be represented by

$$Y^0 = \beta_0 + \epsilon$$
$$Y^1 = \beta_0 + \beta_1 + \epsilon^{Y1} + \epsilon$$

where the ϵs are some individual level random determinants of the potential outcomes and the βs are population-level parameters. This extension of the simple model introduces an interesting new behavioral twist: the random component ϵ^{Y1}. Since Y^1 and Y^0 are simply the outcomes if one does or does not participate in a program, the most ready interpretation of ϵ^{Y1} is the presence of some unobserved determinant of $\{Y^1, Y^0\}$ that influences program impact. Thus, there are unobservables with a common effect on $\{Y^1, Y^0\}$ captured by ϵ but also some unobservables that effect program impact $Y^1 - Y^0$ captured by ϵ^{Y1}. Program impact at the individual level is now $Y^1 - Y^0 = \beta_1 + \epsilon^{Y1}$: program impact can now vary between individuals to the extent that the random component ϵ^{Y1} does.

Observed Y is

$$Y = P \cdot Y^1 + (1 - P) \cdot Y^0$$
$$= P \cdot \left(\beta_0 + \beta_1 + \epsilon^{Y1} + \epsilon\right) + (1 - P)(\beta_0 + \epsilon)$$
$$= \beta_0 + \beta_1 \cdot P + \epsilon^{Y1} \cdot P \cdot + \epsilon$$

Combining terms, we have

$$Y = \beta_0 + \left(\beta_1 + \epsilon^{Y1}\right) \cdot P \cdot + \epsilon$$

P has an interesting coefficient: $(\beta_1 + \epsilon^{Y1})$. This can be thought of as comprising a population parameter (β_1) and an individual-specific random shift off of that term (ϵ^{Y1}) reflecting the role of the unobservable that introduces some channel for differences in program impact between individuals. Such a specification is called a **random coefficients model**.

If we have a sample for which we observe $\{Y_i, P_i\}$ for a sample of $i = 1, ..., N$ individuals, we can regress Y on P in an attempt to estimate average program impact. The estimator of the coefficient on P would be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(P_i - \overline{P}\right) \cdot Y_i}{\sum_{i=1}^{N} \left(P_i - \overline{P}\right)^2}$$

Substituting in

$$Y_i = \beta_0 + \left(\beta_1 + \epsilon_i^{Y1}\right) \cdot P_i + \epsilon_i$$

we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(P_i - \overline{P}\right) \cdot \left(\beta_0 + \left(\beta_1 + \epsilon_i^{Y1}\right) \cdot P_i + \epsilon_i\right)}{\sum_{i=1}^{N} \left(P_i - \overline{P}\right)^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot P_i \cdot \epsilon_i^{Y1}}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2} + \frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\cdot \epsilon_i}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}$$

The unbiasedness of the estimator $\hat{\beta}_1$ thus depends critically not only on the independence or mean independence of P and ϵ but also the independence or mean independence of P and ϵ^{Y1} (similarly, the consistency of $\hat{\beta}_1$ depends critically not only on P and ϵ being uncorrelated but also on P and ϵ^{Y1} being uncorrelated).

But are any of these assumptions regarding independence, mean independence or lack of correlation involving either P and ϵ^{Y1} or P and ϵ reasonable? To examine this, we take another step beyond the simple framework introduced at the beginning of this section: we now consider how program participation is determined. A simple model of program participation might suggest that individuals participate if

$$Y^1 - Y^0 \geq 0$$

In other words, individuals participate if they experience a better outcome by doing so (sometimes this is loosely referred to in the impact evaluation literature as a **"Roy-Type Model"**, after Roy (1951)).

However, this ignores a key concern most individuals likely face when enrolling in a program: doing so likely involves a cost.[34] To capture this, we introduce a cost function for participation along the lines of:[35]

$$C = \gamma_0 + \epsilon^C$$

This model thus characterizes fully the potential benefits $(Y^1 - Y^0)$ and costs (C) of program enrollment. We can now pursue a richer model under which an individual participates (i.e. $P = 1$) if

$$Y^1 - Y^0 - C \geq 0$$

In other words, the individual participates in the program if the gain from participating is at least as large as the cost of doing so.[36]

Inserting the equations for Y^1, Y^0 and C, we have

$$\beta_0 + \beta_1 + \epsilon^{Y1} + \epsilon - \beta_0 - \epsilon - \gamma_0 - \epsilon^C \geq 0$$

[34]Participation even in ostensibly free programs might involve costs, such as time costs, child care costs (to cover children while parents participated, lost wages, etc.).

[35]This is a single equation specification, which may trouble some since the potential outcomes framework is two equations. For instance, there might be costs involved with *not* participating in some types of programs. The U.S. Affordable Care Act, with its penalty for not obtaining health insurance, would seem to be a timely example. This can be easily reconciled with the single equation approach simply by assuming that the single cost equation in the main body of the text is the difference of two underlying cost functions:

$$C^j = \gamma_{j0} + \epsilon_j^C$$

for $j = 0, 1$. The single equation is then generated by the net cost

$$C^1 - C^0 = \gamma_{10} + \epsilon_1^C - \gamma_{00} - \epsilon_0^C = \gamma_0 + \epsilon^C$$

where $\gamma_0 = \gamma_{10} - \gamma_{00}$ and $\epsilon^C = \epsilon_1^C - \epsilon_0^C$.

[36]One detail we have ignored is that $\{Y^1, Y^0\}$ might not be in the same metric (i.e. unit of measurement) as C. For instance, costs could be in monetary terms and the potential outcomes might involve some channel of health. This could be easily resolved either by multiplying by $\{Y^1, Y^0\}$ some price or converting C into health terms, perhaps by attaching a monetary value to $\{Y^1, Y^0\}$ and dividing C by that price. While this can be a fascinating intellectual exercise in its own right, it is not an important detail for the present discussion. We therefore assume that $\{Y^1, Y^0\}$ and C are in the same metric.

or, re-arranging, program participation would equal 1 if

$$\epsilon^{Y1} \geq -\beta_1 + \gamma_0 + \epsilon^C$$

Notice that, all other things (by which we mean ϵ^C) being equal, this condition is more likely to hold the larger is ϵ^{Y1}. In other words, P is more likely to equal 1 at larger values of ϵ^{Y1} and more likely to equal 0 at smaller values of ϵ^{Y1}: in other words, P is likely to be correlated with ϵ^{Y1}.

We have learned something important from this: the random coefficient estimator $\hat{\beta}_1$ is likely to be biased and inconsistent if $\epsilon^{Y1} \neq 0$. The intuition behind this is fairly straightforward. If there are unobserved determinants of Y (in our model, ϵ^{Y1}) that influence program impact and play a role in the program participation decision, large values of that unobservable become more likely when $P = 1$ than when $P = 0$. Thus, when we try to estimate average program impact by comparing the average of Y^1 for those for whom $P = 1$ with the average of Y^0 for those for whom $P = 0$, the former exaggerates what non-participants (i.e. those for whom $P = 0$) would have experienced by way of program impact had they participated since the value of ϵ^{Y1} for participants is generally bigger for participants than non-participants.

Assuming that $\epsilon^{Y1} = 0$ is the same as assuming that there is no unobservable that influences program impact. The random coefficient component of the regression model then drops out and we are back to the straightforward regression model

$$Y = \beta_0 + \beta_1 \cdot P + \epsilon$$

We have learned something very important from this: in the presence of unobservables that influence program impact and participation, it is not possible to obtain an unbiased estimate of average program impact, even if that unobservable plays no role in determining the cost of participation (remember, we have said little to this point about ϵ^C). Thus, the assumption that P is independent, mean independent or uncorrelated with ϵ^{Y1} is tantamount to assuming no unobservable that influences program impact in our simple model of program participation.

Nor do the complications involved with introducing our simple participation decision model end with this. Assuming that $\epsilon^{Y1} = 0$ and re-arranging, the individual decides to participate if

$$\beta_1 - \gamma_0 \geq \epsilon^C$$

Clearly, the smaller is ϵ^C the more likely the individual participates. This is sensible: it essentially says that the lower are the costs of participation (or, in this case, that component of costs determined by unobserved random factors) the more likely is the individual to participate. But this also means that ϵ^C will generally be greater among non-participants than participants, or that participation P is (negatively) correlated with ϵ^C.

This could create problems as there are unobservables that occur in both ϵ^C and ϵ. For instance, some unobserved types of individuals might tend to have low cost and higher values of $\{Y^0, Y^1\}$. If this is the case, ϵ^C and ϵ could be correlated. However, since P and ϵ^C are correlated, this means that P and ϵ would be correlated, rendering the regression estimate of program impact biased and inconsistent.

This can be seen intuitively by considering the estimation of program impact by comparing the average outcomes Y among participants ($P = 1$, for whom $Y = Y^1$) and non-participants ($P = 0$, for whom $Y = Y^0$). Certain unobserved types are more likely to have lower costs of participation. Those types are thus more likely to participate: they are concentrated among the sample for whom $P = 1$. If those same unobserved types are likely to have different values of Y^1 and Y^0 then some of the difference in average Y between participants and non-participants will be picking up the

imbalance of these unobservables between them, as opposed to simply reflecting the impact of the program.

So what does all of this mean? Essentially, it means that if we wish to obtain an unbiased or consistent estimate of average program impact under this behavioral model via regression there cannot be an unobservable that either:

1. Influences program impact and program participation;
2. Influences the potential outcomes $\{Y^0, Y^1\}$ in the same way but also influences program participation (which we have already seen in the discussion in the preceding section).

Thus, even this extremely simple and basic model of the program participation decision has generated two distinct pathways by which unobserved random determinants of the potential outcomes might possibly undermine our ability to estimate average program impact via regression (or even just simple comparison of mean outcomes between participants and non-participants)! From this simple model, an obvious lesson is thus emerging: using selection on observables models essentially precludes the possibility of unobservables that influence both potential outcomes and program participation.

4.1.5 Standard Errors

In this section we briefly digress to consider the subject of estimating the sampling variation of regression estimates. For instance, consider the basic departure point of the discussion of regression, which was the following simple model:

$$Y = \beta_0 + \beta_1 \cdot P + \epsilon$$

Through regression analysis we seek estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively. As we have seen, if we observe $\{Y_i, P_i\}$ for a sample of $i = 1, ..., N$ individuals, the least squares regression estimates for that sample are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot (Y_i - \overline{Y})}{\sum_{i=1}^{N}(P_i - \overline{P})^2} = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \cdot \overline{P}$$

where

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

and

$$\overline{P} = \frac{\sum_{i=1}^{N} P_i}{N}$$

Notice that the value of these depends on the specific set of values for $\{P_i, Y_i\}$ for $i = 1, ..., N$ observed in the particular sample of size N.

However, this means that the particular values for $\hat{\beta}_0$ and $\hat{\beta}_1$ are sample dependent. Suppose, for instance, that we randomly selected a sample of size N from some population of interest. We might then dutifully interview them to obtain information $\{P_i, Y_i\}$ for each of the $i = 1, ..., N$ individuals in the sample and then regress Y on P with the observations in this sample in order to form estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Suppose that we then repeated the process of random selection of a sample of size N, interview and regression. It is not likely that the set of values for $\{P_i, Y_i\}$ across

the $i = 1, ..., N$ observations would be exactly the same as in the first sample: the individuals selected for this second sample would probably not have experienced quite the same mix of values for the outcome of interest because they experienced a different set of random shocks ϵ. This means that the value of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ would not be quite the same as in the first sample.

Clearly we would like to know something about this sampling variation: it tells us something about the reliability or precision of an estimate in the sense of how much it randomly fluctuates from sample to sample. Returning to the regression example, suppose that $E(\hat{\beta}_1) = \beta_1$ (in other words, suppose that $\hat{\beta}_1$ is an unbiased estimator of β_1). This tells us that $\hat{\beta}_1$ is *right on average* as an estimator of our estimand, β_1. However, we would likely still have more confidence in a particular estimate $\hat{\beta}_1$ (calculated from a particular sample) if we knew that the estimates tended to vary little from sample to sample. We would be less confident in that estimate if we knew that the estimates varied a lot from sample to sample. Indeed, we could even imagine a circumstance where we might prefer a slightly biased estimator that did not vary much from sample to sample to an unbiased one that varied a lot from sample to sample.

Fortunately, the two regression methods we have seen thus far (least squares and logit) allow for relatively straightforward estimation of the sample to sample variation in the estimates generated by them. In this subsection we will focus simply on deriving the estimated variance of the estimator $\hat{\beta}_1$ and performing simple statistical tests with the result.[37]

To begin with, the variance of any random variable (we'll use 'Z' as a reference) is simply

$$var\left(Z\right) = E\left(\left(Z - E\left(Z\right)\right)^2\right)$$

The variance of Z is simply its expected squared difference from its expectation. It is a measure of how much Z varies around its expected value. The square root of the variance is the standard deviation. Two key properties of the variance will be important for the derivation to follow. Suppose that c is a constant. Then

$$var\left(c\right) = 0$$

$$var\left(c \cdot Z\right) = c^2 \cdot var\left(Z\right)$$

The first result simply reflects the fact that c is a constant and not a variable. Hence it does not vary at all. The second result is easily derived:

$$var\left(c \cdot Z\right) = E\left(\left(c \cdot Z - E\left(c \cdot Z\right)\right)^2\right) = E\left(\left(c \cdot Z - c \cdot E\left(Z\right)\right)^2\right)$$

$$= E\left(c^2 \cdot \left(Z - E\left(Z\right)\right)^2\right) = c^2 \cdot E\left(\left(Z - E\left(Z\right)\right)^2\right)$$

$$= c^2 \cdot var\left(Z\right)$$

These two properties will be invoked to derive $var\left(\hat{\beta}_1\right)$.

At this point it is also useful to introduce the concept of "degree of freedom". In a sample (for instance of N individuals) variance is usually estimated by

$$\frac{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2}{N-1}$$

[37] In general, the focus of this manual is the properties of estimators of program impact mainly in terms of unbiasedness and consistency. We digress to discuss sampling variation and testing in the main body of the manual only when necessary.

where

$$\overline{Z} = \frac{\sum_{i=1}^{N} Z_i}{N}$$

Note that the denominator of the variance estimator is $N-1$. $N-1$ is the "degrees of freedom" for the estimation of the variance. This is basically a fancy way of indicating the number of observations truly free to contribute to the calculation of the variance. The reason it is not simply N is that the estimate of the variance involves use of an estimate (i.e. $\overline{Z}$). However, a key property of the estimated mean is that

$$\sum_{i=1}^{N} \left(Z_i - \overline{Z} \right) = 0$$

In other words, the differences between the actual values of Z_i and the estimated mean $\overline{Z}$ should cancel out across the sample. But this means that one observation needs to be used simply to insure that this condition holds. That observation then cannot provide truly independent information for the estimation of variance. This leaves us with $N-1$ truly independent observations with which to estimate variance.

We now consider the task of estimating the variance of $\hat{\beta}_1$. We begin with the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot Y_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

The second line relies on exactly the same math used earlier to develop the expectation of $\hat{\beta}_1$, but without the focus on the expectations operator $E(\cdot)$. Then

$$var\left(\hat{\beta}_1\right) = var\left(\beta_1 + \frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

$$= var\left(\frac{\sum_{i=1}^{N}(P_i - \overline{P}) \cdot \epsilon_i}{\sum_{i=1}^{N}(P_i - \overline{P})^2}\right)$$

because as a population parameter β_1 is constant and thus does not vary. For the next step, we treat

$$\frac{\sum_{i=1}^{N}(P_i - \overline{P})}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

as if it were a constant. Then, exploiting the property $var(cZ) = c^2 var(Z)$, we have

$$var\left(\hat{\beta}_1\right) = \frac{\sum_{i=1}^{N}(P_i - \overline{P})^2}{\left(\sum_{i=1}^{N}(P_i - \overline{P})^2\right)^2} var\left(\epsilon_i\right)$$

$$= \frac{var\left(\epsilon_i\right)}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

The variance of the estimate $\hat{\beta}_1$ is thus increasing in the variance of ϵ and decreasing in the variance of P.

The only remaining question within the context of this simple derivation is how to estimate $var(\epsilon_i)$. This is typically estimated from the predicted residual

$$\hat{\epsilon_i} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i$$

The estimated variance of ϵ is then simply

$$\frac{\sum_{i=1}^{N}\left(\hat{\epsilon}_i\right)^2}{N-2}$$

The division by $N-2$ again reflects the degrees of freedom available for the calculation of the variance of $\hat{\epsilon}_i$ (since two parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$ are in the mean function $E(Y|P)=\hat{\beta}_0+\hat{\beta}_1$ used to calculate $\hat{\epsilon}$, leaving $N-2$ parameters truly free to calculate the estimated sample variance of the predicted error).

A smaller variance to $\hat{\beta}_1$ indicates greater precision to the estimate in the sense of less sampling variation. This can be also be seen through a test statistic for significance and the p-value associated with that statistic. The former is typically generated automatically by most commercial statistical packages (such as STATA) as part of the standard regression output. It is basically a test statistic formed under the assumption that the true value of β_1 is zero. The value of the statistic provides some indication of how likely it is that that assumption is true. This is even more directly expressed by the p-value, which can interpreted as the probability that the assumption is true given the value of the test statistic.

Behind this standard regression output test statistic is a basic hypothesis test. That test involves a null hypothesis (denoted H^0) which, if true, would imply a particular probability distribution. Test statistics can then be crafted per that distribution. The null hypothesis would be rejected if the value of that test statistic seems unlikely for that distribution. It would be rejected in favor of an alternative hypothesis (H^a). The specific hypotheses behind the typical automatic regression output are:

$$H^0 : \beta_k = 0$$

$$H^a : \beta_k \neq 0$$

where k indicates the k^{th} regression parameter. This hypothesis is usually tested for each of the $k=1,...,K$ regression parameter estimates. So, for instance, for the basic least squares regression of Y on P it would be tested for $\hat{\beta}_0$ and $\hat{\beta}_1$.

The test statistic on which we focus for now (because it is most relevant to the discussion in the next subsection) is the t-statistic. The t-statistic under the null hypothesis that $\beta_1=0$ is

$$\frac{\hat{\beta}_1-0}{\sqrt{var\left(\hat{\beta}_1\right)}}=\frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)}$$

where $se(\cdot)$ indicates "standard error". This may look to some readers like the standardization common for converting a normally distributed random variable with mean zero into a standard normal random variable. However, it would not be appropriate to view this as a standard normal random variable since the denominator is an estimate. Instead, it is more appropriate to approach this as a statistic to be assesses under **Student's t-distribution** (see Tool Boxes on The Normal and χ^2 Distributions and Student's t-distribution). In particular, this is referred to as a t-statistic, one with $N-2$ degrees of freedom.[38] The t-distribution has an overall shape similar to the normal distribution but with somewhat fatter tails. In formal terms, it is **leptokurtic**, a fancy way of saying that a distribution has more weight in the tails than the normal distribution. This becomes less true as the degrees of freedom increases, until eventually the t-distribution becomes essentially

[38] In general, the basic t-statistic for testing the significance of regressors has $N-K$ degrees of freedom where k is the number of regressors, including the constant.

indistinguishable from the normal distribution (though to be sure there was not that much difference to begin with).

Tool Box: The Normal and χ^2 Distributions

The Normal Distribution (sometimes referred to as the Gaussian distribution) is a workhorse of statistics. If x is a normally distributed random variable with mean μ and variance σ^2, its probability density is given by

$$f(x = c) = \frac{1}{\sigma\sqrt{2 \cdot \pi}} \cdot e^{-\frac{(c-\mu)^2}{2 \cdot \sigma^2}}$$

The fact that x is normally distributed with mean μ and variance σ^2 is typically indicated $x \sim \text{N}(\mu,\sigma^2)$. In Figure 4.4 we illustrate the normal distribution with various values for the mean μ and variance σ^2. N (0,1) refers to a particular, and particularly popular, case of the normal distribution known as the **standard normal distribution**. Notice that the normal distribution is always single peaked (or, more formally, **uni-modal**) and symmetric around that mode. Any variable x that is distributed N (μ,σ^2) can be converted into a standard normal distributed random variable via the conversion

$$\frac{x - \mu}{\sigma}$$

One density directly derived from the normal is the χ^2 (pronounced "chi-squared" as in the Greek letter "chi" or χ). Specifically, if z is a standard normal random variable, the $w = Z^2$ is a chi-squared random variable, denoted χ^2_1. If $\{w_1, w_2, w_3, ..., w_k\}$ are all χ^2_1 distributed random variables, then

$$w_1 + w_2 + w_3 + ... + w_k$$

is distributed χ^2_k (in other words, it is a random variable that is chi-squared distributed with k degrees of freedom). By definition χ^2 random variables must be non-negative. Figure 4.5 graphs the χ^2 distribution for various degrees of freedom.

To test the hypothesis outlined above, we need to decide on the significance of the test. The degree of significance is the probability of a **Type-I error**. A Type-I error involves falsely rejecting a true null hypothesis. From the t-distribution we can find critical values t^c and $-t^c$ (where $t^c = |-t^c|$) such the probability of a Type-I error is α (in other words, different critical values would yield a different probability of a Type-I error). The chosen probability α of a Type-I error is called the **significance** of the test.

Tool Box: Student's t-Distribution

If Z is a standard normal distribution (i.e. $Z \sim \mathrm{N}(0,1)$) and w is a χ^2 random variable with N degrees of freedom, then

$$\frac{z}{\sqrt{\frac{w}{N}}}$$

follows a t-distribution with N degrees of freedom. In practice, the t-distribution looks for the most part like the standard normal distribution, though it is a bit fatter in the tails (particularly at lower degrees of freedom). The t-distribution is thus the ratio of two random variables. It is used for testing in the present context because the denominator of the test statistic for significance is an estimate, and hence the normal distribution (for which the denominator σ is the actual, as opposed to estimated, standard error) would not be an appropriate distributional choice for testing. The t-distribution is often referred to as "Student's t-distribution" because it was introduced in a *Biometrika* paper in 1908 written under the pseudonym "Student". The real identity of the author was William Gossett, who worked at the time for the Guinness Brewery in Dublin, Ireland. He had to write under a pseudonym owing to company policy forbidding employees publishing in scientific papers.

This is illustrated in Figure 4.6. The curve in the figure illustrates the t-distribution under the

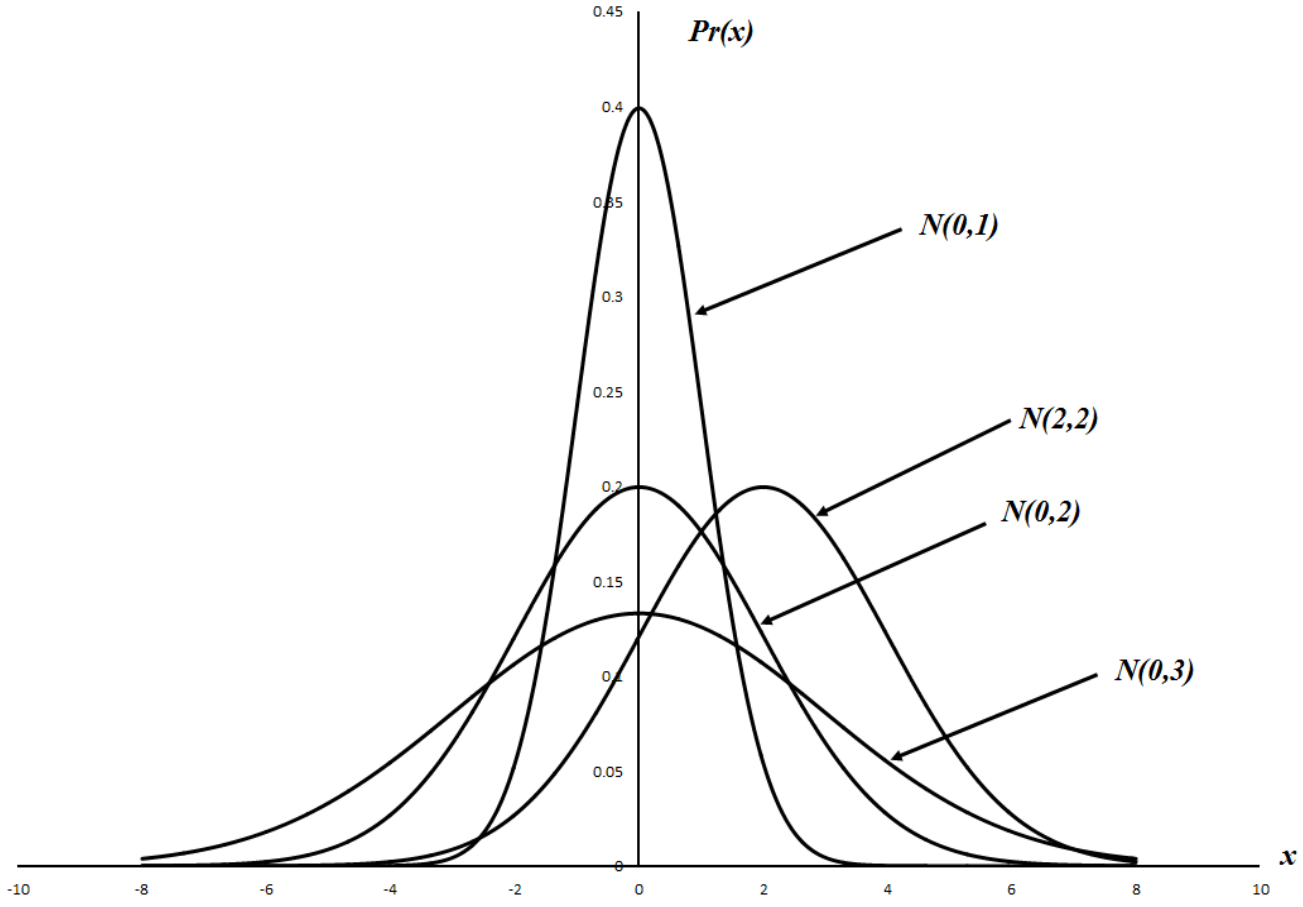


Figure 4.4: Normal Densities

null hypothesis that $\beta_1 = 0$. In particular, note that it is centered on zero. In other words, if the null hypothesis is true, the most likely value for the t-statistic is 0. As you move farther to the left or right from origin at 0 the t-values are becoming less and less likely under the null hypothesis that $\beta_1 = 0$. The critical values $\{-t^c, t^c\}$ provide the critical values beyond which (i.e. to the left of $-t^c$ and the right of t^c) the probability of observing a t-statistic value under the null hypothesis is α.

If α is the significance level of the test, we would accept the null hypothesis if the t-statistic for the sample is greater than $-t^c$ and less that t^c (in other words, if the t-statistic is $\hat{t}$ we would accept the null hypothesis that $\beta_1 = 0$ if $-t^c \leq \hat{t} \leq t^c$). We would reject the null hypothesis (and instead accept the alternative hypothesis that $\beta_1 \neq 0$) if the test statistic is greater that t^c or less that $-t^c$. But we would be doing so knowing that there is a probability of α that the test statistic could be greater than t^c or less than $-t^c$ and the null hypothesis of $\beta_1 = 0$ still be true.

Acceptance and rejection of the null hypothesis are also illustrated in Figure 4.7. In that figure, two potential test statistic values, $\hat{t_1}$ and $\hat{t_2}$, are considered. To fix ideas, one can think of these t-statistics as emerging from two different samples, sample 1 and sample 2 (hence the subscripts on

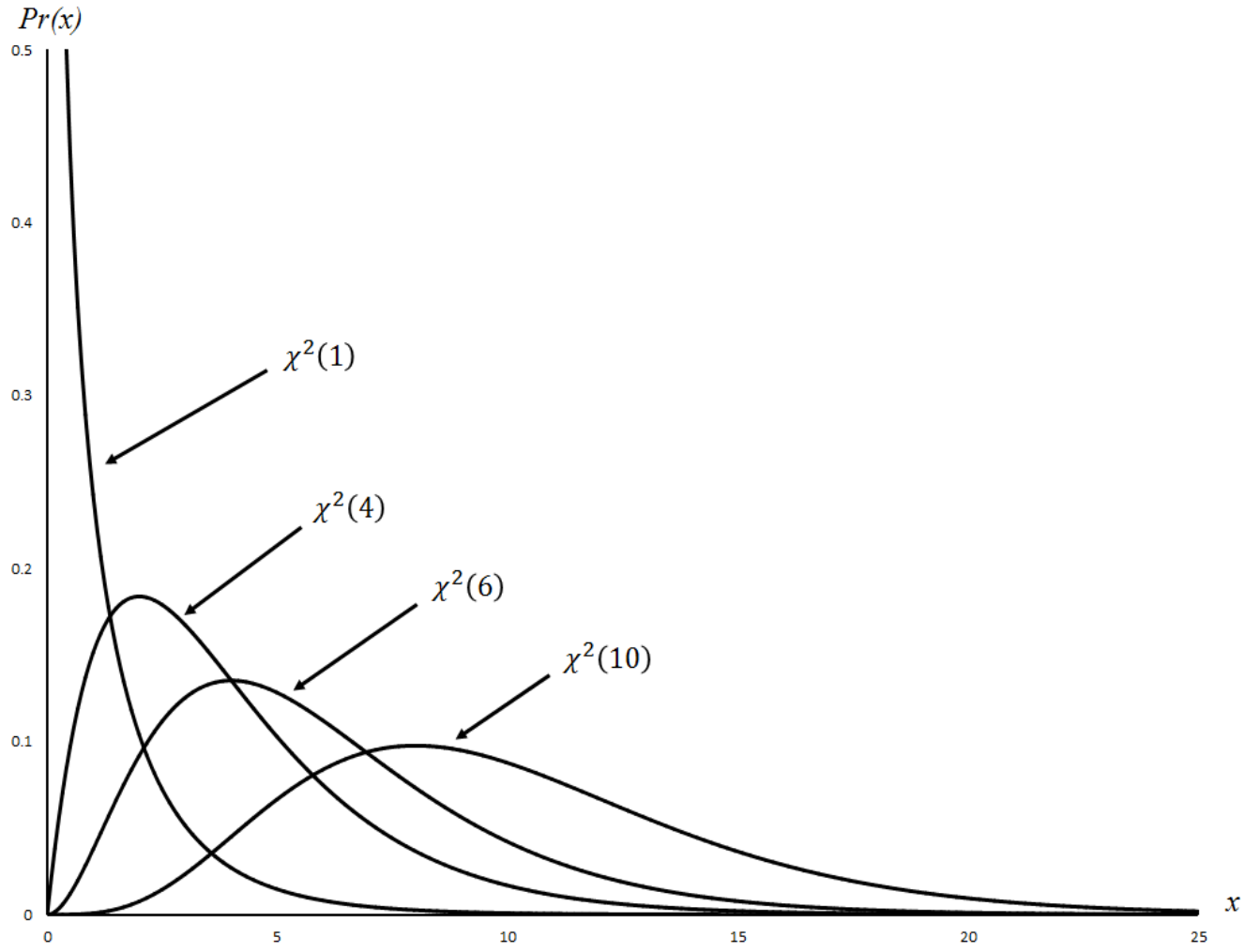


Figure 4.5: Chi-Squared Densities

the t-statistics). Presented with the t-statistic $\hat{t_1}$ we would accept the null hypothesis at significance level α because $-t^c \leq \hat{t_1} \leq t^c$. On the other hand, we would reject the null hypothesis at the α level because $\hat{t_2} \leq -t^c$ (we also would have rejected if $\hat{t_2}$ had been greater than t^c.

Finally, in Figure 4.8 we consider the p-value associated with a given t-statistic value $\hat{t}$. The p-value is the significance level that would be required to reject the null hypothesis that $\beta_1 = 0$ given the value of the test statistic $\hat{t}$. $\hat{t}$ is to the left of the origin (i.e. negative) but the p-value is formed on both sides of the origin. In other words, it is effectively the probability that the t-statistic could be less that $\hat{t}$ or greater than $-\hat{t}$ and the null hypothesis still be true. We have to

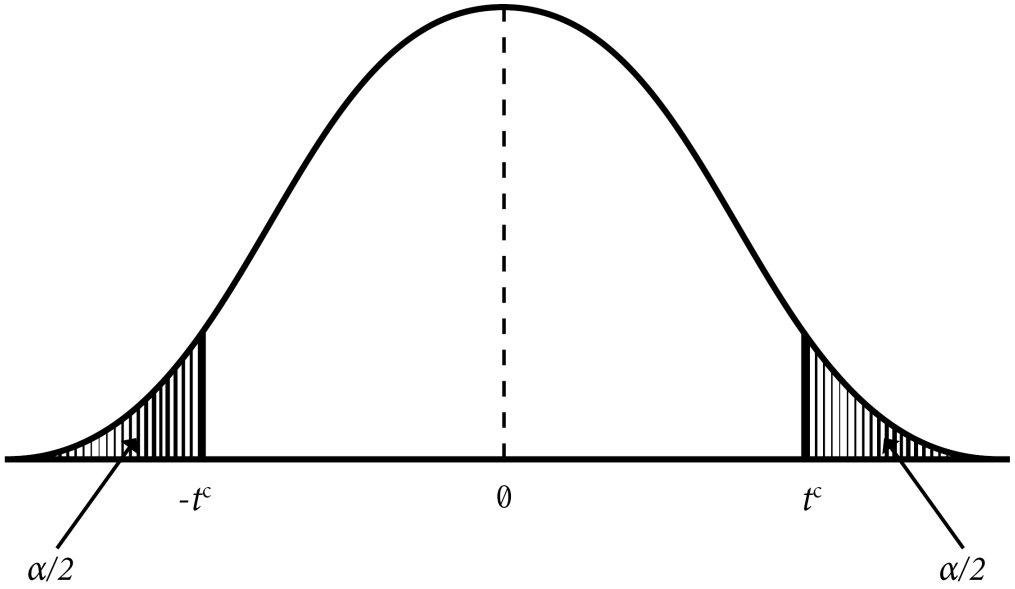


Figure 4.6: The Probability of a Type-I Error

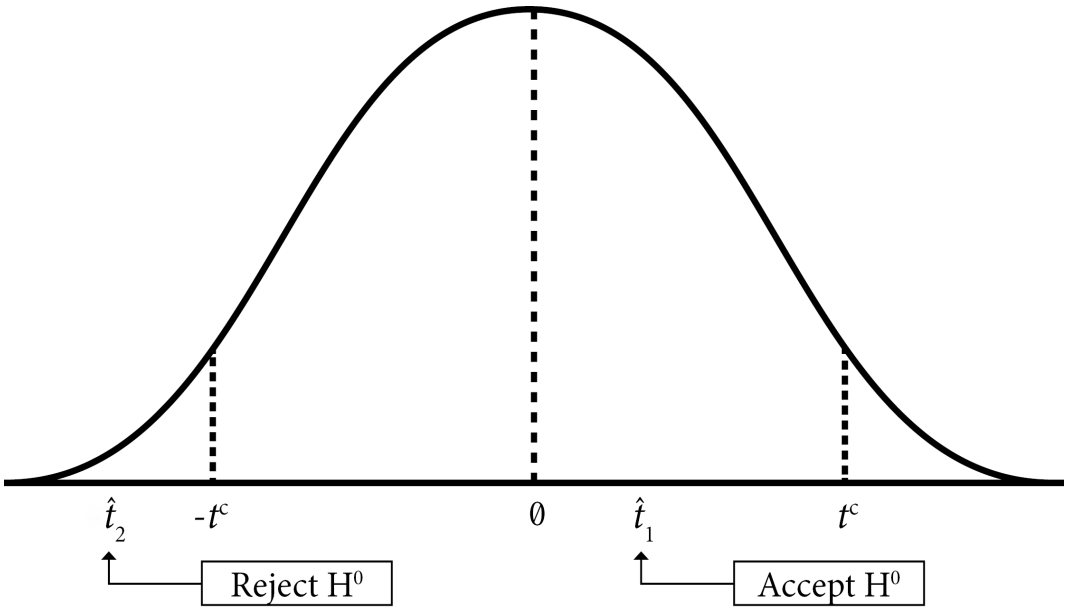


Figure 4.7: Accepting and Rejecting The Null Hypothesis

consider both tails of the t-distribution because the alternative hypothesis indicates that this is a **two-sided test**: the alternative hypothesis is that $\beta_1 \neq 0$, which could happen either if $\beta_1 < 0$ or $\beta_1 > 0$.[39]

Notice that as the test statistic value gets closer and closer to zero the p-value would grow increasingly large until it reached 1. In other words, a t-statistic value of 0 would imply that the null hypothesis that $\beta_1 = 0$ could be rejected only in the face of a 100 percent chance of a Type-I error. A small p-value, on the other hand, indicates that the null hypothesis could be rejected with only a small probability of a Type-I error.

The conventional significance levels for the basic hypothesis test

$$H^0 : \beta_1 = 0$$

$$H^a : \beta_1 \neq 0$$

are 10, 5 and 1 percent. Thus, a p-value greater than .05 but less than or equal to .1 would indicate that the null hypothesis would be rejected at the 10 percent significance level. A value greater than .01 but less than or equal to .05 would indicate that the null hypothesis would be rejected at the 5 percent significance level. Finally, a p-value at or less than .01 would indicate rejection of the null hypothesis at the 1 percent level.

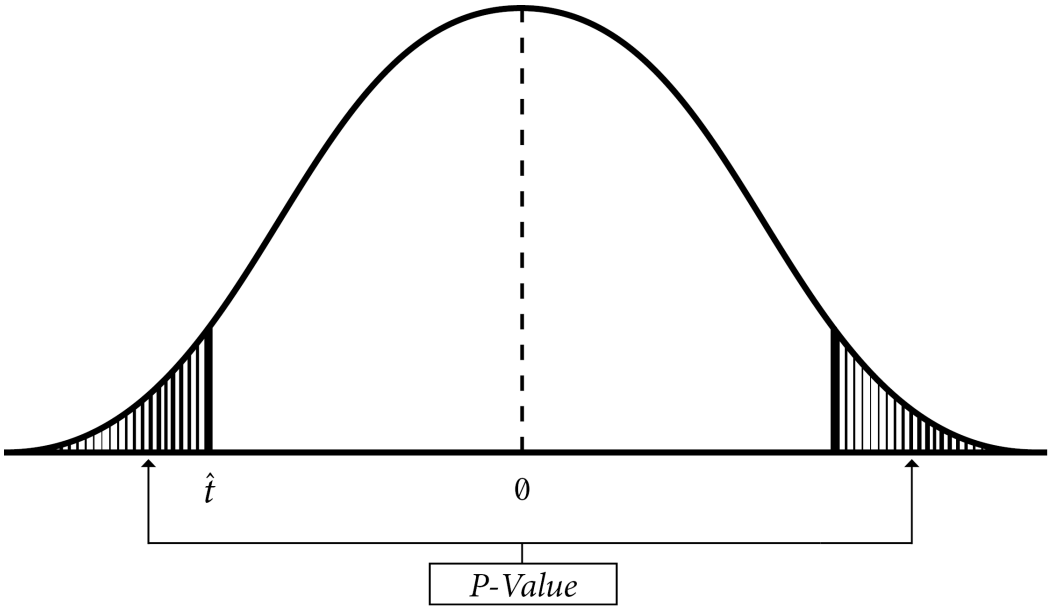


Figure 4.8: The P-value of a Test Statistic $\hat{t}$

We conclude by recalling the t-statistic formula under the null hypothesis that $\beta_1 = 0$:

$$\frac{\hat{\beta}_1 - 0}{\sqrt{var\left(\hat{\beta}_1\right)}} = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)}$$

The more precisely that β_1 is estimated the smaller will be the standard error $se(\hat{\beta}_1)$. Other things being equal, this will make the value of the t-statistic larger. This implies a smaller and smaller p-value for the test statistic: we could, given the test statistic value, reject the null at smaller and smaller levels of significance.

[39] By comparison, a one-sided test would consider only one of these two possibilities for the alternative hypothesis.

4.1.6 Regression and Randomized Data

We conclude with a very brief discussion of the application of multiple regression to program impact evaluation in experimental samples. The reader would be reasonable in asking what could be gained from this. Since program participation is randomized in an experimental sample, it is thus independent of other variables such as individual background characteristics that might otherwise serve as confounders in the non-experimental data setting. In that context, it is not immediately obvious what is to be gained by regression of the outcome of interest Y on program participation P *and* a bunch of controls for potential confounders. It might seem sufficient simply to regress Y on P alone or simply calculate the difference in average outcomes between those for whom P=1 and P=0.

And indeed it is sufficient from the standpoint of recovering an unbiased and consistent estimate of program impact. However, it would also be preferable if the estimate was as precise as possible. For instance, suppose we were to regress Y on P alone (evoking the basic model $Y = \beta_0 + \beta_1 \cdot P + \epsilon$). If this is done with an experimental sample (wherein program participation P is completely randomly determined) then the least squares estimator $\hat{\beta}_1$ is unbiased (and consistent). But unbiased simply means that the *estimator* $\hat{\beta}_1$ is correct on average. The values for the *estimates* $\hat{\beta}_1$ from different samples can vary a lot around actual program impact. Thus for any given sample the estimate might be quite misleading. It would thus be preferable if the sample by sample variation in the estimates $\hat{\beta}_1$ was as small as possible.

It is this consideration that motivates an argument for multiple regression, even with experimental data. In the last subsection, we saw that

$$var\left(\hat{\beta}_1\right) = \frac{var\left(\epsilon_i\right)}{\sum_{i=1}^{N}(P_i - \overline{P})^2}$$

Clearly, the smaller is the variance of ϵ the smaller will be the sampling variation in $\hat{\beta}_1$. Intuitively, one way to reduce the variance of ϵ is to simply reduce its relative share in the total variation in Y. This can be done by adding more explanatory variables. By adding them, we essentially take them out of the residual term ϵ, reducing ϵ's role in the overall variation in Y. This can lead to more precise (i.e. less variable) estimates of $\hat{\beta}_1$, even if program participation P is independent of those explanatory variables.

STATA Output 4.22 (4.5.do)

```
. reg Y P
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  1,  1998) =    3.06
       Model |  49.8180723     1  49.8180723           Prob > F      =  0.0802
    Residual |  32490.895   1998  16.2617092           R-squared     =  0.0015
-------------+------------------------------           Adj R-squared =  0.0010
       Total |  32540.713   1999  16.2784958           Root MSE      =  4.0326

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    .315671   .1803534     1.75   0.080    -.0380294    .6693715
       _cons |   1.020916   .1268258     8.05   0.000     .7721912     1.26964
------------------------------------------------------------------------------
```

We illustrate with a numerical example (provided in the STATA do-file 4.5.do). For this example, we generate 2,000 observations for Y, P and forty explanatory variables $x_1, x_2, ..., x_{40}$.

To begin with, program participation is independently and randomly determined on the basis of whether draws from the standard normal distribution exceed 0 or not (i.e. P is experimentally determined). We then defined Y as

$$Y_i = 1 + .5 \cdot P_i + \psi_1 \cdot x_1 + \psi_2 \cdot x_2 + \psi_3 \cdot x_3 + ... + \psi_{40} \cdot x_{40} + \epsilon$$

where the xs are independently drawn from the standard normal distribution, and the ψs are drawn from the uniform distribution on the interval [0,1] and then multiplied by -1 or 1 according to whether that is an even (e.g. x_4) or odd (x_{21}) numbered variable. Finally, $\epsilon \sim$ N (0,4). True program impact is thus .5.

STATA Output 4.23 (4.5.do)

```
. reg Y P x1
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  2,  1997) =    4.14
       Model |  134.36782     2  67.1839101            Prob > F      =  0.0161
    Residual |  32406.3452  1997  16.2275139           R-squared     =  0.0041
-------------+------------------------------           Adj R-squared =  0.0031
       Total |  32540.713   1999  16.2784958           Root MSE      =  4.0283

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .3251667   .1802117     1.80   0.071    -.0282559    .6785894
          x1 |  -.2038953   .0893258    -2.28   0.023    -.3790769   -.0287137
       _cons |    1.01554   .1267142     8.01   0.000     .7670336    1.264046
------------------------------------------------------------------------------
```

STATA Output 4.24 (4.5.do)

```
. reg Y P x1-x5
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  6,  1993) =   47.29
       Model |  4055.33589     6  675.889315            Prob > F      =  0.0000
    Residual |  28485.3772  1993  14.2927131           R-squared     =  0.1246
-------------+------------------------------           Adj R-squared =  0.1220
       Total |  32540.713   1999  16.2784958           Root MSE      =  3.7806

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .3468251   .1691585     2.05   0.040     .0150791    .6785712
          x1 |  -.2457859    .083874    -2.93   0.003    -.4102758    -.081296
          x2 |   .1870881   .0860391     2.17   0.030     .0183521    .3558241
          x3 |  -1.237739   .0827848   -14.95   0.000    -1.400093   -1.075386
          x4 |   .5545533   .0836324     6.63   0.000     .3905373    .7185693
          x5 |  -.0256608   .0852108    -0.30   0.763    -.1927723    .1414508
       _cons |   1.053226   .1189661     8.85   0.000     .8199152    1.286537
------------------------------------------------------------------------------
```

We consider the role of explanatory variables in determining the precision of an experimental estimate of program impact by regressing Y on P alone, and then regressing Y on increasingly large numbers of controls (i.e. the xs). We begin with Output 4.22, for which the results of simple regression of Y on P are shown. The estimate of program impact (.315671) is only around 60 percent of the true program impact. The estimate is also not particularly precise, as evidenced by a standard error (the square root of the variance) that is roughly 57 percent of the value of the

estimate, a modest t-statistic (1.75) and a p-value of 0.08 (indicating that the null hypothesis for the standard significance test would be accepted at the 10 but not 5 or 1 percent levels).

STATA Output 4.25 (4.5.do)

```
. reg Y P x1-x10
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F( 11,  1988) =   50.55
       Model |  7112.82651    11  646.620592            Prob > F      =  0.0000
    Residual |  25427.8865  1988  12.7906874            R-squared     =  0.2186
-------------+------------------------------           Adj R-squared =  0.2143
       Total |   32540.713  1999  16.2784958            Root MSE      =  3.5764

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .3395683   .1600808     2.12   0.034     .0256245    .6535121
          x1 |  -.2339935   .0794026    -2.95   0.003    -.3897146   -.0782725
          x2 |    .177439   .0814993     2.18   0.030      .017606    .337272
         ... |       ...        ...      ...     ...         ...         ...
          x9 |   -.257163   .0796385    -3.23   0.001    -.4133468   -.1009793
         x10 |   .6646639   .0802806     8.28   0.000     .5072208    .8221069
       _cons |   1.021846   .1126557     9.07   0.000       .80091    1.242781
------------------------------------------------------------------------------
```

STATA Output 4.26 (4.5.do)

```
. reg Y P x1-x20
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F( 21,  1978) =   63.41
       Model |  13092.9821    21  623.475338            Prob > F      =  0.0000
    Residual |  19447.7309  1978  9.83201766            R-squared     =  0.4024
-------------+------------------------------           Adj R-squared =  0.3960
       Total |   32540.713  1999  16.2784958            Root MSE      =  3.1356

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .3489339     .14089     2.48   0.013     .0726254    .6252424
          x1 |  -.2022207   .0698535    -2.89   0.004    -.3392148   -.0652266
          x2 |    .170864   .0715161     2.39   0.017     .0306092    .3111188
         ... |       ...        ...      ...     ...         ...         ...
         x19 |  -.1280954   .0719818    -1.78   0.075    -.2692635    .0130727
         x20 |   .6558725   .0709883     9.24   0.000     .5166528    .7950921
       _cons |   1.009012   .0989782    10.19   0.000     .8148997    1.203125
------------------------------------------------------------------------------
```

In Outputs 4.23-4.27 we add successively greater numbers of controls x to the regression. In Output 4.23 we add just one control (x_1) but nonetheless see a slight improvement in precision, as evidenced by a tiny increase in the t-statistic and tiny decrease in the p-value. A more substantial gain in precision comes with the addition of five controls ($x_1, ..., x_5$) in Output 4.24. The t-statistic and p-value for P have improved to the point where P is actually significant at the 5 percent level.

In Outputs 4.25, 4.26 and 4.27 we see 10, 20 and 40 regressors, respectively, added (to conserve space, only partial output is shown for each case). As this process plays out the standard error of estimated impact falls (both in absolute terms and as a percentage of either the estimate or the true program impact), to the point where, with 40 regressors added, the estimate of program impact is, at .4630212, not far off from the truth. More to the point (from a sampling variation standpoint)

the standard error is only around half of its original value (i.e. .0911341/.1803534=.50530847) and the p-value indicates significance at the 1 percent level.

STATA Output 4.27 (4.5.do)

```
. reg Y P x1-x40
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F( 41,  1958) =  148.40
       Model |  24618.1777    41  600.443359            Prob > F      =  0.0000
    Residual |  7922.53531  1958  4.04623867            R-squared     =  0.7565
-------------+------------------------------           Adj R-squared =  0.7514
       Total |   32540.713  1999  16.2784958            Root MSE      =  2.0115

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .4630212   .0911341     5.08   0.000     .2842912    .6417513
          x1 |  -.2225914   .0451185    -4.93   0.000    -.3110767   -.1341061
          x2 |    .182892   .0461205     3.97   0.000     .0924416    .2733424
         ... |       ...        ...      ...     ...          ...         ...
         x39 |  -.0475312    .045285    -1.05   0.294    -.1363431    .0412807
         x40 |   .7019262   .0458434    15.31   0.000     .6120192    .7918333
       _cons |   .9432708   .0637802    14.79   0.000     .8181866    1.068355
------------------------------------------------------------------------------
```

This is admittedly a somewhat contrived (in the sense that there is no guarantee that the improvement in precision seen is indicative of typical improvement) example, but nonetheless provides a powerful illustration of the gains from multiple regression even in the experimental context. Moreover, in a true experimental setting one need not worry about the bad control problem (even if one of the controls x introduced to improve precision is endogenous). The reason is that the random determination of P precludes a relationship between P and any control x, which is necessary for the bad control problem to arise.

4.2 Matching

We now turn to matching, the other major estimation tradition within the "selection on observables" framework. The basic idea of matching is quite simple: to estimate what would have happened to someone under the counterfactual state (i.e. the alternative program participation status to the one they chose) look to what happened to someone just like them who actually experienced that counterfactual state. For instance, to estimate program impact for a participant (for whom we do not observe Y^0, the outcome in the absence of program participation), the matching approach forms an estimate of that participant's outcome in the absence of participation Y^0 using the outcome observed for a similar non-participant (or outcomes observed for similar non-participants). An estimate of impact at the population level can then be formed by appropriately averaging the estimated program impact across the individuals observed in a random sample from that population. This basic idea has inspired an enormous methodological literature and led to a range of different impact evaluation techniques (some of which can only be loosely described as matching). One thing should, however, already be clear: you can only tell if one individual is "like" another in terms of characteristics that are *observed*. Even at this very initial, tentative moment the reader may thus have some sense of why this is thus considered a "selection on observables" based impact evaluation approach.

4.2.1 Matching: The Basics

We once again begin with a simple model motivated by the potential outcomes framework. Specifically, we have the potential outcome equations

$$Y^0 = \beta_0 + \beta_2 \cdot x + \epsilon$$

$$Y^1 = \beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \epsilon^{Y1} + \epsilon$$

where x is an observed characteristic of the individual and $\{\epsilon, \epsilon^{Y1}\}$ are unobserved characteristics of the individual (in other words, $\{x, \epsilon, \epsilon^{Y1}\}$ take on different values for different individuals but only the variation in x is actually observed across individuals). Program impact is then

$$Y^1 - Y^0 = \beta_1 + \beta_3 \cdot x + \epsilon^{Y1}$$

This is thus a framework where program impact varies across individuals. It does so because individuals have different observed (captured by the term $\beta_3 \cdot x$) and unobserved (captured by ϵ^{Y1}) determinants of program impact.

We also specify the cost of participation as

$$C = \gamma_0 + \gamma_1 \cdot x + \epsilon^C$$

Thus, the cost of participation now depends on the observed variable as well as an unobserved characteristic represented by ϵ^C. We adopt the earlier condition for determining participation status. Specifically, the individual participates (i.e. $P = 1$) if

$$Y^1 - Y^0 - C \geq 0$$

or, inserting the functions we have proposed,

$$\beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \epsilon^{Y1} + \epsilon - \beta_0 - \beta_2 \cdot x - \epsilon - \gamma_0 - \gamma_1 \cdot x - \epsilon^C \geq 0$$

or, removing terms,

$$\beta_1 + \beta_3 \cdot x + \epsilon^{Y1} - \gamma_0 - \gamma_1 \cdot x - \epsilon^C \geq 0$$

Re-arranging, we have,

$$\epsilon^{Y1} + (\beta_3 - \gamma_1) \cdot x - \epsilon^C \geq -\beta_1 + \gamma_0$$

Whether this inequality holds clearly depends on the values of ϵ^{Y1}, x and ϵ^C. Thus, this is a framework where program impact varies across individuals and certain observed and unobserved types of individuals are more common among participants than non-participants.

Before proceeding it is worth briefly considering the regression specification implied by this model. Observed Y is

$$Y = P \cdot Y^1 + (1 - P)\, Y^0$$

$$= P \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \epsilon^{Y1} + \epsilon\right) + (1 - P) \cdot (\beta_0 + \beta_2 \cdot x + \epsilon)$$

$$= \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \beta_3 \cdot P \cdot x + P \cdot \epsilon^{Y1} + \epsilon$$

Thus, x has some effect common to the two potential outcomes (captured by the term $\beta_2 \cdot x$) but also influences program impact through the term $\beta_3 \cdot P \cdot x$. The manner in which x enters the regression specification in some sense provides some insight into the genesis of the term ϵ^{Y1}: there are some unobservables that influence both potential outcomes (their common effect is subsumed in ϵ which effectively summarizes the role of the various unobservables that have a common influence

on the potential outcome) and have implications for program impact. The net effect of these is captured by the term $P \cdot \epsilon^{Y1}$.[40]

In our earlier discussion of random coefficients regression, a similar specification for the unobservables arose. Then, we found that estimating average program impact in an unbiased fashion required two assumptions in the face of a purposeful program participation decision along the lines proposed then and in the current example:

1. That ϵ^{Y1} plays no role in the participation decision (which in this context is tantamount to assuming that $\epsilon^{Y1} = 0$);

2. That the random component of the potential outcomes (ϵ) is uncorrelated with the random component of cost ϵ^C.[41]

What remains to be seen is whether matching requires similar assumptions to recover unbiased estimates of program impact.

Matching estimates program impact for each individual by finding a similar individual who experienced the counterfactual outcome. For a participant the counterfactual outcome is Y^0 while for a non-participant it is Y^1. For each individual, their counterfactual outcome is estimated by the outcome experienced by a similar person for whom that counterfactual is observed. Thus the counterfactual for a non-participant would be the outcome Y $(= Y^0)$ for a similar participant. In this manner, an estimate of $Y^1 - Y^0$ can be formed for each observed individual. It is an estimate because the value of either Y^1 or Y^0 will have been estimated for each individual according to their participation status. With the estimates of program impact $Y^1 - Y^0$ so formed for each observed individual, forming an estimate of average impact for whatever population the observed individuals represent simply involves suitably averaging across them.

To bring focus to this, suppose that you have a representative (of some population of interest) sample of N individuals and, for each, information on $\{Y_i, P_i, x_i\}$ for $i = 1, ..., N$. In other words, we observe for each individual in the sample their outcome of interest Y, program participation status P and a single observed characteristic x.[42]

More formally, x is an observed characteristic, and for any given individual j that random variable takes on the specific value $x_j = x^v$ (for instance, if x is income measured in increments of \$1,000, x^1 might be \$0, x^2 might be \$1,000, x^3 might be \$2,000, etc.). For simplicity we assume

[40]To fix ideas, one might adopt an "error components" structure along the lines of $\epsilon = \epsilon_1 + \epsilon_2$, where ϵ_2 is an unobservable that has the same effect on Y^1 and Y^2 and ϵ_1 is an unobservable that has a common influence on the two potential outcomes but also a particular influence on Y^1 (we speak of a single unobservable behind each for simplicity; in reality many different unobservable characteristics could contribute to their value). Then we have

$$\begin{aligned} Y^0 &= \beta_0 + \beta_2 \cdot x + \beta_5 \cdot \epsilon_1 + \beta_6 \cdot \epsilon_2 \\ &= \beta_0 + \beta_2 \cdot x + \epsilon \end{aligned}$$

and

$$\begin{aligned} Y^1 &= \beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \beta_4 \cdot \epsilon_1 + \beta_5 \cdot \epsilon_1 + \beta_6 \cdot \epsilon_2 \\ &= \beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \epsilon^{Y1} + \epsilon \end{aligned}$$

where $\epsilon^{Y1} = \beta_4 \cdot \epsilon_1$ and $\epsilon = \beta_5 \cdot \epsilon_1 + \beta_6 \cdot \epsilon_2$.

[41]In the regression context, an unbiased estimate actually required independence or mean independence of ϵ and ϵ^C while a consistent one required $corr(\epsilon, \epsilon^C) = 0$. We often focus on the correlation condition (e.g. the possibility that $corr(\epsilon, \epsilon^C) \neq 0$) because in the context of the behavioral models we typically consider the program participation condition would potentially generate a correlation between ϵ and ϵ^C (in the absence of assumptions), which would also undermine independence or mean independence.

[42]The discussion will soon consider observation of many characteristics, but for now it is simplest to consider just one observed characteristic.

that x is discrete (in other words, it takes on a fixed number of discrete values) with V finite possible values.[43] The different possible values of x have associated probabilities of occurring in the population:

$$Pr\left(x = x^v\right)$$

for $v = 1, ..., V$. (Another way of thinking about this is that $Pr(x^v)$ is the frequency of the v^{th} type of individual in the population.) Furthermore

$$\sum_{v=1}^{V} Pr\left(x^v\right) = 1$$

In other words, each individual in the population must have one of the V possible values for x. If the sample is indeed representative of the population of interest, the distribution of the values of x at the population level can be easily estimated. For instance, if there are N_v individuals in the sample for whom $x = x^v$,

$$\hat{Pr}\left(x = x^v\right) = \frac{N_v}{N}$$

is an unbiased estimator of $Pr(x^v)$. We can thus easily estimate from our sample the distribution of observed types of individuals across the population of interest. The remaining question before us is how we estimate average program impact for each observed type of individual, $E(Y^1 - Y^0|x = x^v)$.

The average treatment effect for the population is given by

$$ATE = \sum_{v=1}^{V} E\left(Y^1 - Y^0 | x = x^v\right) \cdot Pr\left(x = x^v\right)$$

In other words, the average treatment effect is the sum of the expectations of the treatment effect for the various types (as captured by x) of individuals in the population, with each of those expectations weighted by the frequency with which the type (i.e. $x = x^v$) on which it conditions occurs in the population.

Confronted with the sample design described, matching estimators typically estimate the average treatment effect in four basic steps:[44]

1. Estimate program impact $Y^1 - Y^0$ for each individual in the sample with characteristics $x = x^v$;

2. Average the resulting individual-level estimates of program impact across each individual with characteristic $x = x^v$;

3. Weight this average by the estimated probability of the type $x = x^v$ occurring in the population;

4. Sum these weighted averages across the V types of observed individual.

[43] Since income could in principle be infinite, this would mean some kind of 'topcoding' of income at some value would be necessary. For instance, income might be recorded in increments of $1,000 from $0 5o $100,000, with all incomes above $100,000 coded into the 102^{nd} category of v (i.e. $v = 102$).

[44] As you begin reading about different matching strategies and applications, it may seem as if there is no consensus way of combining the individual level estimates of program impact formed in the first step. For instance, some in some applications the simple average of the individual-level program estimates is computed across the sample. However, this is essentially equivalent to the four steps described below but simply collapses the last three steps into one. Thus, most matching estimators implicitly involve these steps, even if some of the steps below are collapsed together or even expanded into sub-steps.

While various particular matching exercises may differ in terms of smaller details, they all broadly follow these steps. Clearly, the foundation of the matching approach is the estimation of impact at the individual level: if that cannot be done in a fashion that is "right on average" for those for whom $x = x^v$, then the overall estimate of the average treatment effect at the population level would be biased.

Let us begin with the first step. Consider the j^{th} individual in the sample. Suppose that the j^{th} individual is a participant (i.e. $P_j = 1$ and $Y_j = Y_j^1$) with observed characteristic $x_j = x^v$. The unobserved counterfactual for that individual is their outcome had they not participated in the program (i.e. Y_j^0).

Matching would estimate the j^{th} individual's outcome in the absence of participation (Y_j^0) with the outcome actually experienced by a similar non-participant. This means that, ideally, that non-participant used to form the estimate of Y_j^0 would also exhibit the value x^v for x. Suppose, for instance, that x is income and that income is recorded in \$1,000 increments. If the j^{th} individual's recorded income was \$83,000, matching would seek to form an estimate of Y_j^0 from the observed outcome Y $(= Y^0)$ of non-participant with similar income. Ideally, there would be an individual with a recorded income of \$83,000 among the observed sample of non-participants. If so, that person's observed outcome could form the estimate of Y_j^0. If there was more than one individual with that recorded income, one of their observed outcomes could be randomly selected to serve as the estimate of Y_j^0. Alternatively, the observed outcomes of all of the non-participants with an observed income of \$83,000 could be averaged to form an estimate of Y_j^0.[45] Similarly, had the j^{th} individual been a non-participant with income equal to \$83,000, matching would have sought to form an estimate of Y_j^1 using the outcomes experienced by a participant with an income of \$83,000.

Note that these estimates of the counterfactual outcome for each individual would not equal precisely the counterfactual outcome that that individual actually would have experienced. The reason is that while we can find a similar individual in terms of the observable x, we cannot determine who is similar in terms of ϵ (or, in the case of forming a match for a non-participant, ϵ and ϵ^{Y1}). Once an estimate have been formed for all individuals, be they participants or non-participants, for whom $x = x^k$ (in our specific example, income equals \$83,000), the next step is to compute the average of these estimates. This average would be

$$\hat{ATE}_v = \sum_{z=1}^{N_v} \frac{\hat{TE}_z}{N_v}$$

where z indexes the N_v individuals for whom out $x = x^v$ of the overall sample of N individuals. Furthermore,

$$\hat{TE}_z = \hat{Y}_z^1 - Y_z^0$$

if the z^{th} individual for whom $x = x^v$ is a non-participant (because matching estimates Y^1 for each non-participant) while

$$\hat{TE}_z = Y_z^1 - \hat{Y}_z^0$$

if the z^{th} individual for whom $x = x^v$ is a participant (because matching estimates Y^0 for each participant). The two final steps involve weighting that average by an estimate the $Pr(x = x_v)$ and summing the weighted averages across the V possible values that x can take.

[45] For simplicity this example considered an income measure involving \$1,000 increments. Were income to be recorded as precisely as possible, the individual might have a much more particular observed income, such as \$82,619. It is quite possible that there would be no observed non-participant with that particular income. Typical solutions when matching to estimate Y_j^0 are to find the non-participant with the closest income to that figure (the "nearest neighbor") or average the observed outcomes of several non-participants with incomes in the near neighborhood of \$82,619.

What remains is to determine the conditions in the context of our model under which this estimator will provide an unbiased estimate of the average treatment effect. To do this, we will consider just one component of the estimator (that estimating the average treatment effect for non-participants for whom $x = x^v$). First, let N_v^1 be the number of participants and N_v^0 be the number of non-participants out of the N_v sampled individuals for whom $x = x^v$ (hence, $N_v^1 + N_v^0 = N_v$). The ATE among those for whom $x = x^v$ can then be written

$$\hat{ATE}_v = \sum_{z=1}^{N_v} \frac{\hat{TE}_z}{N_v} = \sum_{w=1}^{N_v^1} \left(\frac{\hat{TE}_w}{N_v} \right) + \sum_{q=1}^{N_v^0} \left(\frac{\hat{TE}_q}{N_v} \right)$$

where w indexes the N_v^1 participants out of the N_v individuals for whom $x = x^v$ and q does the same for non-participants. We can then write

$$\begin{aligned}
\hat{ATE}_v &= \left(\frac{N_v^1}{N_v^1} \right) \cdot \sum_{w=1}^{N_v^1} \left(\frac{\hat{TE}_w}{N_v} \right) + \left(\frac{N_v^0}{N_v^0} \right) \cdot \sum_{q=1}^{N_v^0} \left(\frac{\hat{TE}_q}{N_v} \right) \\
&= \left(\frac{N_v^1}{N_v} \right) \cdot \sum_{w=1}^{N_v^1} \left(\frac{\hat{TE}_w}{N_v^1} \right) + \left(\frac{N_v^0}{N_v} \right) \cdot \sum_{q=1}^{N_v^0} \left(\frac{\hat{TE}_q}{N_v^0} \right) \\
&= \left(\frac{N_v^1}{N_v} \right) \cdot \sum_{w=1}^{N_v^1} \left(\frac{Y_w^1 - \hat{Y}_w^0}{N_v^1} \right) + \left(\frac{N_v^0}{N_v} \right) \cdot \sum_{q=1}^{N_v^0} \left(\frac{\hat{Y}_q^1 - Y_q^0}{N_v^0} \right)
\end{aligned}$$

The last expression reflects the fact that, for participants, we observe Y^1 and it is Y^0 (their unobserved counterfactual outcome) that matching estimates for them. Similarly, for non-participants we observe Y^0 and matching provides an estimate of their outcome had they participated, Y^1.

Notice that

$$\frac{N_v^1}{N_v}$$

is just the proportion of the sample of N_v who are participants while

$$\frac{N_v^0}{N_v}$$

is just the proportion of the same who are non-participants. Behind all of this math is a simple idea: the estimate of the average treatment effect for those for whom $x = x^v$ is just the weighted average of the estimated treatment effects for participants and non-participants for whom $x = x^v$, where the weights are the proportions of the sample for whom $x = x^v$ who are participants and non-participants, respectively. These two components of the estimate of the average treatment effect for those for whom $x = x^v$ are the basic building blocks of the overall average treatment effect for the entire sample. If they cannot be estimated in an unbiased fashion, then the matching-based estimate of the population (that the overall sample represents) average treatment effect will not be unbiased.

To focus our thinking a bit, we will consider the conditions required for matching to yield an unbiased estimate of the average treatment effect for non-participants for whom $x = x^v$. This means in practice considering the conditions under which

$$\sum_{q=1}^{N_v^0} \left(\frac{\hat{Y}_q^1 - Y_q^0}{N_v^0} \right)$$

is an unbiased estimator of average program impact for non-participants for whom $x = x^v$. Since we actually observe Y^0 for non-participants (and hence matching does not estimate it), we can refine our focus even further by asking the conditions under which the matching estimator

$$\sum_{q=1}^{N_v^0} \left(\frac{\hat{Y}_q^1}{N_v^0} \right)$$

provides an unbiased estimate of Y^1 for non-participants for whom $x = x^v$. Earlier we noted that the matching estimator cannot yield the exact counterfactual outcome for each individual. In some sense the unbiasedness of the matching estimator rests on the idea that the difference between the actual counterfactual outcome Y_q^1 that the non-participant would have experienced and the estimate of that counterfactual outcome $\hat{Y}_q^1$ generated through matching is completely random. Hence such differences between actual (Y_q^1) and estimated ($\hat{Y}_q^1$) counterfactual outcomes should cancel out on average across the non-participants for whom $x = x^v$.

Since the outcome under participation, Y^1, is determined by

$$Y^1 = \beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \epsilon^{Y1} + \epsilon$$

the question is really whether the matching estimate of Y^1 for each non-participant for whom $x = x^v$ correctly captures Y^1, as specified, on average for those non-participants. Since the match for each non-participant is a randomly selected participant with characteristic $x = x^v$, we can be assured that the component

$$\beta_0 + \beta_1 + \beta_2 \cdot x$$

is captured correctly (since we are fixing x at x^v for the non-participant and their participant match and β_0 and β_1 are parameters that do not vary in value between participants and non-participants). The question then becomes whether the values for ϵ^{Y1} and ϵ for the typical participant with $x = x^v$ are identical for those of the typical non-participant with the same values for x.

To answer this question, we turn to the program participation decision determination condition

$$\epsilon^{Y1} + (\beta_3 - \gamma_1) \cdot x - \epsilon^C \geq -\beta_1 + \gamma_0$$

If ϵ^{Y1} plays a role in the program participation decision (i.e. if $\epsilon^{Y1} \neq 0$) then the participants would, other things being equal, be those with larger values for ϵ^{Y1}. By dint of this participants would be individuals that, for a given value of income x, experienced a different average value of Y^1 than non-participants would have had they participated: a participant with characteristic $x = x^k$ would have a larger expected value for Y^1 than a non-participant with same characteristic because those with larger ϵ^{Y1} tend to be more likely to participate. We can thus see that ϵ^{Y1} playing a role in the program participation decision (which would be the case if $\epsilon^{Y1} \neq 0$) would preclude unbiased estimation of Y^1 for non-participants using matched participants: it will tend to overestimate Y^1 for non-participants given for each given value of x.

Interestingly, having $\epsilon^{Y1} \neq 0$ would not create problems for estimating Y^0 for participants from the outcomes experienced by matched non-participants. Thus this aspect of this particular model would not necessarily create problems for estimating, for example, the average effect of treatment on the treated (i.e. $E(Y^1 - Y^0|P = 1)$). We observe $E(Y^1|P = 1)$ from our participant sample. Hence we need only to find matches for participants among non-participants to form the estimate of $E(Y^0|P = 1)$. That said, a trivial extension of the model could introduce a similar complication for estimating Y^0 for participants. For instance, the potential outcomes framework

$$Y^0 = \beta_0 + \beta_2 \cdot x + \epsilon^{Y0} + \epsilon$$

$$Y^1 = \beta_0 + \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x + \epsilon^{Y1} + \epsilon$$

would create an analogous problem for estimating Y^0 for participants: the non-participant group would tend to have larger values for ϵ^{Y0}, implying that their observed experiences in the absence of the program would serve as a misleading indication of what participants would have experienced had they not participated. In general, therefore, matching estimators presume the absence of factors like ϵ^{Y0} and ϵ^{Y1} in the potential outcomes (or, put differently, matching estimators will provide an unbiased estimate of program impact only if there are no factors such as ϵ^{Y0} and ϵ^{Y1}).

The other major assumption in the random coefficients setting was that ϵ was unrelated to ϵ^C. In other words, the unobservables governing the potential outcomes could not be related to those determining the costs of enrollment. This assumption is also necessary to view matching estimates as unbiased. The reasoning is quite straightforward. Clearly the program participation decision condition implies that, other things being equal, some types in terms of their value for ϵ^C are more likely to be participants than other types. In other words, the value of ϵ^C is likely to differ between participants and non-participants. In particular, those with lower ϵ^C are more likely to be participants. If ϵ, which contains the unobservables from the potential outcomes equations, was related to ϵ^C, this could result in different types in terms of ϵ among participants compared with non-participants.

For instance, suppose that ϵ and ϵ^C are positively correlated. Therefore those with higher potential outcomes tend to have higher costs of enrollment. Since participants will tend to have lower values for ϵ^C, this means that they would also have lower values for ϵ. By the same logic, non-participants would tend to have larger values for ϵ. This would complicate estimation of program impact by matching considerably. For a given value of x, the outcomes observed among participants ($Y = Y^1$) would tend to understate the outcomes that non-participants with the same values for x would have experienced had they participated. By the same token, the outcomes that non-participants with a given value for income x experienced ($Y = Y^0$) would tend to overstate the outcomes that participants with the same value for x would have experienced had they not participated. The end result is that for each value of x the average treatment effect would be underestimated.

These assumptions (that $\epsilon^{Y1} = 0$ and that ϵ and ϵ^C are unrelated) insure that for every type of individual (as defined by the observed characteristic x) the average values of Y^1 and Y^0 are the same among participants and non-participants. In other words, across the entire sample the only reason that the average values of Y^1 and Y^0 might be different between participants and non-participants is that the average value of the observed characteristic x might differ between them.

In other words, the average values of Y^1 and Y^0 should be the same between participants and non-participants with the same value for the characteristic x. There are a number of ways that this is commonly expressed. Two particularly popular ways of asserting this are

$$E\left(Y^1, Y^0 | x, P\right) = E\left(Y^1, Y^0 | x\right)$$

and

$$Y^0, Y^1 \perp P | x$$

The first states that, conditional on x, program participation status reveals no information about the expected values of Y^1 and Y^0 (i.e. conditional on x they have the same expected values across participants and non-participants). In other words, conditional on the observed characteristic x, Y^1 and Y^0 are *mean independent* of program participation P. The second condition states that Y^1 and Y^0 are **orthogonal** to program participation P conditional on observed x. The precise

meaning of orthogonality in practice has, in the authors' experience, exhibited a certain elasticity across the various statistical and econometric manuscripts we have read. In some cases it seems to have meant independence, in others that a lack of correlation (e.g. Green (2000) p. 229) or even something in some sense in between the two (e.g. Rodgers et al. 1984).

To fix ideas, let us focus on the mean independence assumption $E(Y^1, Y^0|x, P) = E(Y^1, Y^0|x)$. The potential outcomes are, as modelled above, essentially linear functions of the observable x and a random component ϵ. When we say that we assume that the potential outcomes are mean independent of program participation after factoring in the role of the observable x, this is tantamount to assuming that the unobservable ϵ is mean independent of program participation P.

Notice that this is the assumption required to insure the ordinary least squares estimator is unbiased. *Thus we see that unbiased estimation of program impact using the matching estimators rely on essentially the same assumption as was required to assume that linear regression was yielding unbiased estimates of program impact.*[46] In the next section, we will see that the similarity extends even further: regression can be interpreted as a kind of matching estimator.

4.2.2 Regression as Matching

We have seen that the matching estimator for the average treatment effect is

$$\hat{ATE} = \sum_{v=1}^{V} \left(\left(\sum_{zv=1}^{N_v} \frac{\hat{TE}_{zv}}{N_v} \right) \cdot \hat{Pr}\left(x = x^v\right) \right)$$

where

$$\hat{TE}_{zv} = \hat{Y}^1_{zv} - Y^0_{zv}$$

for in the event that the zv^{th} individual in the sample is a non-participant and

$$\hat{TE}_{zv} = Y^1_{zv} - \hat{Y}^0_{zv}$$

in the event that the zv^{th} individual in the sample is a participant.

This is a natural approach to estimation when the estimand (i.e. parameter we wish to estimate with our sample) is

$$ATE = \sum_{v=1}^{V} E\left(Y^1 - Y^0 | x = x^v\right) \cdot Pr\left(x = x^v\right)$$

Thus the matching estimator is simply attempting to estimate the average treatment effect ATE by forming estimates of impact at the individual level and then averaging them, where the averaging process is effectively broken down into two steps with the matching estimator.

We will now demonstrate that the least squares regression estimator can be characterized as a sort of matching estimator that places slightly different weights on the estimated average treatment effects for each given value of the observable x.[47] For the purposes of this demonstration, we also temporarily simplify our model somewhat, so that we now assume the framework

$$Y^0 = \beta_0 + \beta_2 \cdot x + \epsilon$$

[46] Depending on the specific meaning of othogonality considered, the assumption $Y^0, Y^1 \perp P|x$ is consistent either with that required to yield unbiased estimates of program impact in the regression context or simply consistent estimates. Most of the definitions of othogonality that we have seen would align with the assumptions required for unbiasedness in the regression context.

[47] The discussion to follow draws heavily on and is a simplification of p.71-77 of Angrist and Pischke (2009). What follows can be viewed as a sort of poor man's version of their more sophisticated discussion of this topic. We highly recommend their treatment of it to the interested reader.

$$Y^1 = \beta_0 + \beta_1 + \beta_2 \cdot x + \epsilon$$

This is a modification of the framework introduced in the last section in that it drops the term $\beta_3 \cdot x$ (in other words, it eliminates the possibility that program impact depends on x). This is an unnecessary complication for present purposes. The regression specification implied by this is

$$Y = P \cdot Y^1 + (1 - P) \cdot Y^0$$

$$= P \cdot (\beta_0 + \beta_1 + \beta_2 \cdot x + \epsilon) + (1 - P) \cdot (\beta_0 + \beta_2 \cdot x + \epsilon)$$

$$= \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

This is essentially the simple specification that we encountered at the outset of the discussion of multiple regression in the last section.

As we saw in the regression anatomy discussion in the regression section, the estimate $\hat{\beta}_1$ can also be recovered by regressing Y on $P^{\hat{RES}}$, where $P^{\hat{RES}}$ is the predicted residual

$$P_i^{\hat{RES}} = P_i - \hat{P}_i$$

where $\hat{P}_i$ is predicted program participation from the estimated (or "fitted" model) yielded by regression of P on x. In other words, the ordinary least squares estimate $\hat{\nu_1}$ of ν_1 from

$$Y_i = \nu_0 + \nu_1 \cdot P_i^{\hat{RES}} + \vartheta_i$$

is equal to $\hat{\beta}_1$.

The ordinary least squares estimator of ν_1 is, of course,

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} P_i^{\hat{RES}} \cdot Y_i}{\sum_{i=1}^{N} {P_i^{\hat{RES}}}^2} = \hat{\beta}_1$$

Substituting in

$$P_i^{\hat{RES}} = P_i - \hat{P}_i$$

we have

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i\right) \cdot Y_i}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i\right)^2}$$

$$= \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i\right) \cdot \hat{Y}_i}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i\right)^2}$$

where $\hat{Y}_i$ is the predicted Y emerging from the regression of Y on P and x.

This last step might not seem particularly intuitive. It is based on the fact that regression of Y on P and x is equivalent to regression of $\hat{Y}_i$ on P and x in terms of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that would be obtained. The reason for this is actually fairly intuitive: regression yields the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that captures all of the variation in Y that can be captured by P and x. The rest of the variation in Y not captured by the fitted regression

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot P + \hat{\beta}_2 \cdot x$$

is essentially ignored in forming $\hat{Y}$. In other words, since $\hat{Y}$ contains all of the variation in Y that can be explained by P and x (at least in the context of the linear regression model derived from

the behavioral model in this subsection) it is also all of the variation required by ordinary least squares to form estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

The next step requires a bit of trickery regarding $\hat{Y}_i$. Specifically,

$$\hat{Y}_i = \hat{Y_i^0} + P_i \cdot \left(\hat{Y_i^1} - \hat{Y_i^0} \right)$$

Inserting this into

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right) \cdot \hat{Y}_i}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

we have

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right) \cdot \left(\hat{Y_i^0} + P_i \cdot \left(\hat{Y_i^1} - \hat{Y_i^0} \right) \right)}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

$$= \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right) \cdot \hat{Y_i^0}}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2} + \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right) \cdot P_i \cdot \left(\hat{Y_i^1} - \hat{Y_i^0} \right)}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

The term

$$\frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right) \cdot \hat{Y_i^0}}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

is equal to zero (because $\hat{Y_i^0}$ is a function of x alone and hence uncorrelated with the regression residuals from the regression of P on x[48]). This leaves us with

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right) \cdot P_i \cdot \left(\hat{Y_i^1} - \hat{Y_i^0} \right)}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

$$= \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2 \cdot \left(\hat{Y_i^1} - \hat{Y_i^0} \right)}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

We have now established that the least squares estimator of ν_1 is akin to a kind of matching estimator. Specifically, it involves a weighted average of the estimates of the estimated individual-level program impacts

$$\hat{Y_i^1} - \hat{Y_i^0}$$

The weights differ from those used by the matching estimator discussed in the last subsection, but they are very similar estimators in that they offer weighted averages of the estimated individual-level program impacts.

The regression weights are effectively related to the conditional (on x) variance of P. This can be seen more clearly with this simple modification:

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2 \cdot \left(\hat{Y_i^1} - \hat{Y_i^0} \right)}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2}$$

[48] The predicted residuals from any least squares regression are by construction uncorrelated with the regressors.

$$= \frac{\sum_{i=1}^{N} \frac{\left(P_i - \hat{P}_i\right)^2}{N-2} \cdot \left(\hat{Y_i^1} - \hat{Y_i^0}\right)}{\frac{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2}{N-2}}$$

The regression thus gives more weight to those whose values of x are associated with larger conditional variance of P. Since P is a binary variable (and its variance is thus equal to $\hat{P} \cdot (1-\hat{P})$[49]), the regression estimator given the largest weight to values of x at which the treatment probability is closest to 0.5 (i.e. cells where there is the most balance between participants and non-participants).[50]

Previously, we have seen that unbiased estimation of program impact by matching typically requires assumptions like those required to hold for unbiased estimation of program impact in the multiple regression setting. We have just seen that regression can be viewed as a species of matching with a particular set of weights attached to the K estimates of expected program impact conditional on the observable $x = x^v$,

$$E\left(Y^1 - Y^0 | x = x^v\right)$$

In the next section we will round out the discussion of similarities between matching and regression by exploring one more way that either can yield a biased impact of average program impact at the population level.

4.2.3 The Failure of Common Support

We now discuss one other way that either regression or matching could yield a biased estimate of average program impact at the population level. To approach it, let us return to the program participation decision rule from the model we introduced in the last subsection. Specifically, an individual will participate if

$$\epsilon^{Y1} + (\beta_3 - \gamma_1) \cdot x - \epsilon^C \geq -\beta_1 + \gamma_0$$

Re-arranging, we have

$$x \geq \frac{1}{(\beta_3 - \gamma_1)} \cdot \left(-\beta_1 + \gamma_0 + \epsilon^C - \epsilon^{Y1}\right)$$

By the same token, the individual would not participate if

$$x < \frac{1}{(\beta_3 - \gamma_1)} \cdot \left(-\beta_1 + \gamma_0 + \epsilon^C - \epsilon^{Y1}\right)$$

[49] Binary variables have a variance of $P \cdot (1 - P)$, which happens to reach its maximum at $P = .5$

[50] This might seem to be in some sense a vacuous example since

$$\hat{Y_i^1} - \hat{Y_i^0} = \hat{\beta_1}$$

(where $\hat{\beta_1}$ is the estimate of β_1 from regression of Y on P and x) for all individuals $i = 1, ..., N$ and potential values of x. Hence one could slip $\hat{\beta_1}$ out of the summation and be left with

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2 \cdot \left(\hat{Y_i^1} - \hat{Y_i^0}\right)}{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2} = \frac{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2 \cdot \hat{\beta_1}}{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2}$$

$$= \hat{\beta_1} \cdot \frac{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2}{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2} = \hat{\beta_1}$$

This is to a certain extent a fair point, but the model was chosen simply to provide the simplest route to illustrating how the regression estimator can be re-cast as a sort of matching estimator.

There is no guarantee that these conditions will hold for at least *some* individuals at every value of x.

Suppose again that we have a sample of N individuals for whom we observe $\{Y_i, P_i, x_i\}$ for $i = 1, ..., N$. In other words, we observe the outcome of interest, program participation and a single observed characteristic. Once again, it would be useful to return to our earlier conceptualization of x, under which it took on V finite and distinct values (as when we imagined that it represented income recorded in increments).

Consider one particular value for x, $x = x^v$. The individuals with this value for x are indexed from $j = 1, ..., N_v$. Furthermore, let us suppose that individual z among these N_v persons for whom $x = x_v$ has the smallest value for

$$\frac{1}{(\beta_3 - \gamma_1)} \cdot \left(-\beta_1 + \gamma_0 + \epsilon^C - \epsilon^{Y1}\right)$$

If

$$x_j < \frac{1}{(\beta_3 - \gamma_1)} \cdot \left(-\beta_1 + \gamma_0 + \epsilon_z^C - \epsilon_z^{Y1}\right)$$

then no one among those observed with $x = x^v$ would choose to participate in the program. (By similar logic one can imagine values for x where everyone choose to participate and so no one with that value for x would be a non-participant.)

This is typically referred to in the literature on matching as the **Failure of Common Support**. Recall that the support of a random variable is those values for which is has a probability greater than zero. The Failure of Common Support is so called because it implies that there are values for x for which the probability that x take on that value is zero for either participants or non-participants.

There is no theoretical reason that this cannot happen and it should be immediately obvious that it presents a fundamental complication for matching: there is no one among the individuals for whom $x = x^v$ that can be used to form an estimate of Y^1. In other words, referencing the estimand (i.e. what we seek to estimate), no estimate can be formed for

$$E\left(Y^1 - Y^0 | x = x^v\right)$$

because there is no instance in the sample where Y^1 is observed when $x = x^v$.

The failure of common support is actually a potential source of bias to estimates of average treatment effects in regression models as well, although the reason why is not as immediately obvious as was the case with matching. To see this, first let us focus on a regression model "saturated" in x:

$$Y = \beta_0 + \beta_1 \cdot + \sum_{v=1}^{V-1} (\beta_2 v \cdot D^v) + \epsilon$$

where D^v is a dummy variable that, for a given individual, equals 1 if $x = x^v$ for that individual and zero otherwise. A model is saturated in some respect if it contains a dummy variable for every possible value combination that can occur in that respect. Thus, the model above is saturated in x because it has a dummy variable for every value of x (with one dummy excluded to avoid perfect collinearity with the constant term). If there were two observed variables, x_1 and x_2, then a model saturated in x_1 and x_2 would include dummy variables for each value of the two variables, as well as the full set of possible interactions of those dummy variables (see Angrist and Pischke (2009) for more on saturated models and their possibilities).

In the last subsection, we saw that the regression estimator of average program impact could be recast as a sort of matching estimator:

$$\hat{\nu_1} = \frac{\sum_{i=1}^{N} \left(P_i - \hat{P}_i\right)^2 \cdot \left(\hat{Y_i^1} - \hat{Y_i^0}\right)}{\sum_{i=1}^{N} \left(P_i - \hat{P}_i\right)^2}$$

Assume as well that the regression of P on x (yielding $\hat{P}$) is fully saturated in x. With the saturated model, we have introduced a great deal of flexibility in terms of how x influences P. Indeed, the specification can capture potentially complex non-linearity to the effect of x on P.

Recall, however, that since P is binary the regression weights, which essentially involve the variance of P conditional on x can also be written

$$\hat{P}_i \cdot \left(1 - \hat{P}_i\right)$$

But this will be equal to zero under the regression of P on x saturated in x at any value of x where everyone or no one is a participant. Hence no weight is given to those whose values of x are those for which everyone or no one is a participant. While it is true that this makes little difference in a framework where everyone has the same program impact (of β_1) it could begin to make a significant difference if impact varied across individuals and, in particular, impact depended on the value of x.[51]

In many applications, failures of common support have effectively been hidden and are not apparent on first glance. For one thing, it is not typical practice in applied empirical work for regression models to be fully saturated. If, for instance, we had just included x as originally conceived (income measured in $1,000 increments) as a regressor instead of discretizing it and fully saturating the regression model in the resulting dummy variables, the contribution of cells with only or no participants would remain by extrapolation.[52] In matching it is common to use various gimmicks, such as combining cells (i.e. values for x in this context) with small numbers of observations, which are among those more likely to contain only participants or non-participants, in some sense disguising the problem. However, while many applications may disguise the problem, that does not mean that it isn't present. One should always think of the task of estimating the fully saturated model with reasonably discretized observed covariates to get some sense of how whether the failure of common support is likely an issue.

4.2.4 The Propensity Score

To this point, our discussion of matching has focused on forming matches based on a *single* observed characteristic x. In many samples, numerous characteristics of the individuals contained within them might be observed. Matching on these as well allows one to form more and more precise matches for estimating the counterfactual.[53] Indeed, the more precisely one can match the more plausible is becomes that a characteristic influencing both cost and potential outcomes has not been omitted and thus left in ϵ and ϵ^C, in the process creating an avenue for correlation between them and thus introducing bias to the matching estimate.

Furthermore, to a point such matching is quite straightforward. Extending the example above, suppose that we now observe two characteristics of the individual, income (x_1) and age (x_2).

[51] This can be easily demonstrated by thinking about using the regression anatomy formula to estimate separately β_1 and $\{\beta_{21}, \beta_{22}, ..., \beta_{2V-1}\}$.

[52] Indeed, in our example of income measured in increments of $1,000 was top-coded, papering over many possible failures of common support had we continued discretizing based on increments of $1,000 above the top code value.

[53] Though matching on a "bad control" presents the same danger as adding a control did in the regression context.

Suppose as well that income continues to be observed with V finite values (i.e. $x_1 = x^v$, $v = 1, ..., V$). Age is observed in W categories (perhaps representing W possible 5 year age ranges), which we denote by $x_2 = x^w$, $w = 1, ..., W$.

Matching not simply involves finding estimates of the counterfactual outcomes for each the $k \cdot W$ possible combinations of the values of x_1 and x_2. The estimand for this is straightforward:

$$ATE = \sum_{w=1}^{W} \left(\sum_{v=1}^{V} E\left(Y^1 - Y^0 | x_1 = x^v, x_2 = x^w\right) \cdot Pr\left(x_1 = x^v, x_2 = x^w\right) \right)$$

The matching estimator matches in form this estimand in the same sense as is the case with one observable x.

It is not hard to see, however, how adding observables in this fashion can quickly result in a kind of "curse of dimensionality". For instance, suppose that for each individual in our sample we observed income (coded into 5 categories), region of residence (coded into 5 categories), religion (coded into 4 categories), age (coded into 8 categories), gender (coded into 2 categories) and education (coded into 5 categories). This implies 5 X 5 X 4 X 8 X 2 X 5 = 8,000 potential observed "types" of individuals in our data for whom we might need to find a match! (And for the purposes of many applications, these characteristics don't even really tell us that much about an individual!) It is easy to see how this can get out of control (consider this: the addition of just one more observed characteristic coded into 6 categories would have left us with nearly *50,000* types).

Moreover, as the number of potential combinations of observed categories on which we match grows, so too does the likelihood of not being able to form counterfactual matches because everyone in a particular category either always participated or never participated. This can happen for two reasons. First, there is always the possibility that the theoretical failure of common support described in the last subsection might occur. Second, in practice matching is usually done with representative samples. If a certain type of individual occurred rarely enough in the population, this might mean that most reasonable sized representative sample might not provide a match. Suppose, for instance, that some potential combination of observed characteristics represented 0.1 percent of the American population, and that one in five such individuals participate in a program. That means that across the United States roughly 300,000 individuals fall into that category and 60,000 of them are program participants. But in a random sample of 10,000 Americans we might expect on average only 10 sample individuals to be of this type, and only 2 of them to be participants. It is not hard to envision many representative samples that include no individuals of this particular type who are participants.[54]

The solution to this is to estimate the counterfactual for each individuals in a sample by matching them to an individual who experienced the counterfactual outcome and had a similar probability of participation conditional on observed characteristics $\{x_1, x_2, ...\}$. The probability of treatment conditional on K observed covariates,

$$Pr\left(P = 1 | x_1, x_2, ..., x_K\right)$$

is called the **Propensity Score**. Matching on the propensity score is probably[55] the most common current approach to matching.

[54] Another possibility, of course, is that no individual of a given type, participant or non-participant, is observed at all for a given theoretical combination of observed characteristics. This is a potential problem in terms of bias if that particular combination of characteristics does in fact occur in the population.

[55] This is a subjective assessment by the authors; an attempt at estimation of the relative popularity of propensity scores would be non-trivial.

The basic steps in propensity score estimation are straightforward. Assume again a familiar general data setup: we observe for a sample of N individuals an outcome of interest and program participation $\{Y_i, P_i\}$ $(i = 1, ..., N)$. Furthermore, we assume that we observe K individual-level characteristics captured in the variables $x_1, x_2, ..., x_k$. For each of the $i = 1, ..., N$ individuals in the sample, we thus also observe $\{ x_{1i}, x_{2i}, ..., x_{ki}\}$.

The first step is to decide on a specification for the propensity score model. The dependent variable is the program participation decision P. One must then decide on the explanatory variables out of $x_1, x_2, ..., x_k$ to include. In practice, many adopt an "everything and the kitchen sink" approach and include everything for fear of omitting some factor that differed between participants and non-participants and influenced outcomes. The obvious alternative is a more parsimonious strategy that seeks to identify those controls out of $x_1, x_2, ..., x_k$ that really do influence both participation and the outcome of interest. Most of these rely (implicitly or explicitly) on appeal to a model of the participation decision that attempts to identify which variables matter for participation and which do not do so. A danger with the latter route is that the assumed model of the participation decision might fail to recognize some observed variable associated with the participation decision. A danger with the former is the possibility of including a bad control. Some authors (e.g. Rosenbaum and Rubin 1984) propose iterative approaches whereby controls are added incrementally, and with increasingly rich specifications (e.g. interactions, polynomial terms, etc.) until participants and non-participants appear similar at various values for the propensity score.

A second question regarding the specification is how to handle any continuous variables that are included as explanatory variables. There does not seem to be a widely accepted answer to this question at this point. Some appear to suggest inclusion of polynomial terms for these variables (e.g. Angrist and Pischke 2009). In practice such variables are often discretized, so that a given continuous variable enters into the regression as a series of dummies that equals one if the continuous variable was in some range and zero otherwise.

The final question is how to estimate the regression of P on the selected control variables. In principle, any binary regression model should be adequate. Some claim that there are advantages to using the logit regression model. However, this is not really empirical "settled law" at this point.

One the propensity score is estimated, a whole new question arises: what to do with it. This subsection is focused on matching on the propensity score.[56] The idea behind doing so is that those with similar propensity score values will have similar background characteristics.

This is the **balancing property** of the propensity score. The basic idea is that, for a given value of the propensity score, participants and non-participants should have similar distributions of background characteristics. Many tests of this property are essentially tests of whether the propensity score satisfies this property. We will briefly, and perhaps more intuitively than deeply technically, describe such tests.[57] The tests generally take several approaches to the balancing property. Most obviously, some tests simply compare whether the mean of the various covariates is the same between participants and non-participants at a given propensity score. One could, for instance, focus on a standardized difference along the lines of

$$\frac{\overline{x^P} - \overline{x^{NP}}}{\sqrt{\frac{s_P^2 + s_{NP}^2}{2}}}$$

where x is one of the included controls in the propensity score, $\overline{x^P}$ and $\overline{x^{NP}}$ are the means of x among participants and non-participants, and the ss are their corresponding standard errors

[56] In the final subsection of this section, we will discuss some other applications of the propensity score.

[57] See Austin (2008), Austin (2009) and Austin (2011), and work cited within them, for much more in depth, interesting discussion of tests of the balancing property.

(Austin 2011). This would be computed at various values for the propensity score (for instance, one could identify strata of the value of the propensity score based on noticeable mass points for the score) to see if average values for x appeared to be the same at various values for the propensity score. Other tests involve examining differences in higher moments of controls at various value for the propensity score, the c-statistic for the propensity score,[58] etc. The balancing property should be checked not only among the full sample (among whom the propensity score model is estimated) but also any final sample of matches formed on the propensity score for estimating program impact.

We refer the reader to sources such as Austin (2011) for deeper consideration of the specifics of specification of the propensity score, and content ourselves for now with this very brief overview. First, there has been so much methodological work in this area that a careful treatment would consume the manual. Second, there is no real "pay off" in the end in that there is no real received consensus about best practice (something to bear in mind before becoming to rigidly defensive about one's own approach in a particular application or too harshly critical of that adopted by others). This extends beyond the question of getting the least biased estimate of program impact but also the most efficient one (i.e. one with the least sampling variation).[59]

Matching on the propensity score is in principle a rather simple affair: for each individual in the sample for whom an estimate of program impact needs to be formed, find an individual or individuals experiencing the counterfactual outcome with a similar propensity score. Use the counterfactual outcome(s) that that individual (or those individuals) experienced as an estimate of the outcome the first individual, for whom we wish to form an estimate of program impact, would have experienced under the counterfactual state. For example, we would estimate the outcome a non-participant would have experienced had he or she participated by the outcome or outcomes experienced by a participant or participants with a similar propensity score.

Unfortunately, beyond this simple overall logic lies what can seem a bewildering set of alternative proposed practices for forming a match. Briefly, these might include:

- "Nearest Neighbor" matching: matches each individual for whom a program impact estimate must be formed with the individual with the closest propensity score value experiencing the counterfactual outcome;
- Caliper matching: involves finding matches within a certain distance from the propensity score value for the individual requiring a match;
- Kernel and local linear matching: uses weighted contributions from all individuals in the counterfactual state to form an estimate of the counterfactual outcome.

This is in no sense a complete list of the alternatives. Indeed, even the alternatives listed above frequently involve further choices. For instance, for caliper matching, how wide does one set the search distance? Does one select the closest match within the caliper, average across all of the matches within the caliper distance, or form some sort of weighted average of the cases with propensity scores within the caliper distance? The considerations are not limited to identification of matching methods either. For instance, one must decide whether to match with replacement or

[58]The c-statistic is, in the context, basically a measure of the ability of the propensity score model to discriminate between participants and non-participants. For instance, in many instances, the c-statistic is the fraction of pairs of observations where one participates and one does not among which the propensity score indeed predicted a higher probability of participation for the participant. Austin (2011), citing various studies, reasonably question how this reveals much about whether the specification of the propensity score model is correct in the sense that it needs to be for the purpose of covariate balancing between participants and non-participants.

[59]See Hahn (1998), Hirano, Imbens and Ridder (2003) and Angrist and Hahn (2004) for key econometric papers focused on efficiency and the propensity score.

not. A related consideration is "greedy" versus "optimal" matching. Under greedy matching, one would select, for instance, the nearest neighbor to each successive individual requiring a match, even if the individual providing the match might have served as a closer match for a subsequent individual on the list requiring matches.

There is, to our knowledge, no consensus regarding best practice across these many possibilities, either in general or for any particular application. At this point, we make a gross generalization that is somewhat editorial in nature and likely would not hold in many applications, but is nonetheless probably a reasonable useful rule of thumb: the choice of matching method probably does not matter much. In our own work, we have often found that the differences in results from alternative matching methods were modest compared with, for instance, the difference of any or all of them with the estimate of program impact yielded by simple comparison of average outcomes between participants and non-participants. Certainly one should examine the robustness of the estimate of program impact to alternative matching methods.

As a final note, matching on the propensity score is just as subject to concerns regarding failures of common support. Specifically, there may be some ranges for the predicted propensity score that appear only among participants or non-participants. For instance, it is not uncommon that, in a propensity score model that does appear capable of discriminating between participants and non-participants (as evidenced by, for instance, the c-statistic), the maximum propensity score value observed among participants well exceeds that for non-participants. In such cases, matches may not be possible. For instance, participants with propensity score values in the upper ranges not seen among non-participants do not have a match among non-participants with comparable score. Similarly, the lower range of scores for non-participants may well fall below the minimum score values of participants, making it impossible to form matches for the non-participants with these low scores.

If one's goal was to form an estimate of the average treatment effect for the population of interest represented by the sample, these failures to match can present a significant problem. Program impact at the individual level can be computed only for those for whom a match can be formed. This subsample is likely non-random, and hence not necessarily (in fact, probably not) representative of the population of interest as a whole. It is thus difficult to know for whom program impact is being estimated.

4.2.5 Other Propensity Score Strategies

There are a few other ways in which the propensity score is applied to achieve "balancing" of observed types between participants and non-participants. First, in what is probably best viewed as only a slight variation on matching itself, the sample is sometimes stratified according to the various values for the propensity score. For instance, it is common practice to stratify individuals in a sample according to quantiles of the propensity score, with quintiles as a popular choice (Austin, 2011). The average outcomes for participants and non-participants are then averaged in each of these strata and their difference represents program impact within that strata. The idea behind this is that within each of these strata the participants and non-participants will have approximately similar background characteristics. The strata specific estimates of program impact are then averaged across strata. Rosenbaum and Rubin (1984) have found that this basic approach removes most of the bias associated with observed confounders. One can also compute a weighted average of the estimates from the various strata. For instance, when the contribution of each strata is weighted by the number of participants in each can yield an estimate of the average effect of treatment on the treated (Imbens 2004).

Another increasingly popular application of the propensity score is to weight the sample ac-

cording to the inverse of the probability of participation. The general idea behind this is to use the weights as a means of correcting for the imbalance of observed characteristics between participants and non-participants. To begin with, one would expect that the average propensity score value among participants likely exceeds that among non-participants. The weighting estimator would thus assign lower weight to participants with larger values for the propensity score more distant from those found among non-participants. A similar weighting scheme applies to non-participants. For instance, for the purpose of estimating the average treatment effect, Angrist and Pischke (2009) present the following simple expression:

$$E\left(Y_i^1 - Y_i^0\right) = E\left(\frac{Y_i \cdot P_i}{PS_i} + \frac{Y_i \cdot (1 - P_i)}{1 - PS_i}\right)$$

where PS_i is the propensity score for the i^{th} individual. In practice one would compute the sample analog by simply averaging

$$\frac{Y_i \cdot P_i}{\hat{PS}_i} + \frac{Y_i \cdot (1 - P_i)}{1 - \hat{PS}_i}$$

(where $\hat{PS}_i$ is the predicted, which is also to say estimated, propensity score) across the sample of $i = 1, ..., N$ individuals. Notice that this will assign rather lower weights to participants with very high propensity scores and non-participants with very small scores. A similar weighting estimator can be derived for the average effect of treatment on the treated. A major advantage of these estimators is that they sidestep all of the tedious (and somewhat arbitrary) decisions regarding matching methods.

Finally, the propensity score is sometimes used as a regressor to summarize the effect of confounders. The classic basic model is as follows

$$Y_i = \beta_1 + \beta_2 \cdot P_i + \beta_3 \cdot \hat{PS}_i + \epsilon_i$$

where $\hat{PS}_i$ is the estimated propensity score for individual i. Variations on this include specifications with polynomial terms for the estimated propensity score. Some authors (e.g. Wooldridge 2001) have questioned whether this approach is truly superior simply to including the regressors (specified richly using discretization or polynomial terms for continuous regressors, interactions, etc.) used in forming the propensity score rather than the propensity score itself.

4.2.6 An Empirical Example

In this subsection we provide an empirical illustration (provided in the STATA do-file 4.6.do) of the performance of various estimators involving the propensity score. These are compared with results from naive (in the sense of ignoring possible bias from confounding variables) estimation of program impact by comparison of average outcomes between participants and non-participants and regression of the outcome of interest on an indicator of program participation. They are also compared with results from regression of the outcome of interest on program participation and the confounders.

Once again, we simulate our data. Our sample size this time is 5,000 individuals. We begin by parameterizing a framework for the determination of program participation and outcomes of interest. We consider 5 potential confounders $x_1, ..., x_5$. The basic framework for determining the potential outcomes and cost of enrollment is:

$$Y^0 = 2 + 1.5 \cdot x_1 - 1.5 \cdot x_2 + 1.2 \cdot x_3 + 2.4 \cdot x_4 + 3 \cdot x_5 + \epsilon^Y$$

$$Y^1 = 4 + 1.5 \cdot x_1 - 1.5 \cdot x_2 + 1.2 \cdot x_3 + 2.4 \cdot x_4 + 3 \cdot x_5 + \epsilon^Y$$

$$C = 1 - .35 \cdot x_1 + .15 \cdot x_2 - .25 \cdot x_3 + x_4 - .5 \cdot x_5 + \epsilon^C$$

where the ϵs are independently drawn from the normal distribution with variance 100 and the xs are drawn from the standard normal distribution. An individual participates in the program if $Y^1 - Y^0 - C > 0$.

Summary statistics for this sample are provided in Outputs 4.28 and 4.29. Just over half the sample (57.48 percent) wind up as participants in the program. ϵ^Y has roughly the same value among participants and non-participants, while the average value of ϵ^C is notably higher among non-participants. The former reflects the fact that ϵ^Y will cancel out in the term $Y^1 - Y^0$ and hence play no role in program participation while ϵ^C clearly is a determinant of participation. The difference in the average value of ϵ^C suggest that those facing higher unobserved costs of participation are less likely to participate. There appear to be differences in the average values of the xs between the two samples, reflecting the systematic role that they play in shaping cost.

STATA Output 4.28 (4.6.do)

```
. sum P
```

Variable	Obs	Mean	Std. Dev.	Min	Max
P	5000	.5748	.4944227	0	1

STATA Output 4.29 (4.6.do)

```
. by P, sort: sum epsilony epsilonc x_1 x_2 x_3 x_4 x_5 y0 y1
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
epsilony	2126	.0015076	.9870313	-3.186304	3.263425
epsilonc	2126	.8940831	.6064059	-.4481313	3.188653
x_1	2126	-.0493891	.9775606	-3.357434	3.393645
x_2	2126	.0243869	1.003398	-3.328795	4.167702
x_3	2126	-.0708252	1.021268	-3.579602	2.998349
x_4	2126	-.1914471	.9939377	-3.204001	3.135751
x_5	2126	-.0683721	.9766107	-3.224039	3.700645
y0	2126	1.147294	6.640788	-21.94452	23.39046
y1	2126	3.147294	6.640788	-19.94452	25.39046

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
epsilony	2874	.0113242	.9967888	-3.81654	3.641588
epsilonc	2874	-.6759093	.6708789	-3.918931	.7832473
x_1	2874	.043291	1.005641	-3.15619	4.076474
x_2	2874	-.0300237	.9875699	-3.379955	3.25495
x_3	2874	.033678	1.000967	-3.113853	3.630332
x_4	2874	.1354334	.9809225	-3.150622	3.915892
x_5	2874	.0735013	.9943433	-3.794222	3.449271
y0	2874	2.752551	6.710933	-20.50695	25.65636
y1	2874	4.752551	6.710933	-18.50695	27.65636

We begin by considering true program impact. This is simply the average of $Y^1 - Y^0$ across

the entire sample. Normally one does not observe true impact with real world data (if nothing else, with real world data, for any given individual in a sample one can observe Y^1 or Y^0 but not both). True program impact across the sample is 2. Against this benchmark, we first estimate program impact simply by comparing the average value of the observed outcome Y between participants and non-participants and regressing Y on P alone (with the result in Output 4.30). Both exercises suggest that an estimation of program impact that ignores the differences between participants and non-participants in the average values of the xs yields a program impact estimate of 3.605257, far above true average program impact.

STATA Output 4.30 (4.6.do)

```
. reg Y P
      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  1,  4998) =  355.83
       Model |  15883.7256     1  15883.7256           Prob > F      =  0.0000
    Residual |  223102.857  4998  44.6384268           R-squared     =  0.0665
-------------+------------------------------           Adj R-squared =  0.0663
       Total |  238986.583  4999   47.806878           Root MSE      =  6.6812

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    3.605257   .1911236    18.86   0.000      3.23057    3.979943
       _cons |    1.147294   .1449015     7.92   0.000     .8632238    1.431365
------------------------------------------------------------------------------
```

We next consider estimation of the propensity score. To do this, we perform logit regression of P on $x_1, x_2, ..., x_5$. Since our model makes clear that all observed exogenous variables x are indeed confounders, we are able to sidestep the nettlesome task of deciding which covariates to include in the specification (a process for which, as we have explained, there is no clear, accepted practice or guidance). The outcome of this regression, as well as the computation of the actual propensity score (which, in this setting, is the predicted logit probability of P conditional on $x_1, x_2, ..., x_5$) is provided in Output 4.31. The propensity score has a (slightly) higher average value among participants than non-participants, a reassuring result that suggests that the model is at least roughly correct in that it appears that those that it predicts would be more likely to participate are indeed so (one can think of this as a "poor man's c-statistic"[60]). Note, however, that there is modest failure of common support: there are values for the propensity score seen only among (at the lower range) non-participants or (at the upper range) participants.

We next consider regression of Y on P and $x_1, x_2, ..., x_5$ (in other words, on program participation *and* all of the potential confounders). The results are displayed in Output 4.32. Control for the confounders in a regression has resulted in a remarkably accurate estimate of program impact at 2.033182. We also consider regression of Y on P and the propensity score, with the results displayed in Output 4.33. Once again the program impact estimate, at 2.030544, is quite close to the truth. Next we attempt propensity score weighting estimator, with the results displayed in Output 4.34. It too yields a very good estimate of program impact at 2.042988.

We now turn to explicit propensity score matching. In STATA, this was traditionally accomplished with user-written commands, such as `psmatch2` (Leuven and Sianesi 2003) or `pscore` (Becker and Ichino 2002).[61] We report results from the new matching command under STATA's recently introduced package of treatment effect commands. We consider nearest neighbor matching

[60] We joke, of course.

[61] `nnmatch` (Abadie et al. 2004) is a user-written command that performs covariate matching in STATA.

with based on 1 match (the default) and 7 matches. The results are displayed in Output 4.35. The estimates of the average treatment effect are, at 1.91782 and 2.085387, also quite close to the true average treatment effect. Interestingly, adding more matches did not seem to change performance much in terms of accuracy (at least in the limited context of this one example).

STATA Output 4.31 (4.6.do)

```
. logit P x_1 x_2 x_3 x_4 x_5
Iteration 0:   log likelihood = -3409.5749
Iteration 1:   log likelihood = -3314.2849
Iteration 2:   log likelihood = -3314.0938
Iteration 3:   log likelihood = -3314.0938
Logistic regression                               Number of obs   =       5000
                                                  LR chi2(5)      =     190.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -3314.0938                       Pseudo R2       =     0.0280
------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         x_1 |    .0970201   .0294121     3.30   0.001     .0393735    .1546667
         x_2 |   -.0561588   .0293117    -1.92   0.055    -.1136088    .0012912
         x_3 |    .1144228   .0289201     3.96   0.000     .0577405    .1711051
         x_4 |    .3455699   .0299497    11.54   0.000     .2868697    .4042702
         x_5 |    .1566446   .0296904     5.28   0.000     .0984524    .2148368
       _cons |    .3129883   .0292129    10.71   0.000     .2557321    .3702446
------------------------------------------------------------------------------

. predict PS
(option pr assumed; Pr(P))
. by P, sort: su PS
------------------------------------------------------------------------------
-> P = 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          PS |      2126    .5531868    .0963548   .2650811   .8151766
------------------------------------------------------------------------------
-> P = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          PS |      2874    .5907881     .092297   .3191199   .8678818
```

Finally, we consider the consequences of only partial control for observed confounders. Specifically, we re-estimate the models with controls only for x_1, x_2, x_3. The estimated treatment effects are as follows:

- Regression of Y on P and x_1, x_2, x_3: 3.271049;
- Regression of Y on P and the propensity score based on x_1, x_2, x_3: 3.268763;
- Weighting estimate for propensity score based on x_1, x_2, x_3: 3.265033;
- Nearest neighbor propensity score (single) matching based on x_1, x_2, x_3: 3.141292;

These results illustrate the key challenge with selection on observables estimators: we must control for every confounder, or else they cannot be expected to yield accurate estimates of average program impact. In general, the earlier lesson that the pattern of bias is unpredictable depends heavily on what confounders are and are not controlled for, and as such is essentially impossible to gauge from a specific set of estimates involving partial control for confounders.

STATA Output 4.32 (4.6.do)

```
. reg Y P x_1 x_2 x_3 x_4 x_5
      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  6,  4993) =  784.67
       Model |  115982.991     6  19330.4985            Prob > F      =  0.0000
    Residual |  123003.592  4993  24.6352076            R-squared     =  0.4853
-------------+------------------------------           Adj R-squared =  0.4847
       Total |  238986.583  4999   47.806878            Root MSE      =  4.9634

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   2.033182    .144723    14.05   0.000     1.749461    2.316902
         x_1 |   1.435243   .0706837    20.31   0.000     1.296672    1.573814
         x_2 |  -1.615602   .0706278   -22.87   0.000    -1.754064   -1.477141
         x_3 |   1.147418   .0696116    16.48   0.000     1.010949    1.283887
         x_4 |   2.447409   .0712529    34.35   0.000     2.307722    2.587096
         x_5 |    3.03955    .071203    42.69   0.000     2.899961    3.179139
       _cons |   2.015215   .1088433    18.51   0.000     1.801834    2.228596
------------------------------------------------------------------------------
```

STATA Output 4.33 (4.6.do)

```
. reg Y P PS
      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  2,  4997) = 1603.19
       Model |  93410.4865     2  46705.2433            Prob > F      =  0.0000
    Residual |  145576.096  4997  29.1326989            R-squared     =  0.3909
-------------+------------------------------           Adj R-squared =  0.3906
       Total |  238986.583  4999   47.806878            Root MSE      =  5.3975

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   2.030544   .1573897    12.90   0.000     1.721991    2.339097
          PS |    41.8792   .8118259    51.59   0.000     40.28766    43.47074
       _cons |  -22.01972   .4640971   -47.45   0.000    -22.92956   -21.10989
------------------------------------------------------------------------------
```

STATA Output 4.34 (4.6.do)

```
. g wght=(Y*P)/PS -(Y*(1-P))/(1-PS)
. su wght
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wght |      5000    2.042988    15.21309  -92.91533   43.05652
```

We have thus seen that the various selection on observables methods all performed roughly equally well. When all confounders were indeed controlled for performance was quite good in the sense of yielding an accurate estimate of true average program impact (another way of saying this is that the selection on observables models quite effectively addressed the bias to the estimate of program impact yielded by simple comparison of outcomes between participants and non-participants). However, in some sense this was a somewhat easy challenge since the program impact was constant.

We leave as an exercise for the reader expanding the example in STATA do-file.

STATA Output 4.35 (4.6.do)

```
. teffects psmatch (Y) (P x_1 x_2 x_3 x_4 x_5)
Treatment-effects estimation                    Number of obs      =      5000
Estimator      : propensity-score matching      Matches: requested =         1
Outcome model  : matching                                   min =         1
Treatment model: logit                                      max =         1
------------------------------------------------------------------------------
             |               AI Robust
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           P |
   (1 vs 0)  |    1.91782   .1725876    11.11   0.000     1.579554    2.256085
------------------------------------------------------------------------------

. teffects psmatch (Y) (P x_1 x_2 x_3 x_4 x_5), nn(7)
Treatment-effects estimation                    Number of obs      =      5000
Estimator      : propensity-score matching      Matches: requested =         7
Outcome model  : matching                                   min =         7
Treatment model: logit                                      max =         7
------------------------------------------------------------------------------
             |               AI Robust
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           P |
   (1 vs 0)  |   2.085387   .1478581    14.10   0.000      1.79559    2.375184
------------------------------------------------------------------------------
```

4.3 Some Closing Thoughts

This chapter has considered models that assume only selection on observables into program participation status. In other words, they assume that we can "measure all that matters" in terms of characteristics that shape both the outcome of interest and the participation decision. It should be immediately obvious that this is an incredibly strong assumption. Indeed, in many instances it is simply not credible.

The discussion was still important, however, for two reasons. First, the selection on observables models provide a framework through which one can quite usefully approach many of the other quasi-experimental estimators considered in this manual. Put simply, without a solid foundation in the mechanics, properties and implications of the selection on observables models, it is difficult to develop a solid grasp of the other quasi-experimental methods we consider. Second, as a practical matter selection on observables approaches are used in real-world impact evaluations: in some instances their core assumption is more palatable, while in other cases there is simply no other practical option than to pursue a selection on observables estimation strategy.

A implicit theme of this chapter has been a kind of skepticism about the importance of the differences between the alternative selection on observables models. This is a point that can be easily pushed too far. Indeed, there are abundant examples in the impact evaluation literature where different selection on observables estimators have yielded meaningful differences in impact estimates, performed very differently in terms of things like sampling variation, etc.

Nonetheless, the authors remain struck more by the central importance of the assumption common to all selection on observables models than the specific ways in which they approach estimation

differently. Our own gross casual empiricism has suggested to us that often the differences between the estimates generated by the various selection on observables estimators are probably of second order importance compared with how the estimates generated by these models differ either from those that ignore selection altogether (such as simple comparison of average outcomes between participants and non-participants) or those that rely on alternative quasi-experimental identification strategies. We take this opportunity, however, to concede that there are many very smart people who would probably vehemently disagree with this rather breezy assessment.

Chapter 5

Within Estimators

We now turn our attention to a broad class of quasi-experimental estimators often referred to as "within" estimators. These estimators have essentially two common characteristics. First, they typically assume that for a given unit of observation (e.g. the individual) the source of bias with simple estimation of program impact (for instance, by straightforward comparison of average outcomes between participants and non-participants or simple regression of the outcome on an indicator of program participation) does not vary for that unit of observation. Consider, for instance, the possibility for variation over time. Some individual characteristics, such as income, might vary over time. Others, such as genetic endowment, are essentially fixed for a given individual and hence cannot vary over time for them. These models assume that the unobserved variable driving bias in simple estimation of program impact is so fixed.

The other major feature of within models is that they rely fundamentally on variation *within* units of observation, as opposed to variation *across* units of observation (for instance at a point in time). For instance, these models might identify program impact by considering how variation in program participation over time for individuals is associated with variation over time in the outcome of interest for those individuals. This is as opposed to what we might refer to as "cross sectional" identification, whereby program impact is identified by considering how variation in program participation across units of observation (e.g. individuals) is associated with variation in outcome levels across those units of observation.

Cross sectional identification hence involves asking whether those who participate in a program have different average outcomes levels from those who do not do so. Within estimation asks whether changes in participation status appear to be associated with changes in outcome levels.

These models are often referred to as "fixed effects" estimators. A motivation for this terminology is the recognition that the potential confounding unobservable is fixed for the unit of observation (e.g. family background is fixed over time for the individual, genetic endowment is fixed across pairs of twins, etc.). However, there are other models that typically carry different labels but share this basic characteristic with the so-called fixed effects estimators. "Within" estimators are hence a broader class than "fixed effects" estimators.

5.1 Classic Models

5.1.1 The First Differences Estimator

We begin our discussion with an estimator that many would likely not consider a classic fixed-effects estimator: the first differences estimator. We start with this estimator primarily because it provides a particularly clear and convenient framework for examining how within models estimate

program impact. There is a heavy degree of commonality to the fixed effects and first differences approaches: the two estimators rely on an identical assumption about the source of bias from straightforward estimation of program impact evaluation, offer very similar remedies and carry similar data requirements.

Let us start, as has become our custom, with a model of potential outcomes and the costs of participation. Although within estimators can be operationalized in a number of senses,[1] the usual conceptualization of the variation is with respect to *time*. We will focus on variation over time for individuals from a sample of individuals observed at more than one point in time. We therefore need to introduce a time dimension to our variables. Thus, for instance, the outcome of interest is Y_{it}, signifying that this is the outcome of interest observed for individual i at time t.

With this modification in place, we consider the following potential outcome and cost equations:

$$Y_{it}^0 = \beta_0 + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

$$Y_{it}^1 = \beta_0 + \beta_1 + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

where x_{1it} is a time-varying observed individual characteristic, x_{2i} is a fixed observed individual characteristic, μ_i is a fixed unobserved individual characteristic and ϵ_{it}^Y is time-varying unobservable. Note that we have eschewed certain features considered in the last chapter. For instance, the observables do not necessarily influence program impact. The within-framework can accommodate possibilities such as this, but it is not essential to the core points that we need to establish in this chapter.

The cost of participation is given by

$$C_{it} = \gamma_0 + \gamma_1 \cdot x_{1it} + \gamma_2 \cdot x_{2i} + \gamma_3 \cdot \mu_i + \epsilon_{it}^C$$

At any given time period t individual i choose to participate (i.e. P_{it} will equal 1) if

$$Y_{it}^1 - Y_{it}^0 - C_{it} > 0$$

or

$$\beta_1 - \gamma_0 - \gamma_1 \cdot x_{1it} - \gamma_2 \cdot x_{2i} - \gamma_3 \cdot \mu_i - \epsilon_{it}^C > 0$$

There are a few important features to note about this. First, the observed individual characteristics x_{1it} and x_{2i} influence participation and hence we would expect that, other things being equal, the average values of these characteristics should differ between participants and non-participants. There will thus clearly be selection into program participation based on observables. Second, the same is true of the fixed unobservable μ_i: it influences the participation decision and hence would be expected to differ in terms of average value between the participants. Third, by dint of the fact that they are fixed, fixed characteristics such as x_2 and μ cannot explain variation in program participation *over time* for the i^{th} individual. They cannot explain variation in outcomes or participation *over time* for an individual.

Let us now derive a regression specification. The observed outcome is

$$Y_{it} = P_{it} \cdot Y_{it}^1 + (1 - P_{it}) \cdot Y_{it}^0$$

$$= P_{it} \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y\right)$$

$$+ (1 - P_{it}) \cdot \left(\beta_0 + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y\right)$$

[1] For instance, one might consider variation within a household, community, etc.

$$= \beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

It is already clear that unbiased estimation of program impact (i.e. unbiased estimation of β_1) might be a bit tricky. Unbiased estimation would require the independence or mean independence of program participation P_{it} and μ_i and the independence or mean independence of P_{it} and ϵ_{it}^Y. Per the latter condition, since ϵ_{it}^Y played no role in determining program participation, the only readily plausible way that the independence or mean independence of P_{it} and ϵ_{it}^Y could be violated if there was some sort of relationship between ϵ_{it}^Y and ϵ_{it}^C. We rule out this possibility by assumption.

The sticky wicket is then presented by μ_i. It clearly plays a role in shaping the program participation decision and hence should be correlated with P_{it}. Thus, to recover an unbiased estimate of β_1, μ_i must either be controlled for or somehow purged from the regression specification that we actually estimate. Since μ_i is unobserved, explicitly controlling for it is clearly not an option.

Within estimators take the alternative route: they purge μ from the regression equation that we will actually estimate. Suppose that we observed a sample of N individuals at two points in time, $t = 1$ and $t = 2$. Specifically, suppose that we observed $\{Y_{it}, P_{it}, x_{1it}, x_{2i}\}$ for $i = 1, ..., N$ individuals at times $t = 1, 2$.

To fix ideas, let us focus on regression modelling for the i^{th} individual out of the N individuals in our sample. Fixed effects would essentially "subtract" (or, more precisely, "difference") the regression models from the two time periods for that i^{th} individual in the sample:

$$\begin{array}{rccccccccccccc}
 & Y_{i2} & = & \beta_0 & + & \beta_1 \cdot P_{i2} & + & \beta_2 \cdot x_{1i2} & + & \beta_3 \cdot x_{2i} & + & \beta_4 \cdot \mu_i & + & \epsilon_{i2}^Y \\
- & Y_{i1} & = & \beta_0 & + & \beta_1 \cdot P_{i1} & + & \beta_2 \cdot x_{1i1} & + & \beta_3 \cdot x_{2i} & + & \beta_4 \cdot \mu_i & + & \epsilon_{i1}^Y \\
\hline
 & \Delta Y_i & = & 0 & + & \beta_1 \cdot \Delta P_i & + & \beta_2 \cdot \Delta x_{1i} & + & \beta_3 \cdot 0 & + & 0 & + & \Delta \epsilon_i^Y
\end{array}$$

yielding a final first difference specification to be estimated of

$$\Delta Y_i = \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta x_{1i} + \Delta \epsilon_i^Y$$

where "Δ" indicates "change in".[2] In the context of this new regression model, we form a new dependent variable, which is the change in the value of Y between time periods $t = 1$ and $t = 2$ for individual i, or ΔY_i. For concreteness we have focused on the i^{th} of the N individuals, but first differencing would involve this subtraction exercise for all N individuals in the sample, yielding the final first differencing specification, which would be estimated by regressing ΔY on ΔP and Δx_1 using the information on the N individuals in our sample.

There are several really important things to recognize about this new regression specification. First, and by far most importantly, the troublesome unobserved confounder μ_i does not appear in the new specification. It played a role *across time* in determining the *average level* of Y for individual i, but because it was fixed over time really could not explain variation in Y *over time* for that individual. For instance, it might have explained why Y was persistently high over time for that individual, but could not explain why Y changed for that person from time period 0 to 1.

Within estimators sweep out potentially unobserved confounders that do not vary over time, shifting focus fundamentally from identification of the determinants of levels (i.e. identifying program impact as the statistical relationship between program participation and levels of the outcome of interest *across* individuals) to identification of the determinants of changes (i.e. how, across individuals, changes in program participation status appear to be associated with changes in the outcome of interest). The differencing exercise has created a new regression specification that focuses on variation "within" each individual. By doing so, it has swept out the influence of all "fixed effects": factors shaping the outcome that do not vary for the individual (for instance across time).

[2]Therefore, for any given variable W, ΔW means "change in W".

It has thus removed or purged the influence of any confounding unobservable that does not vary for the individual over time. Of course, it has also removed or purged the influence of any *observed* characteristic that does not vary over time.[3]

Notice that the time-varying unobservable ϵ^Y has <u>not</u> been purged out of the regression specification. This would be extremely problematic if ϵ^Y had been correlated with ϵ^C, leading to an extremely important caveat about within models such as the first differenced regression: they do not address any endogeneity bias associated with time-varying unobservables (i.e. time-varying unobserved characteristics). More broadly (since these models do not always involve exploiting change over time[4]) the unobserved confounder must be fixed within the units of observation for which the subtraction exercise above is performed. Any remaining error term that varies within the unit of observation (e.g. as ϵ^Y_{it} varies over time as captured by $\Delta\epsilon^Y_i$) must be independent of ΔP_i. This is the critical identifying assumption of within models: any unobserved confounder (i.e. an unobserved variable not independent of participation) associated with the observed regressor of interest (in our case program participation P) must be fixed across the units of observation (e.g. fixed across time for the individual).

If ϵ^Y and P are independent (or at least mean independent) then so are $\Delta\epsilon^Y$ and ΔP. Estimation of β_1 by least squares regression of ΔY on ΔP and Δx_1 (more generally, regression of the change in the outcome of interest on the change in program participation and observed, time-varying characteristics) will then yield an unbiased estimate $\hat{\beta}_1$ of β_1.[5] If, however, ΔP and $\Delta\epsilon^Y$ were somehow correlated then the within estimate of $\hat{\beta}_1$ would be biased and inconsistent. Within approaches such as first differencing thus involve the potentially strong assumption that any confounding unobserved characteristic is fixed (i.e. does not vary) across the observations for a given unit of observation.

Another major feature of within models already evident from this specification is a loss of information for estimation. For the hypothetical i^{th} individual that we observe at $t = 1$ and $t = 2$, we had two observations (one from time period $t = 1$ and one from time period $t = 2$) for that individual to contribute to the estimation of

$$Y_{it} = \beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon^Y_{it}$$

However, the act of differencing the regression specification between time period $t = 2$ and $t = 1$ required two observations but yielded only one observation to contribute toward estimating the parameters of

$$\Delta Y_i = \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta x_{1i} + \Delta\epsilon^Y_i$$

In general, estimation of within models involves a loss of observations toward estimation equal to the number of units of observations.

Suppose we observe 10,000 individuals for several time periods. Pursuit of a within estimation strategy such as first differencing involves the loss of 10,000 observations. Suppose, for instance, that we observe Y_{it} and P_{it} for 10,000 individuals at three points in time, $t = 1, 2, 3$. For the

[3]Some have faulted within estimators for removing the influence of fixed observed covariates. The argument is typically that they preclude consideration of the role of these fixed observed variables in shaping outcomes. While this can be a legitimate concern, we would emphasize that the ultimate goal in this setting is evaluation of program impact. Therefore the unbiased estimation of program impact, rather than developing a broader understanding of the overall determinants of Y, should be the priority.

[4]For instance, one could imagine that t indexed two members of family i, for which μ_i was an unobserved family-level characteristic that commonly influenced the program participation decisions and outcomes of the members of household i.

[5]We leave the proof of this as an exercise for the reader. It involves a trivial extension of the derivation of the conditions for unbiasedness introduced in the last chapter.

purposes of simply regressing Y on P we have at our disposal 30,000 observations. However, for regressing ΔY on ΔP we have only 20,000 observations. The reason is that we can difference twice, once by differencing Y and P between $t = 3$ and $t = 2$ and again by differencing the same between $t = 2$ and $t = 1$. So while Y and P are observed three times for each individual, ΔY and ΔP are observed only twice for each individual.

The loss of information associated with within estimation is a frequent criticism of the approach. One must recognize a kind of trade-off to this business. Adopting the within approach purges the effect of fixed confounding unobserved characteristics. To the extent that the central assumption of the within approach (i.e. that those fixed unobservables were the source of endogeneity bias to the estimate of program impact from simply regressing Y on P) is correct this strategy yields an unbiased estimate of program impact where one would not obtain one otherwise. However, it does so at the cost of a lower number of observations and hence a larger variance to the estimate of program impact, other things being equal. It is possible that within estimators can deliver estimates that, while unbiased (assuming the key assumption supporting the fixed effects estimator holds), are much further from true program impact than what one might obtain simply by regressing Y on P with the full set of observations.

Our first difference specification has been motivated by the circumstance where one observes data on a sample of N individuals for two time periods ($t = 1$ and $t = 2$). We derived a basic first difference regression specification using as a specific motivating example the i^{th} individual out of the N persons in our sample. Thus, the unit of observation in this motivating example is the individual, and the within variation utilized by the first difference regression model is the change over time in the outcome of interest (ΔY), program participation (ΔP) and in any time-varying observed characteristic of the individual (e.g. Δx_1).

Since within estimation such as first differencing relies on variation over time at the unit of observation (e.g. the individual) it is necessary to observe the outcome of interest, program participation and key time-varying characteristics for those units of observations at more than one point in time. Such a data framework, wherein units of observation are observed more than once, is variously referred to as **longitudinal**, **repeated measures** and **panel data**.

Specifically, suppose that we have a sample of individuals who were selected for a baseline round of data collection and then re-interviewed at various intervals over time. The various time-period specific samples are usually referred to as panels. When they are the same size for each time period (i.e. each individual is interviewed in each round of data collection), the samples from the different time periods are referred to as **balanced panels**.[6]

This is not the only way that within variation could have been conceptualized. For instance, this same basic data structure (i.e. a unit of observation observed twice) might have involved a sample of N pairs of monozygotic twins. First differencing (within pairs of twins) would then sweep out any unobserved confounding variable common to the two twins (such as genetic endowment).

In general, within estimation is a data intensive strategy in that it requires more than one observation per unit of observation. Without this it is impossible to construct variables (e.g.

[6]Sometimes one of these three terms (longitudinal, repeated measures, panel data) would seem to fit more naturally a given sample design. This is not necessarily important from a substantive estimation standpoint, but is still useful for thinking precisely about alternative data designs. The designs can certainly come in various forms. For instance, for a sample of adult women, their fertility and family planning participation over several years could, at two extremes, be captured either by interviewing them annually in each of those years or retrospectively asking them about their fertility and program participation at the conclusion of that time frame. All three terms fit the first design well, while the latter design is perhaps less usefully described as panel because a panel of women was not followed over time for it. For the purpose of estimating many models of program impact, there is no difference in principle between the two designs. In practice, however, issues such as attrition in the panel approach or recall bias under the recall approach may dictate a preferred data design.

ΔY) that somehow reflect within (i.e. within the unit of observation) variation. Typically, this extensive data requirement is from a practical standpoint the main impediment to implementing this approach to program impact estimation. Many datasets involve just one observation per any meaningful conceptualization of the unit of observation.[7]

A less than desirable feature of our particular regression specification,

$$\Delta Y_i = \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta x_{1i} + \Delta \epsilon_i^Y$$

is that it lacks a constant term. Were one to estimate this model (by regressing ΔY on ΔP and Δx_{1i} with no constant term) there are all sorts of complications.[8]

Fortunately a natural extension of the model generates a constant.[9] Consider this simple extension of the potential outcomes framework:

$$Y_{it}^0 = \gamma + \beta_0 \cdot t + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

$$Y_{it}^1 = \gamma + \beta_0 \cdot t + \beta_1 + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

The main differences between these potential outcome equations and those introduced at the outset of this section is that the constant from the original potential outcome framework is now represented by γ while the term $\beta_0 \cdot t$ has been introduced. $\beta_0 \cdot t$ captures a possible time trend to the potential outcomes.

This modified potential outcomes framework leads to a new regression specification. We proceed as we have in the past:

$$\begin{aligned} Y_{it} &= P_{it} \cdot Y_{it}^1 + (1 - P_{it}) \cdot Y_{it}^0 \\ &= P_{it} \cdot \left(\gamma + \beta_0 \cdot t + \beta_1 + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y\right) \\ &+ (1 - P_{it}) \cdot \left(\gamma + \beta_0 \cdot t + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y\right) \\ &= \gamma + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y \end{aligned}$$

The time trend to the potential outcomes thus translates into a time trend term in the regression specification motivated by the potential outcome equations.

Suppose again (for simplicity) that there are only two time periods, $t = 1$ and $t = 2$. Subtracting the regression terms at $t = 1$ from those at $t = 2$ (i.e. taking the first difference between the regression terms in the two time periods), we have

$$\begin{array}{rccccccccccccccc} & Y_{i2} & = & \gamma & + & \beta_0 \cdot 2 & + & \beta_1 \cdot P_{i2} & + & \beta_2 \cdot x_{1i2} & + & \beta_3 \cdot x_{2i} & + & \beta_4 \cdot \mu_i & + & \epsilon_{i2}^Y \\ - & Y_{i1} & = & \gamma & + & \beta_0 \cdot 1 & + & \beta_1 \cdot P_{i1} & + & \beta_2 \cdot x_{1i1} & + & \beta_3 \cdot x_{2i} & + & \beta_4 \cdot \mu_i & + & \epsilon_{i1}^Y \\ \hline & \Delta Y_i & = & 0 & + & \beta_0 & + & \beta_1 \cdot \Delta P_i & + & \beta_2 \cdot \Delta x_{1i} & + & 0 & + & 0 & + & \Delta \epsilon_i^Y \end{array}$$

yielding a final specification to be estimated of

$$\Delta Y_i = \beta_0 + \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta x_{1i} + \Delta \epsilon_i^Y$$

Thus, a fairly trivial (and frankly behaviorally reasonable[10]) extension of the potential outcomes specification yields a first difference regression specification with a constant term.

[7] Although within variation can sometimes be captured by recall questions such as fertility histories.

[8] For instance, measures of the explanatory power of the model are considerably complicated.

[9] We did not introduce this extension at the outset to avoid sowing confusion by trying to illustrate too many things at once.

[10] The time trend can be thought of as the cumulative or overall net effect of time-varying determinants of the potential outcome that are subject to some sort of trend.

In general it is a good idea to include a constant term in a within regression estimation.[11] The reason is fairly simple: a time trend could represent a kind of omitted variable. For instance, if there is a trend to y and x, when one regresses y on x alone x will serve two empirical roles: as control for itself and as a proxy for the role of time in shaping y.

At this point we pause to consider a numerical example. This example is found in STATA do-file 5.1.do. The departure point is the following model of potential outcomes and costs of participation for 5,000 individuals observed at two points in time ($t = 0$ and $t = 1$):

$$Y_{it}^0 = 2 + 1.5 \cdot x_{1it} + 2 \cdot x_{2i} + 3 \cdot t + \mu_i + \epsilon_{it}^Y$$

$$Y_{it}^1 = 4 + 1.5 \cdot x_{1it} + 2 \cdot x_{2i} + 3 \cdot t + \mu_i + \epsilon_{it}^Y$$

$$C_{it} = 1 + 1.5 \cdot x_{1it} - 1 \cdot x_{2i} + \mu_i + \epsilon_{it}^C$$

where the xs and μ are independently normally distributed with mean 0 and variance 4 (i.e. x, $\mu \sim$ N(0,4), ruling out any endogeneity to x) and the ϵs are independently normally distributed with mean 0 and variance 9 (i.e. $\epsilon \sim$ N(0,9), ruling out any kind of relationship between ϵ^Y and ϵ^C).[12]

There are a few important things to note about this setup. First, since $Y_{it}^1 - Y_{it}^0 = 2$, the cost equation is essentially determining the role that the xs and μ play in shaping the participation decision. Second, the fact that $Y^1 - Y^0 = 2$ for every individual in the sample at each of the two points in time means that true average program impact is 2. Third, there is a time-varying individual-level observed characteristic (x_{1it}) and a fixed (with respect to time) individual-level observed characteristic (x_{2i}). Fourth, there is a fixed individual characteristic (μ_i). Finally, there is a time trend to the potential outcome equation (thus creating scope for a constant term to the eventual first differenced regression specification).

We begin with some summary statistics, displayed in Outputs 5.1, 5.2 and 5.3. Output 5.1 presents participation patterns. 57.41 percent of the sample participated in the program. Given the focus of fixed effects models on *within* variation (in this case such variation would be over time for each individual), we also consider the degree to which the participation decision changed over time. To do this, we create the variable `avgP`, which is the average of the program participation variable `P` for each individual across time periods $t = 0$ and $t = 1$. If the individual never participated, `avgP` =0 since $P_{i0} = P_{i1} = 0$. Similarly, `avgP` =1 if the participated in both, time periods. However, if they participated in one time period but not the other (i.e. if they participated half of the time), then `avgP`=.5. We can see from Output 5.1 that 38.5 percent of the individuals in the sample changed their participation status between time periods $t = 0$ and $t = 1$, while 23.34 percent never participated and 38.16 participated throughout the two time periods.

Turning to Output 5.2, we consider the averages of the key variables by participation status. First, and perhaps most importantly, notice that the fixed unobserved characteristic μ has a very different average between participants and non-participants (-0.5097 against 0.7596, respectively). Participants and non-participants thus differ by a factor that influences outcomes and would not be

[11] STATA and other commercial statistical analysis packages typically estimate a regression constant by default. In STATA, suppression of the estimation of the constant in ordinary least squares regression via the `regress` command is achieved with the `noconstant` option, as in

```
reg y x, nocon
```

Since the default is typically to include the constant, the analyst usually does not need to do anything explicit to insure its estimation.

[12] As usual, the parameters of this simulation are chosen fairly randomly. We encourage the reader to experiment with the code, changing parameters.

observable were this real-world data.[13] They also differ by their average values for the observables x, which is not a big surprise given the role that these play in systematically influencing the cost of participation, the main driver of the participation decision (unsurprisingly, non-participants are those facing higher costs of participation).

STATA Output 5.1 (5.1.do)

```
. * Basic summary statistics: participation
. tab P
```

P	Freq.	Percent	Cum.
0	4,259	42.59	42.59
1	5,741	57.41	100.00
Total	10,000	100.00	

```
. ta avgP if t==0
```

avgP	Freq.	Percent	Cum.
0	1,167	23.34	23.34
.5	1,925	38.50	61.84
1	1,908	38.16	100.00
Total	5,000	100.00	

In Output 5.3 we consider potential correlations among the unobservables and between any of them and program participation P. Unsurprisingly, P is highly correlated with μ and ϵ^C: these are two direct determinants of the cost of participation. It is essentially uncorrelated with ϵ^Y, which should be the case since ϵ^Y cancels out of $Y^1 - Y^0$ and thus plays no role in the program participation decision. At 0.0155, ϵ^Y is essentially uncorrelated with ϵ^C (and even that figure is just an artifact of the comparatively small sample size and would tend toward 0 were one to increase the sample size).

We now turn to regression estimation of program impact. Specifically, we consider regressing Y on P, the observed characteristics (x_1 and x_2) and time t. Analysis of the correlations between program participation P and the unobservables suggests that one avenue (correlation between ϵ^Y and P) for endogeneity bias to the regression estimate of program impact is not much of a concern. However, another (a correlation between μ and P) might be a real issue. We thus find ourselves in essentially exactly the circumstance for which fixed effects regression was designed: a potential confounding unobservable fixed across the observations for the unit of observation. In Output 5.4 we report results from regressing Y on P, t and the observables (x_1 and x_2) across the entire sample of 10,000 person years.

Plainly, the estimate of program impact ($\hat{\beta}_1$ =.1176989) is only a small fraction of the true impact value of 2. The estimate of β_1, $\hat{\beta}_1$, obtained from this regression exercise is biased. As we explained in Chapter 2 and reminded the reader in Chapter 4, results such as those in Output 5.4 are not proof of bias. Bias is the condition where $\hat{\beta}_1$ is systematically off in the sense that it is "wrong on average" or

$$E\left(\hat{\beta}_1\right) \neq \beta_1$$

To establish biasedness in the context of this numerical example, we would thus need to simulate

[13] We assume that μ is unobservable. Technically, we can observe it (because we simulated the data, including the observations for μ) but we assume that it would be unobservable were this a real world sample. It represents the fixed unobserved factors not observable in real world data.

many samples in the same fashion, estimate program impact via regression for each, and average the program impact estimates across the many simulated samples. If that average did not equal β_1 (or 2, in our example) this would be evidence of bias. Nonetheless, the specific estimate in our example, which is well wide of the mark in terms of true program impact, is indicative of endogeneity bias.

STATA Output 5.2 (5.1.do)

```
. * Basic summary statistics: variable means
.
. by P, sort: summarize Y y1 y0 c x* mu epsilon*
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	4259	4.334862	6.289426	-19.78382	27.70201
y1	4259	6.334862	6.289426	-17.78382	29.70201
y0	4259	4.334862	6.289426	-19.78382	27.70201
c	4259	5.799442	2.931449	2.000205	18.66917
x_0	4259	1	0	1	1
x_2	4259	-.7789652	1.918122	-7.832516	6.227447
x_1	4259	1.093233	1.778053	-6.753155	8.420812
mu	4259	.7596287	1.869981	-6.420361	7.176875
epsilony	4259	-.0218311	3.000047	-9.903225	9.97731
epsilonc	4259	1.620998	2.671566	-7.540022	11.37924

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	5741	4.753798	6.330902	-17.66687	26.38768
y1	5741	4.753798	6.330902	-17.66687	26.38768
y0	5741	2.753798	6.330902	-19.66687	24.38768
c	5741	-2.42331	3.206232	-18.52968	1.998094
x_0	5741	1	0	1	1
x_2	5741	.4933008	1.885407	-6.450185	6.910412
x_1	5741	-.7917423	1.756657	-8.078415	5.318276
mu	5741	-.5096796	1.91944	-8.141914	6.990262
epsilony	5741	-.0242761	2.978858	-10.44104	10.15385
epsilonc	5741	-1.232716	2.684475	-11.58986	7.975534

The question is the source of the bias. One possibility with which we quickly dispense is the possibility that it is somehow a result of the longitudinal nature of the data. This could be plausible in the event that time t was somehow a "bad control". This could emerge if program participation was somehow correlated with time and time was somehow correlated with a confounding unobservable. However, the framework for our example does not introduce either possibility. Moreover, the program impact estimates generated when Y is regressed on P, x_1 and x_2 for each time period separately are also poor. For instance, Output 5.5 provides the regression for just the subsample at time period $t = 0$.

Given that there is no correlation between program participation P and the time-varying component of the unobservable, ϵ^Y, the only plausible avenue for omitted variable bias is the correlation between P and μ. In other words, endogeneity bias is driven by a fixed (with respect to time) unobserved characteristic. The framework thus indeed reflects exactly the circumstance required for first differencing (and, more generally, within approaches) to yield an unbiased estimate of program impact.

STATA Output 5.3 (5.1.do)

```
. * Correlations among unobservables
.
. corr P mu epsilony epsilonc
(obs=10000)
```

	P	mu	epsilony	epsilonc
P	1.0000			
mu	-0.3139	1.0000		
epsilony	-0.0004	-0.0209	1.0000	
epsilonc	-0.4661	-0.0060	0.0155	1.0000

STATA Output 5.4 (5.1.do)

```
. * Cross sectional regression
. reg Y P x_1 x_2 t
```

Source	SS	df	MS		
Model	278115.263	4	69528.8157		
Residual	120808.106	9995	12.086854		
Total	398923.369	9999	39.8963265		

```
Number of obs =   10000
F(  4,  9995) = 5752.43
Prob > F      =  0.0000
R-squared     =  0.6972
Adj R-squared =  0.6970
Root MSE      =  3.4766
```

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.1176989	.0851456	1.38	0.167	-.0492036	.2846013
x_1	1.290954	.020018	64.49	0.000	1.251715	1.330193
x_2	2.170245	.0186083	116.63	0.000	2.133769	2.206721
t	3.010955	.0695422	43.30	0.000	2.874638	3.147272
_cons	3.093416	.0700131	44.18	0.000	2.956176	3.230656

STATA Output 5.5 (5.1.do)

```
. * Cross sectional regression, first time period only
. reg Y P x_1 x_2 if t==0
```

Source	SS	df	MS		
Model	128127.316	3	42709.1054		
Residual	60360.6217	4996	12.0817898		
Total	188487.938	4999	37.7051286		

```
Number of obs =    5000
F(  3,  4996) = 3535.00
Prob > F      =  0.0000
R-squared     =  0.6798
Adj R-squared =  0.6796
Root MSE      =  3.4759
```

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.2413616	.1207878	2.00	0.046	.0045645	.4781587
x_1	1.310641	.0287164	45.64	0.000	1.254345	1.366938
x_2	2.148506	.0262096	81.97	0.000	2.097124	2.199888
_cons	3.020308	.0861149	35.07	0.000	2.851485	3.189131

To explore this alternative, we compute

$$\Delta Y_i = Y_{i1} - Y_{i0}$$

$$\Delta P_i = P_{i1} - P_{i0}$$

and

$$\Delta x_{1i} = x_{1i1} - x_{1i0}$$

to estimate the fixed-effects specification

$$\Delta Y_i = \beta_0 + \beta_1 \cdot \Delta P_i + \beta_2 \cdot \Delta x_{1i} + \Delta \epsilon_i^Y$$

This is thus the specification that purges the confounding unobservable μ, as well as the fixed (with respect to time) observable x_2.

Results for the regression of ΔY on ΔP and Δx_1 are presented in Output 5.6. The first differenced specification has yielded the program impact estimate 1.841112. This is in the near neighborhood of the true value of 2. We can thus see that the differencing estimate is far closer to the truth than the estimate yielded by straightforward "cross-sectional" regression of Y on P, x_1 and x_2 with the full set of 10,000 individual/time observations or the 5,000 observations from either time period.

STATA Output 5.6 (5.1.do)

```
. * The basic first differenced regression
. reg deltaY deltaP deltax_1 if tt<`T´
      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  2,  4997) = 1875.21
       Model |  67089.3154     2  33544.6577            Prob > F      =  0.0000
    Residual |  89388.6453  4997  17.8884621            R-squared     =  0.4287
-------------+------------------------------           Adj R-squared =  0.4285
       Total |  156477.961  4999  31.3018525            Root MSE      =  4.2295

------------------------------------------------------------------------------
      deltaY |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      deltaP |   1.841112   .1131854    16.27   0.000     1.619219    2.063005
    deltax_1 |   1.469093   .0249797    58.81   0.000     1.420122    1.518065
       _cons |   3.031345   .0598335    50.66   0.000     2.914045    3.148645
------------------------------------------------------------------------------
```

The results in Output 5.6 are based on the manual calculation of the first differences. However, STATA has a command, `xtivreg` with the `fd`, option, that automates the calculation of the first difference model. Estimation results using this command are provided in Output 5.7. This command might be a preferred route for users uncomfortable with computing first differences (which is quite understandable with more complex datasets and many time periods). One complication of the syntax for this command is driven by the fact that it was written with an entirely different estimation method (instrumental variables, which will be discussed in the next chapter) in mind. That is the reason for the seemingly strange line of code `(t=z)`. This is necessary because the command requires the structure of instrumental variables (which will become clearer in the next chapter). To satisfy this requirement, simply choose a variable that is unnecessary to the regression and, in parentheses, set it equal to a variable set to 1 (i.e. z was generated with the code `generate z=1`). We chose t for this purpose because, as we have seen, it is captured by the constant term in any case.

STATA Output 5.7 (5.1.do)

```
. xtset ID t
       panel variable:  ID (strongly balanced)
        time variable:  t, 0 to 1
                delta:  1 unit
. xtivreg Y P x_1 (t=z), fd reg
note: z omitted because of collinearity
First-differenced IV regression
Group variable:    ID                            Number of obs      =      5000
Time variable:     t                             Number of groups   =      5000
R-sq:  within  =      .                          Obs per group: min =         1
       between = 0.2222                                        avg =       1.0
       overall = 0.2222                                        max =         1
                                                  Wald chi2(2)       =   3750.42
corr(u_i, Xb)  = 0.0426                           Prob > chi2        =    0.0000
```

D.Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t						
D1.	0	(omitted)				
P						
D1.	1.841112	.1131854	16.27	0.000	1.619273	2.062951
x_1						
D1.	1.469093	.0249797	58.81	0.000	1.420134	1.518053
_cons	3.031345	.0598335	50.66	0.000	2.914073	3.148616
sigma_u	5.4190635					
sigma_e	4.2294754					
rho	.62144596	(fraction of variance due to u_i)				

```
Instrumented:   t
Instruments:    P x_1 t
```

The extension to more than two time periods is fairly trivial. For instance, with 5 time periods, 4 differences can be calculated (the difference in variables values between the 1st and 2nd, 2nd and 3rd, 3rd and 4th and, finally, 4th and 5th time periods). In other words, where there were 5 observations per individual for the cross-sectional regression, there are 4 for the first-differenced regression. Thus, the loss of observations evident in the two time period case extends to the many time period case: one time period's observation is lost for each individual.

5.1.2 Linear Fixed Effects

Thus far, we have discussed the first differencing approach to dealing with fixed unobservables. This has involved essentially taking the pairwise difference of the key variables (the outcome Y, program participation P and any time-varying observables x_1) over the two time periods in our example, though the extension to many time periods is trivial. We started with the first differencing approach mainly for expositional simplicity: it is perhaps the most intuitively straightforward way to demonstrate how the within approach flushes out fixed confounders such as μ. Moreover, most of the implications of the first differencing approach generalize to alternative within approaches.

In this subsection we discuss two classic linear "fixed effects"[14] estimation approaches: "de-

[14]The label "fixed effects" is probably popular as much as anything to differentiate so called estimators from first

meaning" and use of dummy variables to control for unobserved fixed characteristics.[15] By "linear" we mean that the estimator is typically applied in the context of linear regression.[16]

The first alternative we consider is often referred to as demeaning. Put simply, to demean the key covariates (the outcome Y, program participation P and time-varying covariates x_1) each observation for each of the key covariates has the mean of that variable for the unit of observation subtracted from it.

To fix ideas, let us return to the problematic (from the standpoint of obtaining an unbiased estimate of program impact β_1) basic cross-sectional specification

$$Y_{it} = \beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

Define $\overline{Y_i}$ as follows:

$$\overline{Y_i} = \sum_{t=1}^{T} \frac{Y_{it}}{T}$$

where T is the number of time periods over which individual i is observed. In other words, $\overline{Y}_i$ is the mean value of the outcome Y for individual i across the T time periods over which that individual is observed. For simplicity, assumed that all $i = 1, ..., N$ individuals in our hypothetical sample are observed for each of time period captured in the sample. In other words, assume that we have balanced panels in our sample.

$\overline{Y_i}$ is simply the average of Y for the i^{th} individual in the sample. For instance, in the setting of two time periods considered in the last subsection, $t = 1, 2$, $\overline{Y_i}$ is

$$\overline{Y_i} = \frac{Y_{i1} + Y_{i2}}{2}$$

We can similarly imagine averages for all of the right hand side variables in the specification: $\overline{P_i}$, $\overline{x_{1i}}$, $\overline{x_{2i}}$, $\overline{\mu_i}$, and $\overline{\epsilon_i}^Y$. Because x_2 and μ do not vary for each individual, note that

$$\overline{x_{2i}} = x_{2i}$$

and

$$\overline{\mu_i} = \mu_i$$

The remaining variables do vary over time for the individual, and hence will not equal their averages.

Let us now consider the basic regression specification, but applied to the de-meaned variables:[17]

$$Y_{it} - \overline{Y_i} = \beta_0 \cdot (1-1) + \beta_1 \cdot \left(P_{it} - \overline{P_i}\right) + \beta_2 \cdot (x_{1it} - \overline{x_{1i}}) + \beta_3 \cdot (x_{2i} - \overline{x_{2i}}) + \beta_4 \cdot (\mu_i - \overline{\mu_i}) + \epsilon_{it}^Y - \overline{\epsilon_i}^Y$$

differencing.

[15] There are yet other "fixed effects" approaches, such as the use of "sufficient statistics", which we discuss briefly in a subsequent subsection.

[16] However, one of these estimators is sometimes applied to address fixed unobservables in the context of non-linear models.

[17] This new specification can easily be motivated in a manner analogous to first differencing. First, beginning with

$$Y_{it} = \beta_0 \cdot 1 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

we have

$$\overline{Y_i} = \beta_0 \cdot 1 + \beta_1 \cdot \overline{P_i} + \beta_2 \cdot \overline{x_{1i}} + \beta_3 \cdot \overline{x_{2i}} + \beta_4 \cdot \overline{\mu_i} + \overline{\epsilon_i^Y}$$

Subtracting the second equation from the first yields the specification in the main body of the text:

$$\begin{array}{rccccccccccc}
 & Y_{it} & = & \beta_0 \cdot 1 & + & \beta_1 \cdot P_{it} & + & \beta_2 \cdot x_{1it} & + & \beta_3 \cdot x_{2i} & + & ... \\
- & \overline{Y_i} & = & \beta_0 \cdot 1 & + & \beta_1 \cdot \overline{P_i} & + & \beta_2 \cdot \overline{x_{1i}} & + & \beta_3 \cdot \overline{x_{2i}} & + & ... \\
\hline
- & Y_{it} - \overline{Y_i} & = & \beta_0 \cdot (1-1) & + & \beta_1 \cdot \left(P_{it} - \overline{P_i}\right) & + & \beta_2 \cdot (x_{1it} - \overline{x_{1i}}) & + & \beta_3 \cdot (x_{2i} - \overline{x_{2i}}) & + & ...
\end{array}$$

However, since

$$\overline{x_{2i}} = x_{2i}$$

and

$$\overline{\mu_i} = \mu_i$$

we know that

$$\mu_i - \overline{\mu_i} = 0$$

and

$$x_{2i} - \overline{x_{2i}} = 0$$

Therefore, the de-meaned regression specification reduces to

$$Y_{it} - \overline{Y_i} = \beta_0 \cdot (1-1) + \beta_1 \cdot \left(P_{it} - \overline{P_i}\right) + \beta_2 \cdot (x_{1it} - \overline{x_{1i}}) + \epsilon_{it}^Y - \overline{\epsilon_i}^Y$$

Finally, the term $\beta_0 \cdot (1-1)$ is motivated by the idea of demeaning the constant term, which is based on a column of 1s (and hence has a mean of 1, across individuals or even the entire sample). This leads to our final specification

$$Y_{it} - \overline{Y_i} = \beta_1 \cdot \left(P_{it} - \overline{P_i}\right) + \beta_2 \cdot (x_{1it} - \overline{x_{1i}}) + \epsilon_{it}^Y - \overline{\epsilon_i}^Y$$

This specification has purged the confounding unobservable μ, as well as the non-time varying observable x_2.

The unbiasedness of this estimator hinges on the same assumption as in the first differencing case: that the source of endogeneity bias to the estimate of program impact from straightforward regression of Y on P, x_1 and x_2 is the fixed (i.e. non-time varying) unobserved confounding characteristic μ. In other words, it assumes that P and any time-varying unobservable such as ϵ^Y are independent (i.e. that ϵ^Y is not an unobserved confounder).

Most of the implications of the first difference model carry over to this model as well. For instance, the model still involves the loss of the information from one time period's worth of observations, though the reason can seem a bit less straightforward in this case. To illustrate, let us focus on the regressor

$$Y_{it} - \overline{Y_i}$$

By definition,

$$\sum_{t=1}^{T} \left(Y_{it} - \overline{Y_i}\right) = 0$$

In other words, for individual i the deviations of their values of Y from the average over time for Y sum to zero. This means, however, that one observation for individual i does not really provide truly independent information since it must insure that this condition holds. Focusing on the T^{th} observation (i.e. assuming that $\{Y_{i1}, Y_{i2}, ..., Y_{iT-1}\}$ are independently determined and hence provide $T-1$ truly independent units worth of information about variation in Y), it must be the case that

$$Y_{iT} = \overline{Y_i} - \sum_{t=1}^{T-1} \left(Y_{it} - \overline{Y_i}\right)$$

Therefore, Y_{iT} does not, from this perspective, provide truly independent information about variation in Y: its value is locked down by the need to insure that the condition

$$\sum_{t=1}^{T} \left(Y_{it} - \overline{Y_i}\right) = 0$$

holds. However, such a loss of one truly independent observation will occur for each individual, resulting in the loss of N units of information (i.e. one time period worth of information) from demeaning.

Estimation of this demeaned model is relatively simply. Define

$$\dot{Y_{it}} = Y_{it} - \overline{Y_i}$$

$$\dot{P_{it}} = P_{it} - \overline{P_i}$$

and

$$\dot{x_{1it}} = x_{1it} - \overline{x_{1i}}$$

These can be calculated by straightforward means: simply subtract from the value of the variable for individual i at time t that variable's mean over time for individual i. Estimation is then simply a question of regressing $\dot{Y}$ on $\dot{P}$ and $\dot{x_1}$ while suppressing the constant term.

To illustrate this estimator in action, we continue the two time period example from the last section. As a reference, recall that the estimate of program impact from the basic first-differenced model was 1.841112. To begin with, in Output 5.8, we report results from regression of $\dot{y}$ on $\dot{P}$ and $\dot{x_1}$ with no constant (because our model as posed does not necessarily imply the presence of a constant). (For reference, the variable `ydot` in the output represents $\dot{y}$, `Pdot` represents $\dot{P}$, etc.) The estimate of program impact is, at 1.70862, close to that from first differencing but not the same.

STATA Output 5.8 (5.1.do)

```
. * Simple demeaned fixed effects with no time trend or constant
. reg ydot Pdot x_1dot, nocon
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9998) = 2411.38
       Model |  32633.3214     2  16316.6607           Prob > F      =  0.0000
    Residual |  67651.8475  9998  6.76653806           R-squared     =  0.3254
-------------+------------------------------           Adj R-squared =  0.3253
       Total |  100285.169 10000  10.0285169           Root MSE      =  2.6013

------------------------------------------------------------------------------
        ydot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Pdot |     1.70862   .0984207    17.36   0.000     1.515695    1.901544
      x_1dot |    1.441792   .0217219    66.38   0.000     1.399213    1.484372
------------------------------------------------------------------------------
```

An obvious possible reason for the discrepancy between the fixed effects estimation by demeaning and the first difference estimator is the absence of a time trend in the specification we have attempted thus far. The addition of a constant term is a bit less straightforward in the classic fixed effects model. One possibility would be simply to add demeaned time as an explanatory variable (while still suppressing estimation of the constant). Results for this are presented in Output 5.9.

The estimate of program impact, at 1.84112, now matches that yielded by first differencing. There is a reason for this: with only two time periods, first differencing and the demeaned fixed effects estimator are equivalent. This can be seen through the prism of any of the variables considered in the regression specification. For instance,

$$\overline{Y_i} = \frac{Y_{i1} + Y_{i2}}{2}$$

Then,

$$Y_{i1} - \overline{Y_i} = Y_{i1} - \frac{Y_{i1} + Y_{i2}}{2} = \frac{2 \cdot Y_{i1} - Y_{i1} - Y_{i2}}{2} = -\frac{1}{2}\left(Y_{i2} - Y_{i1}\right)$$

Similar math can be applied to Y_{i2}, x_{1i1}, x_{1i2}, P_{i1} and P_{i2}. The point is that the variables in the demeaned estimation are simply those present in the first differenced model multiplied by $-(1/2)$.

STATA Output 5.9 (5.1.do)

```
. * Simple demeaned fixed effects with time trend
. reg ydot  Pdot x_1dot t_dot, nocons
```

Source	SS	df	MS		
Model	55590.8463	3	18530.2821	Number of obs	= 10000
Residual	44694.3227	9997	4.4707735	F(3, 9997)	= 4144.76
				Prob > F	= 0.0000
Total	100285.169	10000	10.0285169	R-squared	= 0.5543
				Adj R-squared	= 0.5542
				Root MSE	= 2.1144

ydot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Pdot	1.841112	.0800222	23.01	0.000	1.684252	1.997972
x_1dot	1.469093	.0176607	83.18	0.000	1.434475	1.503712
t_dot	3.031345	.0423023	71.66	0.000	2.948424	3.114266

This straightforward math, and hence the equality of the first difference and demeaned fixed effects estimators, breaks down with more than two time periods. Hence, while they are both unbiased estimators of program impact under broadly similar assumptions (namely, that P and μ might be correlated but P and ϵ^Y are independent) the actual values of the estimates of program impact will typically differ between them in samples where the taking of differences or demeaning involves more than two observations per unit of observation (e.g. more than two time periods observed per individual).

In Outputs 5.10 and 5.11 we provide the regression estimates from first differencing and demeaning, respectively, when the empirical example is extended to three time periods.[18] The first differenced and demeaned estimates are no longer equal (at 2.053201 and 2.009387, respectively). Interestingly, both estimates are also much closer to the true value of 2. This largely reflects a precision gain from the increase in sample size.

Many commercial statistical packages have purpose-built fixed effects estimation routines. For instance, in STATA the demeaned model can be estimated via the `xtreg` command with the `fe` option. Typically, these packages handle all of the steps in fixed effects regression, including demeaning.

To explore STATA's packaged fixed effects estimation routine, we first estimate the regression of demeaned Y on demeaned P, x_1 and t *without* suppression of estimation of the constant (in doing so we continue with the three time period extension of our basic example). The results are reported in Output 5.12. There are two major things to note about the results in Output 5.12. First, the coefficient estimates for $\dot{P}$, $\dot{x}_1$ and $\dot{t}$ (represented in Output 5.12 by `Pdot`, `x_1dot` and

[18] To generate 3 time periods worth of output, simply change line 17 of the STATA do-file 5.1.do to

```
loc T=2
```

to

```
loc T=3
```

By changing the value of `T` one can consider any number of scenarios in terms of numbers of time periods.

t_dot, respectively) are essentially the same as in Output 5.10. In other words, the addition of a constant has not really influenced these estimates. Second, the estimate of the constant is tiny and insignificant (it is not everyday in applied empirical work that you see a t-statistic of 0.00 or, by extension, a p-value of 1.000).

STATA Output 5.10 (5.1.do)

```
. * The basic first differenced regression
. reg deltaY deltaP deltax_1 if tt<`T´
```

Source	SS	df	MS		Number of obs = 10000
					F(2, 9997) = 3869.25
Model	138924.79	2	69462.3948		Prob > F = 0.0000
Residual	179470.407	9997	17.9524264		R-squared = 0.4363
					Adj R-squared = 0.4362
Total	318395.196	9999	31.8427039		Root MSE = 4.237

deltaY	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
deltaP	2.053201	.0796707	25.77	0.000	1.89703	2.209371
deltax_1	1.513142	.0177598	85.20	0.000	1.47833	1.547955
_cons	3.007354	.0423764	70.97	0.000	2.924287	3.09042

STATA Output 5.11 (5.1.do)

```
. * Simple demeaned fixed effects with time trend
. reg ydot  Pdot x_1dot t_dot, nocons
```

Source	SS	df	MS		Number of obs = 15000
					F(3, 14997) = 8976.49
Model	161309.426	3	53769.8087		Prob > F = 0.0000
Residual	89833.1105	14997	5.99007205		R-squared = 0.6423
					Adj R-squared = 0.6422
Total	251142.536	15000	16.7428358		Root MSE = 2.4475

ydot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Pdot	2.009387	.0648643	30.98	0.000	1.882245	2.136528
x_1dot	1.504437	.01454	103.47	0.000	1.475937	1.532937
t_dot	3.007264	.0244816	122.84	0.000	2.959277	3.055251

In Output 5.13 we estimate the demeaned fixed effects model using STATA's xtreg package with the fe option. The coefficient estimates for program participation P, the time-varying characteristic x_1 and time t and are essentially identical to those for $\dot{P}$, $\dot{x_1}$ and $\dot{t}$ (represented in Output 5.11 by Pdot, x_1dot and t_dot, respectively) in Outputs 5.10 and 5.11. Note, however, that the t-statistics for those estimates generally run noticeably smaller than the corresponding t-statistics in Outputs 5.10 and 5.11. The reason is that in the regressions in Outputs 5.11 and 5.12 the dependent variable and regressors were demeaned before estimation, which then proceeded, both in terms of calculation of parameter estimates and their standard errors, by means of standard ordinary least squares estimation. Crucially, standard ordinary least squares estimation assumes that each observation provides 1 observation's worth of new and unique information regarding those variables. However, we have just seen that this is not actually the case with demeaned variables: for each unit of observation with the respect to which demeaning occurs, 1 observation is not

truly independent. Naive estimation by ordinary least squares, as in the case of Outputs 5.11 and 5.12, thus produces standard error estimates based on an exaggerated sense of the information for parameter estimation actually contained in the data. This will lead to standard errors biased downward. Purposeful fixed effects packages such as `xtreg` explicitly recognize this information loss in computing standard errors.[19]

STATA Output 5.12 (5.1.do)

```
. * Simple demeaned fixed effects with time trend
. reg ydot  Pdot x_1dot t_dot
```

Source	SS	df	MS		
Model	161309.426	3	53769.8087	Number of obs	= 15000
Residual	89833.1105	14996	5.99047149	F(3, 14996)	= 8975.89
				Prob > F	= 0.0000
				R-squared	= 0.6423
				Adj R-squared	= 0.6422
Total	251142.536	14999	16.743952	Root MSE	= 2.4475

ydot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Pdot	2.009387	.0648665	30.98	0.000	1.88224	2.136533
x_1dot	1.504437	.0145405	103.47	0.000	1.475936	1.532938
t_dot	3.007264	.0244824	122.83	0.000	2.959276	3.055253
_cons	2.14e-08	.0199841	0.00	1.000	-.0391713	.0391713

STATA Output 5.13 (5.1.do)

```
. xtset ID
       panel variable:  ID (balanced)
. * STATA Xtreg, time covariate
. xtreg Y P x_1 t, fe
```

```
Fixed-effects (within) regression               Number of obs      =     15000
Group variable: ID                              Number of groups   =      5000

R-sq:  within  = 0.6423                         Obs per group: min =         3
       between = 0.1571                                        avg =       3.0
       overall = 0.3393                                        max =         3

                                                F(3,9997)          =   5983.73
corr(u_i, Xb)  = 0.0408                         Prob > F           =    0.0000
```

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	2.009386	.0794462	25.29	0.000	1.853656	2.165117
x_1	1.504437	.0178087	84.48	0.000	1.469528	1.539345
t	3.007264	.0299852	100.29	0.000	2.948487	3.066041
_cons	1.914701	.0600506	31.88	0.000	1.79699	2.032412
sigma_u	4.7890736					
sigma_e	2.9976669					
rho	.71849437	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(4999, 9997) =     9.53          Prob > F = 0.0000
```

The other interesting thing about Output 5.13 is the estimate of the constant term. It is

[19] Note that this issue does not arise with first differencing because that explicitly leads to the loss of one observation per unit of observation.

now highly significant (compared with the result in Output 5.12) despite the specification yielding parameter estimates for the other regressors highly similar to those in Output 5.11 (which was the best analogy to the manner in which a constant term was introduced to the first differenced model and for which constant estimation was suppressed altogether) or, for that matter, to those in Output 5.12 (for which constant estimation was not suppressed). This has several implications. First, control for time (properly demeaned) might be useful practice in demeaned models. It appears to do little harm (except perhaps in very small datasets for which the information in each observation is dear and parsimony[20] becomes an important consideration) and, to the extent that there is a time trend to both the outcome and program participation, might avoid omitted variable bias.

Second, the estimation of a classic constant term in the demeaned model is a less clear-cut business than in the first differenced case. The purpose-built fixed effects estimation commands in different commercial statistical packages can also take different approaches to constant estimation and hence yield different results. To get some flavor for how particular the process can be, we consider the estimation of the constant term under STATA's `xtreg` command with the `fe` option.

To begin with, consider the foundational model

$$Y_{it} = \gamma + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

This is essentially the departure point model for the chapter to this point minus the term related to the fixed regressor x_2 (which is not important for the present discussion). Demeaning the regressors yields

$$Y_{it} - \overline{Y_i} = \gamma \cdot (1-1) + \beta_0 \cdot (t - \bar{t}) + \beta_1 \cdot \left(P_{it} - \overline{P_i}\right) + \beta_2 \cdot (x_{1it} - \overline{x_{1i}}) + \beta_4 \cdot (\mu_i - \overline{\mu_i}) + \epsilon_{it}^Y - \overline{\epsilon_1}^Y$$

or

$$\dot{Y_{it}} = \beta_0 \cdot \dot{t} + \beta_1 \cdot \dot{P_{it}} + \beta_2 \cdot \dot{x_{1it}} + \dot{\epsilon_{it}}^Y$$

However, from the foundational model the following is also true:

$$\overline{Y} = \gamma + \beta_0 \cdot \bar{t} + \beta_1 \cdot \overline{P} + \beta_2 \cdot \overline{x_1} + \beta_4 \cdot \overline{\mu} + \overline{\epsilon}^Y$$

where

$$\overline{Y}, \bar{t}, \overline{P}, \overline{x_1}, \overline{\mu}, \overline{\epsilon}^Y$$

are the means of these variables across the entire sample. Adding this specification to the demeaned specification yields

$$\dot{Y_{it}} + \overline{Y} = \gamma + \beta_0 \cdot (\dot{t} + \bar{t}) + \beta_1 \cdot \left(\dot{P_{it}} + \overline{P}\right) + \beta_2 \cdot (\dot{x_{1it}} + \overline{x_1}) + \beta_4 \cdot \overline{\mu} + \dot{\epsilon_{it}}^Y + \overline{\epsilon}^Y$$

which is another way of saying

$$\dot{Y_{it}} + \overline{Y} = \gamma + \beta_0 \cdot (\dot{t} + \bar{t}) + \beta_1 \cdot \left(\dot{P_{it}} + \overline{P}\right) + \beta_2 \cdot (\dot{x_{1it}} + \overline{x_1}) + \beta_4 \cdot \overline{\mu} + \textit{Some Random Error}$$

`xtreg` with the `fe` option estimates this model under the restriction that

$$\overline{\mu} = 0$$

The estimand for the constant term from the `xtreg` command with `fe` option is thus γ, the constant term from the original model. This can be seen by estimating the original regression

[20] In statistics, **parsimony** is the quality of limiting the number of parameters to be estimated. Of course, as we have seen in the last chapter that can actually generate biased estimates of program impact.

model from this chapter without omitted variable bias by including μ as a regressor, results for which are presented in Output 5.14.[21]

STATA Output 5.14 (5.1.do)

```
. * The correct specification of the original model
. reg Y P x_1 x_2 t mu
      Source |       SS       df       MS              Number of obs =   15000
-------------+------------------------------           F(  5, 14994) =11393.63
       Model |  519090.578     5  103818.116           Prob > F      =  0.0000
    Residual |  136624.456 14994  9.11194188           R-squared     =  0.7916
-------------+------------------------------           Adj R-squared =  0.7916
       Total |  655715.034 14999  43.7172501           Root MSE      =  3.0186

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    1.981741   .0650937    30.44   0.000      1.85415    2.109333
         x_1 |    1.496318   .0145315   102.97   0.000     1.467835    1.524802
         x_2 |    1.991945   .0132706   150.10   0.000     1.965933    2.017957
           t |    3.007309   .0301917    99.61   0.000      2.94813    3.066488
          mu |    .9712205   .0133681    72.65   0.000     .9450173    .9974237
       _cons |    1.997153   .0544501    36.68   0.000     1.890424    2.103882
------------------------------------------------------------------------------
```

We conclude by mentioning another popular linear fixed effects method commonly applied: the dummy variable estimator. It is simple enough in conceptualization: it attempts to control for μ_i from

$$Y_{it} = \gamma + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

by estimating it for each of the $i = 1, ..., N$ individuals in the sample. This is accomplished by regressing Y_{it} on P_{it}, t, x_{1it}, and a dummy variable for each individual. The regression model actually estimated is thus on the lines of

$$Y_{it} = \delta + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it} + \sum_{j=2}^{N} \phi_j \cdot d_j + \epsilon_{it}^Y$$

where d_j is a dummy that equals 1 if $i = j$ (for instance, $d_{550} = 1$ if $i = 550$). The ϕs are parameters to be estimated in the regression. They estimate the "fixed effect" of being each individual in the sample. That fixed effect captures both observed but fixed (with respect to time) individual characteristics and unobserved fixed characteristics. The former are not explicitly included in the regression because the effect of variables like x_{2i}, which do not vary over time for the individual, on Y_{it} cannot be separately identified from the effect of d_i on Y_{it}. Put differently, ϕ_i cannot be separately estimated from β_3. Instead, x_2 is effectively subsumed in the fixed effect. In other words,

$$\phi_i = \beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i$$

A dummy variable is not offered for the first individual in the sample $i = 1$ because a fixed effect for that individual could not be separately identified from a constant term. Rather, the constant term

[21] Just after deriving the approach behind `xtreg`, we discovered that STATA had helpfully laid it out in a discussion at

http://www.stata.com/support/faqs/statistics/intercept-in-fixed-effects-model/

(retrieved December 4, 2013).

δ captures that individual's fixed effect, and the estimates of the fixed effect (i.e. the ϕs) for every other individual in the sample can be interpreted as the difference between their own fixed effect and the first individual's fixed effect. This is the reason that the constant term is now labelled δ and not γ: in the dummy variable model we are not estimating the original constant γ.

Estimation for the dummy variable model are provided in Output 5.15. For practicality, the estimates of ϕ are provided for only the 2^{nd} through 6^{th}, and 4,996th to 5,000th individuals. The most important thing to recognize is that the estimated coefficients for the key variables (including program impact, the coefficient on P) are identical to those obtained from either the manually (i.e. by use) demeaned model or `xtreg` with the `fe` option. Interestingly, the standard errors and t-statistics for those key variables are the same as those yielded by the `xtreg` command (see Output 5.13). The reason for this is that the dummy variable model explicitly adds 4,999 parameters (in this example), thus insuring that standard errors are estimated based on the correct understanding of the true number of independent observations (to see this, review the discussion of degrees of freedom in the last chapter). The dummy variable estimator is thus equivalent in terms of parameter estimates for the key covariates to the demeaned estimator and in terms of standard error estimates to any implementation of the demeaned estimator that correctly recognizes the number of independent observations (i.e. degrees of freedom) as with `xtreg` with the `fe` option.

STATA Output 5.15 (5.1.do)

```
. * The dummy variable fixed effects model
. reg Y P x_1 t i.ID
```

Source	SS	df	MS	Number of obs = 15000
				F(5002, 9997) = 12.59
Model	565881.924	5002	113.131132	Prob > F = 0.0000
Residual	89833.1105	9997	8.98600685	R-squared = 0.8630
				Adj R-squared = 0.7945
Total	655715.034	14999	43.7172501	Root MSE = 2.9977

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	2.009386	.0794462	25.29	0.000	1.853656	2.165117
x_1	1.504437	.0178087	84.48	0.000	1.469528	1.539345
t	3.007264	.0299852	100.29	0.000	2.948487	3.066041
ID						
2	-4.807608	2.447981	-1.96	0.050	-9.606144	-.0090715
3	-1.476874	2.448031	-0.60	0.546	-6.275507	3.32176
4	-.9128602	2.447719	-0.37	0.709	-5.710882	3.885162
5	-2.823543	2.447782	-1.15	0.249	-7.621688	1.974602
6	3.224894	2.447763	1.32	0.188	-1.573214	8.023002
...	...	...	...	...	...	...
4996	-5.342831	2.447689	-2.18	0.029	-10.14079	-.5448671
4997	-8.186501	2.448126	-3.34	0.001	-12.98532	-3.387681
4998	-3.084011	2.44818	-1.26	0.208	-7.882936	1.714914
4999	-2.466617	2.447702	-1.01	0.314	-7.264606	2.331372
5000	.3235599	2.447739	0.13	0.895	-4.474502	5.121622
_cons	4.716826	1.731343	2.72	0.006	1.323045	8.110608

In terms of popularity, the demeaned fixed effects model is probably most commonly employed, followed by the dummy variable fixed effects model and, more distantly, the first difference model. The first difference and fixed effects models differ primarily in the precision with which they estimate the parameters of the key covariates, including program impact. This is manifested by smaller standard errors on the estimated parameters for these key covariates. Specifically, when the ϵ^Ys

are correlated over time (i.e. $corr(\epsilon^Y_{it}, \epsilon^Y_{it+1}) \neq 0$) the first difference estimator is generally more efficient. Although the fixed effects models are generally more popular the basis for that relative popularity is unclear. As this discussion has made clear, however, with proper specification their differences in terms of the point estimates for program impact are not particularly important compared with the difference between any of them and the program impact estimate yielded by simple regression of Y_{it} on P_{it}, t, x_{1it} and x_{2i}.

5.1.3 Measurement Error

We now briefly discuss one potential pitfall of within estimators: they tend to worsen bias from mismeasurement of regressors. This statement has two implications: first, that mismeasurement of regressors introduces bias to estimates generated by regression and, second, that this bias is often worse in within regression models. The first logical step in a discussion of this is to explain measurement error bias in regressions per se. For the purpose of doing so, as well as examining the implications of measurement error in the context of within models, we begin with the simple model

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

This is the true data generating process behind the outcome y. We assume the independence of x and ϵ.

This is essentially a model involving cross-sectional variation in the outcome and regressor. In other words, it shows how levels of y depend on levels of x. Ordinary least squares regression of y on x would identify estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively, by considering how variation in levels of y across individuals appeared to be linearly associated with variation in levels of x across the same individuals. Since x is independent of ϵ, those estimates would also be unbiased (i.e. $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$).

Suppose, however, that we actually observe not x but a mismeasured version of x that we will call $\tilde{x}$. Specifically, suppose we observe $\{y_i, \tilde{x}_i\}$ for a sample of $i = 1, ..., N$ individuals where

$$\tilde{x}_i = x_i + \upsilon_i$$

Thus, we observe not x_i, but x_i mis-measured by a random margin represented by υ_i. For simplicity, assume that the error υ_i is uncorrelated with x_i.

Let us now consider what we estimate when we observe the mismeasured regressor $\tilde{x}$ instead of the correctly measured regressor x. Substituting in, we have

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot x_i + \epsilon_i \\ &= \beta_0 + \beta_1 \cdot (\tilde{x}_i - \upsilon_i) + \epsilon_i \\ &= \beta_0 + \beta_1 \cdot \tilde{x}_i - \beta_1 \cdot \upsilon_i + \epsilon_i \\ &= \beta_0 + \beta_1 \cdot \tilde{x}_i - \xi_i \end{aligned}$$

where

$$\xi_i = -\beta_1 \cdot \upsilon_i + \epsilon_i$$

is the regression error when the mismeasured regressor $\tilde{x}_i$ is considered. Given the discussion of omitted variable bias in the last chapter, regression of y_i on $\tilde{x}_i$ will yield a biased estimate of β_1 because

$$corr\left(\tilde{x}_i, \xi_i\right) \neq 0$$

In other words, when we regress y_i on $\tilde{x}_i$

$$E\left(\hat{\beta}_1\right) \neq \beta_1$$

because $\tilde{x}_i$ is correlated with the error term via the omitted variable υ_i ($\tilde{x}_i$ is clearly correlated with υ_i via the relationship $\tilde{x}_i = x_i + \upsilon_i$).

Recall that the ordinary least squares estimate of β_1, $\hat{\beta}_1$, is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}\left(\tilde{x}_i - \overline{\tilde{x}}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N}\left(\tilde{x}_i - \overline{\tilde{x}}\right)^2}$$

The probability limit of this can be viewed as

$$plim\left(\hat{\beta}_1\right) = \frac{cov\left(y, \tilde{x}\right)}{var\left(\tilde{x}\right)}$$

where $cov(\cdot)$ and $var(\cdot)$ indicate covariance and variance, respectively. This can be expanded:

$$plim\left(\hat{\beta}_1\right) = \frac{cov\left(y, \tilde{x}\right)}{var\left(\tilde{x}\right)} = \frac{cov\left(\beta_0 + \beta_1 \cdot \tilde{x}_i - \beta_1 \cdot \upsilon_i + \epsilon_i, \tilde{x}\right)}{var\left(\tilde{x}\right)}$$

$$= \frac{\beta_1 \cdot var\left(\tilde{x}\right) - \beta_1 \cdot cov\left(\tilde{x}, \upsilon\right)}{var\left(\tilde{x}\right)}$$

$$= \beta_1\left(1 - \frac{cov\left(\tilde{x}, \upsilon\right)}{var\left(\tilde{x}\right)}\right) = \beta_1\left(1 - \frac{\sigma_\upsilon^2}{var\left(\tilde{x}\right)}\right)$$

However,

$$var\left(\tilde{x}\right) = \left(\sigma_x^2 + \sigma_\upsilon^2\right)$$

Therefore,

$$plim\left(\hat{\beta}_1\right) = \beta_1\left(1 - \frac{\sigma_\upsilon^2}{var\left(\tilde{x}\right)}\right) = \beta_1 \cdot \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2}\right)$$

$\hat{\beta}_1$ from the regression of y on $\tilde{x}$ is thus biased (we established this earlier simply by noting the correlation between $\tilde{x}$ and ξ) and inconsistent.

There are a number of interesting insights to be gleaned from the probability limit of $\hat{\beta}_1$,

$$plim\left(\hat{\beta}_1\right) = \beta_1 \cdot \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2}\right)$$

First, the inconsistency takes the form of a tendency to underestimate the true magnitude of β_1. Put more simply, the direction of the inconsistency of the estimate $\hat{\beta}_1$ of β_1 is toward 0. For this reason, measurement error bias is often referred to as **attenuation bias**. In a multiple regression mismeasured regressors still imply biased and inconsistent parameter estimates, but the situation gets a little more complicated in the sense that it becomes hard to generalize about the direction of the bias. Nonetheless, a tendency toward attenuation typically arises. Second, the degree of inconsistency depends on the proportion of the variation in the mismeasured variable $\tilde{x}$ driven by variation in the true value x. Put differently, the larger is the relative contribution of measurement error υ to the variation in the observed variable $\tilde{x}$ the worse will be the attenuation bias.

Measurement error bias can be worse, sometimes far worse, in the context of a within model than in the corresponding cross sectional regression model. To see this, let us now consider a

circumstance of balanced panels of N individuals for two time periods. Suppose that we observe $\{Y_{it}, \tilde{x}_{it}\}$ for a sample of $i = 1, ..., N$ observed at time periods $t = 1, 2$. We assume that the true values for the x (x_{i1} and x_{i2} for each individual i) are correlated over time (with correlation given by ρ) but that the measurement errors (υ) are uncorrelated over time and uncorrelated with the true values for x.

Starting from the cross sectional data

$$y_{it} = \beta_0 + \beta_1 \cdot x_{it} + \epsilon_{it}$$

we have the corresponding first differenced regression model

$$\Delta Y_i = \beta_1 \cdot \Delta x_i + \Delta \epsilon_i$$

where

$$\Delta Y_i = Y_{i2} - Y_{i1}$$

and

$$\Delta x_i = x_{i2} - x_{i1}$$

Since we observe $\tilde{x}_{it} = x_{it} + \upsilon_{it}$ rather than x_{it}, the regression function in terms of what we observe is

$$\begin{aligned} \Delta Y_i &= \beta_1 \cdot \Delta \tilde{x}_i - \beta_1 \cdot \Delta \upsilon_i + \Delta \epsilon_i \\ &= \beta_1 \cdot \Delta \tilde{x}_i + \Delta \xi_i \end{aligned}$$

Clearly the estimate of β_1, $\hat{\beta}_1$, will be biased since $\Delta \tilde{x}_i$ is correlated with $\Delta \xi_i$ because $\Delta \tilde{x}_i$ depends in part on $\Delta \upsilon_i$.

The probability limit of the estimate of β_1, $\hat{\beta}_1$, from the first difference regression of Δy on Δx is

$$\begin{aligned} plim\left(\hat{\beta}_1\right) &= \frac{cov\left(\Delta y, \Delta \tilde{x}\right)}{var\left(\Delta \tilde{x}\right)} \\ &= \frac{cov\left(\beta_1 \cdot \Delta \tilde{x}_i - \beta_1 \cdot \Delta \upsilon_i + \Delta \epsilon_i, \Delta \tilde{x}\right)}{var\left(\Delta \tilde{x}\right)} \\ &= \beta_1 - \beta_1 \cdot \frac{cov\left(\Delta \upsilon_i, \Delta \tilde{x}\right)}{var\left(\Delta \tilde{x}\right)} \\ &= \beta_1 \cdot \left(1 - \frac{2 \cdot \sigma_\upsilon^2}{2 \cdot (1 - \rho) \cdot \sigma_x^2 + 2 \cdot \sigma_\upsilon^2}\right) \\ &= \beta_1 \cdot \left(\frac{(1 - \rho) \cdot \sigma_x^2}{(1 - \rho) \cdot \sigma_x^2 + \sigma_\upsilon^2}\right) \end{aligned}$$

where $\rho = corr(x_{i1}, x_{i2})$.[22]

[22] The math of some of the final stages might not be immediately obvious. Most obviously

$$var\left(\Delta \tilde{x}\right) = var\left(\tilde{x}_2 - \tilde{x}_1\right) = var\left(x_2 + \upsilon_2 - x_1 - \upsilon_1\right)$$

where we suppress the i subscript for notational simplicity. Now, we assume that the measurement error terms υ are uncorrelated over time, but the true values x might be correlated over time. Then we have

$$\begin{aligned} var\left(\Delta \tilde{x}\right) &= var\left(x_2 + \upsilon_2 - x_1 - \upsilon_1\right) = var\left(x_2 - x_1\right) + var\left(\upsilon_2\right) + var\left(\upsilon_1\right) \\ &= var\left(x_2 - x_1\right) + 2 \cdot var\left(\upsilon\right) = var\left(x_2 - x_1\right) + 2 \cdot \sigma_\upsilon^2 \end{aligned}$$

The probability limit of the first difference estimator

$$plim\left(\hat{\beta}_1\right) = \beta_1 \cdot \left(\frac{(1-\rho)\cdot\sigma_x^2}{(1-\rho)\cdot\sigma_x^2 + \sigma_\upsilon^2}\right)$$

is in form very similar to the probability limit for the corresponding cross sectional model:

$$plim\left(\hat{\beta}_1\right) = \beta_1 \cdot \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2}\right)$$

Hence, many of the conclusions formed in the cross sectional setting carry over to within models. In particular, the greater is the variance of the measurement error υ compared with that of the true regressor x, the greater is the measurement error bias.

The most obvious *difference* in structure with the cross sectional case is the presence of the term $(1-\rho)$. By the addition, we see the most important truth about measurement error in within models: in within models it is generally worse than in a corresponding cross sectional model to the extent that the true time-varying regressors are correlated. The reason is rather simple: if x is highly correlated over time, it exhibits less true within variation than if it had been less highly correlated. Measurement error thus makes a proportionally larger contribution to the within variation than it did to the cross sectional variation.

There is a kind of intuitive quality to this result. To see this, let us focus on a simple example where we wish to estimate the impact of income on health care demand. Thus, we observe some channel of health care demand, HD, and income Y. Assume that there are no unobserved factors that influence HD and are associated with Y (a whopping assumption, but let's run with it for present purposes). To assess the impact of income on health care demand, we might regress HD on Y. Of course, for any number reasons Y is likely to be mismeasured, preventing us from observing the true income for individuals in our sample. If nothing else, individuals might not know their

$$= var\left(x_2\right) + var\left(x_1\right) - 2\cdot cov\left(x_{1i}, x_{2i}\right) + 2\cdot\sigma_\upsilon^2$$

where the third and last steps rely on the fact that, for any two variables Z and W and constants a and b

$$var\left(a\cdot Z + b\cdot W\right) = a^2\cdot var\left(Z\right) + b^2\cdot var\left(W\right) + 2\cdot a\cdot bCov\left(Z, W\right)$$

Continuing, we have

$$\begin{aligned} var\left(\Delta\tilde{x}\right) &= var\left(x_2\right) + var\left(x_1\right) - 2\cdot cov\left(x_1, x_2\right) + 2\cdot\sigma_\upsilon^2 \\ &= \sigma_x^2 + \sigma_x^2 - 2\cdot var\left(x\right)\cdot\frac{cov\left(x_1, x_2\right)}{var\left(x\right)} + 2\cdot\sigma_\upsilon^2 \\ &= \sigma_x^2 + \sigma_x^2 - 2\cdot var\left(x\right)\cdot\frac{cov\left(x_1, x_2\right)}{\sqrt{\left(var\left(x\right)\right)^2}} + 2\cdot\sigma_\upsilon^2 \end{aligned}$$

We assume that the variance of x does not evolve over time. Therefore $var(x) = var(x_1) = var(x_2)$ and

$$var\left(\Delta\tilde{x}\right) = \sigma_x^2 + \sigma_x^2 - 2\cdot var\left(x\right)\cdot\frac{cov\left(x_1, x_2\right)}{\sqrt{var\left(x_1\right)\cdot var\left(x_1\right)}} + 2\cdot\sigma_\upsilon^2$$

However,

$$\frac{cov\left(x_1, x_2\right)}{\sqrt{var\left(x_1\right)\cdot var\left(x_1\right)}}$$

is the correlation of x_1 and x_2, which we represent by ρ. Therefore

$$var\left(\Delta\tilde{x}\right) = \sigma_x^2 + \sigma_x^2 - 2\cdot\sigma_x^2\cdot\rho + 2\cdot\sigma_\upsilon^2 = 2\cdot(1-\rho)\cdot\sigma_x^2 + 2\cdot\sigma_\upsilon^2$$

precise income,[23] they might deliberately provide a distorted report of their income, etc. Thus there is likely to be some measurement error bias to the regression estimate of the impact of Y on HD.

Clearly, income might vary from time period to time period as the economy is buffeted by shocks, different individuals meet different fates in terms of their investments, etc. However, it seems reasonable to suspect that income Y is highly correlated over time: those with high income tend to have persistently higher than average income, while many of the poor seem essentially stuck in a low income trap. Indeed, it seems likely to suspect that, in many instances, the variation in income *across* individuals (i.e. the cross-sectional variation in income) dwarfs the variation in income *within* individuals over time. This means, however, that measurement error is likely to be a much larger portion of within than cross sectional variation in income, suggesting that within regression estimators of the impact of income Y on health care demand HD likely suffer from far worse measurement error bias.

An open question is how big a problem measurement error bias might be in the context of program impact evaluation. The straightforward and simple single regressor measurement error bias story would require somehow imperfectly observing program participation status. It is possible that participation could be mismeasured, though the authors must admit that it seems somewhat implausible in the context of some of the programs that they have evaluated. That said, in most instances there was no ready way for us to test for mismeasurement of participation (for instance, by spot checking program enrollment rolls against individual survey respondent's reports of their participation status). Of course, the likelihood of mismeasurement may depend on the conceptualization of participation. Nonetheless, measurement error of other regressors that serve as controls in a program impact evaluation regression is always a possibility. This could introduce a bad control problem.

Finally, we note that there is at least one circumstance where within estimators may actually yield unbiased estimates in the face of measurement error when the corresponding cross-sectional estimator would not: when the measurement error υ is correlated over time, a possibility we have thus far ruled out by assumption. To see this, consider the extreme case where υ is perfectly correlated over time (i.e. $corr(\upsilon_{it}, \upsilon_{it+1}) = 1$). In that case $\upsilon_{it} = \upsilon_{it+1} = \upsilon_i$. However, this means that the confounding unobservable when we observe $\tilde{x}$ instead of x is fixed over time: exactly the circumstance within models are designed to address. Within estimators thus actually purge measurement error bias when the bias is perfectly correlated over time.[24]

When the correlation in the measurement error over time is imperfect (i.e. less than 1) but not 0 the situation gets a bit murkier. Within estimators will likely reduce variation in measurement error to the extent that they purge out the correlated portion. To see this, imagine that the measurement error has an error components structure:

$$\upsilon_{it} = \upsilon_i^1 + \upsilon_{it}^2$$

In other words, the measurement error has a persistent component υ^1 and a time-varying component υ^2. Clearly, the larger is the relative contribution of υ^1 to the overall variation in υ, the greater the reduction in overall variation in measurement error achieved with a within model. Whether that can outweigh the loss in variation in the true values of the regressors when those true values are correlated is unclear in any given application.

[23] Indeed, as this manual was written, North Carolina tax law changed, requiring all residents, including the authors, to submit new paperwork for determining deductions. The author of this chapter was, despite actually being an economist, embarrassingly unsure of his precise income.

[24] Of course, to be sure of enjoying this unequivocal benefit one would need to assume that the measurement error is indeed perfectly correlated over time, a very strong assumption in many applications.

5.1.4 Nonlinear Models

The discussion of within estimators has until now focused on linear regression specifications, such as

$$Y_{it} = \gamma + \beta_0 \cdot t + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{1it}\beta_3 \cdot x_{2i} + \beta_4 \cdot \mu_i + \epsilon_{it}^Y$$

This model is linear in the sense that the outcome Y_{it} is continuous and all of the observed and unobserved regressors (t, P_{it}, x_{1it}, x_{2i} and μ_i) have a linear effect on the dependent variable Y_{it}. Linearity has made it relatively straightforward to convincingly remove any confounding influence from fixed unobservables such as μ_i by first differencing, demeaning, etc.

We now briefly consider nonlinear regression models. Things become a bit trickier when we shift the discussion to these non-linear models. First, we clarify what we mean by a non-linear model. To begin with, we might consider a general regression model along the general lines of

$$Y_{it} = f\left(\beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{it} + \mu_i + \epsilon_{it}^Y\right)$$

where $f(\cdot)$ is some nonlinear function. For instance, we might have

$$f(a) = exp(a)$$

where $exp(a)$ indicates the exponential function

$$f(a) = exp(a) = e^a \approx 2.718281828^a$$

This gives rise to the regression model

$$Y_{it} = exp\left(\beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{it} + \mu_i + \epsilon_{it}^Y\right)$$

This is essentially along the lines of a classic non-linear regression model for a continuous outcome Y (although the dependent variable Y might be in some sense limited due to the limited range of the functional form governing the regressors[25]).

The problem with a non-linear regression model such as this should be immediately apparent. Consider the earlier de-meaning exercise. We would now be left with

$$Y_{it} - \overline{Y_i} = exp\left(\beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{it} + \mu_i + \epsilon_{it}^Y\right)$$

$$-exp\left(\beta_0 \cdot 1 + \beta_1 \cdot \overline{P_i} + \beta_2 \cdot \overline{x_i} + \overline{\mu_i} + \overline{\epsilon_i^Y}\right)$$

Clearly, the confounding unobservable μ_i is not going to drop out of this specification. A similar result obtains with first differencing. Hence the first difference estimator and the de-meaning fixed effects estimator do not purge the confounding unobservable μ. In some instances the dummy variable estimator might represent a viable solution for controlling for the fixed unobservable, though as we will see this approach often introduces a new peril in the nonlinear context.

Probably more often in health and related fields where human welfare is studied at the micro level (e.g. at the individual level) non-linear regression models arise in the explicitly[26] limited dependent variable context. Suppose that Y is a binary outcome (e.g. whether the individual uses

[25] In our example we have the general functional form $y = e^a$, which places no real restriction on the domain (i.e. the values a can take on) but does imply restrictions on the range (i.e. the values y can take on), namely that y cannot take on negative values.

[26] As opposed to inadvertently due to the range of the nonlinear functional form employed.

or does not use contraception, visits or does not visit a particular health care provider, etc.) and we are interested in the effect of the regressors on the probability that Y equals 1:

$$Pr\left(Y_{it}=1|P_{it},x_{it},\mu_i\right)=f\left(P_{it},x_{it},\mu_i\right)$$

where primary interest in the program impact context is with the participation regressor P_{it} and some determinants of Y_{it} (namely μ_i in the present context) may not even be observed. Starting with the latent variable model

$$Y_{it}^{*}=\beta_0+\beta_1\cdot P_{it}+\beta_2\cdot x_{it}+\mu_i+\epsilon_{it}^{Y}$$

if we assume that ϵ_{it}^{Y} is the difference of two Type-I Extreme Value variables we have the logit choice probability:[27]

$$Pr\left(Y_{it}=1|P_{it},x_{it},\mu_i\right)=\frac{exp\left(\beta_0+\beta_1\cdot P_{it}+\beta_2\cdot x_{it}+\mu_i\right)}{1+exp\left(\beta_0+\beta_1\cdot P_{it}+\beta_2\cdot x_{it}+\mu_i\right)}$$

Clearly, as with the basic non-linear regression above, de-meaning or differencing would not purge μ_i. Indeed, it is not even obvious how to meaningfully demean or first-difference Y_{it}.

This problem extends to other limited dependent variable models as well. Suppose, for instance, that Y is a **count outcome**, such as number of children born, number of visits to a health care provider, number of arrests, etc. There are many different regression models for count outcomes (see Cameron and Trivedi (1998) for a comprehensive review of count outcome regression models). The classic count data model is Poisson regression, whereby

$$P\left(Y_{it}=n|P_{it},x_{it},\mu_i\right)=\frac{exp\left\{-\left(\beta_0+\beta_1\cdot P_{it}+\beta_2\cdot x_{it}+\mu_i\right)\right\}\cdot\left(\beta_0+\beta_1\cdot P_{it}+\beta_2\cdot x_{it}+\mu_i\right)^{n}}{n!}$$

for $n=0,1,2,....$ Thus we have a very different kind of outcome, and a very different kind of distributional assumption, but it should already be evident that we have the same basic problem: first differencing and demeaning will not purge the fixed unobservable μ_i.

As a gross generalization, most limited dependent variable models lack a consistent within estimator developed for that model. For instance, we are unaware of any consistent within estimator for the probit model. One oft-cited exception is a fixed effects estimator for the logit usually called the **Chamberlain model** or **Chamberlain logit model** (see, for instance, Maddala (1982) for discussion of the Chamberlain model).

Fixed effects models purpose-built for non-linear regression frameworks often take a somewhat novel approach to the control for the fixed confounding unobservables. The Chamberlain logit model is a good example. Suppose that we have a sample of N individuals across T time periods, and observe an outcome variable Y and program participation indicator P (i.e. $\{Y_{it},P_{it}\}$) for each of the $i=1,...,N$ individuals for each of the $t=1,...,T$ time periods.[28] Individual i's contribution to the likelihood at time period t is

$$L\left(Y_{it}|\mu_i\right)=\left(\frac{exp\left(\beta_0+\beta_1\cdot P_{it}+\mu_i\right)}{1+exp\left(\beta_0+\beta_1\cdot P_{it}+\mu_i\right)}\right)^{Y_{it}}$$

$$\cdot\left(\frac{1}{1+exp\left(\beta_0+\beta_1\cdot P_{it}+\mu_i\right)}\right)^{1-Y_{it}}$$

[27] Had we assumed that ϵ_{it}^{Y} is normal we would have had the probit model.

[28] This is, among other things, tantamount to assuming balanced panels, which we do without loss of generality.

Individual i's contribution to the likelihood function across all T time periods would then just be the product of their contribution from each individual time period:

$$L\left(Y_{i1}, Y_{i2}, ..., Y_{iT}|\mu_i\right) = \prod_{t=1}^{T}\left\{\left(\frac{exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)}{1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)}\right)^{Y_{it}} \cdot \left(\frac{1}{1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)}\right)^{1-Y_{it}}\right\}$$

$$= \frac{exp\left(\sum_{t=1}^{T} Y_{it} \cdot \left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)\right)}{\prod_{t=1}^{T}\left(1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)\right)}$$

$$= \frac{exp\left(\mu_i \sum_{t=1}^{T} Y_{it}\right) \cdot exp\left(\sum_{t=1}^{T} Y_{it} \cdot \left(\beta_0 + \beta_1 \cdot P_{it}\right)\right)}{\prod_{t=1}^{T}\left(1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)\right)}$$

This clearly depends on μ_i, which we must somehow purge.

The Chamberlain model involves the insight that

$$\sum_{t=1}^{T} Y_{it}$$

is a **sufficient statistic** for μ_i. Basically, a statistic z is sufficient for a parameter Θ if the probability of an outcome conditional on z does not depend on Θ. So saying that

$$\sum_{t=1}^{T} Y_{it}$$

is a sufficient statistic for μ_i is tantamount to saying that once we condition on

$$\sum_{t=1}^{T} Y_{it}$$

the probability of $\{Y_{i1}, Y_{i2}, ..., Y_{iT}\}$ no longer depends on μ_i. The proof of this is a bit tedious (the reader can skip this if desired; we discuss it mainly to make clear just how complex the motivation for such models can be), but runs as follows. To begin with, assume that

$$\sum_{t=1}^{T} Y_{it} = q$$

In other words, for this individual the binary outcome Y takes on the value 1 q times between $t = 1$ and $t = T$.

$$L\left(Y_{i1}, Y_{i2}, ..., Y_{iT}|\sum_{t=1}^{T} Y_{it} = q\right) = \frac{Pr\left(Y_{i1}, Y_{i2}, ..., Y_{iT}, \sum_{t=1}^{T} Y_{it} = q\right)}{Pr\left(\sum_{t=1}^{T} Y_{it} = q\right)}$$

This simply reflects Baye's Rule.[29] The next step reflects the fact that knowing that

$$\sum_{t=1}^{T} Y_{it} = q$$

[29]There are several versions of Baye's Rule. We invoke

$$Pr\left(A|B\right) = \frac{Pr\left(A, B\right)}{Pr\left(B\right)}$$

tells us nothing about the value of any particular Y_{it}. We have

$$L\left(Y_{i1}, Y_{i2}, ..., Y_{iT} | \sum_{t=1}^{T} Y_{it} = q\right) = \frac{Pr\left(Y_{i1}, Y_{i2}, ..., Y_{iT}, \sum_{t=1}^{T} Y_{it} = q\right)}{Pr\left(\sum_{t=1}^{T} Y_{it} = q\right)} = \frac{Pr\left(Y_{i1}, Y_{i2}, ..., Y_{iT}\right)}{Pr\left(\sum_{t=1}^{T} Y_{it} = q\right)}$$

By the math in the last paragraph, the next step involves recognizing that the numerator is equal to

$$\frac{exp\left(\mu_i \sum_{t=1}^{T} Y_{it}\right) \cdot exp\left(\sum_{t=1}^{T} Y_{it} \cdot (\beta_0 + \beta_1 \cdot P_{it})\right)}{\prod_{t=1}^{T} \left(1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)\right)}$$

Notice that from this

$$\frac{exp\left(\mu_i \sum_{t=1}^{T} Y_{it}\right)}{\prod_{t=1}^{T} \left(1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)\right)}$$

does not depend on the particular sequence $\{Y_{i1}, Y_{i2}, ..., Y_{iT}\}$ generating the result that

$$\sum_{t=1}^{T} Y_{it} = q$$

The denominator involves the sum of the probabilities for all of the ways (i.e. possible sequences $\{Y_{i1}, Y_{i2}, ..., Y_{iT}\}$ by which

$$\sum_{t=1}^{T} Y_{it} = q$$

Of course, the term

$$\frac{exp\left(\mu_i \sum_{t=1}^{T} Y_{it}\right)}{\prod_{t=1}^{T} \left(1 + exp\left(\beta_0 + \beta_1 \cdot P_{it} + \mu_i\right)\right)}$$

can be factored out from the probability of each possible sequence under which the Y_{it} is equal to 1 q times. This leaves the conditional likelihood

$$L\left(Y_{i1}, Y_{i2}, ..., Y_{iT} | \sum_{t=1}^{T} Y_{it} = q\right) = \frac{exp\left(\sum_{t=1}^{T} Y_{it} \cdot (\beta_0 + \beta_1 \cdot P_{it})\right)}{\sum_{d=1}^{D} \left\{exp\left(\sum_{t=1}^{T} Y_{it}^{d} \cdot (\beta_0 + \beta_1 \cdot P_{it})\right)\right\}}$$

where D is the number of possible sequences such that

$$\sum_{t=1}^{T} Y_{it} = q$$

For instance, if there are two time periods and Y equals 1 for individual i just once, the possible sequences of values for $\{Y_{i1}, Y_{i2}\}$ are {0, 1} and {1,0}.[30]

[30]Notice that an individual does not make a meaningful contribution to the likelihood if either

$$\sum_{t=1}^{T} Y_{it} = T$$

(in which case they always choose $Y = 1$) or

$$\sum_{t=1}^{T} Y_{it} = 0$$

(in which case they always choose $Y = 0$). In other words, individuals make a contribution to the likelihood only if $0 < q < T$, or if they sometimes chose $Y = 1$ and sometimes chose $Y = 0$. In some samples, it is possible that a large proportion of the individuals within them always chose $Y = 1$ or $Y = 0$ over time, implying a possibly large loss of observations when estimating the Chamberlain model.

This discussion of the Chamberlain model has been quite an intense little mathematical detour, one which the reader can be forgiven for finding a bit overwhelming or tedious (indeed, on editing the authors found that they had made a huge number of mistakes in composing such detailed mathematical derivations!). That said, it may not be that important that the reader follows it fully. There are probably two important things to retain from this experience. First, it should be clear enough that derivation of consistent fixed effects estimators for limited dependent variable models is often quite complex compared with the (fairly straightforward) linear "within" approaches. Second, per the particulars of the Chamberlain model, it can involve an enormous loss of observations over time since it identifies treatment effects only with those individuals who did not consistently make the same outcome choice over time.

Despite the fact that the Chamberlain model is well-known, it does not appear (in our admittedly limited experience) to be employed often in the program impact evaluation context.[31] Generally speaking, two approaches appear to be popular for addressing fixed unobservables when estimating limited dependent variable models with panel data (or any circumstance where there is "within" variation in the outcome and key regressors).

First, the dummy variable fixed effects estimator is often employed. However, when contemplating a fixed effects estimation strategy along these lines, one must always be mindful of the **incidental parameters problem**. This complication has been the focus of a rather large (extending back at least to Neyman and Scott (1948)) and ongoing (see, for instance, Greene (2004) for a recent example) literature that has examined many facets of the problem, the details of which are far beyond the scope of this manuscript. The basic idea of the incidental parameters problem is that in nonlinear models small sample inconsistency of estimates of the dummy variables designed to control for unobserved heterogeneity (whether at the individual level, community level, etc.) generally contaminates estimates of the parameters that are the object of interest (β_1 in the above example).

Whether, in light of the incidental parameters problem, the dummy variable manifestation of the fixed effects model can be applied in any particular application with a nonlinear model is often a judgment call. The basic incidental parameters problem involves the contamination of estimates of parameters of interest by the small sample inconsistency of the dummy variables designed to control for the unobserved heterogeneity. What this means in practice is that the fewer the observations at the level of the hypothesized confounding unobservables (e.g. the fewer the observations for each individual when the focus is controlling for an individual-level fixed confounding unobservable along the lines of μ_i) the more likely that the estimate of the dummy variable control for that unobserved heterogeneity will be inconsistent, yielding inconsistent estimates.

A good example of the dummy variable approach with limited dependent variable models is Angeles et al. (1998). Angeles et al. (1998) sought to estimate the impact of a family planning program on fertility. They had panel data on fertility involving samples of women from various villages over many years. The program was implemented at the village level over time. There was thus within variation in the outcome (because women had children in some but not all of the time periods in which they were observed) and program participation (because the program was gradually introduced across villages over time, so that in a given village the program might have been operating in some but not all of the time periods). Since the key outcome, giving birth in a given year, is binary, Angeles et al. modeled it with a logit regression.

Their concern was with potential endogeneity of program placement owing to community-level unobservables that might influence both the reproductive choices of women and program placement.

[31] Implementing it in STATA is trivial via the `xtlogit` command with `fe` option. Most commercial statistical packages of which we are aware offer an estimation command or routine for the Chamberlain model in one fashion or another.

For instance, some villages might have cultures more receptive to fertility regulation, a prospect that might make them lobby for the program and render the women within such villages less likely to give birth. In a simple regression of fertility (i.e. whether a woman gave birth in any given year) on program participation (i.e. whether the family planning program was present in the village in that given year), the presence of such unobserved village level factors would likely lead to an exaggerated estimate of program impact. Of course, this prior regarding the direction of bias is based on just one scenario. Perhaps villages resistant to fertility control would be prioritized, likely leading to underestimation of true program impact.

Under the assumption that these confounding village-level unobservables were fixed over time, Angeles et. al (1998) used dummy variables for village to control for community level permanent heterogeneity in their logit regression of fertility on program participation. This may have been reasonable because, cumulatively, they had many woman-year observations for each village. This somewhat alleviates concerns about the incidental parameters problem. Unfortunately, how many observations are sufficient for the unit of observation to overcome the incidental parameters problem is generally not clear.

The other possible approach is to try to use a linear model. For instance, with a binary outcome one could employ the linear probability model. This has the advantage of being amenable to standard linear within models and not suffering from unfortunate properties such as the incidental parameters problem. Some may object to this by arguing that such linear models offer inferior fit to the data. However, this consideration must be balanced against some recognition of the overall goal, to model average program impact, for which such linear models are likely often perfectly adequate.

5.1.5 Hausman-Type Tests

We now briefly discuss a class of endogeneity tests that are typically referred to as **"Hausman tests"**, though a variety of names are attached to them (Hausman, Hausman-Wu, Hausman-Taylor, etc.) and have in one fashion or another been developed in numerous manuscripts (Hausman (1978), Wu (1973), Durbin (1954), etc.). Hausman-type tests test for the endogeneity of regressors in a regression model. More specifically, they test for correlation between regressors and the true regression error. Typically, Hausman tests actually test for the possibility that some (not necessarily all) of the regressors are correlated with the regression error.

The basic idea of all Hausman tests is quite simple. The test involves comparing "consistent" regression estimates with "efficient" ones. Large differences in the estimates are then taken as evidence of correlation between at least some of the regressors and the error term.

In this context, a "consistent" estimate is one that recognizes and corrects for *potential* correlation between regressors and the true regression error. The within estimators we have just studied are the canonical example: they recognize that regressors may be correlated with the true "cross-sectional" regression error term due to the presence of fixed confounding unobservables in the error term, and proceed to purge them. Consider, for instance, the following regression model

$$Y_{it} = \beta_0 + \beta_1 \cdot P_{it} + \beta_2 \cdot x_{it} + \mu_i + \epsilon_{it}^Y$$

Within estimators recognize the possible correlation between either P_{it} or x_{it} and the fixed (over time for individual i) unobservable μ_i and deal with it by means we have discussed (first differencing, demeaning, introduction of dummy variables for individual i).

We have seen that these strategies come at a very heavy cost in terms of truly independent observations available for estimation (i.e. degrees of freedom). In other words, they entail a loss of information that manifests itself, other things being equal, in terms of higher standard errors to

estimates of parameters (such as the estimate $\hat{\beta}_1$ of program impact β_1) compared with, for instance, plain old ordinary least squares estimation of Y_{it} on P_{it} and x_{it}. To obtain unbiased estimates of program impact β_1 and other parameters (β_2 being the only other regression parameter in our simple model) ordinary least squares regression of Y_{it} on P_{it} and x_{it} requires that the observed regressors P_{it} and x_{it} be independent (or at least mean independent) of any fixed unobservables such as μ_i. However, even if this assumption is correct (i.e. that P_{it} and x_{it} are independent or at least mean independent of any fixed unobservables such as μ_i) standard ordinary least squares estimation would yield incorrect standard error estimates. The reason is that the basic formulas used to estimate standard errors for ordinary least squares regression estimates such as $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ assume that the error terms are all completely independent. However, in the presence of a persistent unobservable (even one unrelated to P_{it} and x_{it}) this is clearly not the case: the errors for a given individual i will be correlated across time owing to the persistent presence of μ_i in their error term.

There are several solutions to this problem that yield "more correct" (i.e. likely closer to the truth) standard error estimates than the basic ordinary least squares formulas that assume complete independence of errors. One in particular, random effects estimation, is often used to produce the "efficient" estimates against which Hausman tests compare the "consistent estimates". Random effects are the topic of a fairly vast literature, with many different facets, proposed estimators, etc., but the core idea is rather simple: to model the error term as correlated for each individual over time. Typically this involves some sort of explicit distributional assumption about the error term (normality being an extremely popular choice). For instance, one might assume that the ϵs are independently (across individuals and within each individual over time) normally distributed while each individual receives a draw for μ that is independent across individuals but constant over time

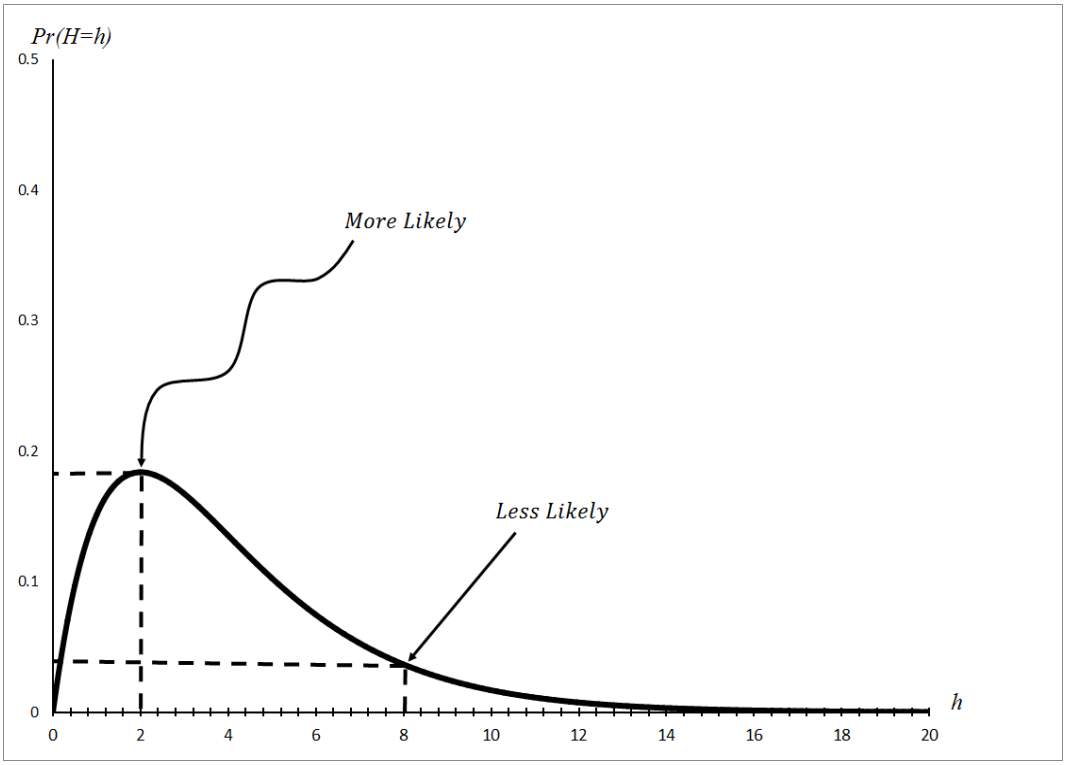


Figure 5.1: More and Less Likely χ^2 Values

for a given individual. Estimation then often proceeds by maximum likelihood.

To summarize, the consistent estimator delivers consistent estimates but typically at a high cost in terms of loss of information and thus higher standard errors. The efficient estimator delivers estimates with lower standard errors, but under the assumption that the regressors are unrelated to the error term (in other words, the estimates of parameters such as β_1 or β_2 could be inconsistent if this assumption is wrong). Hausman tests conclude that the regressors are probably independent of the errors if, after correcting for the differences in standard errors of the two approaches, the two sets of estimates ("consistent" and "efficient") appear similar. If, however, they are very different, this is taken as evidence that some regressors are indeed correlated with the error term.

Hausman tests typically involve the simultaneous testing of all regressors, and the result of the test can be interpreted as indicating whether at least some of the tested regressors are correlated with the error term (i.e. endogenous). Because they involve testing all of the regressors, Hausman test statistics are usually expressed in terms of linear algebra, which is ideally suited for the vectors of parameter estimates and matrices of estimates for the variances and covariances of the estimates involved in the test statistic. However, to give some flavor for the statistic, we roughly cast it in terms of one parameter, β_1. We emphasize that this is not a true Hausman statistic, but instead designed to provide the reader with a very rough sense of the structure of the Hausman statistic in terms of just one regressor.

Specifically, define $\hat{\beta}_1^C$ as the consistent estimate of β_1 generated by the within estimator (typically, the default in most ready-made Hausman tests offered by commercial statistical packages such as STATA is the demeaned fixed effects estimator). Let $\hat{\beta}_1^E$ be the efficient estimate of β_1 from the random effects model. The Hausman statistic would be roughly of the form

$$\frac{\left(\hat{\beta}_1^C - \hat{\beta}_1^E\right)^2}{\hat{var}\left(\hat{\beta}_1^C\right) - \hat{var}\left(\hat{\beta}_1^E\right)}$$

where the $\hat{var}(\hat{\beta}_1^C)$ is the estimated variance of the within estimate of β_1 and $\hat{var}(\hat{\beta}_1^E)$ is the same for the random effects estimate of β_1. Notice that, other things being equal, the larger the difference in the estimate of β_1 generated by the within and random effects models, the larger is the value of this test statistic. The null hypothesis in a Hausman-type test is typically the exogeneity of the regressors, and the larger is the value of the test statistic, the more likely we will reject the null and conclude that there is evidence for correlation between at least some of the regressors and the regression error term.

We re-emphasize that this is not a real Hausman statistic. A real Hausman statistic simultaneously tests all of the parameter estimates for the regressors, taking into account the variances and covariances of all of the estimates. Nonetheless, this rough example hopefully provides some intuitive sense of the structure of the Hausman statistic. Formally, the Hausman statistic is

$$H = \left(\vec{\hat{\beta}^C} - \vec{\hat{\beta}^E}\right)^T \left(\hat{var}^C - \hat{var}^E\right)^{-1} \left(\vec{\hat{\beta}^C} - \vec{\hat{\beta}^E}\right)$$

where

$$\vec{\hat{\beta}^C}$$

$$\vec{\hat{\beta}^E}$$

are, respectively, vectors of the coefficient estimates from a consistent (such as a within estimator) and an efficient (such as random effects) estimator. If there are K regressors common to the two models, then these vectors are $(1XK)$ in dimension. $\hat{var}^C$ and $\hat{var}^E$ are the (KXK) matrices

containing the variance and covariance estimates for estimates of the βs from the two models. The term

$$(\cdot)^{-1}$$

refers to the inverse (in the linear algebra sense) of whatever matrix lies within the parentheses.[32] Hausman (1978) shows that this statistic has a χ^2 statistic under the null hypothesis that the regressors are uncorrelated with the error term.

We reject the null hypothesis of exogenous regressors if the statistic is very large. To get some idea of the logic of this, Figure 5.1 provides an illustration of a χ^2 distributed random variable. As the Figure makes clear, this is a distribution for which only non-negative values have positive probability. Further, smaller values (i.e. those closer to the origin) are more likely than larger values under the χ^2 distribution. For instance, as the Figure illustrates, a test statistic value of 2 has a far higher probability under the χ^2 than a value such as 8. Figure 5.2 illustrates the p-value associated with the statistic value of 8 (as indicated by the shaded area under the curve). Once again, the p-value is the probability of a Type-I error (rejecting a true null hypothesis).

The logic of the Hausman test decision is rather simple. If the null hypothesis is correct, then the test statistic H follows a χ^2 distribution. We reject the null hypothesis if the test statistic H is sufficiently large that it is unlikely to have occurred under the assumed χ^2 distribution (as evidenced, for instance, by a small p-value).

In Outputs 5.16 through 5.18 we illustrate a Hausman test using the three time period extension of the simulated empirical example we have considered thus far in this chapter. We do so using STATA's Hausman test command. The steps in performing the test (both in general and in STATA) are fairly simple. First, estimate the consistent and efficient models. Doing so yields the statistics

$$\hat{\vec{\beta}}^C$$

$$\hat{\vec{\beta}}^E$$

$$\hat{var}^C$$

and

$$\hat{var}^E$$

that are the building blocks of the Hausman test statistic. The final step is, of course, the actual computation of the statistic itself from these components.

Output 5.16 shows the results for the consistent model. This is the fixed effects model as implemented by STATA's `xtreg` command with the `fe` option. The key step following estimation is to save the estimation results. This is done in the line

[32] The inverse in a Hausman test can be a bit complicated, and the crude but illustrative single parameter estimate test statistic example we provided above makes clear a potential problem. That statistic was

$$\frac{\left(\hat{\beta}_1^W - \hat{\beta}_1^R\right)^2}{\hat{var}\left(\hat{\beta}_1^W\right) - \hat{var}\left(\hat{\beta}_1^R\right)}$$

The assumption is that

$$\hat{var}\left(\hat{\beta}_1^W\right) > \hat{var}\left(\hat{\beta}_1^R\right)$$

However, there is no absolute guarantee that this will be the case in a particular estimation sample, resulting in a possibly undefined or negative test statistic because the denominator is zero or negative (both of which are not supported by the χ^2 distribution), with particular concern over the negative possibility. In matric algebra terms, there is the possibility that

$$\hat{var}^C - \hat{var}^E$$

is not positive definite. In that case an alternative approach such as the Moore-Penrose pseudoinverse is usually pursued.

```
estimates store f1
```

This stores in memory the results of the last estimation (in this case the `xtreg , fe` regression) under the name `f1` (which was chosen arbitrarily: it could have been called virtually anything). Output 5.17 does the same for the efficient estimator, in this case STATA's implementation of random effects. This is estimated with the `xtreg` command with the `re` option. In this case, the results are stored under the name `r1`. Finally, Output 5.18 provides results for the Hausman test. The large test statistic value (of 971.61) and corresponding p-value of essentially 0 strongly suggests rejection of the null hypothesis. We can conclude that the evidence (as manifested in the Hausman test statistic) strongly suggests that at least some of the regressors are indeed endogenous. This is a result that should come as no surprise, since our empirical example was engineered to deliver endogeneity (based on a relationship between program participation P and a fixed unobserved characteristic μ relegated to the regression error term).

We conclude with a caveat about Hausman-type tests that is in some sense logically self-evident. The test assesses the evidence for endogeneity by comparing estimates from a consistent estimator and an efficient one, correcting for differences in the variance of the estimates produced by the two. The logic behind the test is essentially that if any regressors are endogenous this will be revealed by the consistent estimates deviating substantially (i.e. beyond what efficiency differences as manifested by different variances for parameter estimates can explain) from the efficient estimates. In other words, the test essentially relies on the consistent estimates revealing endogeneity by deviating from the efficient estimates. But this is a reliable test logic only if the consistent estimates are indeed consistent estimates. For instance, suppose that there was also a time-varying individual level unobserved confounding characteristic. Then the fixed effects estimator (which correct only for endogeneity from fixed confounding unobservables such as μ) would <u>not</u> provide

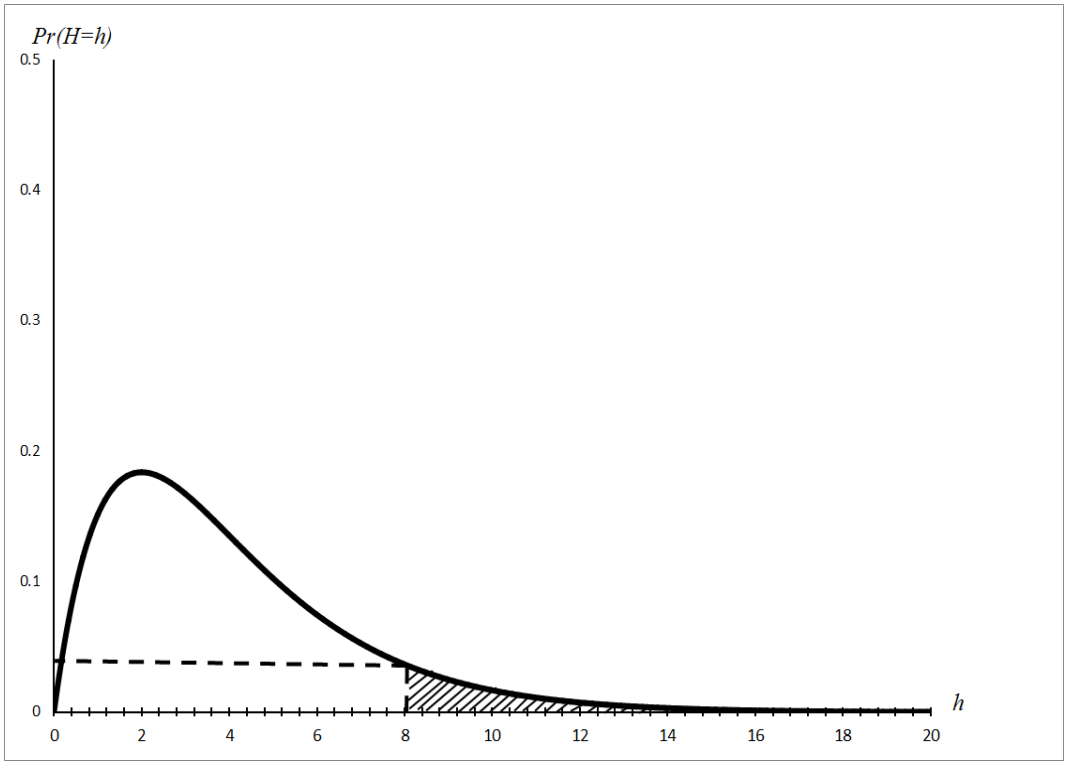


Figure 5.2: The P-value (shaded area) of a χ^2 Test Statistic of 8

consistent estimates.

Acceptance of the Hausman test null hypothesis of no endogeneity thus rests on the assumption that the consistent estimator is indeed consistent. If this is not the case, then the results of the test may not be particularly reliable. In particular, one could conclude from a Hausman test that regressors are exogenous when in fact they are not.

STATA Output 5.16 (5.1.do)

```
. xtreg Y P x_1 x_2 t, fe
note: x_2 omitted because of collinearity
Fixed-effects (within) regression               Number of obs      =     15000
Group variable: ID                              Number of groups   =      5000

R-sq:  within  = 0.6423                         Obs per group: min =         3
       between = 0.1571                                        avg =       3.0
       overall = 0.3393                                        max =         3

                                                F(3,9997)          =   5983.73
corr(u_i, Xb)  = 0.0408                         Prob > F           =    0.0000

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   2.009386   .0794462    25.29   0.000     1.853656    2.165117
         x_1 |   1.504437   .0178087    84.48   0.000     1.469528    1.539345
         x_2 |           0  (omitted)
           t |   3.007264   .0299852   100.29   0.000     2.948487    3.066041
       _cons |   1.914701   .0600506    31.88   0.000      1.79699    2.032412
-------------+----------------------------------------------------------------
     sigma_u |  4.7890736
     sigma_e |  2.9976669
         rho |  .71849437   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4999, 9997) =     9.53          Prob > F = 0.0000
. estimates store f1
```

STATA Output 5.17 (5.1.do)

```
. xtreg Y P x_1 x_2 t, re
Random-effects GLS regression                   Number of obs      =     15000
Group variable: ID                              Number of groups   =      5000

R-sq:  within  = 0.6335                         Obs per group: min =         3
       between = 0.7687                                        avg =       3.0
       overall = 0.7169                                        max =         3

                                                Wald chi2(4)       =  34353.45
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .7600145   .0685936    11.08   0.000     .6255735    .8944555
         x_1 |    1.35674   .0159516    85.05   0.000     1.325475    1.388005
         x_2 |   2.094289   .0179129   116.92   0.000      2.05918    2.129397
           t |    3.00084   .0308106    97.40   0.000     2.940453    3.061228
       _cons |   2.738357    .060991    44.90   0.000     2.618817    2.857897
-------------+----------------------------------------------------------------
     sigma_u |  1.6081443
     sigma_e |  2.9976669
         rho |   .2234789   (fraction of variance due to u_i)
------------------------------------------------------------------------------
. estimates store r1
```

STATA Output 5.18 (5.1.do)

```
. hausman f1 r1, equations(1:1)
                 ---- Coefficients ----
                |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
                |       f1           r1         Difference          S.E.
----------------+----------------------------------------------------------------
              P |    2.009386     .7600145        1.249372        .0400826
            x_1 |    1.504437      1.35674        .1476968         .007918
              t |    3.007264      3.00084        .0064239               .
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      971.61
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

5.2 The Difference-in-Differences Model

We now turn to one particular model generally regarded as fitting most naturally into the family of within estimators: the difference-in-differences estimator. At their intuitive core, difference-in-differences models identify program impact as the difference in the changes in an outcome experienced between participants and non-participants across an interval of time over which a program was introduced. In other words, they identify program impact as the difference in outcome trend between the two groups across an interval of time during which the program was introduced. They are typically categorized as within estimators because they rely, ultimately, on the assumption that any potential unobserved confounding variables are fixed with respect to time and identify program impact with variation in outcomes and participation over time.

5.2.1 The Basic Model

The basic "bare bones" difference-in-differences (DID for brevity) estimator is very simple. At a minimum, one must observe participants and non-participants at two points in time, before a program is introduced and after it is introduced (we will refer to these two points as time periods $t = 1$ and $t = 2$, respectively). Although we observe participants and non-participants before and after the initiation of a program, it is critical that they can be empirically differentiated even before the program commences.[33] Define

$$\overline{Y}_t^P$$

as the average (across some suitable representative sample) of an outcome of interest for participants at time period t (where $t = 1, 2$ in our simple example). Define

$$\overline{Y}_t^N$$

[33] We should also mention that "program commencement" need only occur in the population of reference for the impact evaluation using the difference-in-differences estimator. For instance, the program might be long running but access to it expands due to changes in eligibility requirements, geographic coverage, etc. such that for the population subject to the expansion distinct pre- and post-program availability periods can be defined.

analogously for non-participants. The DID estimate of program impact is then

$$\overline{Y}_2^P - \overline{Y}_1^P - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

From this expression the origin of the term "difference-in-differences" should be evident: the estimator is the difference in the differences over time (between the pre- and post-program implementation phases) for the participant and non-participant groups.

The DID estimator of program impact is thus the difference in changes over time in outcomes between participants and non-participants. In other words, any difference in changes over time between participants and non-participants can be ascribed to the program. But this implies that the changes over time experienced by non-participants are indicative of the changes participants would have experienced had they not begun participating after time period 1. The assumption that the changes in the outcome over time among non-participants indicate the changes that would have occurred among non-participants had they not participated after time period 1 (or, put more simply, that had they not participated participants would have experienced the same average change in the outcome over time that non-participants did experience) is called the **parallel trend assumption**.

The basic logic of this estimator is illustrated in Figures 5.3-5.6. We begin with Figure 5.3, which illustrates the trajectory of an outcome Y (specifically the trajectory of the average value of Y) between two points in time, $t = 1$ and $t = 2$, for participants and non-participants. Between $t = 1$ and $t = 2$ the program is introduced. Hence when we distinguish between participants and non-participants we do so with reference to the participation decision made *after* the introduction of the program; no one can participate *before* the program is introduced. Notice that the non-participant sample experienced a lower average outcome over time than participants, but there was some upward trend in the average outcome even for non-participants (in other words, there was some upward trend in the average outcome for non-participants, even though they did not avail themselves of the program after its introduction).

Figure 5.4 makes more explicit the average outcomes for participants and non-participants over time. Following the discussion above, the average outcome is denoted

$$\overline{Y}_t^k$$

where $t = 1, 2$ are the two time periods and $k = P, N$ denote participants and non-participants, respectively. A simple and straightforward estimator of program impact might be

$$\overline{Y}_2^P - \overline{Y}_2^N$$

This estimator is simply the comparison of average outcomes between participants and non-participants at time $t = 2$. We can see from this Figure that this might be a problematic estimator of program impact. Given that participants and non-participants exhibited different average outcomes at time period $t = 1$ (with $\overline{Y}_1^P > \overline{Y}_1^N$) it is not at all clear that the average level of Y non-participants experienced at time period $t = 2$ are indicative of what participants would have experienced at $t = 2$ in the absence of the program. There is thus no reason to believe that this is an unbiased estimator of the average effect of treatment on the treated.[34] Moreover, to the extent that the experiences of participants might not be indicative of what non-participants would have experienced had they participated at time period $t = 2$, it is likely not an unbiased estimator of the average treatment effect[35] either.

[34] In this context, the average effect of treatment on the treated would be meaningful at time $t = 2$ and hence be $E(Y_2^1 - Y_2^0 | P = 1)$, where Y^1 and Y^0 are defined as in the beginning of this chapter and all preceding chapters: as the outcome when individuals (in this case participants at time period $t = 2$) participate and do not do so, respectively.

[35] Which is given by $E(Y^1 - Y^0)$.

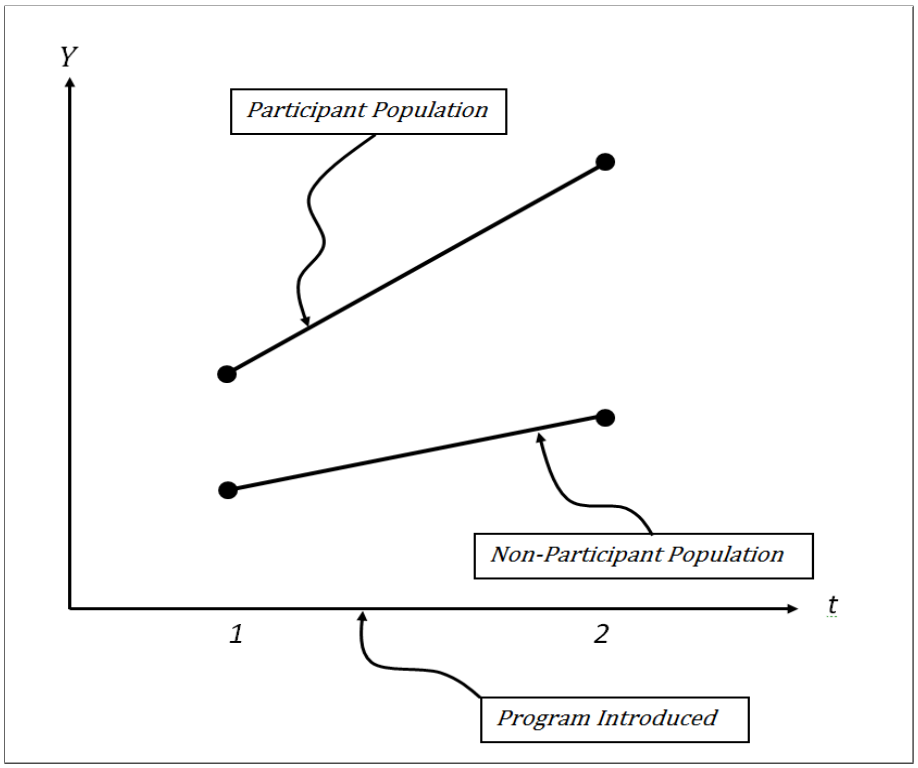


Figure 5.3: Outcome Trajectories, Participants and Non-Participants

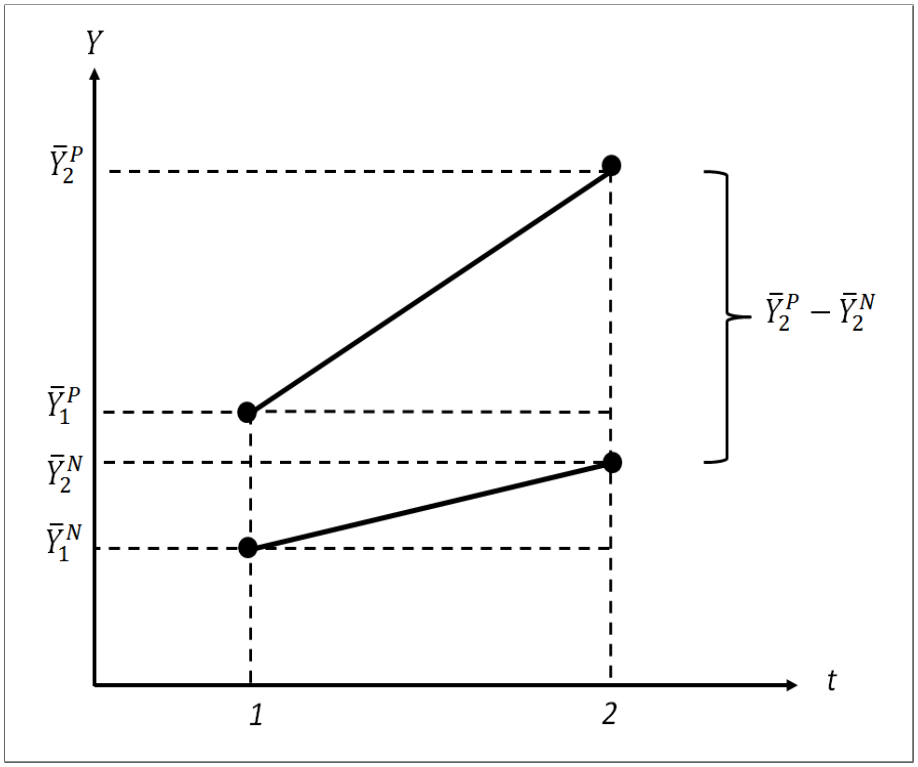


Figure 5.4: Changes Over Time, Participants and Non-Participants

Figure 5.5 illustrates the parallel trend assumption. As we have seen, the solid lines between the dots represent the actual evolution of the average outcome for participants and non-participants. The double dotted line illustrates the parallel trend assumption. This is the assumed evolution of the outcome for participants had they not participated. Notice that it is parallel to the line indicating the actual change in the average outcome for non-participants. Thus it assumes that the *change over time* for participants would, had they not begun participating after time period 1, have been the same as that for non-participants. The parallel nature of the lines is also the origin of the expression "parallel trend".

The change over time represented by the parallel trend assumption is represented in Figure 5.5 by Δ. The parallel trend assumption is that

$$\Delta = \overline{Y}_2^N - \overline{Y}_1^N$$

In other words, the changes implied by the parallel trend assumption equal those that non-participants actually did experience.

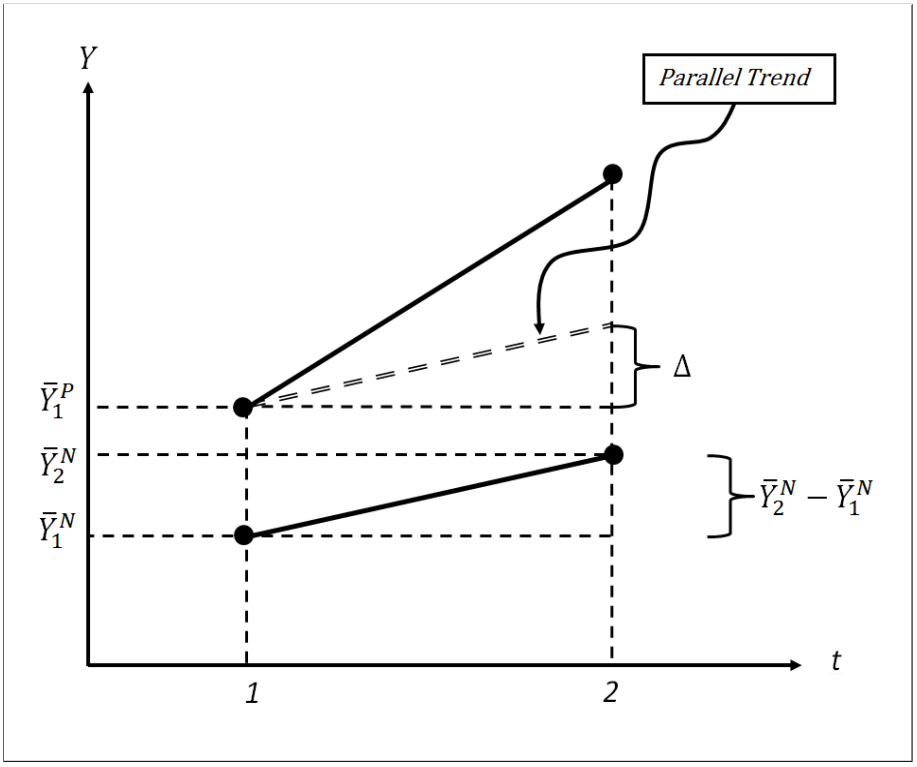


Figure 5.5: The Parallel Trend Assumption

Finally, Figure 5.6 rounds out the graphical discussion. The DID estimator is

$$\overline{Y}_2^P - \overline{Y}_1^P - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

However, the parallel trend assumption tells us that

$$\Delta = \overline{Y}_2^N - \overline{Y}_1^N$$

Hence

$$\overline{Y}_2^P - \overline{Y}_1^P - \left(\overline{Y}_2^N - \overline{Y}_1^N\right)$$

$$\overline{Y}_2^P - \overline{Y}_1^P - \Delta$$

Thus, the DID estimate of program impact is the change over time that participants experienced minus an estimate of the change participants would have experienced in the absence of the program. The parallel trend assumption allows us to form an estimate of that counterfactual change from the change over time that non-participants actually did experience.

Typically, a regression version of the DID model is implemented. To fix ideas, let us continue to assume that we observe participants and non-participants at only two time period, $t = 1$ and $t = 2$. The classic regression model is

$$Y_{it} = \beta_0 + \beta_1 \cdot P_i \cdot d_t + \beta_2 \cdot P_i + \beta_3 \cdot d_t + \epsilon_{it}$$

where d_t equals 1 if $t = 2$ and 0 if $t = 1$. P_i equals 1 if individual i is *ever* a participant in the program (which in the context of this example means if they participate at time $t = 2$, after the program is introduced). ϵ_{it} is an error term, independently and identically distributed across individuals and within each individual across time. As we will see, the consequences of violations of this assumption (particularly in the form of correlation of the errors over time for the same individual) can carry rather severe consequences for the performance of the DID model.

In the context of this simple regression model, the regressor P_i (as it enters the regression specification through the term $\beta_2 \cdot P_i$) controls for fixed average differences between participants and non-participants. The regressor d_t (as it enters the regression specification through the term $\beta_3 \cdot d_t$) controls for the common time trend in the outcome Y between participants and non-participants. Finally, the interaction of P_i and d_t (as it enters through the regression specification through the term $\beta_1 \cdot P_i \cdot d_t$) is program impact. Basically, the regressor formed by the interaction $P_i \cdot d_t$ can be viewed as capturing the difference in observed trend in the outcome Y between the participant and

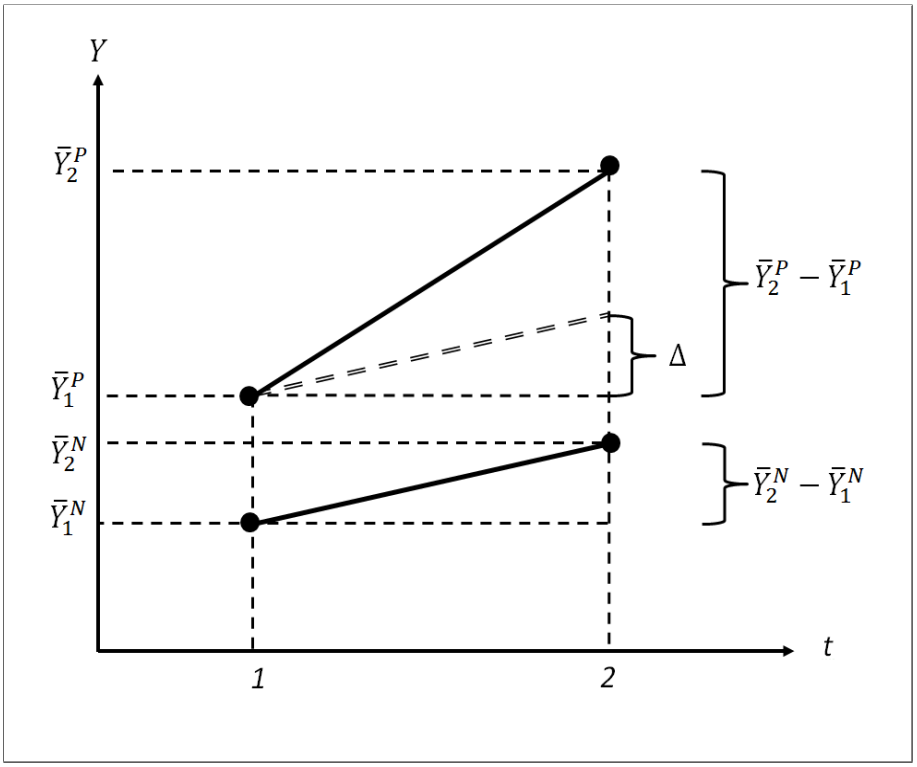


Figure 5.6: The Difference-in-Differences

non-participant samples. However, if the parallel trend assumption holds then the participant and non-participant subpopulations would have experienced the same trends (i.e. the same differences over time) in the outcome Y. Therefore, if the parallel trend assumption is reasonable, then any difference in observed trends over time between the participant and nonparticipant groups (as manifested by a non-zero estimate $\hat{\beta}_1$ of β_1) must reflect average program impact.

STATA do-file 5.2.do provides a simulated empirical illustration of the DID model in action. The basis for the data generating process behind the simulation are the following equations for the potential outcomes (Y^1 and Y^0) and cost of enrollment (C):

$$Y_t^1 = 6 + 2.5 \cdot x_1 + t + \mu + \epsilon_t^Y$$

$$Y_t^0 = 4 + 2.5 \cdot x_1 + t + \mu + \epsilon_t^Y$$

$$C_t = 1 - 2.5 \cdot x_1 - 2 \cdot \mu + \epsilon_t^C$$

where x_1 and μ are $\sim$ N (0,4) are time-invariant characteristics (x_1 is "observable" while μ is not). The ϵs are $\sim$ N(0,9) variables drawn independently across individuals and time. We draw 20,000 longitudinal observations. DID models are typically estimated with longitudinal data since it is often only in the context of such data that we can differentiate who will and will not become participants *before* a program is introduced.[36] Once again, these specific values for the parameters of this system are essentially randomly chosen (and we encourage the reader to experiment with them in the .do file).

We consider two time periods, $t = 1$ and $t = 2$. The program is not available at time $t = 1$ (i.e. no one can participate at time period $t = 1$). Hence the observed outcome Y_t at time $t = 1$, Y_1 equals Y_1^0 for all individuals (regardless of their participation decision in time period $t = 2$).

STATA Output 5.19 (5.2.do)

```
. * Basic summary statistics: participation
. by t, sort: tab P
```

```
-> t = 1
          P |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      8,996       44.98       44.98
          1 |     11,004       55.02      100.00
------------+-----------------------------------
      Total |     20,000      100.00
```

```
-> t = 2
          P |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      8,996       44.98       44.98
          1 |     11,004       55.02      100.00
------------+-----------------------------------
      Total |     20,000      100.00
```

The program becomes available between time periods 1 and 2, and individuals participate at time period 2 if

$$Y_2^1 - Y_2^0 - C_2 > 0$$

[36] Longitudinality is not, however, absolutely essential per se to the identification of the DID model. Below we will discuss briefly the identification of the DID model with non-longitudinal data.

Individuals can only participate at time period $t = 2$. No one participates in the program at time $t = 1$ while some elect to do so at time $t = 2$. However, to set up the variable to control for participant status, P, we let events in time period 2 dictate the value assigned to P in both time periods. Thus $P = 1$ if the individual *ever* decides to participate once the program becomes available to them, which in the context of this example means if they do so at time $t = 2$. $P = 0$ if the individual *never* decides to participate.

In Output 5.19 we present basic summary statistics regarding participation. 55.02% elect to participate at time period $t = 2$. However, as the variable P is constructed, this means that 55.02% have the value $P = 1$ at time period $t = 1$ as well.

Output 5.19 provides summary statistics by time t and participant status P. Had the program been available at time $t = 1$ it is not necessarily the case that the same individuals that elected to participate at time $t = 2$ would have done so at $t = 1$ as well.

If nothing else, a given individual's draws for ϵ_t^C might be such that they would have elected to participate in the program at one of but not both time periods. Nonetheless there are two factors, x_1 and μ, that influence the participation decision and are not time-varying. Hence we would expect some differences between participants and non-participants.

Such differences are evident in Output 5.20. The participants have, unsurprisingly, larger average values for the time invariant determinants of potential outcomes and cost x_1 and μ (this is unsurprising, of course, because larger values for these drive down the cost of participation). This contributes to higher average values to the potential outcomes for participants. Participants are *persistently* different from non-participants, despite the fact that participation is a meaningful choice only at time period $t = 2$.

This is due to the role of the persistent characteristics x_1 and μ in shaping the cost of participation and hence the participation decision. In other words, though the exact subset participating might have differed had the participation decision been made at $t = 1$ (due to the aforementioned fluctuations in ϵ_t^C between $t = 1$ and $t = 2$) the same kinds of individuals on average will tend to participate regardless of the time period in which the participation decision is made. Finally, note that the average values of the potential outcomes Y^1 and Y^0 are increasing over time. This is due to the time trend.

In Output 5.21 we consider the average value of the observed outcome Y over time. Clearly, it is rising (from 4.96787 to 7.12163) as events progress from $t = 1$ to $t = 2$.

We have already seen one possible reason: the positive time trend to the potential outcomes Y^1 and Y^0 that are behind the observed outcome Y. However, it is also possible that this has to do with some individuals deciding to participate at time period $t = 2$ and avail themselves of a positive return from treatment.

In this case, that true positive return from treatment (i.e. program participation) at time period $t = 2$ is

$$Y_2^1 - Y_2^0 = 6 + 2.5 \cdot x_1 + 2 + \mu + \epsilon_2^Y - \left(4 + 2.5 \cdot x_1 + t + \mu + \epsilon_2^Y\right) = 2$$

Although it is not shown in an Output table, the average individual-level program impact at time period $t = 2$ in our simulated sample is indeed just about 2. We thus have our benchmark for thinking about the results offered for alternative program impact estimators.

In Output 5.22 we turn to actual estimation of average program impact. First, we take the average of the observed variable at the various possible permutations of P and t. In other words, we take the average of the outcome at every possible combination of values for P and T: $\{P = 1, T = 2\}$, $\{P = 0, T = 2\}$, $\{P = 1, T = 1\}$ and $\{P = 0, T = 2\}$. We then consider two estimators of program impact.

STATA Output 5.20 (5.2.do)

```
. * Basic summary statistics
. by t P, sort: summarize Y y1 y0 c x_1 mu epsilon*
```

-> t = 1, P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	8996	.864926	4.798989	-19.51024	19.01579
y1	8996	2.864926	4.798989	-17.51024	21.01579
y0	8996	.864926	4.798989	-19.51024	19.01579
c	8996	6.111167	5.274775	-11.20045	29.50878
x_1	8996	-1.242994	1.619078	-8.141914	4.645814
mu	8996	-1.014855	1.774254	-7.726572	6.37245
epsilony	8996	-.0127333	2.931612	-10.977	10.89786
epsilonc	8996	-.0260286	3.025866	-10.36658	12.19946

-> t = 1, P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	11004	8.322113	4.987734	-8.958962	28.28016
y1	11004	10.32211	4.987734	-6.958962	30.28016
y0	11004	8.322113	4.987734	-8.958962	28.28016
c	11004	-3.182799	5.403687	-24.35852	15.77308
x_1	11004	1.018218	1.678373	-4.997754	8.420812
mu	11004	.8050956	1.801293	-5.138058	7.586161
epsilony	11004	-.0285286	3.006486	-11.73894	12.81559
epsilonc	11004	-.0270622	2.992002	-11.01307	11.64383

-> t = 2, P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	8996	1.887782	4.782555	-18.25732	17.64407
y1	8996	3.887782	4.782555	-16.25732	19.64407
y0	8996	1.887782	4.782555	-18.25732	17.64407
c	8996	7.308511	4.082521	2.000105	31.73855
x_1	8996	-1.242994	1.619078	-8.141914	4.645814
mu	8996	-1.014855	1.774254	-7.726572	6.37245
epsilony	8996	.0101228	2.944227	-12.33223	11.07838
epsilonc	8996	1.171316	2.790233	-10.25706	11.07769

-> t = 2, P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	11004	11.40041	4.946977	-9.651225	31.53767
y1	11004	11.40041	4.946977	-9.651225	31.53767
y0	11004	9.40041	4.946977	-11.65123	29.53767
c	11004	-3.988187	4.39947	-27.70955	1.999446
x_1	11004	1.018218	1.678373	-4.997754	8.420812
mu	11004	.8050956	1.801293	-5.138058	7.586161
epsilony	11004	.0497687	3.011935	-11.02399	11.3459
epsilonc	11004	-.8324501	2.803536	-11.03448	9.127626

The first simple comparison of average outcomes between participants and non-participants at

time period $t = 2$ in our simulated sample is the sample analog to the population-level estimator:[37]

$$E\left(Y_2|P=1\right) - E\left(Y_2|P=0\right)$$

In the context of our earlier graphical discussion of the DID estimator, this is equivalent to

$$Y_2^P - Y_2^N$$

The estimate of average program impact from this simple (and naive) estimator if 9.512628. Given that true average impact is 2, this represents a wild overestimate of impact.

STATA Output 5.21 (5.2.do)

```
. * Basic summary statistics
. by t, sort: summarize Y
```

-> t = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	20000	4.96787	6.148853	-19.51024	28.28016

-> t = 2

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	20000	7.12163	6.793191	-18.25732	31.53767

The other estimate presented in Output 5.22 is the sample analog to the basic population DID estimator:

$$E\left(Y_2|P=1\right) - E\left(Y_1|P=1\right) - \left(E\left(Y_2|P=0\right) - E\left(Y_1|P=0\right)\right)$$

In the context of the earlier graphical discussion of the DID model, this is equivalent to

$$Y_2^P - Y_1^P - \left(Y_2^N - Y_1^N\right)$$

The DID estimate of program impact is 2.0554412, which is extremely close to the true value of 2.

With Output 5.23 we begin considering program impact through the lens of regression. Specifically, Output 5.23 reports results from a simple regression of Y on P using the subsample from time period $t = 2$.[38] The estimate of program impact is, at 9.512628, the same as that yielded by the simple, naive estimator just considered (i.e. simple comparison of average outcomes between participants and non-participants at time $t = 2$). The most obvious reason that this estimate and the simple comparison based one with which it is numerically identical deviate so greatly from the true value (of 2) is due to the different types by x_1 and μ in the participant and non-participant

[37] That statement is something of a mouthful, but the basic idea is quite simple. We wish to learn about population-level parameters, which are defined across the population and, hence, expectations $E(\cdot)$ across the population are the most appropriate and direct way of approaching them. Estimation attempts to learn something about these population-level parameters from samples from those populations, which are typically (and essentially by definition) smaller than the population itself. In these samples we form estimates of expectations by forming averages across the sample. Hence our objective is an expectation, but we estimate it with a sample average, which is the analog of a population expectation in the sample context.

[38] It would not really be beneficial to add the subsample from time period $t = 1$. This is a regression that essentially considers cross-section (i.e. across individuals) variation in levels of Y and P. There is no meaningful cross sectional variation in program participation at time $t = 1$ since participation was not an option at that time.

subsamples. In Output 5.24 we add the observable x_1 as a control, but the resulting estimate of program impact is, at 4.75097, 240 percent of the true value of 2. Addition of the one fixed determinant of potential outcomes that we can observe thus cuts some of the distance between estimated and actual average program impact, but does not close that gap altogether.

STATA Output 5.22 (5.2.do)

```
. * Basic IE estimate and DID estimate
. mean Y if P==1&t==2
Mean estimation                     Number of obs    =   11004
------------------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Y |   11.40041    .047159      11.30797    11.49285
------------------------------------------------------------------------------

. loc YP2=_b[Y]
. mean Y if P==1&t==1
Mean estimation                     Number of obs    =   11004
------------------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Y |   8.322113   .0475475      8.228911    8.415314
------------------------------------------------------------------------------

. loc YP1=_b[Y]
. mean Y if P==0&t==2
Mean estimation                     Number of obs    =    8996
------------------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Y |   1.887782   .0504238       1.78894    1.986624
------------------------------------------------------------------------------

. loc YN2=_b[Y]
. mean Y if P==0&t==1
Mean estimation                     Number of obs    =    8996
------------------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Y |    .864926    .050597      .7657443    .9641077
------------------------------------------------------------------------------

. loc YN1=_b[Y]
. loc basic=`YP2´-`YN2´
. di `basic´
9.512628
. loc DID=`YP2´- `YP1´ - (`YN2´-`YN1´)
. di `DID´
2.0554412
```

Output 5.25 provides results from the DID regression. The estimate of program impact is the interaction of the program participation indicator and a dummy variable indicating whether an observation is from time period $t = 2$ or not. In the STATA output, this is the estimate associated with the term `Pdt` (as in P times a dummy variable for time $t = 2$). At 2.055441 that estimate is exactly the same as that generated by the simple DID estimate presented in Output 5.22. The regression based approach is in this instance essentially the same as that associated with difference-in-differences in averages associated with the various permutations of the values of time t and program participation P.

STATA Output 5.23 (5.2.do)

```
. * Basic IE regression
.
. reg Y P if t==2
      Source |       SS       df       MS              Number of obs =   20000
-------------+------------------------------           F(  1, 19998) =18856.11
       Model |  447889.683     1  447889.683           Prob > F      =  0.0000
    Residual |  475012.951 19998  23.7530228           R-squared     =  0.4853
-------------+------------------------------           Adj R-squared =  0.4853
       Total |  922902.634 19999  46.1474391           Root MSE      =  4.8737

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    9.512628   .0692747   137.32   0.000     9.376844    9.648412
       _cons |    1.887782   .0513848    36.74   0.000     1.787064    1.988501
------------------------------------------------------------------------------
```

STATA Output 5.24 (5.2.do)

```
. * Basic IE multiple regression
.
. reg Y P x_1 if t==2
      Source |       SS       df       MS              Number of obs =   20000
-------------+------------------------------           F(  2, 19997) =29603.57
       Model |  689893.562     2  344946.781           Prob > F      =  0.0000
    Residual |  233009.072 19997  11.6522014           R-squared     =  0.7475
-------------+------------------------------           Adj R-squared =  0.7475
       Total |  922902.634 19999  46.1474391           Root MSE      =  3.4135

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |     4.75097   .0587015    80.93   0.000      4.63591     4.86603
         x_1 |    2.105798    .014612   144.11   0.000     2.077158    2.134439
       _cons |    4.505278   .0403131   111.76   0.000     4.426261    4.584295
------------------------------------------------------------------------------
```

5.2.2 Extensions and Complications

The Parallel Trend Assumption

In many applications the parallel trend assumption might be questionable. The premise of the DID model is that, while there might be fixed differences between participants and non-participants, both groups would experience the same changes in the outcome over time. This is tantamount to assuming that any time-varying unobserved determinants of potential outcomes Y^1 and Y^0 either do not exhibit a discernible trend (e.g. are identically distributed over time) or that any trend that they do exhibit is the same among participants and non-participants.

This might not be a realistic assumption. For instance, suppose that income influences both program participation and the potential outcomes. To fix ideas, suppose as well that those who elect to participate in a program after its introduction tend to be poor. Suppose as well that income growth is particularly slow for the poor. Then the non-participants might have experienced more

rapid income growth over time, and hence different trends in potential outcomes. If we cannot observe and hence control for income over time this could present a big problem.

STATA Output 5.25 (5.2.do)

```
. * Basic DID regression
.
. g dt=t-1
. ta dt
```

dt	Freq.	Percent	Cum.
0	20,000	50.00	50.00
1	20,000	50.00	100.00
Total	40,000	100.00	

```
.
. g Pdt=P*dt
.
. reg Y P dt Pdt
```

Source	SS	df	MS	Number of obs = 40000
Model	769521.898	3	256507.299	F(3, 39996) =10732.60
Residual	955897.531	39996	23.8998282	Prob > F = 0.0000
				R-squared = 0.4460
				Adj R-squared = 0.4459
Total	1725419.43	39999	43.1365641	Root MSE = 4.8887

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	7.457187	.0694884	107.32	0.000	7.320988	7.593386
dt	1.022856	.0728933	14.03	0.000	.8799835	1.165729
Pdt	2.055441	.0982715	20.92	0.000	1.862827	2.248056
_cons	.864926	.0515434	16.78	0.000	.7638998	.9659522

The parallel trend assumption is essentially untestable with only two waves of observations (one before and one after program introduction). If more than one wave of observations was observed before the introduction of the program, one could test whether the change in average outcomes over pre-introduction time periods is the same between participants and non-participants. Of course, this is not an exact test of the parallel trend assumption as required by the DID estimator since there is still the possibility that the trends in the absence of the program might have begun to diverge after the last wave of observations observed before the data was introduced. For instance, a program might be introduced because it has been projected that two populations that had kept apace of each other in terms of outcome trajectory might diverge. On the other end of the observational window, the trouble with testing the parallel trend assumption with post-enrollment data is that we might be picking up fundamental trends or the gradually mounting impact of the program (remember that the effect of many human resource-related programs might be only very gradually felt in the period after enrollment).

Another possible wrinkle is that one might elect to participate because the interval before participation begins was somehow exceptional. A case in point is **Ashenfelter's dip**, whereby participants in adult job training programs have been observed to experience a dip in earnings prior to their enrollment (Ashenfelter and Card (1985)). Individual enrollment in a program might be driven in part by transitory shocks. This makes it extremely difficult to be sure that the pre-participation outcomes observed for participants are indicative of what might have happened in the absence of participation, after the temporary circumstances of the pre-enrollment shock have

subsided.

Data Structure

In our numerical example involving the DID model, the simulated data was longitudinal (i.e., in the context of that example, each individual was observed twice). Longitudinal data is not necessarily required to operationalize the basic DID estimator. A DID model could in principle be estimated with either a longitudinal sample or repeated cross sectional samples from the population for whom program impact is to be estimated. What is required is simply that one has representative samples from that population before and after program implementation and participants and non-participants can be differentiated *even before participation begins*. The latter condition is often difficult to satisfy outside of the context of longitudinal data and probably the major reason that longitudinal data features so prominently in evaluations using the DID approach.

One advantage of longitudinal data is that it can provide some further flexibility in terms of modelling. For instance, it allows for the estimation of a full "fixed effects" version of the DID regression specification. A typical specification is along the lines of

$$Y_{it} = \beta_0 + \beta_1 \cdot P_i \cdot d_t + \sum_{j=2}^{N} \phi_j \cdot d_j + \beta_3 \cdot d_t + \epsilon_{it}$$

where

$$\sum_{j=2}^{N} \phi_j \cdot d_j$$

is defined as in the dummy variable version of the straightforward fixed effects regression model discussed earlier in this chapter. Below we will encounter an instance where this estimator might be useful.

Many Time Periods and Programs

The DID regression model is easily extended to many time periods, and many programs. We dispense with the latter first. Consider two programs, program A and program B. Participation in these is captured by the dummy variables P^A and P^B. A DID specification that allows for consideration of the impact of each of these programs (against the alternative of not participating) is

$$Y_{it} = \beta_0 + \beta_{1A} \cdot P_i^A \cdot d_t + \beta_{1B} \cdot P_i^B \cdot d_t + \beta_{2A} \cdot P_i^A + \beta_{2B} \cdot P_i^B + \beta_3 \cdot d_t + \epsilon_{it}$$

It is also trivial to evaluate the possibility that the impact of A depends on participation in B by adding the term

$$\beta_{1AB} \cdot P_i^A \cdot P_i^B \cdot d_t$$

If one wanted simply to estimate the impact of participation in A compared with what would have occurred in B, then the original DID specification would suffice, with those participating in B treated in much the same fashion as non-participants were approached in that original specification.

The extension to many time periods is also quite straightforward. To fix ideas, define d_t^j to be a dummy variable that equals 1 if $t = j$ and 0 otherwise. For instance, consider four time periods, $t = 1, 2, 3, 4$. The define three dummy variables d_t^2, d_t^3 and d_t^4 such that, for example, d_t^3 equals 1 if an observation is drawn from time period 3 (i.e. $t = 3$) and 0 otherwise (i.e. if $t = 1$, $t = 2$ or

$t = 4$). Suppose as well that the program is introduced between time periods $t = 2$ and $t = 3$. A DID specification might be along the lines of

$$Y_{it} = \beta_0 + \beta_{12} \cdot P_i \cdot d_t^2 + \beta_{13} \cdot P_i \cdot d_t^3 + \beta_{14} \cdot P_i \cdot d_t^4 + \beta_2 \cdot P_i + \beta_{32} \cdot d_t^2 + \beta_{33} \cdot d_t^3 + \beta_{34} \cdot d_t^4 + \epsilon_{it}$$

Since the program had not yet been introduced at time period $t = 2$ we would not expect there to be any difference in the trend in the outcome between participants and non-participants at that point under the parallel trend assumption. Thus, a significance test on $\hat{\beta}_{12}$ (specifically, a test of whether the estimate is significantly different from zero) is tantamount to a test of the parallel trend assumption. To the extent that the estimate is not statistically significantly different from zero, the test result is consistent with the parallel trend assumption (though, as discussed above, it does not preclude violation of it).

A test of whether the estimates $\hat{\beta}_{13}$ and $\hat{\beta}_{14}$ differ significantly[39] has interesting implications: assuming that all who participate make the decision to do so after $t = 2$ but before $t = 3$, this is basically a test of whether program impact is evolving. If one is certain that the treatment effect would not be cumulative (or declining) with time but instead a discrete, one-time increase in Y, one could estimate the more **parsimonious**[40] specification

$$Y_{it} = \beta_0 + \beta_{12} \cdot P_i \cdot d_t^2 + \beta_{134} \cdot P_i \cdot d_t^{34} + \beta_2 \cdot P_i + \beta_{32} \cdot d_t^2 + \beta_{34} \cdot d_t^3 + \beta_{34} \cdot d_t^4 + \epsilon_{it}$$

where d_t^{34} equals 1 if $t = 3$ or $t = 4$.

In the last example, everyone who participated initiated participation at the same point in time (between time periods $t = 2$ and $t = 3$). DID regression can also accommodate the possibility that the participant subsample initiates participation at various points in time. One possibility would be to categorize participants according to their point of initiation of participation. For instance, suppose that waves of data were collected at four points in time: 2009, 2010, and 2011. Suppose as well that the program was introduced at the end of 2009 but participants initiated participation gradually, with some doing so immediately and some waiting until 2011. One could simply define dummy variables P_i^j where $j = 2010, 2011$ to identify participants who initiated participation in a given year indicated by the value of j. A possible specification would then be

$$Y_{it} = \beta_0 + \beta_{1,2010,2010} \cdot P_i^{2010} \cdot d_t^{2010} + \beta_{1,2010,2011} \cdot P_i^{2010} \cdot d_t^{2011} + \beta_{1,2011,2011} \cdot P_i^{2011} \cdot d_t^{2011}$$

$$+\beta_{2,2010} \cdot P_i^{2010} + \beta_{2,2011} \cdot P_i^{2011} + \beta_{3,2010} \cdot d_t^{2010} + \beta_{3,2011} \cdot d_t^{2011} + \epsilon_{it}$$

where, for instance, d_t^{2011} equals 1 if the observation is drawn from the year $t = 2011$ and zero otherwise.

[39] Formally, the null hypothesis would be

$$H^0 : \beta_{13} = \beta_{14}$$

while the alternative hypothesis would be

$$H^a : \beta_{13} \neq \beta_{14}$$

This test is a simple extension of the basic significance test discussed in the preceding chapter.

[40] In statistics, one specification or model is more parsimonious than another if it presents fewer parameters to be estimated. Often a more parsimonious specification is, as in the case under discussion, somehow a restricted version of the less parsimonious alternative. Parsimony has the advantage of, for instance, possibly more precise (i.e. with lower standard errors) estimates as the information contained in the observations is focused on precise estimation of fewer parameters. On the other hand, invalid restrictions can involve serious consequences, including the possibility of bias where none existed in the less parsimonious model.

Serial Correlation

A recent paper (Bertrand et al. (2004)) has called attention to a potentially important but generally overlooked flaw in most conventional applications of the DID estimator. Bertrand et al. essentially take for granted the unbiasedness of the estimates of program impact emerging from the standard DID model. Instead, they focus on the estimate for the standard error of that impact, which will be the basis for any inference about whether the program had a statistically significant impact. They find that with serially correlated data (i.e. data where the random shocks ϵ_{it} are correlated over time) DID models have an alarming tendency to deliver standard error estimates that suggest the program had a significant impact even when it did not actually have one (a type-I error). Serially correlated errors are a potentially important feature in many of the samples to which we might apply the DID model. For instance, many human resource outcomes such as health might exhibit a strong persistence in terms of unobserved shocks.

Bertrand et al. (2004) illustrate the severity of this problem with a rather dramatic example using the Current Population Survey. Essentially, they introduce placebo laws (i.e. randomly assigning fictitious laws to various observations within their CPS sample) and find that in an astonishingly large percentage of cases the DID model suggested that the laws had a significant impact on wages! To remedy this problem, the authors offer a number of suggestions. One possibility would be to collapse the data from many time periods before and after implementation to just a simple pre-/post-sample. The potential advantage of this is that a long string of pairwise highly correlated errors might be reduced to two "aggregate" error terms, one from before implementation of the program and one after, that might not be so highly correlated.

However, the authors appear to favor a Huber-White type estimator. A Huber-White type estimator in this context[41] corrects for violations of one of the major assumptions behind the basic formula for estimation of standard errors for linear regression estimates. Specifically, the standard formulas all assume that observations are independent. Correlation of regression errors (the ϵs) clearly violates this assumption. The Huber-White adjustment to standard errors is a *post-estimation* procedure, and so does not change the value of the estimates of regression parameters (for the most part, the βs in the regression discussions to this point) but only the estimates of their standard errors.

Huber-White standard errors are supported by nearly all commercial statistical packages-refer to the documentation of your preferred program for more information. For instance, in STATA they are implemented using the `cluster` option. For instance, for the last regression in the DID simulation the proper syntax would be

```
reg Y P dt Pdt, cluster(ID)
```

where `ID` was the individual identifier. We would not expect the Huber-White clustering adjustment to make much difference in the context of the DID example in this chapter since, in the context of that example, the time-varying regression error was uncorrelated over time.

Another possible solution might involve implementing the dummy variable fixed effects specification of the DID estimator that was discussed in a preceding subsection. This would allow for a persistent component to the error term over time for each individual.

[41] Another purpose to which Huber-White type estimators are put is correction for heteroskedasticity. Heteroskedasticity is the condition where the error variance is not constant across observations (when it is constant, we have the circumstance of homoskedasticity or homoskedastic errors). Homoskedastic errors are the other major assumption behind the formulas that produce the standard error estimates produced by linear regression routines in statistical packages such as STATA.

5.2.3 Experimental Samples

The DID estimator is often applied in the setting of experimental samples within which program participation is (typically purposefully) randomly assigned. The typical study design template involves a sample from a population of interest who are observed before a program is introduced and after it has been introduced with random determination of program participation. The model estimated is then typically

$$Y_{it} = \beta_0 + \beta_1 \cdot P_i \cdot d_t + \beta_2 \cdot P_i + \beta_3 \cdot d_t + \epsilon_{it}$$

where the regressors are defined in the same fashion as was the case when this basic DID regression model was first introduced. A key advantage of this before-after, program-control data design is that when this DID specification is estimated with data from such a design there is a natural test of whether randomization was indeed successful. Specifically, if randomization of participation is successful there should not be average fixed differences between participants and non-participants and hence $\hat{\beta}_2$ should not be statistically significantly different from zero. Alternatively, when it is significantly different from zero (i.e. when randomization may not have been fully successful) the DID model can offer a fall-back position for estimation of program impact by quasi-experimental means.

5.3 Some Closing Thoughts

Within estimators, including the difference-in-differences estimators, have an incredibly attractive feature: they do not require much in the way of explicit modelling of the program participation decision. This is in contrast to approaches such as instrumental variables, which will be discussed in the next chapter. This is a major selling point for within estimators: they offer the potential for identification of program impact while avoiding the possible need for many, possibly elaborate assumptions regarding the "structure" of individual behavior regarding the joint determination of outcomes and participation. By and large, they achieve this by focusing on variation within individuals (or other units of observation, depending on the application). The individual thus in some sense serves as a control for themselves.

However, it must be recognized that the price of this is a somewhat awkward assumption: that there is sufficient within variation in outcomes and participation to make these models practically estimable, but that there is no substantial variation in the potential unobserved confounders. In other words, we assume that outcomes and participation vary a lot over our observational window for the individual, but the things that are unobserved and related to outcomes and participation are relatively static. There is something almost implausible to this notion. Thus, while within models allow one to avoid elaborate modelling of the joint determination of outcomes and participation, their "black box" appeal to a fixed unobservable confounder is not a free ride in the sense that that notion seems inherently incongruous with the requirements (in terms of variation) of these models for what *can* be observed.

Chapter 6

Instrumental Variables

We now turn to the final quasi-experimental estimator of program impact that we consider in this manual: instrumental variables. Instrumental variables is an estimation approach for recovering *consistent* estimates of program impact when program participation might be associated with unobserved characteristics that also influence the outcome of interest, a complication that we have seen yields biased and inconsistent estimates of program impact from simple methods of estimating impact such as comparison of outcomes between participants and non-participants or regression of the outcome of interest on an indicator of program participation. As we will learn in this chapter, instrumental variables is a potentially powerful tool for program impact evaluation, but one that involves strong assumptions and that can be challenging to interpret.

The instrumental variables strategy is based on the insight that there may be different sources or drivers of variation in program participation. Indeed, in our various numerical examples to this point, we typically introduced several different observed and unobserved characteristics of the individual that influence the program participation decision and hence variation in participation status across the population and any representative sample from it.

We can think of each of these characteristics as generating different "channels" of variation in program participation, some of which are associated with the unobserved determinants of the outcome of interest and thus from an endogeneity standpoint taint the overall variation in program participation by that association. However, some of those channels of variation are likely not related to those unobservables. They thus provide variation in program impact free from the endogeneity concerns associated with the overall variation in program participation.

Instrumental variables seeks to identify program impact by focusing on channels of variation in program participation associated with observed characteristics of the individual not correlated with unobservables associated with the outcome of interest. Specifically, it focuses on the variation in program participation driven by observed characteristics that meet three criteria. First, such observed characteristics do not independently and directly influence the outcome. In other words, whatever influence they have on outcome variation is felt through their impact on the variation in program participation. Second, such observed characteristics significantly influence the participation decision. Third, and crucially, these observed characteristics are independent of the unobserved confounding characteristics that influence the outcome of interest. Observed characteristics that fit these criteria are referred to as *instruments*.

Instrumental variables thus involves rather strong assumptions concerning valid instruments. As a program impact evaluation method, additional controversy has arisen over the interpretation of the estimates of program impact yielded by it. Specifically, in a world in which program impact varies between individuals, it turns out that the estimate of program impact yielded by instrumental

variables is a consistent estimate of program impact only for the types of people whose program participation behavior is responsive to variation in the instrument(s) employed. Despite the strong assumptions and interpretational challenges, however, instrumental variables is a popular program impact estimator that is today the focal point of exciting methodological research on numerous fronts.

6.1 Instrumental Variables Basics

In this section, we introduce the classic linear instrumental variables model. We then discuss instrumental variables estimation in the limited dependent variable setting. Finally, we discuss testing of the assumptions related to the instrumental variables model.

6.1.1 The Classic Linear Model

We begin with a behavioral model. Doing so is particularly important in the case of instrumental variables because it can help to clarify which sort of behavioral pathways are permissible under this approach. While our model is certainly not the only behavioral model that can motivate instrumental variables estimation, all models used to do so must adhere to the basic restrictions present in ours.

For present purposes, we will assume a constant program impact (i.e. we assume that program impact does not vary across individuals). The determination of the potential outcomes for representative individual i are as follows:

$$Y_i^0 = \beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

We have an observed (x) and unobserved (μ) individual-level characteristic determining potential outcomes as well as a purely random, idiosyncratic unobserved component ϵ^Y. We assume that the three are independently distributed. Program impact for individual i is

$$\begin{aligned} & Y_i^1 - Y_i^0 \\ &= \beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &\quad - \left(\beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &= \beta_1 \end{aligned}$$

Thus, program impact is constant across individuals.

Note that we have dropped the time subscript t in these potential outcome equations. We will do the same in the cost equation to be introduced shortly. We no longer distinguish between fixed and time-varying observed or unobserved characteristics. Although instrumental variables is sometimes applied in the panel data setting, it is a cross-sectional method in the sense that within variation or even repeated measures per se are not intrinsically necessary to apply it. The logic of identification of program impact (i.e. the manner in which instrumental variables allows for consistent estimation of program impact when program participation is endogenous) that we establish in the cross-sectional case applies to any other setting where we might apply it (such as to

the change in program participation in a panel data setting where one might use both the within and instrumental variables approaches simultaneously[1]).

A regression specification can be derived in much the same fashion as with behavioral models in preceding chapters. The observed outcome is

$$\begin{aligned} Y_i &= P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0 \\ &= P_i \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &+ (1 - P_i) \cdot \left(\beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \end{aligned}$$

Notice that the unobservable μ_i appears in this equation, and would hence be an element of the regression residual.

The cost of program participation is given by

$$C_i = \gamma_0 + \gamma_1 \cdot x_i + \gamma_2 \cdot z_{1i} + \gamma_3 \cdot z_{2i} + \gamma_4 \cdot \mu_i + \epsilon_i^C$$

With one exception (the variables z_1 and z_2) this is, given the potential outcome equations, a fairly standard cost function in the tradition of the behavioral models discussed thus far. Individual i will choose to participate (i.e. their value for the program participation indicator P, P_i, equals 1) if

$$Y_i^1 - Y_i^0 - C_i > 0$$

or

$$\beta_1 - \gamma_0 - \gamma_1 \cdot x_i - \gamma_2 \cdot z_{1i} - \gamma_3 \cdot z_{2i} - \gamma_4 \cdot \mu_i - \epsilon_i^C > 0$$

where we assume that ϵ_i^C is uncorrelated with both ϵ_i^Y and the variables z_{1i} and z_{2i}.

There are several important things to note about this. First, the observed individual characteristic x_i influences participation and hence we would expect that, other things being equal, the average value of this characteristic should differ between participants and non-participants. There will thus clearly be selection into program participation based on observables. Second, the same is true of the unobserved characteristic μ_i. In other words, we have an unobserved characteristic that influences participation as well as the outcome of interest Y (where $Y_i = P_i \cdot Y_i^1 + (1 - P) \cdot Y_i^0$). Finally, we have individual characteristics z_{1i} and z_{2i} that do not influence directly the outcome of interest Y_i. z_{1i} and z_{2i} do influence the program participation decision, and are independent of either of the unobserved determinants of the outcome, μ and ϵ_i^Y (we introduce no avenue for there to be correlation). In other words, the zs meet all of the criteria for an instrument.

It should now be clear that unbiased or consistent estimation of program impact (i.e. unbiased estimation of β_1) by straightforward means (such as regression of Y_i on P_i and x_i) is probably not in the cards. The unobserved characteristic μ influences, and hence is correlated with, program participation. However, it is also a determinant of the outcome Y. We are thus presented with the classic endogeneity challenge as a regressor, P_i, is correlated with the error term $\beta_3 \cdot \mu_i + \epsilon_i^Y$ (in this instance via the correlation between P and μ).

Instrumental variables attempts to recover a consistent estimate of program impact by restricting attention to the variation in program participation driven by variation in the instrumental

[1]The application of the instrumental variables approach in the context of a within model would most likely be motivated by concerns of confounding *time-varying* unobserved characteristics, since the within model itself would address any concerning *fixed* unobserved characteristics.

variables. Specifically, instrumental variables estimates the impact on the outcome of interest of the variation in program participation associated with those instruments.

In Figure 6.1 we lay out the "classical" setting where instrumental variables is applied. As in earlier diagrams of this nature, the solid line arrows indicate causal pathways (as in variation in Z *causes* variation in P) while the dashed lines merely indicate statistical association that may or may not meet any reasonable definition of causality. In Figure 6.1 we have three determinants of an outcome of interest Y: program participation P, an observed characteristic X, and an unobserved characteristic μ. P is associated with (i.e. correlated with) both X and μ. Finally, we have an "instrument" Z that influences program participation P but is unrelated to the other determinants of the outcome Y. Notice that the behavioral model we have laid out is completely consistent with this framework.

Clearly, straightforward regression of Y on P and X with a sample of individuals whose values for $\{Y, P, X\}$ were determined within such a framework would likely not yield an estimate of program impact that actually reflected the true causal effect of program participation on the outcome Y. To review the lessons of Chapters 2 and 4, the reason is that in such a regression, P is essentially being asked to play two empirical roles:

1. As a control for program participation (which is what we want it to do);

2. As a proxy for μ, the unobserved determinant of Y with which program participation P is associated (a role that we <u>do not</u> want participation to play for purposes of estimating the causal impact of participation P on the outcome Y).

The muddled nature of the empirical role of program participation P yields a muddled (or, more formally, biased and inconsistent) estimate of program impact with such a straightforward regression.

Instrumental variables solves this problem by focusing only on the "channel" of variation in program participation driven by the instrument Z. Given that it is related only to variation in the instrument Z, which is itself unrelated to μ in particular, this channel of variation in program participation P should be unrelated to μ. This is the essence of the strategy of instrumental variables for identification of the causal impact of program participation P on the outcome Y: focus only on that part of the variation in program participation out of the overall observed variation in P that is related to a characteristic Z (the instrument) that we can be sure is not related to the confounding unobservables μ with which the *overall* variation in program participation P *is* associated.

The justification for this approach is that the channel of variation in program participation P driven by the instrument Z is "clean" in the sense of being uncorrelated with the otherwise troublesome unobservable μ. An estimate of the impact on Y of variation in this channel of program participation P should thus reflect program impact since that channel of variation is not tainted by association with unobserved characteristics such as μ.

In Figure 6.2 we expand a bit on Figure 6.1, offering a somewhat more "relaxed" instrumental variables framework. Two modifications to Figure 6.1 are offered in Figure 6.2. First, the solid arrow connection Z and P in Figure 6.1 is replaced by a dotted line. Technically, the instrument Z needs only to be correlated with program participation P (i.e. a causal relationship is not essential). One needs to be careful threading this needle, however: under no circumstances can the instrument be correlated with the unobservable μ. Second, a statistical association between the instruments Z and the <u>observed</u> determinants of P is technically allowed.

While all of this may seem theoretically reasonable, the reader could be forgiven for wondering how, in practice, this actually works. How does one pick out the channels of variation in program

participation associated only with the instruments? To make this a bit more concrete, we focus on the case of just one instrument. Specifically, we assume for present purposes that of the variables $\{z_{1i}, z_{2i}\}$ we observe only z_{1i}. The typical estimation approach in this case is two-stage least squares, so named because it involves two stages of estimation. The steps in two-stage least squares estimation are:

1. Estimate by ordinary least squares regression the model:

$$P_i = \kappa_1 + \kappa_2 \cdot z_{1i} + \kappa_3 \cdot x_i + \upsilon_i$$

2. Compute predicted program participation from the fitted model:

$$\hat{P}_i = \hat{\kappa}_1 + \hat{\kappa}_2 \cdot z_{1i} + \hat{\kappa}_3 \cdot x_i$$

3. Regress Y_i on $\hat{P}_i$ and x_i.

The estimated coefficient on $\hat{P}_i$ from the last step is a consistent estimator of β_1. The steps would have been exactly the same if we had used z_{2i} instead of z_{1i} (and we could have used both instruments simultaneously in this fashion, regressing P on x, z_1 and z_2 in step 1).[2]

From steps 1 and 2 we can see how instrumental variables focuses on only that portion of the variation in program participation associated with the instrument. The model

$$P_i = \kappa_1 + \kappa_2 \cdot z_{1i} + \kappa_3 \cdot x_i + \upsilon_i$$

[2]Use of more than one instrument for a single endogenous variable, which is what would occur if we used z_1 and z_2 *simultaneously* as instruments, is called **over-identification**.

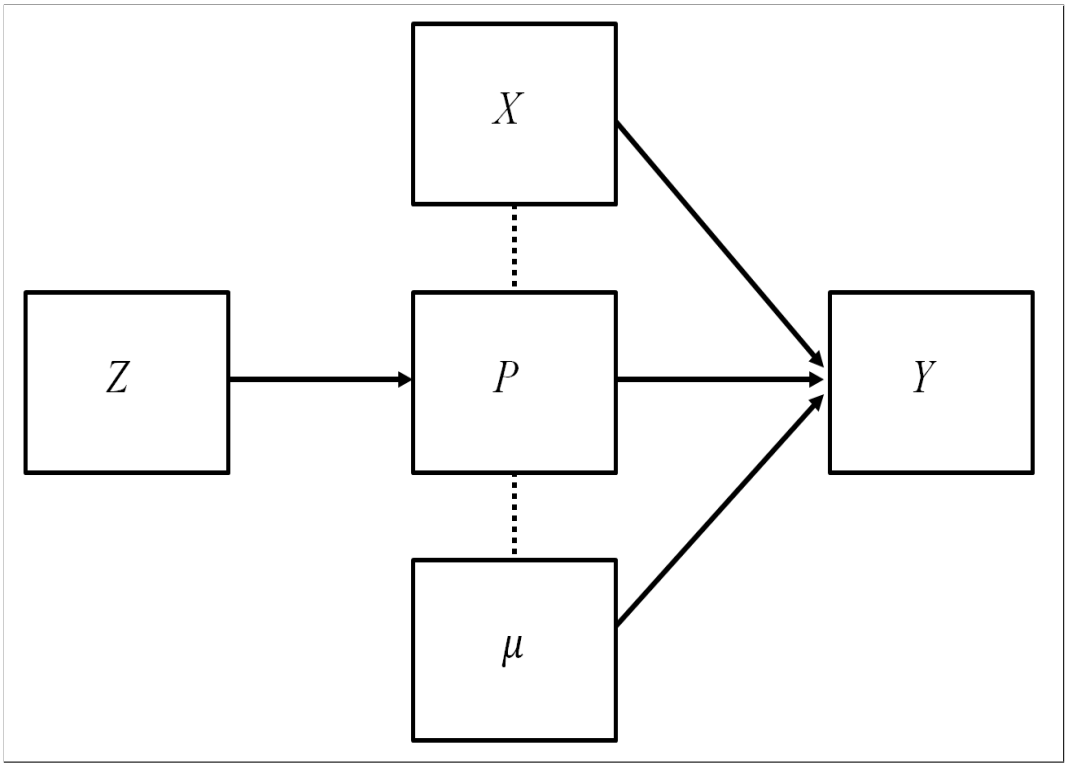


Figure 6.1: Instrumental Variables Schematic

essentially parses the variation in program participation P_i into two parts: that associated with the instrument z_1 and any of the other observed exogenous variables in the model (e.g. x) and that associated with every other determinant of program participation, which is relegated to υ_i. When we predict $\hat{P}_i$ we create a variable the variation in which is driven entirely by variation in the exogenous variables in the model, x_i, and the instrument z_{1i}. When we then use $\hat{P}_i$ as a regressor in the final step, we thus estimate the regression model determining the outcome Y_i with only that variation in program participation P_i associated with the exogenous variable x_i and the instrument z_{1i}.

One might ask why we need an instrument if the exogenous variable x_i appears in the first stage regression. To see why, suppose that we had simply regressed P_i on just the exogenous variable x_i. Predicted program participation would then be

$$\hat{P}_i = \hat{\kappa}_1 + \hat{\kappa}_3 \cdot x_i$$

Notice that this is simply a linear function of x_i and hence its entire variation depends on, and matches, the variation in x_i. Including $\hat{P}_i$ in this instance is thus tantamount to including x_i twice as a regressor. The regression determining the outcome then becomes:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot \hat{P}_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &= \beta_0 + \beta_1 \cdot (\hat{\kappa}_1 + \hat{\kappa}_3 \cdot x_i) + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &= \psi_0 + \psi_1 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \end{aligned}$$

where

$$\psi_0 = \beta_0 + \beta_1 \cdot \hat{\kappa}_1$$

and

$$\psi_1 = \beta_1 \cdot \hat{\kappa}_3 + \beta_2$$

The point is that we cannot separately identify β_1 in this new model. Were one to attempt to estimate this in a commercial package such as STATA (calling the first-stage predicted program participation variable `phat`), they would get an error along the lines of

```
note:  phat omitted because of collinearity
```

What this means is the `phat` was dropped because it was perfectly correlated with x. For all intents and purposes, it carries the same information as x. When one tries to include two perfectly correlated regressors in a regression model, the resulting situation is referred to as **multicollinearity**, under which estimation of the model with both regressors impossible.

This example provides insight into one of the other common terms used to describe instruments: **exclusion restrictions**. The idea behind this term is that instruments are variables associated with program participation[3] P_i but that do not influence Y_i directly and hence are *excluded from* the outcome equation explaining variation in Y_i. As we have seen, an excluded variable is essential for basic identification of a treatment effect (at least in the context of the linear two-stage least squares estimator).

In the parlance of instrumental variables estimation, x_i is usually referred to as an "exogenous variable". This is a term usually reserved for variables that are not determined within the system. In common usage in instrumental variables applications, it often effectively would seem to mean

[3]Or, more generally, the endogenous variable, but wherever possible we frame the discussion in this manual in terms of the program impact estimation challenge.

a variable not determined within the system but not an instrument, since instruments are often referred to separately as instruments. However, strictly speaking, this is a somewhat confusing attempt at drawing a distinction since, as this model is written, both x and the zs fit the classical definitions of exogeneity.

To examine the properties of this simple instrumental variables estimator, we derive an expression (i.e. formula) for it. This formula happens to be the same as that behind the two-stage least squares estimator, though we do not initially derive it directly from the two-stage approach. Instead, we use this derivation exercise as an opportunity to derive an alternative (to the least squares or maximum likelihood methods introduced in chapter 4) method for estimating the parameters of a statistical model: the **method of moments**. The method of moments seeks estimates of the parameters of a model that satisfy "moment conditions". These are essentially things that should be true of the model given the estimate.

To fix ideas, let us briefly consider the simple regression model

$$y_i = \zeta_0 + \zeta_1 \cdot x_i + \vartheta_i$$

This is a simple model relating one variable (y_i) to variation in another (x_i). We assume that x_i is independent of ϑ_i. Assuming that we have a sample of N individuals for whom we observe $\{y_i, x_i\}$ (for each of the individuals $i = 1, ..., N$), simple ordinary least squares regression of y_i on x_i using this sample yields the estimates

$$\hat{\zeta}_1 = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

and

$$\hat{\zeta}_0 = \overline{y} - \hat{\zeta}_1 \cdot \overline{x}$$

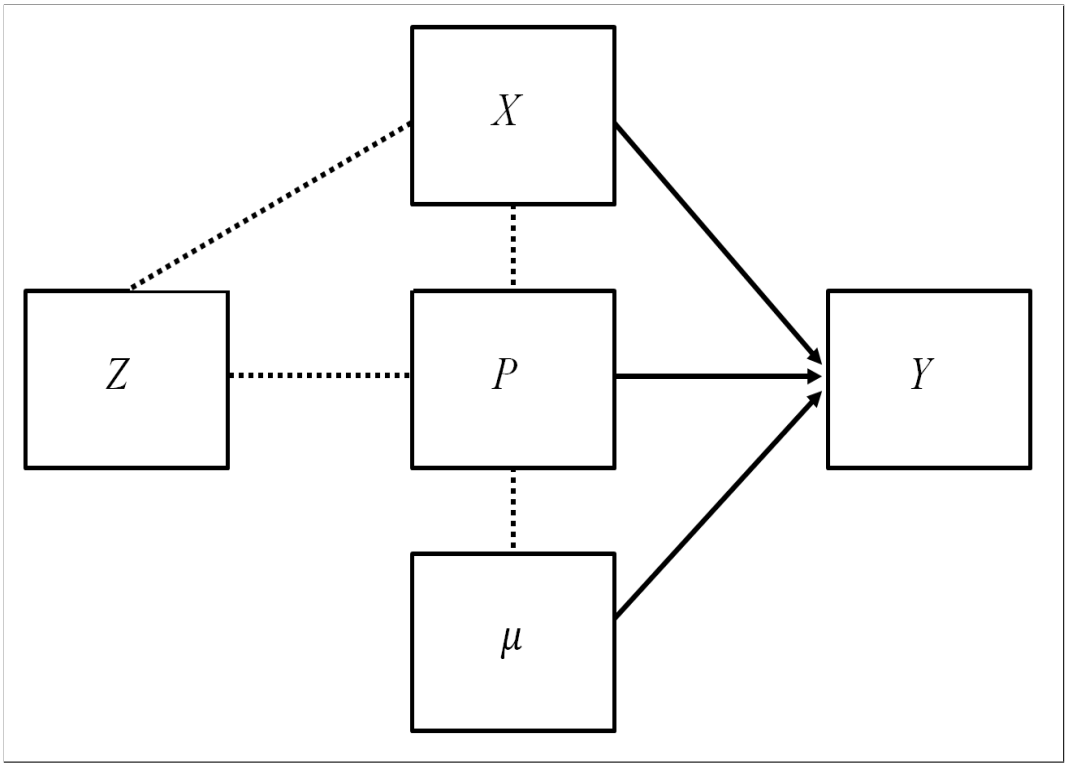


Figure 6.2: "Relaxed" Instrumental Variables Schematic

Thus we have the familiar least squares estimates.

The method of moments starts by asking what should be true of estimated parameters of the model, and solving for estimates that fulfill that requirement. First, we would hope that the estimated errors would be zero on average. Since the estimated error term in this simple model is

$$\hat{\vartheta}_i = y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i$$

the average of the estimated errors across the sample is

$$\overline{\vartheta} = \frac{\sum_{i=1}^{N} \hat{\vartheta}_i}{N}$$

Setting this equal to zero and multiplying both sides by N, we have

$$\sum_{i=1}^{N} \hat{\vartheta}_i = \sum_{i=1}^{N} \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

This is our first moment condition.

The second traditional moment condition applied in this setting is motivated by the idea that the regressor x_i should be independent from the estimated error $\hat{\vartheta}_i$. This moment condition is almost always written

$$\sum_{i=1}^{N} x_i \cdot \hat{\vartheta}_i = \sum_{i=1}^{N} x_i \cdot \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

This condition is sometimes motivated by the notion of finding the parameter estimates $\hat{\zeta}_0$ and $\hat{\zeta}_1$ that sets the covariance of the regressor x_i and the estimated error $\hat{\vartheta}_i$ equal to zero (e.g. Griffiths et al. 1993). Independence implies zero covariance.

The algebraic justification for this is straightforward. The covariance of x and $\hat{\vartheta}$ is

$$\frac{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot \left(\hat{\vartheta}_i - \overline{\hat{\vartheta}} \right)}{N-1}$$

However, the first moment condition guarantees that

$$\overline{\hat{\vartheta}} = 0$$

The covariance then becomes

$$\frac{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot \left(\hat{\vartheta}_i - \overline{\hat{\vartheta}} \right)}{N-1} = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot \left(\hat{\vartheta}_i \right)}{N-1} = \frac{\sum_{i=1}^{N} \left(x_i \cdot \hat{\vartheta}_i - \overline{x} \cdot \hat{\vartheta}_i \right)}{N-1}$$

$$= \frac{\sum_{i=1}^{N} \left(x_i \cdot \hat{\vartheta}_i \right)}{N-1} - \frac{\overline{x} \sum_{i=1}^{N} \hat{\vartheta}_i}{N-1}$$

However, since our first moment condition guarantees that the average of $\hat{\vartheta}_i$ is zero, we have

$$\frac{\sum_{i=1}^{N} \left(x_i \cdot \hat{\vartheta}_i \right)}{N-1} - \frac{\overline{x} \sum_{i=1}^{N} \hat{\vartheta}_i}{N-1} = \frac{\sum_{i=1}^{N} \left(x_i \cdot \hat{\vartheta}_i \right)}{N-1} - \overline{x} \cdot 0 = \frac{\sum_{i=1}^{N} \left(x_i \cdot \hat{\vartheta}_i \right)}{N-1}$$

Setting this equal to zero and multiplying both sides of the equation by $N-1$ we have the second moment condition.

A related conceptual approach to the second moment condition is an **orthogonality condition**. Two vectors (such as $[x_1, x_2, ..., x_N]$ and $[\hat{\vartheta}_1, \hat{\vartheta}_2, ..., \hat{\vartheta}_N]$) are orthogonal if their inner product[4] is equal to zero. This is equivalent to saying that they are linearly independent (that is, they have no linear association), which is the same as saying that they are uncorrelated. It turns out that the left hand side of the second moment condition

$$\sum_{i=1}^{N} x_i \cdot \hat{\vartheta}_i = \sum_{i=1}^{N} x_i \cdot \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right)$$

is the inner product of $[x_1, x_2, ..., x_N]$ and $[\hat{\vartheta}_1, \hat{\vartheta}_2, ..., \hat{\vartheta}_N]$. Therefore, when we set it equal to zero (as with the second moment condition) we effectively make the uncorrelatedness of x and $\hat{\vartheta}$ the second moment condition.

We thus have two moment conditions:

$$\sum_{i=1}^{N} \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

and

$$\sum_{i=1}^{N} x_i \cdot \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

The method of moments condition finds the values for $\hat{\zeta}_0$ and $\hat{\zeta}_1$ that insure that these two moment conditions hold. The solutions that do so are

$$\hat{\zeta}_1 = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

and

$$\hat{\zeta}_0 = \overline{y} - \hat{\zeta}_1 \cdot \overline{x}$$

In other words, for this simple model the method of moments estimators and estimates of $\hat{\zeta}_0$ and $\hat{\zeta}_1$ are the same as those from ordinary least squares estimation.

In this simple example we were able to solve for unique estimates of $\hat{\zeta}_0$ and $\hat{\zeta}_1$. The reason that we were able to do so is that there were as many moment conditions as statistical parameters that we wished to estimate. In classical mathematical terms, we had "as many equations as unkowns". Later, we will learn about a generalization of the method of moments that allows for the possibility of more moment conditions than parameters to be estimated. In general, when there are more parameters to be estimated than moment conditions the model is ***under*-identified**. The earlier discussion about attempting to estimate two-stage least squares with the observed covariate x_i as the only first-stage regressor is an example of this. Effectively, we were attempting to estimate

[4]The inner product is a concept from matrix algebra. If one had two vectors

$$[c_{11}, c_{12}, ..., c_{1K}]$$

and

$$[c_{21}, c_{22}, ..., c_{2K}]$$

The inner product of these two vectors would be

$$c_{11} \cdot c_{21} + c_{12} \cdot c_{22} + ... + c_{1K} \cdot c_{2K} = \sum_{k=1}^{K} c_{1k} \cdot c_{2k}$$

.

two parameters (β_1 and β_2) with just one instrument, x_i.[5] When there are as many parameters to be estimated as moment conditions, the model is ***just*-identified**. Finally, when there are more moment conditions than parameters to be estimated the model is ***over*-identified**. The current simple regression example above is of a just-identified model, because we have two moment conditions

$$\sum_{i=1}^{N} \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

and

$$\sum_{i=1}^{N} x_i \cdot \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

for the parameters to be estimated, ζ_0 and ζ_1.

To conclude this introduction to the method of moments, let us consider the precise semantics of method of moments estimators. For instance,

$$\sum_{i=1}^{N} x_i \cdot \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right) = 0$$

is a **moment condition**, as in "we are imposing some condition on this moment". The **moment** itself is

$$\sum_{i=1}^{N} x_i \cdot \left(y_i - \hat{\zeta}_0 - \hat{\zeta}_1 \cdot x_i \right)$$

Hence the condition we are imposing on the moment (making it a "moment condition") is that it equals zero. The distinction between a moment condition and the moment itself will become important below, when we consider the generalization of the method of moments under which we cannot necessarily find parameter estimates for which moment conditions hold exactly.

We now focus our thinking about the method of moments approach on the instrumental variables problem. To begin with, the regression model for Y that we developed at the outset of this chapter is

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \nu_i$$

where

$$\nu_i = \beta_3 \cdot \mu_i + \epsilon_i^Y$$

Basically, this is the original regression model with the error term collected into one variable for expositional simplicity. For further simplicity[6] we temporarily introduce the restriction that $\beta_2 = 0$,

[5] The equation of interest at that point was

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

Estimation of two-stage least squares with just the regressor x_i in both stages leaves only two moment conditions:

$$\sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 \cdot P_i - \beta_2 \cdot x_i) = 0$$

$$\sum_{i=1}^{N} x_i \cdot (Y_i - \beta_0 - \beta_1 \cdot P_i - \beta_2 \cdot x_i) = 0$$

(The term $\beta_3 \cdot \mu_i$ is subsumed into the regression error since μ_i is not observable.) There are just two moment conditions with which to estimate three parameters, β_0, β_1 and β_2, a clear example of under-identification.

[6] Specifically, to avoid using the more complicated matrix algebra required for multiple regression.

yielding the model:

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \nu_i$$

Thus we have recast this as a single regressor model.[7]

At this stage, we consider just one of the instruments introduced through the cost function, z_{1i} (the choice of which of the two instruments, z_{1i} or z_{2i}, to consider at this stage is random). We now need to characterize the moment conditions appropriate for instrumental variables estimation of β_0 and β_1 (though our primary interest lies with the latter since it is the impact of program participation P_i). To do so, we suppose that we have a sample of N individuals for whom we observe $\{Y_i, P_i, z_{1i}\}$ for each of those $i = 1, ..., N$ individuals. The first and most obvious moment condition from the last example of the method of moments is that average value of the predicted error term equals zero:

$$\frac{\sum_{i=1}^{N} \hat{\nu}_i}{N} = \frac{\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i\right)}{N} = 0$$

Multiplying both sides by N we have the final version of the first moment condition

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i\right) = 0$$

This is a fairly standard moment condition for most regression models the primary purpose of which is to identify the estimate of the constant term, $\hat{\beta}_0$.

For the second moment condition, we look to the required properties of an instrument such as z_{1i}. These are, most obviously, that the instrument play no direct role in the determination of Y_i (a condition guaranteed by the regression specification, which permits only P_i and ν_i to play a direct role in determining Y_i) and that the instrument be uncorrelated with the error term ν_i. This latter requirement suggests a moment condition:

$$\sum_{i=1}^{N} z_{1i} \cdot (\nu_i) = \sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i\right) = 0$$

This can be thought of as either a zero-covariance condition or an orthogonality condition, but the latter language is typically employed.

We now have two moment conditions, the requirement for exact identification since we are solving for estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, of two parameters, β_0 and β_1, respectively. The moment conditions are

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i\right) = 0$$

and

$$\sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i\right) = 0$$

Solving these for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields the instrumental variables estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (z_{1i} - \overline{z}) \cdot \left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N} (z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

[7] This specification is reasonable as long as $\beta_2 = 0$ or $\gamma_1 = 0$, in which case participation and x are not associated. For present purposes, one can assume that either of these holds.

where

$$\overline{z} = \frac{\sum_{i=1}^{N} z_i}{N}$$

In other words, where $\overline{z}$ is defined analogously to $\overline{Y}$ and $\overline{P}$.[8]

Now that we have a formula for the linear instrumental variables estimator (for the case of a single regressor that is endogenous, P_i, and a single instrument z_{1i}) we can proceed to examine the

[8]We have derived a formula for the instrumental variables estimator, which before this point had been introduced through the framework of two-stage least squares estimation. The same basic expression can be derived directly from the two-stage least squares estimator. To simplify things, consider the framework of an outcome y_i, a possibly endogenous variable x_i and an instrument z_i. Furthermore, suppose, that $\overline{y} = \overline{x} = \overline{z} = 0$ (in other words, suppose that all three variables have a mean or average of 0). This leads to the following two-stage system:

$$x_i = \omega \cdot z_i + \varepsilon_i$$

$$y_i = \zeta \cdot x_i + \vartheta_i$$

Notice that restricting the means of all the variables to zero has removed the need for constant terms in the model. The first stage ordinary least squares estimator is

$$\hat{\omega} = \frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2}$$

leading to a predicted endogenous variable value of

$$\hat{x}_i = \frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2} \cdot z_i$$

When we regress y_i on $\hat{x_i}$, the ordinary least squares estimate of ζ is

$$\hat{\zeta} = \frac{\sum_{i=1}^{N} \hat{x}_i \cdot y_i}{\sum_{i=1}^{N} (\hat{x}_i)^2} = \frac{\sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2} \cdot z_i \right) \cdot y_i}{\sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2} \cdot z_i \right)^2} = \frac{\left(\frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2} \right) \cdot \sum_{i=1}^{N} z_i \cdot y_i}{\left(\frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2} \right)^2 \cdot \sum_{i=1}^{N} (z_i)^2} = \frac{\sum_{i=1}^{N} z_i \cdot y_i}{\frac{\sum_{j=1}^{N} x_j \cdot z_j}{\sum_{j=1}^{N} (z_j)^2} \cdot \sum_{i=1}^{N} (z_i)^2}$$

$$= \frac{\sum_{i=1}^{N} z_i \cdot y_i}{\sum_{j=1}^{N} x_j \cdot z_j} = \frac{\sum_{i=1}^{N} z_i \cdot y_i}{\sum_{i=1}^{N} x_i \cdot z_i}$$

which is essentially the expression

$$\frac{\sum_{i=1}^{N} (z_i - \overline{z}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x}) \cdot (z_i - \overline{z})}$$

when all of the variables involved have a mean or average of zero.

Another popular general motivation for the two-stage least squares estimator which we apply to the simple model, begins with the simple covariance between y_i and z_i:

$$cov(y_i, z_i) = cov(\zeta \cdot x_i + \vartheta_i, z_i)$$

However, the covariance of a valid instrument (z_i) and the regression error from the equation of interest, ϑ_i, should be zero. We then have

$$cov(y_i, z_i) = cov(\zeta \cdot x_i + \vartheta_i, z_i) = cov(\zeta \cdot x_i, z_i) = \zeta \cdot cov(x_i, z_i)$$

Dividing both sides by $cov(x_i, z_i)$ we have

$$\zeta = \frac{cov(y_i, z_i)}{cov(x_i, z_i)}$$

which is of the same basic structure as the earlier formula for the two-stage least squares estimator.

properties of the estimator. We begin with that formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)} = \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot Y_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

(the second step is explained in Chapter 4) and insert the true data generating process behind Y_i:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot Y_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)} = \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot (\beta_0 + \beta_1 \cdot P_i + \nu_i)}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

$$= \beta_0 \cdot \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z})}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)} + \beta_1 \cdot \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot P_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)} + \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

$$= \beta_0 \cdot 0 + \beta_1 \cdot \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)} + \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

$$= \beta_1 \cdot 1 + \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

The next logical step would be to take the expectation of $\hat{\beta}_1$ and see if it equals β_1, establishing the unbiasedness of the instrumental variables estimator if it does.

Unfortunately, it doesn't:

$$E\left(\hat{\beta}_1\right) = \beta_1 + E\left(\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}\right) \neq \beta_1$$

We cannot break down the term

$$\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

with expectations. The reason goes back to a basic property of expectations: $E(A \cdot B) = E(A) \cdot E(B)$ only if A and B are independent. Because P_i is correlated with ν_i (via the unobserved characteristic μ_i that is subsumed in ν_i) the term

$$\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z})}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

is not independent of ν_i. We have thus learned an important property of the linear instrumental variables estimator: it is <u>not</u> unbiased. We have heard instrumental variables estimators such as two-stage least squares erroneously described as unbiased in casual discussions (e.g. at seminars) so often that we re-iterate: the instrumental variables estimator is <u>not</u> unbiased.

Hope is not lost, however. Unbiasedness tells us, essentially, whether the estimator correctly estimates the parameter (in this case true program impact β_1) correctly on average. There is still, however, the question of consistency. Consistency is basically concerned with whether, as the sample size becomes increasingly large, the distribution of the estimator (i.e. the various

values it can take on and their associated probabilities of occurring in any given sample) becomes concentrated around the true parameter value (in this case, β_1).

Consistency is typically considered with probability limits (which were introduced in Chapter 4). Probability limits have the convenient quality that

$$plim\left(\frac{A}{B}\right) = \frac{plim\,(A)}{plim\,(B)}$$

Returning to the expression

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}$$

we have

$$plim\left(\hat{\beta}_1\right) = plim\,(\beta_1) + plim\left(\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}\right)$$

$$= \beta_1 + \frac{plim\left(\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i\right)}{plim\left(\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)\right)}$$

$$= \beta_1 + \frac{plim\left(\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{N}\right)}{plim\left(\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}{N}\right)}$$

Notice that

$$\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \nu_i}{N}$$

is simply the covariance between the instrument and the regression error term. However, since we assume that the instrument and error term in the equation of interest are uncorrelated, this means that the probability limit of the covariance is zero. Hence,

$$plim\left(\hat{\beta}_1\right) = \beta_1 + \frac{0}{plim\left(\frac{\sum_{i=1}^{N}(z_{1i} - \overline{z}) \cdot \left(P_i - \overline{P}\right)}{N}\right)} = \beta_1$$

In other words, the instrumental variables estimator may not be unbiased, but it is consistent.

A reasonable question at this point are the implications of these findings for actual empirical practice. We have found that the instrumental variables estimator is not unbiased but is consistent. Use of the estimator thus rests on the finding that it is consistent. Consistency is concerned with the properties of an estimator as sample sizes become very large. There is absolutely no guarantee about the performance of a consistent estimator with smaller sample sizes. Some (e.g. Angrist and Krueger 2001) have as a result suggested caution when applying instrumental variables to smaller sample sizes. This is sage advice.

At this point, it might be useful to pause and consider a numerical example of instrumental variables estimation under this single instrument case. This example is captured in STATA do-file 6.1.do. The departure point is the following model of potential outcomes and cost of participation for 10,000 individuals:

$$Y_i^1 = 4 + 2 \cdot x + \mu + \epsilon_i^Y$$

$$Y_i^0 = 2 + 2 \cdot x + \mu + \epsilon_i^Y$$
$$C_i = 1 + 1.5 \cdot x - .5 \cdot z_1 + .5 \cdot z_2 + \mu + \epsilon_i^C$$

where x and μ are independently normally distributed with mean 0 and variance 4 (i.e. x, $\mu \sim$ N(0,4)), the ϵs are independently normally distributed with mean 0 and variance 9 (i.e. ϵs $\sim$ N(0,9)) and the instruments z_1 and z_2 are independently normally distributed with mean 0 and variance 4 (i.e. x, $\mu \sim$ N(0,4). There is no real avenue for the endogeneity of either x or the instruments z. The parameters of the simulation were, as always, more or less randomly chosen.

This is another example of constant program impact: $Y_i^1 - Y_i^0 = 2$. Thus, the cost equation is determining the role that x and μ play in shaping the participation decision. The instruments influence participation only through their role in cost: they have no independent role in determining the outcome Y beyond their role in shaping participation. Moreover, by design the instruments are independent of the error terms ϵ^C and, especially, ϵ^Y.

Output 6.1 presents the distribution of participation. 58.13 percent participate in the program. Once again, the numerical example involves a relatively popular program. We encourage the reader to fiddle with the simulation code to explore circumstances of even more and far less popular programs.

STATA Output 6.1 (6.1.do)

```
. * Basic summary statistics: participation
. tab P
```

P	Freq.	Percent	Cum.
0	4,187	41.87	41.87
1	5,813	58.13	100.00
Total	10,000	100.00	

In Output 6.2 we consider the means of key variables between participants and non-participants. Participants are individuals with relatively low average values for Y^1 and for Y^0. Behind this finding are differences in the average values of μ and x between the two groups. Specifically, the cost of enrollment is increasing in x and μ. Thus participants should be individuals with smaller values for x and μ. This implies that the participants should have lower values for Y, Y^1 and Y^0, other things being equal. Notice as well that the average values of the instruments z_1 and z_2 differ between participants and non-participants which, if this was a real world sample, would serve as *prima facie* evidence that these candidate instruments might be strong predictors of program participation. Were this a real world example, at this phase we might refer to them as *candidate* instruments because we would as yet have no indication whether they met the criteria for instruments (particularly exclusion from the equation of interest and independence from the errors), an issue which we might address with tests to be discussed in a subsequent subsection. On the other hand, as we will see it is not always possible to test to see if an instrument is indeed uncorrelated with the unobserved determinants of the outcome. In practice, the justification for an instrument often rests on theoretical behavioral arguments (i.e. why the candidate instrument should behaviorally influence participation but otherwise be unrelated to the outcome) rather than an empirical test.

In Output 6.3 we consider the correlations among key covariates. Note that program participation P is highly correlated with the unobservable μ but not with ϵ. Thus program participation is an endogenous variable in a regression of Y on P and x, and for exactly the reasons that the

simulation design intended to engineer. Crucially, the instruments appear essentially uncorrelated with the unobservables μ, ϵ^Y and ϵ^C. They thus appear to meet the exogeneity requirement of an instrument (namely that an instrument not be correlated with the regression error in the equation of interest). The only reason we can assess this in this fashion is that this is a simulated example and hence we can "see" the errors. In the real-world analog to this sample, we would observe only $\{Y, P, x, z_1, z_2\}$.

STATA Output 6.2 (6.1.do)

```
. * Basic summary statistics: variable means
.
. by P, sort: summarize Y y1 y0 c x* z* mu epsilon*
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	4187	4.977852	4.644148	-10.74594	22.10318
y1	4187	6.977852	4.644148	-8.745936	24.10318
y0	4187	4.977852	4.644148	-10.74594	22.10318
c	4187	5.575921	2.770066	2.002451	19.11571
x0	4187	1	0	1	1
x	4187	1.130789	1.721677	-5.620321	7.176875
z1	4187	-.3372715	1.952999	-8.078415	6.060055
z2	4187	.383663	1.959136	-6.330251	6.769236
mu	4187	.7467416	1.903202	-5.616416	7.12235
epsilony	4187	-.0304688	2.972894	-10.58663	10.57696
epsilonc	4187	1.772528	2.597622	-7.321414	11.07838

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	5813	1.748017	4.738986	-14.47957	20.4715
y1	5813	1.748017	4.738986	-14.47957	20.4715
y0	5813	-.2519832	4.738986	-16.47957	18.4715
c	5813	-2.267171	3.135526	-17.07104	1.997842
x0	5813	1	0	1	1
x	5813	-.829659	1.772296	-8.141914	4.701032
z1	5813	.2619722	1.990245	-7.180359	8.420812
z2	5813	-.3029925	1.965626	-6.960695	6.588125
mu	5813	-.5577272	1.92052	-7.726572	7.586161
epsilony	5813	-.034938	2.950077	-10.366	12.81559
epsilonc	5813	-1.182473	2.631718	-11.73894	8.629721

In Output 6.4 we present the results of the obligatory attempt at simple regression of Y on P and x. Given that true program impact is 2, the regression results represent a huge underestimate of true program impact at .263087 (only around 13 percent of true impact). There is a simple and (what should be at this point) familiar explanation for this: program participation P is an endogenous regressor. Specifically, it is correlated with μ in the error term. Straightforward impact evaluation methods that ignore the possibility of endogeneity like this simple regression are thus unlikely to yield unbiased or consistent estimates of program impact.

We next turn to instrumental variables, relying on just one of the two instruments, z_1. We do this randomly and without loss of generality (in other words, we could just as easily have used z_2 for our illustration of the one instrument case). We start with "manual" two-stage least squares, whereby we perform the first stage regression of P on x and z_1, then use the estimates to predict

program participation and then perform the second stage regression of the outcome Y on predicted program participation and x. In Output 6.5 we present results from the "first stage" of this process, the regression of P on x and z_1. Notice that z_1 is a highly significant predictor of program participation. We want an instrument to be a highly significant predictor of the endogenous variable (in this case program participation P) and, at the end of this subsection, we will briefly discuss how critical this really can be. Following the regression estimation, we use the "fitted" model (i.e. the estimated model from this first stage regression) to predict program participation `Phat`.

STATA Output 6.3 (6.1.do)

```
. * Correlations among unobservables
.
. corr P z* mu epsilony epsilonc
(obs=10000)
```

	P	z1	z2	mu	epsilony	epsilonc
P	1.0000					
z1	0.1481	1.0000				
z2	-0.1701	-0.0034	1.0000			
mu	-0.3188	-0.0001	0.0155	1.0000		
epsilony	-0.0007	0.0090	-0.0083	0.0065	1.0000	
epsilonc	-0.4866	0.0071	0.0078	-0.0092	-0.0041	1.0000

STATA Output 6.4 (6.1.do)

```
. * Cross sectional regression
. reg Y P x
```

Source	SS	df	MS		
Model	122730.14	2	61365.0699	Number of obs	= 10000
Residual	123469.862	9997	12.3506914	F(2, 9997)	= 4968.55
				Prob > F	= 0.0000
Total	246200.001	9999	24.6224624	R-squared	= 0.4985
				Adj R-squared	= 0.4984
				Root MSE	= 3.5144

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.263087	.0813785	3.23	0.001	.1035687	.4226052
x	1.781695	.0200693	88.78	0.000	1.742355	1.821035
_cons	2.963129	.0588625	50.34	0.000	2.847747	3.078512

Output 6.6 presents the second stage of the "manual" two-stage least squares process. The estimate of program impact is the coefficient on the predicted program participation, `Phat`. At 2.361004 this estimate of program impact is far closer to the true value of 2 than the naive regression estimate of .263087.

We now briefly digress to consider the consequences of small sample sizes and (from the "glass half full" perspective) the advantages of large sample sizes when conducting instrumental variables estimation. In Outputs 6.7 and 6.8 we report second stage regression results after repeating the manual two-stage least squares, but in this case with sample sizes of 150 and 500, respectively. Neither of these sample sizes are unheard of in applied micro-level empirical work. However, in both cases it is clear that the two-stage least squares estimator is not performing particularly well, providing estimates well short of mark in terms of true program impact of 2. Indeed, at 150

observations (a sample size not unheard of for, for instance, multiple regression or within estimators, with decent results in some cases) the estimate of program impact is only around half of the true value.

STATA Output 6.5 (6.1.do)

```
. * First stage of ``manual´´ two-stage least squares,
. * single instrument case
.
. reg P x z1
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) = 1718.26
       Model |  622.633312     2  311.316656            Prob > F      =  0.0000
    Residual |  1811.26979  9997  .181181333            R-squared     =  0.2558
-------------+------------------------------           Adj R-squared =  0.2557
       Total |  2433.9031  9999  .243414651            Root MSE      =  .42565

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.1192699   .0021278   -56.05   0.000    -.1234408    -.115099
          z1 |   .0367036    .002132    17.22   0.000     .0325245    .0408826
       _cons |   .5798419   .0042566   136.22   0.000      .571498    .5881857
------------------------------------------------------------------------------

.
. predict Phat
(option xb assumed; fitted values)
```

STATA Output 6.6 (6.1.do)

```
. * Second stage of ``manual´´ two-stage least squares,
. * single instrument case
.
. reg Y Phat x
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) = 4982.32
       Model |   122900.4     2  61450.1998            Prob > F      =  0.0000
    Residual |  123299.602  9997  12.3336603            R-squared     =  0.4992
-------------+------------------------------           Adj R-squared =  0.4991
       Total |  246200.001  9999  24.6224624            Root MSE      =  3.5119

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Phat |   2.361004   .4792454     4.93   0.000     1.421587    3.300422
           x |    2.03184   .0597786    33.99   0.000     1.914662    2.149018
       _cons |   1.745816   .2802903     6.23   0.000     1.196391    2.295242
------------------------------------------------------------------------------
```

Some of this probably simply reflects higher ordinary sampling variation associated with smaller sample size. However, there is reason to be skeptical of the idea that this explains all of the discrepancy between the results in, say, Outputs 6.6 and 6.7. As we move from the 10,0000 observation case in Output 6.6 to the 150 observation case in 6.7, standard errors for all parameter estimates widen. However, the degree of increase is much greater for predicted program participation `Phat` than x. Some of this uneven pattern of increase to the relative standard errors of P and x might reflect very imprecise estimates of the relationship between program participation P and the instrument z_1 due to the small sample size. This could lead to a very noisy predicted program participation. While

this is probably generally a factor at smaller sample sizes it might not be the only explanation. For instance, in Output 6.9 we re-ran the model with 10,000 observations for the first stage regression (of P on x and z_1), thus restoring much of the lost precision to the first-stage estimate, but only 150 observations for the second stage. The estimated program is actually even more wide of the mark at -.7863779, and the absolute size of the standard error of P is actually larger in Output 6.9 than in Output 6.7. This suggests that first stage precision isn't everything where the performance of two-stage least squares is concerned.

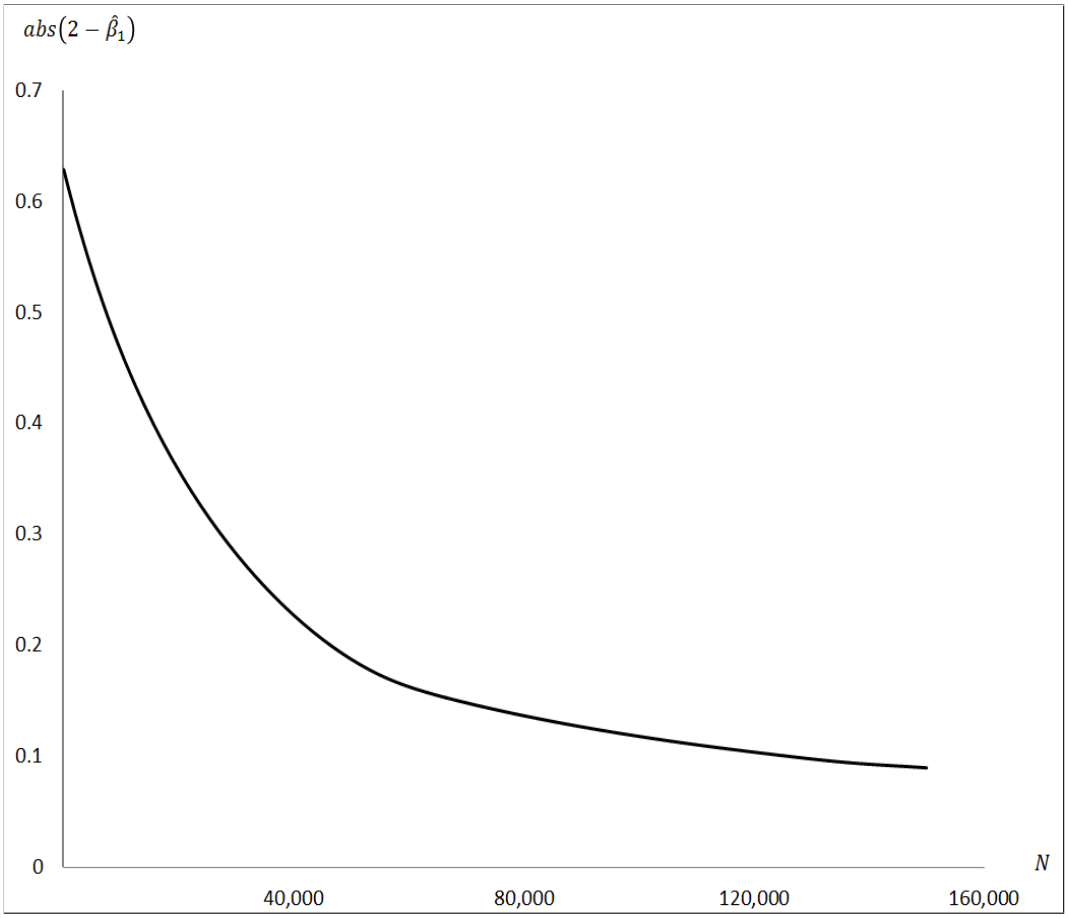


Figure 6.3: TSLS Performance

In general, the poor performance of the two-stage least squares estimator at the smaller sample sizes likely reflects to an extent the admonition to be careful about the distinction between unbiasedness (which does not depend on sample size) and consistency (a property that, when it obtains, does so only at larger sample sizes). It is riskier (from the standpoint of obtaining highly misleading program impact estimates) to rely on impact estimates from two-stage least squares estimation with small sample sizes. In Figure 6.3 we illustrate the performance of the two-stage least squares estimator applied to this example as sample size increases. Specifically, we re-ran the model at different sample sizes ranging from 100 to 150,000 observations in increments of 100 observations. For each sample size, we calculated the absolute value of the difference between the impact estimate under two-stage least squares for that sample size and the true value of 2. Using STATA's `lowess` command, we then performed locally weighted regression of the absolute value of the deviations of the estimates from 2 on the number of observations. Locally weighted regression is beyond the

scope of this manual, but you can usefully (for present purposes) view it as an extremely flexible way of looking at how the central tendency of the absolute deviation of estimated program impact from the truth varies as sample size changes.

Figure 6.3 illustrates the evolution of that central tendency as sample size increases. It is plain from the pattern of the central tendency (per the "predicted" central tendency from local regression of the absolute deviation of the program impact estimate from the truth on the number of observations) that, in general, the two-stage least squares estimator performs increasingly well as sample size increases. Indeed, by 150,000 observations the predicted absolute deviation of the program impact estimate is less than 5 percent of the true value of that estimate. On the other hand, the predicted absolute deviation remains quite high at sample sizes in ranges seen quite frequently in empirical work (e.g. around 20,000 observations).

To be sure, some of this simply reflects the benefits of lower sampling variation from larger sample sizes in any sort of impact evaluation model. However, to get some sense of how plausible that is, in Figure 6.4 we add to Figure 6.3, comparing the performance of our basic two-stage least squares estimator to that of the "correct" program impact least squares regression estimator (i.e. least squares regression of the outcome Y on program participation P, the observed characteristic x and the unobserved characteristic μ) as sample size increases. We focus on this regression because it provides an unbiased and consistent estimate of program impact. Hence, any improvement in it in terms of the predicted absolute deviation of its estimate of program impact from the true value

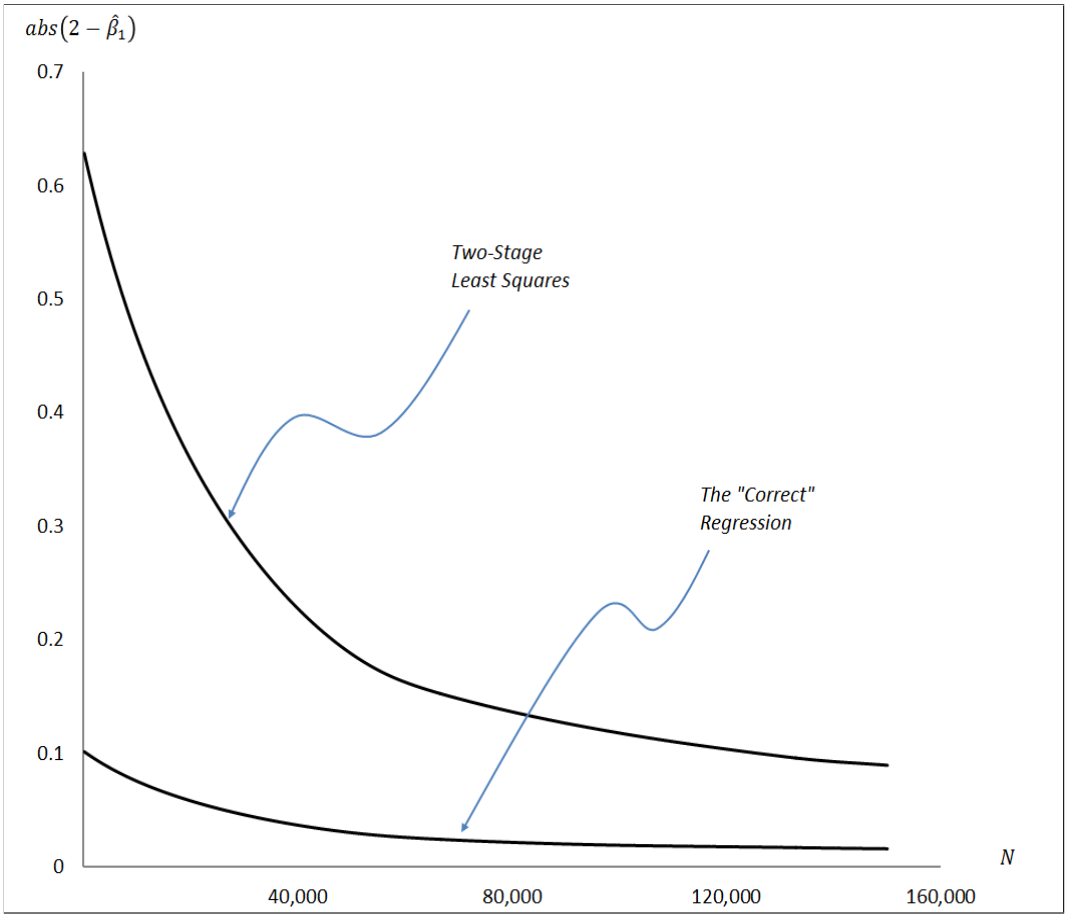


Figure 6.4: TSLS Performance versus "Correct" OLS

of 2 as sample size increases *must* be due solely to precision gains (i.e. reduced sampling variation).

STATA Output 6.7 (6.1.do)

```
. * Second stage of ``manual´´ two-stage least squares,
. * single instrument case
.
. reg Y Phat x
```

Source	SS	df	MS		
Model	1598.83911	2	799.419555	Number of obs =	150
Residual	1914.4531	147	13.0234905	F(2, 147) =	61.38
				Prob > F =	0.0000
Total	3513.29221	149	23.5791424	R-squared =	0.4551
				Adj R-squared =	0.4477
				Root MSE =	3.6088

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Phat	1.016743	3.202113	0.32	0.751	-5.311381	7.344866
x	1.777144	.3959087	4.49	0.000	.9947358	2.559552
_cons	2.753803	2.024577	1.36	0.176	-1.247234	6.754841

STATA Output 6.8 (6.1.do)

```
. * Second stage of ``manual´´ two-stage least squares,
. * single instrument case
.
. reg Y Phat x
```

Source	SS	df	MS		
Model	6096.17393	2	3048.08697	Number of obs =	500
Residual	6236.60312	497	12.5484972	F(2, 497) =	242.90
				Prob > F =	0.0000
Total	12332.777	499	24.7149841	R-squared =	0.4943
				Adj R-squared =	0.4923
				Root MSE =	3.5424

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Phat	1.433846	2.758165	0.52	0.603	-3.985255	6.852948
x	1.897721	.3344124	5.67	0.000	1.240684	2.554757
_cons	2.033478	1.577156	1.29	0.198	-1.065236	5.132192

It is evident that, in terms of predicted absolute deviation of the program impact estimate from the true value of 2 the regression estimator performs better than two-stage least squares at every sample size. Moreover, the overall improvement in the performance of the two-stage least squares estimator is much greater at smaller sample sizes and still noticeably great at the absurdly large sample size (at least for this example) of, say, 120,000. This is likely indicative of something more going on with the two-stage least squares estimator than just reduction in sampling variation as sample size grows. Whatever the case, however, this comparison reflects the basic wisdom that one should exercise caution when applying the two-stage least squares estimator to smaller sample sizes.

In Output 6.10 we present two-stage least squares estimates using STATA's built-in two-stage least squares package (which is implemented through the STATA command `ivregress`). It is

generally better practice to use the two-stage least squares routine that comes with your preferred commercial statistical software than to engage in the manual estimation of two-stage least squares along the lines reported in Outputs 6.5 and 6.6, for reasons that will shortly become clear. With the "`first`" option we have elected to have the output display the first stage results. Notice that the "`First-stage regressions`" results are identical to those in Output 6.5: there is generally no difference in results between the manual approach to two-stage least squares (i.e. estimate the first stage as an ordinary regression, predict program participation and then estimate the second stage as an ordinary, stand-alone regression) and a commercial two-stage least squares package for the first stage regression of program participation P on x and z_1.

STATA Output 6.9 (6.1.do)

```
. * Second stage of ``manual´´ two-stage least squares,
. * single instrument case
.
. reg Y Phat x if _n<151
```

Source	SS	df	MS		
Model	1581.12731	2	790.563657		
Residual	2041.67345	147	13.888935		
Total	3622.80076	149	24.3140991		

Number of obs	= 150
F(2, 147)	= 56.92
Prob > F	= 0.0000
R-squared	= 0.4364
Adj R-squared	= 0.4288
Root MSE	= 3.7268

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Phat	-.7863779	4.320652	-0.18	0.856	-9.324994	7.752238
x	1.829088	.5175141	3.53	0.001	.8063595	2.851817
_cons	3.228184	2.473724	1.30	0.194	-1.660472	8.11684

The difference in the two approaches emerges at the second stage. In Output 6.6, the program impact estimate from manual second stage estimation was 2.361004, with a standard error of .4792454. In Output 6.10 the second stage estimate of program impact is the same, but the estimated standard error of that estimate is larger, at .4951864. This reflects the fact that the second stage standard errors from manual estimation are somewhat naive in that they assume that the second stage is simply a stand alone regression and not part of a multi-stage estimation process. In particular, the manual second stage estimates of the standard error of the program impact estimate do not recognize that there will be some sampling variation in the program impact estimate because there is some sampling variation in *first* stage estimates and hence in predicted program participation. Prepared routines such as STATA's `ivregress` command use correct standard error formulas that take account of this sampling variation emanating from the first stage regression. Although the difference in standard errors was modest in this example, in some cases it can be much larger.

Up to this point we have examined the instrumental variables estimator in terms of just one instrument. However, the model at the beginning of the chapter presented two potential instruments, z_{1i} and z_{2i}. We now consider estimation with more than one instrument.

To begin with, we pick up where the earlier discussion of instrumental variables and method of moments ended. In that discussion, there were two moment conditions

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i \right) = 0$$

$$\sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i\right) = 0$$

We return the regressor x_i to the model (i.e. we no longer assume that $\beta_2 = 0$), resulting in the slightly modified moment conditions

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

$$\sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

The addition of a new parameter to be estimated does not result in an under-identified (i.e. more parameters to be estimated than moment conditions) because the introduction of method of moments (when the method of moments was applied to a simple regression model) suggests a new moment condition associated with the regressor x_i:

$$\sum_{i=1}^{N} x_i \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

We thus have three moment conditions and three parameters to estimate. In other words, we have exact identification.

Adding a new instrument, however, creates identification problems. Extending the logic of our last discussion of instrumental variables and the method of moments, adding the second instrument z_{2i} creates a new moment condition:

$$\sum_{i=1}^{N} z_{2i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

We thus now have four moment conditions:

$$\sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

$$\sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

$$\sum_{i=1}^{N} x_i \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

$$\sum_{i=1}^{N} z_{2i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right) = 0$$

but only three parameters to estimate: β_0, β_1 and β_2. In other words, we have four equations, and three unknowns ($\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$). This is thus a case of over-identification. The problem is that we cannot find a set of estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ at which the moment conditions hold exactly.

However, we can find those estimates at which the conditions come as close to holding as possible. This is the objective of a generalization of the method of moments known, unsurprisingly, as the **generalized method of moments**.

For notational simplicity, let us rename the moments as follows:

$$g_1 = \sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i \right)$$

$$g_2 = \sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i \right)$$

$$g_3 = \sum_{i=1}^{N} x_i \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i \right)$$

$$g_4 = \sum_{i=1}^{N} z_{2i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i \right)$$

Notice that we have dropped the "= 0" part of these moment conditions: we had already concluded that the moment conditions (i.e. that each moment equals 0) could not be made to hold exactly anyway.

STATA Output 6.10 (6.1.do)

```
. * Two stage least squares: STATA command
.
. ivregress 2sls Y x (P=x z1), first
First-stage regressions
```

Number of obs	=	10000			
F(2, 9997)	=	1718.26			
Prob > F	=	0.0000			
R-squared	=	0.2558			
Adj R-squared	=	0.2557			
Root MSE	=	0.4257			

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.1192699	.0021278	-56.05	0.000	-.1234408	-.115099
z1	.0367036	.002132	17.22	0.000	.0325245	.0408826
_cons	.5798419	.0042566	136.22	0.000	.571498	.5881857

Instrumental variables (2SLS) regression

Number of obs	=	10000	
Wald chi2(2)	=	9333.40	
Prob > chi2	=	0.0000	
R-squared	=	0.4652	
Root MSE	=	3.6287	

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.361004	.4951864	4.77	0.000	1.390457	3.331552
x	2.03184	.061767	32.90	0.000	1.910779	2.152901
_cons	1.745816	.2896135	6.03	0.000	1.178184	2.313448

```
Instrumented:  P
Instruments:   x z1
```

These can be assembled into a vector of moments:

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}$$

The generalized method of moments seek estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that essentially minimize a **norm** of this vector. In particular, it seeks the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize[9]

$$\begin{bmatrix} g_1 & g_2 & g_3 & g_4 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix} = (g_1)^2 + (g_2)^2 + (g_3)^2 + (g_4)^2$$

Notice that the closer each of the moments g_j (for $j = 1, ...4$) is to zero, the smaller will be this quantity. Thus, while the generalized method of moments cannot achieve estimates that make the moment conditions $g_j = 0$ hold exactly (which would not be possible given the over-identified nature of the system of moment conditions is over-identified) it finds the estimates that get the moments to be collectively as close to zero as possible.

In practice, generalized method of moments estimators seek to minimize *weighted* moment conditions. To get some idea how this might work, consider the very simple weighting matrix

$$\begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix}$$

where the ws are weights to apply to the moment conditions. Generalized method of moments might seek estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize[10]

$$\begin{bmatrix} g_1 & g_2 & g_3 & g_4 \end{bmatrix} \begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}$$

[9]The expression

$$\begin{bmatrix} g_1 & g_2 & g_3 & g_4 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}$$

shows a matrix operation known as matrix multiplication. It is more important that the reader recognizes the function

$$(g_1)^2 + (g_2)^2 + (g_3)^2 + (g_4)^2$$

that is the result of that matrix multiplication, and for which the generalized methods of moments seeks estimates to minimize its value, than it is to necessarily understand matrix operations. We define the matrix operation simply for the benefit of the interested reader.

[10]The expression

$$\begin{bmatrix} g_1 & g_2 & g_3 & g_4 \end{bmatrix} \begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}$$

simply shows a more complex matrix operation.

$$= w_1 \cdot (g_1)^2 + w_2 \cdot (g_2)^2 + w_3 \cdot (g_3)^2 + w_4 \cdot (g_4)^2$$

With the addition of weights, each moment no longer counts equally for the minimization problem. The result of adding a weighting matrix is usually considerably messier (in appearance) but this simple exercise gives the reader some intuitive idea of what the weights are doing in essence.

In practice the generalized method of moments estimators use more complex weighting matrices, the development and discussion of which is beyond the scope of this manual. As a gross generalization, the weighting matrices used in generalized method of moments estimation tend to give more weight to moments with lower variance which, intuitively, means that they give more weight to moments about which we have more information. The weighting matrix can be imposed by the researcher or, more often (and preferably), determined by the data.

STATA Output 6.11 (6.1.do)

```
. * First stage of ``manual´´ two-stage least squares,
. * two instrument (over-identified) case
.
. reg P x z1 z2
```

Source	SS	df	MS	Number of obs = 10000
				F(3, 9996) = 1333.14
Model	695.526508	3	231.842169	Prob > F = 0.0000
Residual	1738.37659	9996	.173907222	R-squared = 0.2858
				Adj R-squared = 0.2856
Total	2433.9031	9999	.243414651	Root MSE = .41702

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.119577	.0020847	-57.36	0.000	-.1236635	-.1154906
z1	.0365565	.0020887	17.50	0.000	.0324622	.0406508
z2	-.0428673	.0020938	-20.47	0.000	-.0469716	-.0387629
_cons	.5791768	.0041704	138.88	0.000	.5710019	.5873517

```
.
. predict Phat
(option xb assumed; fitted values)
```

We highly recommend that generalized method of moments estimation be performed with commercial statistical programs that offer generalized method of moments estimation as part of their package. STATA is one example of such a program. The generalized method of moments is the subject of a vast literature offering many insights into best practice, and so it might be most efficient for the reader simply to utilize the pre-prepared generalized method of moments routines in many statistical packages.

At this point the reader would be justified in asking why we have taken these detours into the worlds of the method of moments and the generalized method of moments. After all, we could simply have stuck with two-stage least squares (as an earlier footnote made clear, even the formula for the basic single instrument instrumental variables estimator could be derived directly from two-stage least squares estimation).

There are at least two reasons we discuss moment-based methods. First, they are an extremely popular platform for methodological thinking about instrumental variables, and are frequently employed in empirical applications of instrumental variables. To not discuss them would thus leave an important door closed to the reader. Second, one of the most popular tests of instruments, which we will discuss below, emerges from the moment-based methods.

We now turn our attention to an empirical example involving over-identification. To do so,

we simply continue our simulated empirical example from earlier in this subsection, but now use both instruments, z_1 and z_2, as predictors of program participation. Results for the first stage estimation of two-stage least squares via the "manual" approach are presented in Output 6.11. We now include the additional instrument z_2. Notice that the predictive power of the model in terms of R^2 has improved somewhat from the figure of .2558 in Output 6.5 to .2858 in Output 6.11. This is one indication that predicted participation based on the fitted first stage model might now be a bit less noisy.

Output 6.12 presents the results from the second stage of estimation under the manual approach. The estimate of program impact is now a bit closer to the truth than in Output 6.6 (2.361004 in Output 6.6 against 2.108621 below in Output 6.12). The standard error of the program impact estimate has fallen dramatically (around 35 percent from the Output 6.6 estimate of .4792454 to .3117998 in Output 6.12). This would seem to present fairly convincing evidence for improved precision from the inclusion of an additional instrument in the first stage.

STATA Output 6.12 (6.1.do)

```
. * Second stage of ``manual´´ two-stage least squares,
. * two instrument (over-identified) case
.
. reg Y Phat x
```

Source	SS	df	MS		Number of obs = 10000
					F(2, 9997) = 5003.69
Model	123163.926	2	61581.9631		Prob > F = 0.0000
Residual	123036.075	9997	12.3072997		R-squared = 0.5003
					Adj R-squared = 0.5002
Total	246200.001	9999	24.6224624		Root MSE = 3.5082

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Phat	2.108621	.3117998	6.76	0.000	1.49743	2.719811
x	2.001747	.0411059	48.70	0.000	1.921171	2.082323
_cons	1.892261	.1842913	10.27	0.000	1.531013	2.253509

A skeptic might simply wonder, however, whether this difference just reflects random noise. After all, the results of a given estimation exercise in terms of estimates of program impact and of the standard error of that estimate nearly always vary across repeated samples of the same size. It is thus fair to ask whether there really is a systematic gain from over-identification above exact identification (or, more generally, whether there is a gain from more instruments).

To try to get some sense of this, in Figure 6.5 we present results for local regression of predicted absolute deviation of the program impact estimate from the true value of 2 on sample size for the just-identified (i.e. using just z_1 as an instrument) and over-identified models (i.e. using both z_1 and z_2 as instruments). It is clear from the figure that there is indeed a persistent performance advantage in the over-identified case.

In Output 6.13 we present the results from this over-identified case using STATA's two-stage least squares package. The results mirror what was observed in the single instrument case in terms of patterns between manual two-stage least squares estimation and estimation with the `ivregress` command. Namely, first stage results are identical while second stage estimated standard errors are larger with the purpose-built two-stage least squares package.

In Output 6.14 we report results from re-estimation of the over-identified model with `ivregress`, but this time using the `gmm` option rather than the `2sls`. The `2sls` option prompts `ivregress`

to perform two-stage least squares estimation, while the `gmm` option instead directs it toward generalized method of moments estimation. The results from Outputs 6.13 and 6.14 are very similar, with the possible exception that the estimates of the standard errors for the various parameter estimates ("first" and "second" stage) in 6.14 are a bit larger. This reflects the fact that the generalized method of moments implementation in `ivregress` by default pursues a weighting matrix designed to address possibly homoskedastic errors.

The interesting thing about generalized method of moments is that in principle it opens the door to the researcher adding all sorts of moments that they feel are appropriate. Whether, beyond a certain point, it is wise to do so is a separate issue (various published and unpublished work suggests that the excessive addition of moments can significantly degrade the performance of the generalized method of moments estimator). Nonetheless, in Output 6.15 the authors demonstrate how to "manually" perform generalized method of moments estimation in STATA. First, the moment (or at least that part of it motivated by a residual) is constructed and then applied to STATA's `gmm` estimator, with separate specification of instruments.

Econometrics can be a somewhat semantically sloppy discipline (e.g. there are many potential meanings to the word "structural" in econometrics). The terminology implied by STATA's `gmm` command is one popular approach in the generalized method of moments tradition. Specifically, in

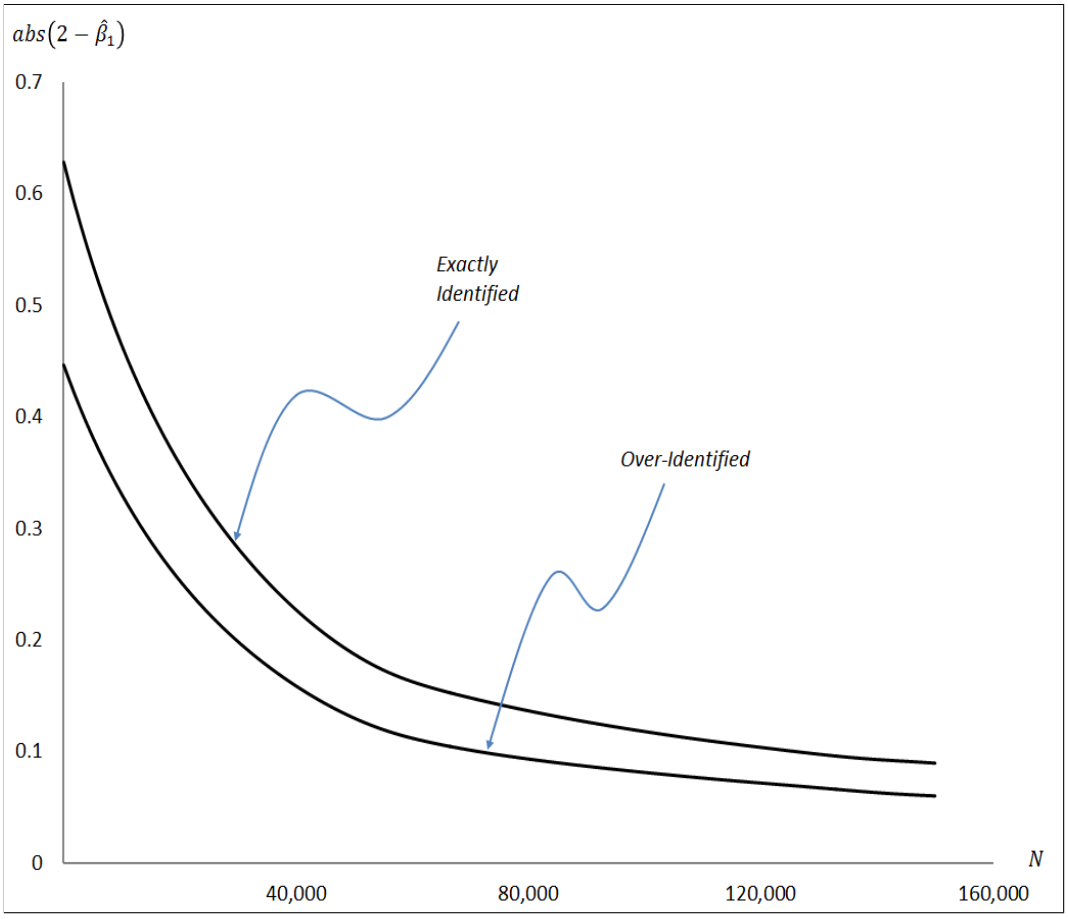


Figure 6.5: TSLS Performance, Over-Identification

this instance the "moment" is simply the residual condition

$$\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right)$$

while the "instruments" are all of the exogenous variables (z_1, z_2 and x).

STATA Output 6.13 (6.1.do)

```
. * Two stage least squares: STATA command
.
. ivregress 2sls Y x (P=x z1 z2), first
First-stage regressions
```

```
                                                Number of obs   =      10000
                                                F(   3,   9996) =    1333.14
                                                Prob > F        =     0.0000
                                                R-squared       =     0.2858
                                                Adj R-squared   =     0.2856
                                                Root MSE        =     0.4170
```

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.119577	.0020847	-57.36	0.000	-.1236635	-.1154906
z1	.0365565	.0020887	17.50	0.000	.0324622	.0406508
z2	-.0428673	.0020938	-20.47	0.000	-.0469716	-.0387629
_cons	.5791768	.0041704	138.88	0.000	.5710019	.5873517

```
Instrumental variables (2SLS) regression              Number of obs =   10000
                                                      Wald chi2(2)  = 9487.14
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.4727
                                                      Root MSE      =  3.6031
```

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.108621	.3202348	6.58	0.000	1.480972	2.736269
x	2.001747	.042218	47.41	0.000	1.919001	2.084493
_cons	1.892261	.1892769	10.00	0.000	1.521285	2.263237

```
Instrumented:  P
Instruments:   x z1 z2
```

We conclude with a brief discussion of the final criteria an instrument must meet, one that we have not discussed thus far: the instrument must have some ability to predict the value of the endogenous variable, in this instance program participation. For a long time, it was not unusual in empirical work to see instrumental variables applications with instruments that were only marginally significant as first-stage predictors of the endogenous variables (in some instances they were not significant at all!).

This pattern of practice eventually provoked scrutiny in the form of a series of famous papers (e.g. Staiger and Stock 1997, Bound et al. 1995, etc.) examining the performance of instrumental variables estimators (particularly two-stage least squares) with "weak" instruments. There are a lot of interesting facets to this work, but the findings can be succinctly summarized as follows: with weak instruments instrumental variables performs poorly. If the instruments are weak enough (as evidenced by very low first stage statistical significance to the estimates of the effect of the instruments on the endogenous variable) the performance of instrumental variables could be quite appalling.

We illustrate this with another local regression exercise, the results of which are presented in Figure 6.6. This time, we compare the predicted absolute deviation of the program impact estimate from local regression of the absolute deviation on sample size using our original example of manual estimation of two-stage least squares with a single instrument but with a modification to generate a "weak" instrument. The weak instrument case is generated simply multiplying the original coefficient on z_1 by 1/10, resulting in the new cost of participation function

$$C_i = 1 + 1.5 \cdot x - .05 \cdot z_1 + .5 \cdot z_2 + \mu + \epsilon_i^C$$

Notice that the coefficient on z_1 is now .05, instead of .5. We have thus reduced drastically the explanatory power of z_1 in terms of program participation. To illustrate, in Output 6.16 we present the first stage equation results with this adjustment. As you can see, the t-statistic of the estimate for z_1 has plunged to 1.45(from 17.22 in Output 6.5): it is no longer a statistically significant predictor of program participation. The second stage estimate of program impact (not reported) is now 6.254036, which is wildly different from the true value of 2.

From Figure 6.6 we can see that across sample sizes the effect of this adjustment in the first stage explanatory power of z_1 has been devastating: the performance of the just-identified two-stage least squares is now terrible at smaller sample sizes (with a predicted absolute deviation of the program impact from the truth on the order of *seven to eight times* the true value of 2). However, the situation only improves to a point as the sample size increases to the extremely generous size

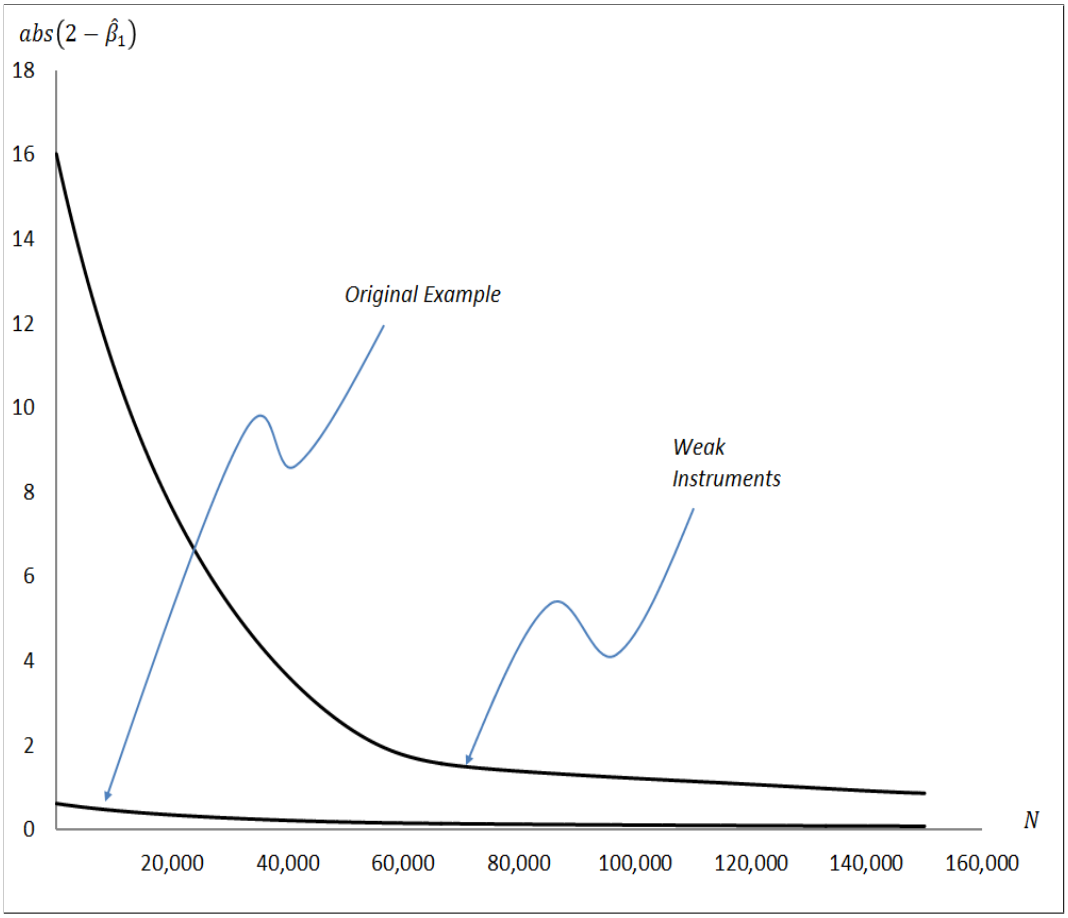


Figure 6.6: TSLS Performance, Weak Instruments

of 150,000: at a sample size of 150,000 the predicted absolute deviation of the program impact estimate is 0.871217, around *ten times* the figure of 0.089567 for the original model at 150,000 observations and larger than the predicted absolute deviation of 0.628647 for the original model at *100 observations*.

STATA Output 6.14 (6.1.do)

```
. * Two stage least squares: STATA command
.
. ivregress gmm Y x (P=x z1 z2), first
First-stage regressions
```

```
                                                Number of obs   =      10000
                                                F(   3,   9996) =    2308.62
                                                Prob > F        =     0.0000
                                                R-squared       =     0.2858
                                                Adj R-squared   =     0.2856
                                                Root MSE        =     0.4170
```

P	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.119577	.0016949	-70.55	0.000	-.1228994	-.1162547
z1	.0365565	.0020593	17.75	0.000	.0325199	.0405931
z2	-.0428673	.0020369	-21.05	0.000	-.04686	-.0388746
_cons	.5791768	.0041754	138.71	0.000	.5709922	.5873614

```
Instrumental variables (GMM) regression         Number of obs =   10000
                                                Wald chi2(2)  = 9351.62
                                                Prob > chi2   =  0.0000
                                                R-squared     =  0.4727
GMM weight matrix: Robust                       Root MSE      =  3.6031
```

Y	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.108944	.324678	6.50	0.000	1.472587	2.745301
x	2.001951	.0430581	46.49	0.000	1.917559	2.086344
_cons	1.893205	.1920943	9.86	0.000	1.516708	2.269703

```
Instrumented:  P
Instruments:   x z1 z2
```

Interestingly, the local predicted absolute deviation of the program impact estimate from the true value of 2 for simple regression of the outcome Y on program participation P and the observed characteristic x (in other words, straightforward regression ignoring the potential endogeneity of P) is fairly steady at around 1.7, falling only incrementally as sample size increases. It is not until around 60,000 observations that the predicted absolute deviation for the weak instrument case falls below this threshold. In other words, unless one has a fairly substantial sample size, it might be better to ignore endogeneity altogether than attempt to address it with instrumental variables estimation with weak instruments. Bartels (1991) actually found that under some (admittedly extreme) circumstances a strong instrument (in terms of first stage ability to predict program participation) correlated with the error term of the equation of interest (i.e. μ or ϵ^Y in our model) might provide more useful estimates of program impact than a weak instrument uncorrelated with the error term. In other words, he found that it is possible that an instrument that violates perhaps the most important theoretical standard for an instrument could be better than a weak instrument.

The most convincing gauge of instrument strength is significance of first stage estimated instrument coefficients, either for single instruments (as manifested by t-statistics and corresponding p-values) or, in the over-identified case, for the instruments collectively (as manifested by the F-statistic[11] from a joint test of the instruments). How much significance is enough remains something of an open question at this point, though some progress has been made. For instance, Staiger and Stock (1997) and Stock and Yogo (2005) essentially suggest a "weak" instrument threshold of an F-statistic value of less 10 for the joint first-stage significance of the instruments.

STATA Output 6.15 (6.1.do)

```
. * GMM manual setup
.
. global XB "{b0}"
. global XB "$XB - {b1}*P"
. global XB "$XB - {b2}*x"
.
. gmm (Y-$XB), instruments(x z1 z2) vce(r)
Step 1
Iteration 0:   GMM criterion Q(b) =  21.929148
Iteration 1:   GMM criterion Q(b) =  .00059405
Iteration 2:   GMM criterion Q(b) =  .00059405
Step 2
Iteration 0:   GMM criterion Q(b) =  .00004626
Iteration 1:   GMM criterion Q(b) =  .00004615
Iteration 2:   GMM criterion Q(b) =  .00004615
GMM estimation
Number of parameters =   3
Number of moments    =   4
Initial weight matrix: Unadjusted                     Number of obs  =   10000
GMM weight matrix:     Robust
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
/b0	1.893205	.1920943	9.86	0.000	1.516708	2.269703
/b1	2.108944	.324678	6.50	0.000	1.472587	2.745301
/b2	2.001951	.0430581	46.49	0.000	1.917559	2.086344

```
Instruments for equation 1: x z1 z2 _cons
```

[11] See the Tool Box on The F-Distribution.

Tool Box: The F-Distribution

Like the normal, Student's t and χ^2 distributions, the F-distribution is a very popular workhorse distribution for hypothesis testing in program impact evaluation and regression analysis more generally. The F-distribution tends to be a "Go To" distribution for hypothesis testing involving linear combinations of estimated parameters from linear regression models. Formally, if w_1 and w_2 are χ^2 distributed random variables with degrees of freedom M and N, respectively, then the distribution of

$$Q = \frac{\frac{w_1}{M}}{\frac{w_2}{N}}$$

is given by the F-distribution with M and N degrees of freedom, which is usually denoted $F_{M,N}$. The F-distribution would be a natural choice for, for instance, testing the *joint significance* of the instruments as determinants of program participation per a linear regression model of program participation.

STATA Output 6.16 (6.1.do)

```
. * First stage of ``manual´´ two-stage least squares,
. * single instrument case
.
. reg P x z1
```

Source	SS	df	MS		
Model	592.466433	2	296.233216	Number of obs	= 10000
Residual	1838.47747	9997	.183902918	F(2, 9997)	= 1610.81
				Prob > F	= 0.0000
Total	2430.9439	9999	.243118702	R-squared	= 0.2437
				Adj R-squared	= 0.2436
				Root MSE	= .42884

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.121639	.0021437	-56.74	0.000	-.1258411	-.1174368
z1	.0031147	.0021479	1.45	0.147	-.0010956	.007325
_cons	.5819928	.0042885	135.71	0.000	.5735864	.5903991

6.1.2 Limited Dependent Variables

The last subsection was concerned mainly with instrumental variables methods most obviously applicable to a linear setting, wherein the outcome of interest Y and the endogenous variable, program participation P in this manual, are treated as continuous variables for estimation purposes. Whether one pursued two-stage least squares or moment-based estimation, the underlying regression models, which serve as the basis for the first and second stage estimation models under two-stage least squares and provide the basis for the moments in the moment-based approaches, were linear in nature.

Clearly, observed program participation as it is typically conceptualized in this manual, and was in the last subsection, actually isn't continuous: it is a binary variable that takes on a value of 1 if the individual participates and 0 if they do not do so. In other words, in the past section, we essentially pretended that observed program participation was a continuous variable for estimation

purposes when it actually was not in the context of the behavioral model of the last subsection. Nor is it typically treated as continuous in real world evaluations: in many, if not most, impact evaluations it is conceptualized as a limited dependent variable that can take on a discrete, finite number of possible values.

In the last subsection we also assumed that the observed outcome of interest was continuous. Sometimes, however, outcomes of interest for which we wish to estimate program impact are in fact limited dependent in nature. In this subsection we discuss instrumental variable models that explicitly recognize the possibility that the outcome, program participation or both are limited dependent variables (in other words, not continuous).

In doing so we enter into a discussion on a topic that is in some sense far more wide ranging and diverse than that associated with linear instrumental variables. There have been many instrumental variables models purpose-built for various circumstances where either the outcome, endogenous variable (in our case program participation) or both are limited dependent in nature. Just these three permutations suggest three distinct modelling circumstances.

There are also all sorts of ways that a dependent variable can be "limited". Typically it is discrete in nature, in which case it could be binary (e.g. participate versus don't participate for program participation; use contraception or do not do so for the outcome), multinomial (e.g. choose from several participation alternatives for the participation variable; utilize any of several alternative health care providers for the outcome) or a count variable (e.g. number of times the individual participated or was "treated", suggesting a dose-response approach to the participation variable; the number of doctor or clinic visits, or number of children born to a woman, as outcome possibilities). Frankly, the possibilities do not end there either.

Just these possibilities for the nature of the limited dependent variables suggest 3X3=9 possible combinations of modelling challenges when both the outcome of interest and program participation are explicitly treated as limited dependent in nature. When one considers that one of the two dependent variables could be continuous, that adds 2X3=6 permutations, for a total of 15 potential modelling circumstances from just this short enumeration of possibilities!

Clearly, this would make for a lengthy discussion of instrumental variables models for limited dependent variables: assuming 3 pages of discussion per possibility (which is conservative) it quickly becomes a mathematically dense and frankly tedious (for the reader and the authors) 45 page journey of uncertain benefit anyway (since it still would not cover every conceivable possibility, but just those briefly enumerated in the past two paragraphs). We therefore discuss just a few models. However, we do so in a manner designed to give the reader some overall feel for typical approaches used for instrumental variables in the limited dependent variable setting as well as a sense of some of the cross-cutting issues associated with those approaches.

We begin by outlining three possible general cases for limited dependent variables involving the two dependent variables considered thus far in this manual (specifically, an outcome of interest and program participation). We use as a departure point the motivating model from the last subsection. We have as an outcome of interest equation

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y \end{aligned}$$

All that we have done is to collect the more involved error term $\beta_3 \cdot \mu_i + \epsilon_i^Y$ into the more expositionally simple error term ξ_i^Y. The individual participates if $Y_i^1 - Y^0 i - C_i > 0$ or, as we derived in the last subsection, if

$$\beta_1 - \gamma_0 - \gamma_1 \cdot x_i - \gamma_2 \cdot z_{1i} - \gamma_3 \cdot z_{2i} - \gamma_4 \cdot \mu_i - \epsilon_i^C > 0$$

The left hand side of this can be thought of as a latent tendency to participate, P_i^*,

$$\begin{aligned} P_i^* &= \beta_1 - \gamma_0 - \gamma_1 \cdot x_i - \gamma_2 \cdot z_{1i} - \gamma_3 \cdot z_{2i} - \gamma_4 \cdot \mu_i - \epsilon_i^C \\ &= \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \end{aligned}$$

The individual then chooses to participate if $P_i^* > 0$. Note that once again we have collected a more involved error term ($-\gamma_4 \cdot \mu_i - \epsilon_i^C$) into a simple summary term (ξ_i^C). Thus we have a continuous outcome of interest and limited dependent observed participation indicator. For the purposes of organizing our discussion, we call this **Case 1**.

This notational approach is common in the program impact evaluation and, more broadly, causal modelling literatures in econometrics and statistics, and essentially gives away which dependent variables are continuous and which are limited in nature. For instance, for the purposes of presenting system of equations such as

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y \\ P_i^* &= \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \end{aligned}$$

The "*" superscript typically indicates a continuous but unobserved latent variable that underlies an observed limited dependent variable. The observed limited dependent variable behind which lies this latent variable is typically indicated by the same variable name sans the "*" superscript. Hence, the latent variable P_i^* lies behind the observed program participation variable P_i. For instance, we might observe program participation as a binary outcome, with $P_i = 1$ if $P_i^* \geq c$ (where c is some constant such as 0) and $P_i = 0$ otherwise. Continuous observed outcomes without restriction on their observed range typically have no such superscript (e.g. Y_i).

Had we instead written

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y \\ P_i &= \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \end{aligned}$$

this would have indicated a limited dependent observed outcome of interest (e.g. a binary outcome of interest) but a continuous program participation outcome. It is not clear whether this case is particularly relevant in the program impact evaluation literature since observed participation is generally designated discretely, typically as a binary variable (as in participated versus did not do so) and sometimes as a count outcome (as in a dose response type model where the number treatments received is the main concern).[12] Less often is it conceptualized in terms of an observed continuous participation measure. Nonetheless, we briefly discuss this circumstance, and refer to it as **Case 2**.

Finally, with the system

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y \\ P_i^* &= \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \end{aligned}$$

we have a latent outcome of interest and latent program participation, indicating that the observed outcome of interest and observed program participation are limited dependent in nature. For instance, we could have a binary indicator of program participation, and wish to know the impact of participation on a binary outcome such as whether the individual uses modern contraception or not. We call this **Case 3**.

We begin with Case 1. Our focus at this point is thus

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

[12] Though as we mentioned other possibilities, such as a multinomial participation variable, are possible.

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

To cut the discussion to a manageable span, we assume that observed P_i is binary in nature, equalling 1 if the individual participates and 0 if they do not. Indeed, as we discuss each successive case, we assume that any observed limited dependent variables are binary in nature.

The reader might ask why one would simply not model P_i with the linear probability model, potentially bringing the instrumental variables estimation within the context of the linear instrumental variables methods discussed in the last subsection. Indeed, this is effectively exactly the route pursued in the last subsection. As we saw, this approach often yielded estimates of program impact very close to the truth.

Some (Angrist and Krueger (2001), for instance) appear to advocate linear instrumental variables techniques whenever either the endogenous variable or outcome of interest is binary in nature. In practice, this amounts to applying the linear probability model to any discrete observed dependent variable. The argument for this effectively rests on the durability of the consistency of linear instrumental variables: specifically, it is a consistent approach to estimation of the average treatment effect even when either the outcome of interest or program participation is binary in nature.[13]

The same cannot be said for estimation of either equation by logit or probit. Since this is Case 1, we focus on the consequences of estimating the "first stage" program participation equation by logit or probit and then plugging the predicted probability of participation from the fitted model into the outcome equation in the place of actual observed participation status. Angrist and Krueger (2001) caution against this strategy: it is consistent only if the assumption regarding the functional form (e.g. probit or logit in the first stage) is exactly right.

Nonetheless, there are alternatives to linear instrumental variables that have been pursued in this circumstance, as well as Cases 2 and 3. There are several reasons for this. First, some object to known shortcomings of the linear probability model (such as its failure to insure that predicted probabilities from fitted models stay within the interval spanning from 0 to 1, inclusive[14]), a deficiency that could in principle become important for some kinds of simulations one might conceivably perform. Second, under some circumstances there are alternatives that might offer more efficient estimates of program impact (i.e. estimates that vary less from sample to sample). Finally, and probably most dubiously, some of the alternatives have on occasion proven attractive for the additional identification that they offer in the face of instruments that might somehow be insufficient.

To get some sense of these alternatives, we work through one or two for each of the three Cases we consider. The idea is not to provide comprehensive coverage of all of the available options, but instead to give the reader a feel for the kinds of alternatives that have been proposed and, in some cases, widely employed. For Case 1 we derive a popular likelihood-based (i.e. involving maximum likelihood estimation) approach to the estimation of this model. This approach involves a specific parametric assumption regarding the joint distribution of the errors $\{\xi^Y, \xi^C\}$. In other words, it involves adopting the assumption that $\{\xi^Y, \xi^C\}$ follow a particular bivariate (i.e. two variable) joint distribution. Specifically, it makes a distributional assumption that has proven extremely popular for application of instrumental variables to limited dependent variables: namely, that the two errors follow the bivariate normal distribution.

To see how the likelihood for this model is constructed, let us consider the problem in general and start by looking at estimation of the two equations separately. Separate estimation of the two equations would be reasonable if the two error terms ξ_i^Y and ξ_i^C were independent. We consider

[13]The consistency of the linear instrumental variables approach of course requires instruments with the properties stipulated in the last subsection.

[14]Probabilities less than 0 or greater than 1 are not meaningful.

them separately at first to give the reader some sense of the logic behind the components of the joint estimation model driven by each equation.

We begin with the program participation equation. To fix ideas, suppose that participation is observed as a binary choice (i.e. P_i can take on the values 1 or 0, indicating participant or non-participant status, respectively). This has in fact been almost without exception our approach to participation to this point in the manual. We assume that the individual participates if

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \geq 0$$

or if

$$-\xi_i^C \leq \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i}$$

The probability that the individual participates is the probability that this expression holds. Given individual i's draws for x, z_1 and z_2, the uncertainty as to whether this holds would revolve around the error term ξ^C.

More specifically, the probability that this expression holds is also the cumulative density of the random variable $-\xi^C$. In most applications, a cumulative density is indicated with a capital F. Then, for a given random variable m and constant c

$$F(c) = Pr(m \leq c)$$

This is in contrast to the probability density $f(\cdot)$, which tells us

$$f(c) = Pr(m = c)$$

Now, a straightforward property of density functions is that

$$F(c) = \int_{-\infty}^{c} f(w)\, dw$$

This is just a very fancy way of saying that the probability that a variable m takes on a value less than or equal to c is equal to the sum of the probabilities of m taking on each of the values from minus infinity through c, for increments to the values of m that are infinitesimally small.[15]

[15] An integral $\int$ can be thought of as summation $\sum$ across the values of a continuous variable. Think of

$$\int_{-\infty}^{c} f(w)\, dw$$

as the sum of all the values $f(w)$ for every tiny incremental change in value w can have in the interval between minus infinity ($-\infty$) and the constant c. To get some perhaps more concrete idea of what the integral in the main text is doing, if m was a discrete variable that had M possible values that it could take on between minus infinity and c. (For instance, if c was 8, then perhaps the only the values m could take below 8 with positive probability might be {-10, -7, -2.3, -0.5, 0, 0.8, 1.9, 4.4 and 7 }, in which case $M = 9$.) Then,

$$Pr(m \leq c) = F(c) = \sum_{w=1}^{M} Pr(m = m_w) = \sum_{w=1}^{M} f(m_w)$$

where m_w is the w^{th} possible value of m. In the specific numerical example we gave in parentheses above, we would have

$$Pr(m \leq 8) = F(8) = Pr(m = -10) + Pr(m = -7) + Pr(m = -2.3) + Pr(m = -0.5) + Pr(m = 0) + Pr(m = 0.8)$$
$$+Pr(m = 1.9) + Pr(m = 4.4) + Pr(m = 7) = f(-10) + f(-7) + f(-2.3) + f(-0.5) + f(0) + f(0.8)$$
$$+f(1.9) + f(4.4) + f(7)$$

The integral in the main text is simply exactly this idea applied to a continuous variable.

With this notation in mind, the probability that individual i participates in the program is

$$\begin{aligned} Pr\left(P_i = 1\right) &= Pr\left(-\xi_i^C \leq \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i}\right) \\ &= F\left(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i}\right) \\ &= \int_{-\infty}^{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}} f\left(w\right) dw \end{aligned}$$

where $f\left(\right)$ is the probability density function for $-\xi^C$ (i.e. $Pr(-\xi^C = c) = f(c)$ and $F(\cdot)$ is its cumulative density). If we assume that $-\xi_i^C$ is the difference in two Type-I extreme value random variables, we can avoid integration altogether since the cumulative density $F(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i})$ is the logistic distribution, giving rise to the logit model.

If we assume that $-\xi^C$ is distributed normally with 0 mean (a standard regression assumption) and variance 1 (that is, that it follows the standard normal distribution), then[16]

$$Pr\left(-\xi^C = c\right) = f\left(c\right) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot exp\left(-\frac{c^2}{2}\right)$$

then, we have

$$Pr\left(P_i = 1\right) = \int_{-\infty}^{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}} \frac{1}{\sqrt{2 \cdot \pi}} \cdot exp\left(-\frac{w^2}{2}\right) dw$$

which is also commonly written

$$Pr\left(P_i = 1\right) = \Phi\left(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i}\right)$$

where $\Phi(\cdot)$ is the cumulative density function for the standard normal distribution.

Before proceeding we note that

$$Pr\left(P_i = 0\right) = 1 - \Phi\left(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i}\right)$$

[16]The assumption that $-\xi^C$ follows the standard normal distribution is standard in probit models. The restriction that $(\sigma^C)^2 = 1$, where σ^C is the standard deviation of $-\xi^C$, can be made without loss of generality since the parameters $\{\delta_0, \delta_1, \delta_2, \delta_3\}$ cannot be separately estimated from the variance of $-\xi^C$ anyway. The variance of $-\xi^C$ is not identified from a fundamental standpoint because we do not observe the full range of variation in the latent variable P^* but instead observe just the discrete outcomes $P = 0$ and $P = 1$. Put differently, the information in the observed variation in P_i generated by the behavioral threshold for participation of

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \geq 0$$

is indistinguishable in terms of what we observe as a result of the participation decision threshold

$$\frac{P_i^*}{\sigma^C} = \frac{\delta_0}{\sigma^C} + \frac{\delta_1}{\sigma^C} \cdot x_i + \frac{\delta_2}{\sigma^C} \cdot z_{1i} + \frac{\delta_3}{\sigma^C} \cdot z_{2i} + \frac{\xi_i^C}{\sigma^C} \geq 0$$

Logit models are subject to a similar need to assume some normalizing value for the variance, though the need is less evident from the derivations shown in this manual (demonstrating this would be a tedious exercise with little profit in terms of the subject at hand). An immediate implication of this is that the parameters $\{\delta_0, \delta_1, \delta_2, \delta_3\}$ that we would estimate under either logit or probit should not be interpreted in any sort of absolute terms. They are not necessarily in the same metric (i.e. to the same scale or units) as the parameters of the original motivating latent variable model

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \geq 0$$

Rather, they are defined only conditional on the scale implied by the assumption that σ^C is equal to some constant (1 in the case of probit).

$$= \int_{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}}^{\infty} \frac{1}{\sqrt{2\cdot\pi}}\cdot exp\left(-\frac{w^2}{2}\right)dw$$

The logic of this is rather straightforward. $-\xi_i^C$ is a continuous random variable that could, in principle, take on any value from minus infinity ($-\infty$) to infinity (∞). We have

$$Pr\left(P_i=1\right) = Pr\left(-\xi_i^C \leq \delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

$$= \int_{-\infty}^{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}} \frac{1}{\sqrt{2\cdot\pi}}\cdot exp\left(-\frac{w^2}{2}\right)dw$$

The last term is just the total probability "mass" of ξ_i^C between $-\infty$ and $\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}$. However, the probability that the individual does not participate is then the probability that ξ_i^C takes on any of the remaining values (those not compatible with participation) from $\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}$ to ∞, or

$$\int_{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}}^{\infty} \frac{1}{\sqrt{2\cdot\pi}}\cdot exp\left(-\frac{w^2}{2}\right)dw$$

Behind all of this math is the simple truth that the probability that the individual participates plus the probability that they don't participate must sum to 1 (because there is no third possibility: either you do or do not participate). Therefore the sum of the total probability of all of the values of ξ_i^C that are compatible with participation and the total probability of all of the values of ξ_i^C not compatible with participation must be 1.

Suppose that we observed $\{P_i, x_i, z_{1i}, z_{2i}\}$ for a sample of N individuals indexed by i ($i = 1, ..., N$). The contribution of individual i to the likelihood function for probit estimation would then be the probability of their observed outcome P_i given $\{x_i, z_{1i}, z_{2i}\}$ and the parameters $\{\delta_0, \delta_1, \delta_2, \delta_3\}$:

$$Pr\left(P_i|\delta_0,\delta_1,\delta_2,\delta_3\right)$$

$$= Pr\left(P=1|\delta_0,\delta_1,\delta_2,\delta_3\right)^{P_i}\cdot Pr\left(P=0|\delta_0,\delta_1,\delta_2,\delta_3\right)^{(1-P_i)}$$

$$= \Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)^{P_i}$$

$$\cdot\left(1-\Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)\right)^{1-P_i}$$

The likelihood for the overall sample would then just be the product of the likelihood functions for each individual:

$$L\left(\delta_0,\delta_1,\delta_2,\delta_3\right) = \prod_{i=1}^{N} Pr\left(P_i|\delta_0,\delta_1,\delta_2,\delta_3\right)$$

Maximum likelihood estimation then seeks those values of $\{\delta_0, \delta_1, \delta_2, \delta_3\}$ that maximize the likelihood function.

We now turn to the outcome equation. This is a linear model for a continuous outcome (Y being continuous under Case 1). Specifically, the regression equation for the outcome is

$$Y_i = \beta_0+\beta_1\cdot P_i+\beta_2\cdot x_i+\xi_i^Y$$

The error term for this is

$$\xi_i^Y = Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i$$

Maximum likelihood estimation of the parameter of those models uses the probability

$$Pr\left(\xi_i^Y = Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i\right) = f\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i\right)$$

All that remains is to make some assumption regarding the functional form of the probability density.

A common assumption is that ξ^Y is normally distributed. The individual's contribution to the likelihood function is the probability for the error draw ξ_i^Y:

$$Pr\left(\xi_i^Y|\beta_0,\beta_1,\beta_2,\sigma^Y\right) = Pr\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i|\beta_0,\beta_1,\beta_2,\sigma^Y\right)$$
$$= \frac{1}{\sigma^Y\cdot\sqrt{2\cdot\pi}}\cdot exp\left(-\frac{1}{2}\cdot\left(\frac{Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i}{\sigma^Y}\right)^2\right)$$

Notice that the standard deviation of ξ^Y, σ^Y, is a parameter to be estimated for this model.[17] The overall likelihood function is then

$$L\left(\beta_0,\beta_1,\beta_2,\sigma^Y\right) = \prod_{i=1}^{N} Pr\left(P_i|\beta_0,\beta_1,\beta_2,\sigma^Y\right)$$

Maximum likelihood estimation seeks values for $\{\beta_0,\beta_1,\beta_2,\sigma^Y\}$ that maximize the likelihood.

We have now gotten some sense of how to separately estimate the regression models for the outcome of interest Y and program participation P. This would be justifiable if ξ_i^Y and ξ_i^C were not statistically related. However, if they were correlated (for instance, via a common unobserved characteristic μ_i present in both ξ_i^Y and ξ_i^C) this would yield biased and inconsistent estimates of program impact.

Joint estimation of the outcome and participation equations (i.e. estimation of the equations together) potentially allows for consistent estimation of program impact. The basic idea behind joint estimation is that we control for the potential endogeneity of program participation in the outcome equation by explicitly capturing any correlation between the error terms in the two equations through the joint estimation process. To jointly estimate we essentially assume that the two error terms follow some bivariate distribution (the bivariate normal distribution would be an example of a specific parametric bivariate distribution) and then use the joint probabilities of the outcome of interest (or, rather, the error term from the outcome of interest) and the individuals observed participation status as their contribution to the likelihood. The bivariate distributional assumption creates an avenue by which the error terms from the outcome and program participation models are correlated.

To get some general sense of how joint estimation is set up, consider the joint probability density function

$$Pr\left(\xi_i^Y = c_1,\xi_i^C = c_2\right) = f\left(c_1,c_2\right)$$

This is simply the bivariate generalization of the univariate densities we considered for separate estimation of the outcome and program participation equations. Combining the logic of the two separate estimations that we developed above, the individual's contribution to the likelihood if they are a participant is

$$\int_{-\infty}^{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}} f\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i,w\right)dw$$

Notice that this basically cannibalizes the logic of the separate maximum likelihood estimation processes that we developed for the participation and outcome equations. It provides the probability that

$$Pr\left(\xi_i^Y = Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i,-\xi_1^C\leq\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

[17] The fundamental reason we can do so is that we observe the full variation in Y.

This is just the joint probability of participation and the outcome value being Y_i, given $\{x_i, z_{1i}, z_{2i}\}$.

Their probability in the event that they do not participate is

$$\int_{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}}^{\infty} f\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i, w\right) dw$$

This follows the logic for the non-participation probability discussed earlier.

The departure point for the particular model for which we examine a simulated example is that ξ_i^Y and ξ_i^C are jointly normally distributed (or more specifically, that they follow the bivariate normal distribution). Joint normality is perhaps the most common parametric assumption for limited dependent variable instrumental variables models. By that we mean that it is the most common specific assumption regarding the particular functional form (i.e. particular distributional assumption) for the joint distribution of ξ_i^Y and ξ_i^C.

Operationalizing this assumption simply involves plugging in the bivariate normal function form for $f(c_1, c_2)$ into

$$\int_{-\infty}^{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}} f\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i, w\right) dw$$

and

$$\int_{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}}^{\infty} f\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i, w\right) dw$$

The bivariate normal density is a bit complex. It is

$$f^B\left(\xi_i^Y, \xi_i^C\right) = \frac{1}{2\cdot\pi\cdot\sigma^Y\cdot\sigma^C\cdot\sqrt{1-\rho^2}}\cdot exp\left(-\frac{1}{2}D\left(\xi_i^Y, \xi_i^C\right)\right)$$

where

$$D\left(\xi_i^Y, \xi_i^C\right) = \frac{1}{1-\rho^2}\left[\left(\frac{\xi_i^Y-\mu_{\xi^Y}}{\sigma^Y}\right)^2+\left(\frac{\xi_i^C-\mu_{\xi^C}}{\sigma^C}\right)^2-2\rho\cdot\frac{\left(\xi_i^Y-\mu_{\xi^Y}\right)\cdot\left(\xi_i^C-\mu_{\xi^C}\right)}{\sigma^Y\cdot\sigma^C}\right]$$

where $f^B(\cdot)$ indicates the bivariate normal density. There are few restrictions that can be imposed on this immediately. First, $\sigma^C = 1$. Further, $\mu_{\xi^Y} = \mu_{\xi^C} = 0$ (because we assume that the error term from the latent variable models have zero mean). With these restrictions in place, we have

$$f^B\left(\xi_i^Y, \xi_i^C\right) = \frac{1}{2\cdot\pi\cdot\sigma^Y\cdot\sqrt{1-\rho^2}}\cdot exp\left(-\frac{1}{2}D\left(\xi_i^Y, \xi_i^C\right)\right)$$

where

$$D\left(\xi_i^Y, \xi_i^C\right) = \frac{1}{1-\rho^2}\left[\left(\frac{\xi_i^Y}{\sigma^Y}\right)^2+\left(\xi_i^C\right)^2+2\rho\cdot\frac{\xi_i^Y\cdot\xi_i^C}{\sigma^Y}\right]$$

Even with the restrictions this is still a very complicated density function.

The probability of the observed Y_i and P_i for individual i is

$$Pr\left(Y_i, P_i|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^Y\right)$$

$$= Prob\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i, P=1|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^Y\right)^{P_i}$$

$$\cdot Prob\left(Y_i-\beta_0-\beta_1\cdot P_i-\beta_2\cdot x_i, P=0|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^Y\right)^{(1-P_i)}$$

$$= \left(\int_{-\infty}^{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}} f\left(Y_i - \beta_0 - \beta_1 \cdot P_i - \beta_2 \cdot x_i, w\right) dw \right)^{P_i}$$

$$\cdot \left(\int_{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}}^{\infty} f\left(Y_i - \beta_0 - \beta_1 \cdot P_i - \beta_2 \cdot x_i, w\right) dw \right)^{(1-P_i)}$$

where the standard normal density function $f^B(\cdot)$ is applied for the density $f(\cdot)$. The adoption of the standard normal density function implies the addition of the parameter ρ, which is the correlation between the error terms ξ_i^C and ξ_i^Y. The overall likelihood function is then just the product of these individual contributions across the sample:

$$L\left(\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho\right) = \prod_{i=1}^{N} Pr\left(Y_i, P_i | \delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho\right)$$

Maximum likelihood estimation seeks values of the parameters $\{\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho\}$ that maximize the likelihood.

STATA Output 6.17 (6.1.do)

```
. etregress Y x, treat(P= x z1 z2)
Iteration 0:   log likelihood =   -31838.6
Iteration 1:   log likelihood =  -31838.25
Iteration 2:   log likelihood = -31838.249
Linear regression with endogenous treatment       Number of obs   =      10000
Estimator: maximum likelihood                     Wald chi2(2)    =    9636.59
Log likelihood = -31838.249                       Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y						
x	1.966486	.0342883	57.35	0.000	1.899282	2.03369
P	1.812895	.245371	7.39	0.000	1.331977	2.293814
_cons	2.063856	.1468008	14.06	0.000	1.776131	2.35158
P						
x	-.4146941	.0088248	-46.99	0.000	-.4319904	-.3973977
z1	.1279286	.0072017	17.76	0.000	.1138135	.1420437
z2	-.149916	.0073103	-20.51	0.000	-.1642439	-.1355881
_cons	.2718587	.0144125	18.86	0.000	.2436106	.3001068
/athrho	-.2882669	.0435902	-6.61	0.000	-.3737021	-.2028317
/lnsigma	1.274525	.0089846	141.86	0.000	1.256915	1.292134
rho	-.2805389	.0401595			-.3572256	-.2000952
sigma	3.577	.032138			3.514562	3.640548
lambda	-1.003488	.1493879			-1.296283	-.7106928

```
LR test of indep. eqns. (rho = 0):   chi2(1) =    41.85   Prob > chi2 = 0.0000
```

This estimator is implemented by STATA's `etregress` command. In Output 6.17, we estimate the over-identified case from the simulation example from the last subsection (in other words, we use both instruments, z_1 and z_2). Obviously, the most important issue is the estimate of program impact, which at 1.81295 is a reasonably accurate estimate of true program impact of 2 (it is not much further from it than the over-identified model estimates presented in the last subsection). This model also provides a test of the correlation of the error terms in the two equations (the outcome equation Y and the participation equation P^*) in the form of the statistical significance

of /athrho, which is simply a function of ρ. It is significant (with a z-statistic of -6.61 and corresponding p-value of 0.000), indicating that the errors are correlated. This is not a surprising result since we created errors correlated by construction via the unobserved characteristic μ common to the two error terms.

Output 6.17 is based on maximum likelihood estimation. In other words, it involves joint estimation of the program participation and outcome equations. This is a relatively straightforward exercise since the bivariate normal density function provides a natural foundation for joint estimation of the two equations via maximum likelihood.[18] There is also a two-step version of this estimator. This is a common alternative to joint estimation in models of this nature assuming normality. Joint estimation is often referred to as "full information" likelihood estimation, because by joint estimating it takes into account all of the information available from the sample in one estimation. By contrast, the two-step estimation is often referred to as "limited information" because actual estimation of the outcome and program participation equations is done separately, insuring that for each estimation process only some of the information in terms of variation related to the system of equations (i.e. the outcome equation and the participation equation) is taken into account. In general full information maximum likelihood has better statistical properties (e.g. smaller standard errors) but can be quite complicated to estimate, depending on the nature of the system of equations.

To get some idea of how the two-step estimation process under the bivariate normality assumption works, we begin with the system of equations for Case 1:

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

To bring this discussion more in line with the classical discussion of this problem (found in Maddala 1982) we introduce new error terms:

$$\xi_i^{YY} = -\xi_i^Y$$

and

$$\xi_i^{CC} = -\xi_i^C$$

Since $\{-\xi_i^Y, -\xi_i^C\}$ follow a bivariate normal distribution (by assumption) so to do $\{\xi_i^{YY}, \xi_i^{CC}\}$. This change of variables just helps to avoid a bunch of confusing issues when determining the sign of the coefficients of some parameters down the line.

With this in hand, we now have

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i - \xi_i^{YY}$$

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} - \xi_i^{CC}$$

Once again, the individual participates if

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} - \xi_i^{CC} \geq 0$$

or

$$\xi_i^{CC} \leq \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i}$$

Therefore, this condition holds when $P_i = 1$.

[18] The STATA command etregress follows common practice by actually reducing the bivariate normal density to a univariate normal density. We omit this step since it is tedious but would not add much to our understanding of the basic mechanics of the model.

Next we consider the expectation of the outcome conditional on participation (i.e. $P_i = 1$):

$$E\left(Y_i|P_i=1\right)=\beta_0+\beta_1+\beta_2\cdot x_i-E\left(\xi_i^{YY}|P_i=1\right)$$

$$=\beta_0+\beta_1+\beta_2\cdot x_i-E\left(\xi_i^{YY}|\xi_i^{CC}\leq\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

This simply invokes the participation condition from the last paragraph. Notice that the expectation of the new error term

$$E\left(\xi_i^{YY}|\xi_i^{CC}\leq\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

is the expectation of one (by assumption) normally distributed random variable (ξ_i^{YY}) conditional on another (by assumption) normally distributed random variable (ξ_i^{CC}) being less than some value ($\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}$). The next step in developing the two-step estimator involves a joint normality assumption between the two equations and then relies on an extremely convenient property of the bivariate normal distribution.

Specifically, if two random variables w and h together follow a bivariate normal distribution, then

$$E\left(w|h<c\right)=\rho\cdot\sigma^W\cdot\frac{-\phi\left(c\right)}{\Phi\left(c\right)}$$

where $\phi(\cdot)$ is the standard normal probability density and $\Phi(\cdot)$ is the standard normal cumulative density. The expression

$$\frac{-\phi\left(c\right)}{\Phi\left(c\right)}$$

is also known as the **Inverse Mill's Ratio**, which is often denoted $\lambda(c)$.[19]

Given this

$$E\left(\xi_i^{YY}|\xi_i^{CC}\leq\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

$$=\rho\cdot\sigma^{\xi^{YY}}\cdot\frac{-\phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}{\Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}$$

[19]This is the bivariate generalization of the Inverse Mill's ratio, which was classically laid out in univariate terms. Specifically, suppose that w is a normally distributed random variable with mean μ and variance σ^2 and c is some constant. Then,

$$E\left(w|w\geq c\right)=\mu+\sigma\cdot\frac{\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}$$

and

$$E\left(w|w\leq c\right)=\mu+\sigma\cdot\frac{-\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}$$

There are actually a number of slightly different versions of the Inverse Mills ratio. For instance, even in the simple univariate case, both

$$\frac{\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}$$

and

$$\frac{-\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}$$

are commonly referred to as the Inverse Mill's ratio. See

http://www.stata.com/support/faqs/statistics/inverse-mills-ratio/

for a rather interesting discussion of the slightly distinct versions of the Inverse Mill's ratio that have been put forth in the bivariate context over the years.

Notice that in principle the Inverse Mill's Ratio could be computed from the fitted probit model from a probit regression of P_i on x_i, z_{1i} and z_{2i}. Plugging the Inverse Mill's ratio back into the original expectation of interest, we have

$$E\left(Y_i|P_i=1\right)=\beta_0+\beta_1+\beta_2\cdot x_i-E\left(\xi_i^{YY}|P_i=1\right)$$

$$=\beta_0+\beta_1+\beta_2\cdot x_i-E\left(\xi_i^{YY}|\xi_i^{CC}\leq\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

$$=\beta_0+\beta_1+\beta_2\cdot x_i+\rho\cdot\sigma^{\xi^{YY}}\cdot\frac{\phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}{\Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}$$

We now have one "branch" of the two step estimator in place, and in developing it have introduced the essential logic of the other branch of the estimator.

The other half of the two-step estimator focuses on the expectation of Y_i conditional on the individual not participating:

$$E\left(Y_i|P_i=0\right)=\beta_0+\beta_2\cdot x_i-E\left(\xi_i^{YY}|P_i=0\right)$$

$$=\beta_0+\beta_2\cdot x_i-E\left(\xi_i^{YY}|\xi_i^{CC}>\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

To develop an expression for

$$E\left(\xi_i^{YY}|\xi_i^{CC}>\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

we note that if two random variables w and h together follow a bivariate normal distribution, then

$$E\left(w|h>c\right)=\rho\cdot\sigma^W\cdot\frac{\phi\left(c\right)}{1-\Phi\left(c\right)}$$

where $\phi(\cdot)$ is the standard normal probability density and $\Phi(\cdot)$ is the standard normal cumulative density. The expression

$$\frac{\phi\left(c\right)}{1-\Phi\left(c\right)}$$

is also referred to as an Inverse Mill's Ratio. Using this result, we have

$$E\left(Y_i|P_i=0\right)=\beta_0+\beta_2\cdot x_i-E\left(\xi_i^{YY}|\xi_i^{CC}>\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)$$

$$=\beta_0+\beta_2\cdot x_i-\rho\cdot\sigma^{\xi^{YY}}\cdot\frac{\phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}{1-\Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}$$

This completes the other "branch" of the two-step estimator.

We thus have the new system

$$E\left(Y_i|P_i=1\right)=\beta_0+\beta_1+\beta_2\cdot x_i+\rho\cdot\sigma^{\xi^{YY}}\cdot\frac{\phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}{\Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}$$

$$E\left(Y_i|P_i=0\right)=\beta_0+\beta_2\cdot x_i-\rho\cdot\sigma^{\xi^{YY}}\cdot\frac{\phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}{1-\Phi\left(\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}\right)}$$

Each of these regressions is referred to as a **censored regression** because it would be estimated with a censored sample.

Censoring and **truncation** refer to two different types of missing data problems. Though much has been written about them, the chief difference is that under censoring we observe the degree of missingness and typically have some information about the incomplete (in the sense that they are missing values for key variables) observations, whereas neither is typically the case with truncation (limiting severely the options for addressing any bias associated with truncation).

The regressions are censored in the sense that in the first case we restrict attention to only the subsample for whom $P_i = 1$, in some sense thus behaving as if Y_i is unobserved for those for whom $P_i = 0$, while for the second equation we proceed as if Y_i could be observed only if $P_i = 0$. Because these two specifications are closely related and the participation decision determines the differences in specification between observations, this basic setup is often referred to as a **switching regression**.

Estimation of the two step model then proceeds as follows:

1. Conduct probit regression of P_i on x_i, z_{1i} and z_{2i};
2. Use the fitted model to calculate the predicted Inverse Mill's Ratios. For participants, this would be
$$\hat{\lambda}_i = \frac{-\phi\left(\hat{\delta}_0 + \hat{\delta}_1 \cdot x_i + \hat{\delta}_2 \cdot z_{1i} + \hat{\delta}_3 \cdot z_{2i}\right)}{\Phi\left(\hat{\delta}_0 + \hat{\delta}_1 \cdot x_i + \hat{\delta}_2 \cdot z_{1i} + \hat{\delta}_3 \cdot z_{2i}\right)}$$
while for non-participants it would be
$$\hat{\lambda}_i = \frac{\phi\left(\hat{\delta}_0 + \hat{\delta}_1 \cdot x_i + \hat{\delta}_2 \cdot z_{1i} + \hat{\delta}_3 \cdot z_{2i}\right)}{1 - \Phi\left(\hat{\delta}_0 + \hat{\delta}_1 \cdot x_i + \hat{\delta}_2 \cdot z_{1i} + \hat{\delta}_3 \cdot z_{2i}\right)}$$
where $\phi(\cdot)$ is then standard normal probability density and $\Phi(\cdot)$ is the standard normal cumulative density;
3. Regress Y_i on P_i, x_i and $\hat{\lambda}_i$.

This multistage approach has the advantage that, in principle, it could be estimated with any statistical software that includes capabilities for linear and probit regression. On the other hand, a package supporting this would have an advantage analogous to that provided by purpose-built linear two-stage least squares: it would produce correct standard errors that account correctly for first stage variation.

The two-step estimator can be implemented in STATA as an option under the `etregress` command. Output 6.18 provides estimates of the two-step version of the model (prompted by the `two` option in the `etregress` command line). The results are more or less the same as in the maximum likelihood case. The coefficient on the predicted Inverse Mills Ratio from the regression of Y_i on P_i, x_i and $\hat{\lambda}_i$ is represented by the coefficient on `lambda` in Output 6.18. A significance test on the coefficient for `lambda` is tantamount to a test of correlation of the error terms $\{\xi_i^Y, \xi_i^C\}$. With a z-statistic of -6.62 and corresponding p-value of 0.000, the coefficient for `lambda` is indeed significant, providing evidence consistent with the error terms being correlated.

Before proceeding to Case 2, we note a major way in which this model departs from the linear instrumental variables framework: it is identified (in the sense that it will at least estimate something) even without exclusion restrictions in the form of variables such as z_1 and z_2 that appear in the program participation equation but not the outcome equation. In the linear two-stage least squares case, the first stage prediction was

$$\hat{P}_i = \hat{\delta}_0 + \hat{\delta}_1 \cdot x_i + \hat{\delta}_2 \cdot z_{1i} + \hat{\delta}_3 \cdot z_{2i}$$

where the $\hat{\delta}$s are the estimated values of the parameters δ from the first stage linear regression of P_i on x_i, z_{1i} and z_{2i}. The presence of the instruments z_{1i} and z_{2i} creates some variation in predicted participation $\hat{P}_i$ not related to x_i. If they had been omitted, the predicted participation would have been

$$\hat{P}_i = \hat{\delta}_0 + \hat{\delta}_1 \cdot x_i$$

Inserting this into

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot \hat{P}_i + \beta_2 \cdot x_i + \epsilon_i^Y \\ &= \beta_0 + \beta_1 \cdot \left(\hat{\delta}_0 + \hat{\delta}_1 \cdot x_i\right) + \beta_2 \cdot x_i + \epsilon_i^Y \\ &= \beta_0 + \beta_1 \cdot \hat{\delta}_0 + \beta_1 \cdot \hat{\delta}_1 \cdot x_i + \beta_2 \cdot x_i + \epsilon_i^Y \end{aligned}$$

Notice that the two terms

$$\beta_1 \cdot \hat{\delta}_1 \cdot x_i$$

and

$$\beta_2 \cdot x_i$$

are not distinguishable from each other in terms of the variation in Y_i that they introduce: they both vary linearly with x_i. Therefore the two terms are not separately identified. This is also referred to as the **multicollinearity** problem in regression.

STATA Output 6.18 (6.1.do)

```
. etregress Y x, treat(P= x z1 z2) two
Linear regression with endogenous treatment     Number of obs      =     10000
Estimator: two-step                             Wald chi2(3)       =  11746.12
                                                Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y						
x	1.989078	.0375077	53.03	0.000	1.915564	2.062592
P	2.002369	.2761537	7.25	0.000	1.461117	2.54362
_cons	1.953914	.1642172	11.90	0.000	1.632054	2.275774
P						
x	-.4150826	.008854	-46.88	0.000	-.4324362	-.3977291
z1	.1269563	.0073346	17.31	0.000	.1125808	.1413318
z2	-.1490121	.0074129	-20.10	0.000	-.163541	-.1344832
_cons	.2721292	.0144512	18.83	0.000	.2438054	.300453
hazard						
lambda	-1.125637	.1699686	-6.62	0.000	-1.45877	-.792505
rho	-0.31327					
sigma	3.59321					

The maximum likelihood and two-step models we have considered for Case 1 that assume that $\{\xi_i^Y, \xi_i^C\}$ follow the bivariate normal distribution do not in principle suffer from this problem: they do not absolutely require exclusion restrictions. The reason is that the assumption of bivariate normality introduces a nonlinear functional form, based on the normal distribution, that allows for identification even in the absence of exclusion restrictions in the form of variables z_{1i} and z_{2i} that appear in the participation equation but not the outcome equation. This is perhaps easiest to see

with the two-step model. The nonlinearity means that the variation generated from a first-stage fitted model involving only x_i will not be linear. The reason is that the normal probability density and cumulative densities $\phi(\cdot)$ and $\Phi(\cdot)$ are not linear. But that means that the variation from the first stage is not exactly the same as the linear variation in x_i already in the outcome equation via the linear term $\beta_2 \cdot x_i$. Thus the nonlinear and linear functions of x can be included in the same regression specification.

To see this, consider a simple regression example involving basic variables W and Q. Starting with the basic, simple model

$$W = \omega_0 + \omega_1 \cdot Q + \zeta$$

it should be straightforward that we cannot separately identify another term linear in Q along the lines of

$$W = \omega_0 + \omega_1 \cdot Q + \omega_2 \cdot Q + \zeta$$

However, we could add and separately identify (i.e. estimate the coefficient parameter for) terms non-linear in Q. Examples might include:

$$W = \omega_0 + \omega_1 \cdot Q + \omega_2 \cdot Q^2 + \zeta$$

$$W = \omega_0 + \omega_1 \cdot Q + \omega_2 \cdot exp\,(Q) + \zeta$$

$$W = \omega_0 + \omega_1 \cdot Q + \omega_2 \cdot ln\,(Q) + \zeta$$

The reason that the parameters ω_1 and ω_2 are technically separately identified (i.e. both estimable) is that the added term does not offer more variation linear in Q but instead non-linear in Q. Identification such as this is **identification by non-linearity**.

STATA Output 6.19 (6.1.do)

```
. etregress Y x , treat(P= x ) two
Linear regression with endogenous treatment     Number of obs      =     10000
Estimator: two-step                             Wald chi2(3)       =  11840.18
                                                Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y						
x	1.9269	.0840427	22.93	0.000	1.762179	2.091621
P	1.480892	.688932	2.15	0.032	.1306097	2.831174
_cons	2.2565	.4013276	5.62	0.000	1.469913	3.043088
P						
x	-.3860353	.0083773	-46.08	0.000	-.4024545	-.3696162
_cons	.2562942	.0139695	18.35	0.000	.2289144	.2836739
hazard						
lambda	-.7343532	.4124184	-1.78	0.075	-1.542678	.0739719
rho	-0.20669					
sigma	3.5529664					

In Outputs 6.19 and 6.20 we repeat the estimation in 6.18 with x_i as the only first and second stage regressor. We begin on a slightly hopeful note with Output 6.19, in which we can see that the estimate of program impact (the most important indicator of performance, after all) is, at

1.480892, in the general neighborhood of the true value of 2. This is certainly an improvement over the estimate of program impact of .263087 from simple regression of Y on P and x reported in Output 6.4. However, to be sure, is not a great estimate. Looking beyond it, there are other somewhat disturbing indications. For instance, the coefficient on `lambda` is certainly significant (with a t-statistic of -1.78 and corresponding p-value of 0.075) but not as strongly as in the model with exclusion restrictions.

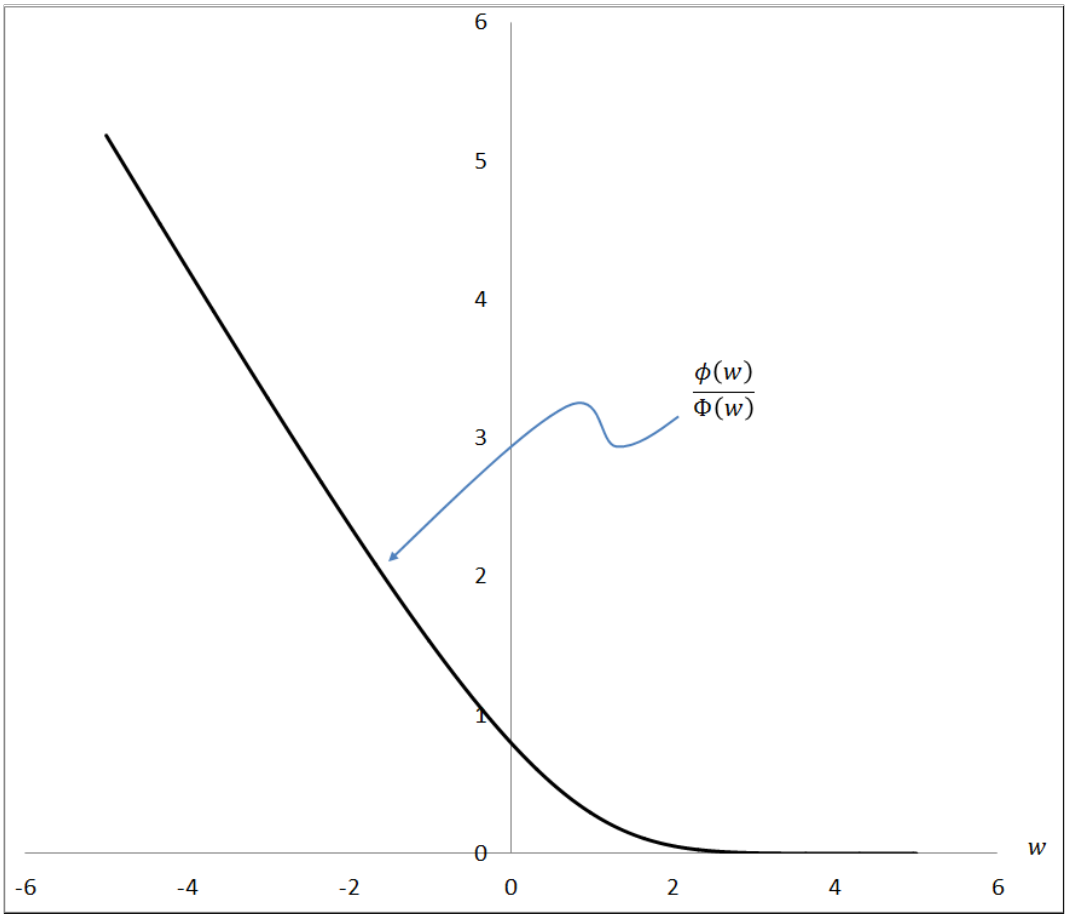


Figure 6.7: The Inverse Mills Ratio

In Output 6.20 we present results from a slight tweaking of the model. Specifically, we reduce the variability in x by a factor of 10. The program impact estimate is now terrible, and the model is no longer capable of picking up the correlation in $\{\xi_i^Y, \xi_i^C\}$, with a t-statistic and p-value for `lambda` of 0.30 and 0.768, respectively. This demonstrates that reasonable performance under identification by non-linearity is not a guarantee and is, indeed, somewhat fragile.

One reason that the reduction in the variation in x_i in the model behind Output 6.20 might have been consequential is that it likely made the variation in the Inverse Mills Ratio far more linear in x_i. The basic problem can be seen in Figure 6.7, which plots the function

$$\frac{\phi(w)}{\Phi(w)}$$

This is essentially one of the Inverse Mills Ratios. Plainly, the Inverse Mills Ratio is nearly linear in w for much of the range of w. An implication of this is that one may need a great deal of

variation in the argument of the Inverse Mills Ratio to be confident that it is truly non-linear in those arguments. By reducing the variation in x_i, we restricted the range of variation in the Inverse Mill's Ratio, potentially to portions largely linear in x_i. This is an important thing to bear in mind when relying on identification by nonlinearity via models that rely on joint normality, particularly their two-stage incarnations relying on the Inverse Mills Ratio: a necessary but not necessarily sufficient condition for good model performance is likely to be substantial enough variation in the first-stage linear prediction[20] to insure that the Inverse Mills Ratio is non-linear enough to insure reasonable identification. One way of assessing this is to graph the predicted Inverse Mills Ratio against the linear first-stage prediction to see how nonlinear in that linear first-stage prediction it appears to be.

STATA Output 6.20 (6.1.do)

```
. etregress Y x , treat(P= x ) two
Linear regression with endogenous treatment     Number of obs      =     10000
Estimator: two-step                             Wald chi2(3)       =    100.19
                                                Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y						
x	1.07308	1.94759	0.55	0.582	-2.744125	4.890285
P	-3.541461	13.17488	-0.27	0.788	-29.36375	22.28083
_cons	5.284988	7.916494	0.67	0.504	-10.23105	20.80103
P						
x	-.3821541	.0636103	-6.01	0.000	-.506828	-.2574802
_cons	.2563553	.0126979	20.19	0.000	.2314678	.2812428
hazard						
lambda	2.415759	8.177005	0.30	0.768	-13.61088	18.44239
rho	0.60623					
sigma	3.9849					

One important potential wrinkle we have not discussed is model performance in the event that the assumption that $\{\xi_i^Y, \xi_i^C\}$ follow the bivariate normal distribution is wrong. In other words, the assumption of bivariate normality provided identification to the model, and it performed well with exclusion restrictions and reasonably well without them but with strong variation in x_i. However, this was in a sample where bivariate normality was not a wholly inappropriate assumption. Indeed, each of the errors was by itself normally distributed.[21] It is reasonable to ask what might have happened had bivariate normality been a less appropriate assumption for this sample. We will briefly discuss this later in this subsection.

We have now worked through a maximum-likelihood instrumental variables model involving the

[20] That is,

$$\hat{\delta}_0 + \hat{\delta}_1 \cdot x_i$$

which is then plugged into the Inverse Mills Ratio.

[21] Each error term was a sum of linear terms involving μ_i and ϵ_i^Y or ϵ_i^C. However, μ and the ϵs were normally distributed by construction of the numerical example, and linear combinations of normally distributed random variables are themselves normally distributed. Hence, if W and S are normally distributed, then so is $R = a \cdot W + b \cdot S$ for any constants a and b.

popular assumption of joint normality of the error terms in the outcome and program participation equations. We have learned three important things:

1. The basics of likelihood construction in this tradition;
2. There is often a two-step variant of the model that relies in some fashion or another on the Inverse Mills Ratio;
3. The model is technically identified by nonlinearity, but the model may not perform well and what performance it can muster might be fragile.

Although this discussion has focused on Case 1, some or all of its lessons carry over to Cases 2 and 3.

We next briefly consider Case 2:

$$Y_i^* = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

$$P_i = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

This is a model that assumes that the outcome is binary but the program participation variable is continuous. As such it is not really a model suitable for many program impact evaluation applications (for which program participation is really a binary choice), but we discuss it because it is an extremely important instrumental variables model that could potentially have some applicability (one might encounter a program participation to which is more or less continuous, as in a limiting case of a dose-response type setup).

Many authors would probably essentially recommend applying linear instrumental variables (e.g. two-stage least squares) to Case 2 on the argument that it will still lead to a consistent estimate of average treatment effects (e.g. Angrist and Krueger (2001), Angrist and Pischke (2009), etc.). In the context, this means that the outcome equation is modelled by the linear probability model. It must be acknowledged, however, that the linear instrumental variables approach in this setting has not met with universal endorsement. For instance, Lewbel et. al. (2012) offer a number of objections to the use of the linear probability model centered mainly on the idea that it is just not a very good fit to the data in many applications, and drive the point home with an example under which the linear probability model might be misleading. It is also worth remembering that our program participation variable is in fact binary. Thus, the estimation of the program participation equation under the assumption that P is continuous is in practice estimation of that equation by the linear probability model. This will become somewhat of an issue when we consider models for Case 2 that assume the bivariate normality of $\{\xi_i^Y, \xi_i^C\}$.

To provide an example of the performance of two-stage least squares (as well as an empirical demonstration of all remaining models in this subsection) we introduce a new simulated sample. The data are generated by STATA do-file 6.2.do, which in terms of data setup is exactly like STATA do-file 6.1.do with one small modification: the potential outcomes, and the observed outcome, are now discrete. Specifically, we constructed the potential outcome equations in exactly the same fashion as before, but now discretized them according to whether or not they exceed zero. Thus, we initially create a continuous potential outcome variable Y^1 exactly as in STATA do-file 6.1.do, but in STATA do-file 6.2.do then re-assign Y^1 the value 1 if the original continuous potential outcome Y^1 exceeds 0, and 0 otherwise. Similarly, we re-assign Y^0 the value 1 if the original continuous potential outcome Y^0 exceeds 0, and 0 otherwise. With both potential outcomes $\{Y^1, Y^0\}$ so discretized, the observed outcome Y (which is just $Y = P \cdot Y^1 + (1 - P) \cdot Y^0$) is also effectively discretized.

STATA Output 6.21 (6.2.do)

```
. * Basic summary statistics: participation
. tab P
```

P	Freq.	Percent	Cum.
0	4,187	41.87	41.87
1	5,813	58.13	100.00
Total	10,000	100.00	

STATA Output 6.22 (6.2.do)

```
. * Basic summary statistics: variable means
.
. by P, sort: summarize Y y1 y0 c x* z* mu epsilon*
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	4187	.8619537	.3449898	0	1
y1	4187	.9343205	.2477506	0	1
y0	4187	.8619537	.3449898	0	1
c	4187	5.575921	2.770066	2.002451	19.11571
x0	4187	1	0	1	1
x	4187	1.130789	1.721677	-5.620321	7.176875
z1	4187	-.3372715	1.952999	-8.078415	6.060055
z2	4187	.383663	1.959136	-6.330251	6.769236
mu	4187	.7467416	1.903202	-5.616416	7.12235
epsilony	4187	-.0304688	2.972894	-10.58663	10.57696
epsilonc	4187	1.772528	2.597622	-7.321414	11.07838

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	5813	.6504387	.4768724	0	1
y1	5813	.6504387	.4768724	0	1
y0	5813	.4747979	.4994074	0	1
c	5813	-2.267171	3.135526	-17.07104	1.997842
x0	5813	1	0	1	1
x	5813	-.829659	1.772296	-8.141914	4.701032
z1	5813	.2619722	1.990245	-7.180359	8.420812
z2	5813	-.3029925	1.965626	-6.960695	6.588125
mu	5813	-.5577272	1.92052	-7.726572	7.586161
epsilony	5813	-.034938	2.950077	-10.366	12.81559
epsilonc	5813	-1.182473	2.631718	-11.73894	8.629721

In Output 6.21 we report basic participation patterns under this slight but important tweak of the original simulation model from STATA do-file 6.1.do. The participation rate is exactly the same as it was before the tweak to discretize Y^1, Y^0 and Y. The reason is that participation was still based on the pre-discretization (i.e. when Y^1 and Y^0 were still continuous) participation decision rule of

$$Y^1 - Y^0 - C \geq 0$$

This can be justified by appealing to the idea that the potential outcomes and cost are in the metric of indirect utility or some related welfare measure from other social sciences.

In Output 6.22 we provide means of the key variables between participants and non-participants. The only real difference between this and the figures in Output 6.2 lies with Y^1, Y^0 and Y, the values of which reflect the discretization process added to STATA do-file 6.2.do. Even then, however, the qualitative patterns to the Ys in Output 6.22 are much the same as in Output 6.2, even if the scale has been changed dramatically by discretization.

Finally (in terms of summary statistics), Output 6.23 provides the correlations among program participation P, the instruments z and the unobservables μ, ϵ^Y and ϵ^C. This table is identical to Output 6.3. The reason for this is that the discretization process has not changed the underlying realities of the latent variable relationships of the model.

STATA Output 6.23 (6.2.do)

```
. * Correlations among observables and unobservables
.
. corr P z* mu epsilony epsilonc
(obs=10000)
```

	P	z1	z2	mu	epsilony	epsilonc
P	1.0000					
z1	0.1481	1.0000				
z2	-0.1701	-0.0034	1.0000			
mu	-0.3188	-0.0001	0.0155	1.0000		
epsilony	-0.0007	0.0090	-0.0083	0.0065	1.0000	
epsilonc	-0.4866	0.0071	0.0078	-0.0092	-0.0041	1.0000

STATA Output 6.24 (6.2.do)

```
. * Cross sectional regression
. reg Y P x
```

Source	SS	df	MS	Number of obs = 10000
Model	522.822468	2	261.411234	F(2, 9997) = 1858.74
Residual	1405.96753	9997	.140638945	Prob > F = 0.0000
				R-squared = 0.2711
Total	1928.79	9999	.19289829	Adj R-squared = 0.2709
				Root MSE = .37502

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P	.0162611	.0086839	1.87	0.061	-.0007612	.0332833
x	.1161857	.0021416	54.25	0.000	.1119877	.1203837
_cons	.7305721	.0062813	116.31	0.000	.7182596	.7428846

True average program impact at the population level is still

$$E\left(Y^1 - Y^0\right)$$

which is estimated in the sample by the average of $Y^1 - Y^0$ across the 10,000 individuals in the sample. Since the Ys are now discretized, this average should be taken across those discretized variables and is interpreted as the change in the probability that the individual makes the binary

choice as the individual switches from non-participant to participant. This average in the sample is .1324. In other words, the program impact is to increase the probability that the individual makes the binary choice by 13.24 percentage points. To put it in more concrete terms, the probability that $Y = 1$ increases by 13.24 percentage points on average as P switches from 0 (i.e. $P = 0$) to 1 (i.e. $P = 1$).

Against this true program impact, Outputs 6.24 and 6.25 provide program impact estimates from simple regression of Y on P and x, in the former case with simple linear regression (i.e. the linear probability model) and in the latter case from logit regression (the results with probit are essentially the same, as the reader can easily confirm with simple modification of STATA do-file 6.2.do to perform probit instead of logit estimation). In both cases the estimate of program impact is far from the truth. The linear probability model provides an estimate of .0162611, only a bit over 10 percent of the true program impact. The figure from the logit model is .0148932, which is in the near neighborhood of the program impact estimate from the linear probability model. Discretization of the Ys has apparently not overcome the deficiencies of the straightforward program impact regressions.

STATA Output 6.25 (6.2.do)

```
. logit Y P x
Iteration 0:   log likelihood = -5741.0029
Iteration 1:   log likelihood = -4269.5901
Iteration 2:   log likelihood = -4124.5278
Iteration 3:   log likelihood = -4122.5123
Iteration 4:   log likelihood =  -4122.512
Logistic regression                               Number of obs   =      10000
                                                  LR chi2(2)      =    3236.98
                                                  Prob > chi2     =     0.0000
Log likelihood =  -4122.512                       Pseudo R2       =     0.2819

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    .112547   .0650359     1.73   0.084    -.014921    .2400149
           x |   .8843894   .0216051    40.93   0.000     .8420442    .9267346
       _cons |   1.504836   .0504358    29.84   0.000     1.405984    1.603688
------------------------------------------------------------------------------

.
. replace P=0
(5813 real changes made)
.
. predict P0
(option pr assumed; Pr(Y))
.
. replace P=1
(10000 real changes made)
.
. predict P1
(option pr assumed; Pr(Y))
.
. g mx= P1-P0
.
. su mx
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mx |     10000    .0148932    .0092435   .0000414   .0281293
```

In Output 6.26 we report the first stage of standard two-stage least squares estimation applied

to Case 2. This is identical to Output 6.11. In both cases we simply apply a linear probability model to the "first stage" participation decision with an over-identified model (i.e. one that includes both instruments, z_{1i} and z_{2i}). As we can see, the two instruments are, even separately considered, re-assuringly statistically significant predictors of program participation.

STATA Output 6.26 (6.2.do)

```
. * First stage of ``manual´´ two-stage least squares,
. * two instrument (over-identified) case
.
. reg P x z1 z2
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  3,  9996) = 1333.14
       Model |  695.526508     3  231.842169            Prob > F      =  0.0000
    Residual |  1738.37659  9996  .173907222            R-squared     =  0.2858
-------------+------------------------------           Adj R-squared =  0.2856
       Total |  2433.9031  9999  .243414651            Root MSE      =  .41702

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.119577   .0020847   -57.36   0.000    -.1236635   -.1154906
          z1 |   .0365565   .0020887    17.50   0.000     .0324622    .0406508
          z2 |  -.0428673   .0020938   -20.47   0.000    -.0469716   -.0387629
       _cons |   .5791768   .0041704   138.88   0.000     .5710019    .5873517
------------------------------------------------------------------------------

.
. predict Phat
(option xb assumed; fitted values)
```

STATA Output 6.27 (6.2.do)

```
. * Second stage of ``manual´´ two-stage least squares,
. * two instrument (over-identified) case
.
. reg Y Phat x
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) = 1867.37
       Model |  524.58987     2  262.294935            Prob > F      =  0.0000
    Residual |  1404.20013  9997  .140462152            R-squared     =  0.2720
-------------+------------------------------           Adj R-squared =  0.2718
       Total |    1928.79  9999   .19289829            Root MSE      =  .37478

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Phat |    .133629     .03331     4.01   0.000     .0683348    .1989233
           x |     .13018   .0043914    29.64   0.000      .121572    .1387881
       _cons |   .6624695   .0196881    33.65   0.000      .623877    .7010621
------------------------------------------------------------------------------
```

Output 6.27 reports the results of the "second stage" regression of the two-stage least squares models. The program impact estimate is .133629. In other words, the two-stage least squares estimate is that program participation (i.e. switching from being a non-participant, $P = 0$, to a participant $P = 1$) increases the probability that $Y = 1$ by 13.3629 percentage points. This is remarkably close to the true impact of a 13.24 percentage point increase in the probability that $Y = 1$ as a result of participation. It would seem that, in this example, the linear instrumental

variables approach as implemented by two-stage least squares performed well. However, the reader would do well to remember per the concerns and example presented in Lewbel et al. (2012) that good performance by the linear probability model is not a guarantee.

We next turn to a maximum likelihood estimator for the system

$$Y_i^* = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

$$P_i = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

The general form for the likelihood for joint estimation is derived from the same basic logic as in Case 1, only in this instance it is the outcome Y_i that is explicitly treated as a binary choice variable while program participation is treated as a continuous variable. The probability for the observed outcome variable values $\{Y_i, P_i\}$ for individual i if their outcome is $Y_i = 1$

$$Pr\left(Y = 1, \xi_i^C | \delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^P\right)$$

$$= Pr\left(Y = 1, P_i - \delta_0 - \delta_1 \cdot x_i - \delta_2 \cdot z_{1i} - \delta_3 \cdot z_{2i} | \delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^P\right)$$

$$= \int_{-\infty}^{\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i} f\left(w, P_i - \delta_0 - \delta_1 \cdot x_i - \delta_2 \cdot z_{1i} - \delta_3 \cdot z_{2i}\right) dw$$

For the $Y_i = 0$ case the relevant probability would be

$$Pr\left(Y = 0, \xi_i^C | \delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^P\right)$$

$$= Pr\left(Y = 0, P_i - \delta_0 - \delta_1 \cdot x_i - \delta_2 \cdot z_{1i} - \delta_3 \cdot z_{2i} | \delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho, \sigma^P\right)$$

$$\int_{\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i}^{\infty} f\left(w, P_i - \delta_0 - \delta_1 \cdot x_i - \delta_2 \cdot z_{1i} - \delta_3 \cdot z_{2i}\right) dw$$

What remains for joint estimation is to make an explicit assumption regarding the functional form of the joint density $f(\cdot, \cdot)$. Once again, the bivariate normal probability density has proven to be an extremely popular choice for $f(\cdot, \cdot)$. This is the key assumption behind many maximum likelihood estimation routines for this framework (continuous endogenous explanatory variable and a binary outcome of interest) available in many commercial statistical programs. For instance, it is the key assumption behind STATA's `ivprobit` command.

Results from maximum likelihood estimation of this system with `ivprobit` are provided in Output 6.28. Since the outcome equation is a probit model, the impact estimate is really the marginal effect of program participation on the probability that $Y = 1$. Computation of this marginal effect is performed manually following estimation. The program impact estimate emerging from this exercise is .1367863, which is quite close to the true value of .1324. The statistical significance of the parameter estimate for `/athrho` is tantamount to a test of whether the equation errors $\{\xi_i^Y, \xi_i^C\}$ are indeed correlated. This parameter is significant, with a z-statistic of -4.13 and an associated p-value of 0.000, indicating error correlation.

There is a tradition for two-step estimation of this model. A particularly popular approach was proposed by Rivers and Vuong (1988) and further explored by Bollen et al. (1995).[22] Essentially, the three steps involved are:

1. Estimate the participation equation by least squares regression of P_i on x_i, z_{1i} and z_{2i};

[22] See as well Terza et. al. (2008) for a very useful discussion of two-step estimation in this context.

2. Use the estimated model to get the predicted residuals $\hat{P}_i^{RES} = P_i - \hat{\delta}_0 - \hat{\delta}_1 \cdot x_i - \hat{\delta}_2 \cdot z_{1i} - \hat{\delta}_3 \cdot z_{2i}$;
3. Perform logit regression of Y_i on P_i, x_i and $\hat{P}_i^{RES}$.

This model has the benefit of simplicity in execution as well as a natural test of error correlation based on the significance of the estimated coefficient on $\hat{P}^{RES}$ in the second stage. We refer to this as a "residual inclusion" estimator but it is also called a **control function** approach.

STATA Output 6.28 (6.2.do)

```
. ivprobit Y x (P=x z1 z2)
Fitting exogenous probit model
Iteration 0:   log likelihood = -5741.0029
Iteration 1:   log likelihood = -4160.7851
Iteration 2:   log likelihood =  -4122.031
Iteration 3:   log likelihood =  -4121.874
Iteration 4:   log likelihood =  -4121.874
Fitting full model
Iteration 0:   log likelihood = -9563.0923
Iteration 1:   log likelihood = -9556.4996
Iteration 2:   log likelihood = -9554.6775
Iteration 3:   log likelihood = -9554.6744
Iteration 4:   log likelihood = -9554.6744
Probit model with endogenous regressors           Number of obs   =      10000
                                                  Wald chi2(2)    =    2379.47
Log likelihood = -9554.6744                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   .5983606   .1291341     4.63   0.000     .3452624    .8514587
           x |   .5562321   .0140512    39.59   0.000     .5286922    .5837721
       _cons |   .5346809   .0879501     6.08   0.000     .3623018      .70706
-------------+----------------------------------------------------------------
     /athrho |  -.2449417   .0592906    -4.13   0.000    -.3611492   -.1287342
   /lnsigma  |  -.8748167   .0070711  -123.72   0.000    -.8886757   -.8609576
-------------+----------------------------------------------------------------
         rho |  -.2401579    .055871                     -.3462259   -.1280277
       sigma |   .4169384   .0029482                      .4111999     .422757
------------------------------------------------------------------------------
Instrumented:  P
Instruments:   x z1 z2
------------------------------------------------------------------------------
Wald test of exogeneity (/athrho = 0): chi2(1) =    17.07 Prob > chi2 = 0.0000
.
. replace P=0
(5813 real changes made)
.
. predict P0, pr
.
. replace P=1
(10000 real changes made)
.
. predict P1, pr
.
. g mx=P1-P0
.
. su mx
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mx |     10000    .1367863    .0783594   2.86e-06   .2351975
```

As developed by Rivers and Vuong (1988), this approach also relies on the joint normality of the error terms $\{\xi_i^Y, \xi_i^C\}$. The reason is admittedly perhaps not obvious. If one assumes that the error terms $\{\xi_i^Y, \xi_i^C\}$ from

$$Y_i^* = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

$$P_i = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

are jointly normal, then an important property of joint normality that the Rivers and Vuong approach exploits is that

$$\xi_i^Y = \theta \cdot \xi_i^C + \eta_i^Y$$

where η_i^Y is also normally distributed. This means that the error term can be broken down additively (i.e. as a sum of terms) into a part explained by the other error term with which it is jointly normally distributed (i.e. ξ_i^C) and a term unrelated to ξ_i^C (i.e. η_i^Y). Note that the correlation of the two error terms $\{\xi_i^Y, \xi_i^C\}$ is captured through the term $\theta \cdot \xi_i^C$.

This property of jointly normal variables is the basic motivation for the Rivers and Vuong (1988) derivation of the residual inclusion estimator. There are three key threads to their approach:

1. $\xi_i^Y = \theta \cdot \xi_i^C + \eta_i^Y$;
2. all correlation between $\{\xi_i^Y, \xi_i^C\}$ is captured through the term $\theta \cdot \xi_i^C$;
3. the term η_i^Y is normally distributed.

For the purposes of estimating the outcome equation, the first two suggest that inclusion of ξ_i^C as a regressor (along with P_i and x_i) should control for any correlation between $\{\xi_i^Y, \xi_i^C\}$ that could render an endogenous variable. The third point, that η_i^Y is normally distributed, allows one to motivate probit regression of the equation of interest. The only remaining detail is to use an estimate of ξ_i^C, $\hat{P}_i^{RES}$, from the first stage least squares regression of program participation P_i on x_i, z_{1i} and z_{2i}.

Strictly speaking, the residual inclusion/control function approach is not really appropriate in this application. The reason is that the estimator is motivated by a *continuous* endogenous variable. We simply do not have a continuous endogenous variable (our endogenous variable, program participation P, is binary). This has several implications. The predicted residuals from the first stage regression simply are not continuous. This means, among other things, that they are not actually normally distributed. More generally, they offer only the roughest, most choppy proxy for the actual error term ξ_i^C underlying what is actually the latent variable relationship (even though under the control function/residual inclusion approach we effectively pretend that it is not) driving observed program participation. It is thus likely that in many applications the predicted first stage residuals from linear probability model estimation of the residuals of an endogenous variable that is in fact binary will just prove too crude to support good model performance.

We also emphasize that this model is not identified by non-linearity alone. Exclusion restrictions in the form of instruments that influence directly program participation P but not the outcome of interest Y are required to estimate the model. Moreover, Bollen et al. (1995) find that decent model performance (i.e. in our context reasonably reliable estimates of program impact) depends on having strong first stage instruments and decent overall explanatory power to the first-stage model determining the endogenous variable (with R^2 values of at least .2 and preferably .3 or more).

Outputs 6.29 and 6.30 report results from manual estimation of the two-step residual inclusion estimator. The first stage is reported in Output 6.29. Despite the fact that participation is binary and not continuous (and hence contrary to the assumption of the model) the remaining

results are mostly reassuring. First, the instruments z_{1i} and z_{2i} are quite statistically significant (with t-statistics of 17.50 and -20.47, respectively). Moreover, overall model explanatory power is reasonable, with an R^2 of 0.2858.

STATA Output 6.29 (6.2.do)

```
. * First Stage, manual residual inclusion
.
. reg P x z1 z2
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  3,  9996) = 1333.14
       Model |  695.526508     3  231.842169            Prob > F      =  0.0000
    Residual |  1738.37659  9996  .173907222            R-squared     =  0.2858
-------------+------------------------------           Adj R-squared =  0.2856
       Total |  2433.9031  9999  .243414651            Root MSE      =  .41702

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.119577   .0020847   -57.36   0.000    -.1236635   -.1154906
          z1 |   .0365565   .0020887    17.50   0.000     .0324622    .0406508
          z2 |  -.0428673   .0020938   -20.47   0.000    -.0469716   -.0387629
       _cons |   .5791768   .0041704   138.88   0.000     .5710019    .5873517
------------------------------------------------------------------------------

.
. predict Phat
(option xb assumed; fitted values)
.
. g Pres=P-Phat
```

Output 6.30 provides the second stage results. The program impact estimate is based on the estimation of the marginal effect of participation on the probability that $Y = 1$ using the fitted model. The impact estimate is .1373265, which is in the near neighborhood of the truth. The predicted residual `Pres` is significant, with a t-statistic of -4.10. This result is indicative of error correlation between the outcome of interest equation and the program participation equation.

There is a two-step version of the STATA command `ivprobit`, invoked with the option `twostep`. It is not exactly the basic residual inclusion approach a la Rivers and Vuong (1988) but is instead based mainly on Newey (1987), a related estimation approach the details of which are not important in present circumstances. The command syntax would, in this case, be

```
ivprobit Y x (P=x z1 z2), twostep
```

Essentially, it relies on a motivating logic similar to the Rivers and Vuong (1988) approach.

Finally, we address Case 3:

$$Y_i^* = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

we have a latent outcome of interest and latent program participation, indicating that the observed outcome of interest and observed program participation are limited dependent in nature. As we have done thus far in this discussion, we assume that $\{Y_i, P_i\}$ are binary in nature.

Clearly, one potential strategy for this model would be linear instrumental variables. We have effectively already considered this in Outputs 6.26 and 6.27. As we saw, in this example at least it produced a fairly good estimate of program impact.

STATA Output 6.30 (6.2.do)

```
. * Second stage, manual residual inclusion
.
. probit Y P x Pres
Iteration 0:   log likelihood = -5741.0029
Iteration 1:   log likelihood =  -4153.322
Iteration 2:   log likelihood = -4113.6263
Iteration 3:   log likelihood = -4113.4563
Iteration 4:   log likelihood = -4113.4563
Probit regression                                 Number of obs   =      10000
                                                  LR chi2(3)      =    3255.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -4113.4563                       Pseudo R2       =     0.2835
------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    .616391   .1397584     4.41   0.000     .3424697    .8903123
           x |   .5730003   .0201965    28.37   0.000     .5334159    .6125848
        Pres |  -.5933588   .1447704    -4.10   0.000    -.8771036    -.309614
       _cons |   .5508058    .082158     6.70   0.000     .3897791    .7118326
------------------------------------------------------------------------------
.
. replace P=0
(5813 real changes made)
.
. predict P0
(option pr assumed; Pr(Y))
.
. replace P=1
(10000 real changes made)
.
. predict P1
(option pr assumed; Pr(Y))
.
. g mx=P1-P0
.
. su mx
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mx |     10000    .1373265    .0830362   5.13e-06   .2420664
```

A general form for the basic full information maximum likelihood based approach[23] to estimating this system is a straightforward extension of what we have developed thus far. We now have four possible combinations of the outcome and participation variables. First, it could be that both equal 1 (i.e. $Y_i = P_i = 1$). This would lead to the general probability

$$Pr\left(Y=1,P=1|\delta_0,\delta_1,\delta_2,\delta_3,\beta_0,\beta_1,\beta_2,\rho\right) =$$

$$= \int_{-\infty}^{\beta_0+\beta_1\cdot P_i+\beta_2\cdot x_i}\int_{-\infty}^{\delta_0+\delta_1\cdot x_i+\delta_2\cdot z_{1i}+\delta_3\cdot z_{2i}} f\left(w,g\right)dgdw$$

where we now have double integration $\int\int$. For $\{Y_i = 0, P_i = 1\}$, the probability would be

$$Pr\left(Y=0,P=1|\delta_0,\delta_1,\delta_2,\delta_3,\beta_0,\beta_1,\beta_2,\rho\right) =$$

[23] We are unaware of a limited information maximum likelihood approach to estimation of this system, but that is probably more of an indication of our ignorance than anything else.

$$= \int_{\beta_0+\beta_1 \cdot P_i+\beta_2 \cdot x_i}^{\infty} \int_{-\infty}^{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}} f(w,g)\, dgdw$$

For $\{Y_i = 1, P_i = 0\}$ we have

$$Pr(Y = 1, P = 0|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho) =$$

$$= \int_{-\infty}^{\beta_0+\beta_1 \cdot P_i+\beta_2 \cdot x_i} \int_{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}}^{\infty} f(w,g)\, dgdw$$

Finally, for $\{Y_i = 0, P_i = 0\}$ the probability is

$$Pr(Y = 0, P = 0|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho) =$$

$$= \int_{\beta_0+\beta_1 \cdot P_i+\beta_2 \cdot x_i}^{\infty} \int_{\delta_0+\delta_1 \cdot x_i+\delta_2 \cdot z_{1i}+\delta_3 \cdot z_{2i}}^{\infty} f(w,g)\, dgdw$$

The probability of the individual's observed outcomes $\{Y_i, P_i\}$ is then

$$Pr(Y_i, P_i|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho)$$

$$= (Pr(Y = 1, P = 1|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho))^{(Y_i \cdot P_i)}$$
$$\cdot (Pr(Y = 0, P = 1|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho))^{((1-Y_i) \cdot P_i)}$$
$$\cdot (Pr(Y = 1, P = 0|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho))^{(Y_i \cdot (1-P_i))}$$
$$\cdot (Pr(Y = 0, P = 0|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho))^{((1-Y_i) \cdot (1-P_i))}$$

The overall likelihood is then simply the product of the individual probabilities

$$Pr(Y_i, P_i|\delta_0, \delta_1, \delta_2, \delta_3, \beta_0, \beta_1, \beta_2, \rho)$$

for each of the $i = 1, ..., N$ individuals in the sample.

What remains for full information maximum likelihood estimation is simply to make some assumption about the density $f(\cdot, \cdot)$. A common assumption is, unsurprisingly, bivariate normality. This gives rise to a model known as the bivariate probit.

In Output 6.31 we present results from the estimation of the bivariate probit in STATA via the `biprobit` command. As with the `ivprobit` command the outcome equation is modelled effectively by probit (specifically, the equivalent to probit arising from the bivariate normal distrution). [24] Therefore, formal program impact must be computed with marginal effects from the fitted model. The estimate of program impact so computed is .1381713, which is not far from the truth. Finally, the z-statistic on `/athrho` is -4.63, with a corresponding p-value of 0.000. This is indicative of a significant underlying bivariate correlation ρ and hence correlation between $\{\xi_i^Y, \xi_i^C\}$.

Thus far, we have introduced three limited dependent variable model scenarios for which one might wish to pursue instrumental variables estimation of program impact. In each we considered linear instrumental variables. We then explored maximum likelihood approaches. Generally this included both full information joint estimation of the output and program impact equations and multiple stage limited information estimation. To operationalize the maximum likelihood approach, some assumption about the distribution of the error terms in the two equations must be made. We focused on joint normality. It is probably the most popular distributional assumption in this setting and involves (relatively) analytically simple likelihood functions.

[24] The participation equation is modelled similarly.

STATA Output 6.31 (6.2.do)

```
. * Biprobit
. biprobit (Y= x P) (P=x z1 z2)
Fitting comparison equation 1:
Iteration 0:   log likelihood = -5741.0029
Iteration 1:   log likelihood = -4160.7851
Iteration 2:   log likelihood =  -4122.031
Iteration 3:   log likelihood =  -4121.874
Iteration 4:   log likelihood =  -4121.874
Fitting comparison equation 2:
Iteration 0:   log likelihood = -6798.6892
Iteration 1:   log likelihood = -5106.9076
Iteration 2:   log likelihood =  -5102.731
Iteration 3:   log likelihood = -5102.7308
Comparison:    log likelihood = -9224.6048
Fitting full model:
Iteration 0:   log likelihood = -9224.6048
Iteration 1:   log likelihood = -9215.0515
Iteration 2:   log likelihood = -9213.1451
Iteration 3:   log likelihood =  -9213.144
Iteration 4:   log likelihood =  -9213.144
Seemingly unrelated bivariate probit              Number of obs   =      10000
                                                  Wald chi2(5)    =    4473.88
Log likelihood =  -9213.144                       Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y						
x	.5566868	.0134387	41.42	0.000	.5303474	.5830262
P	.6019772	.1124713	5.35	0.000	.3815375	.8224169
_cons	.5207007	.0801426	6.50	0.000	.3636241	.6777773
P						
x	-.4155855	.0088425	-47.00	0.000	-.4329165	-.3982545
z1	.1265757	.0072732	17.40	0.000	.1123204	.140831
z2	-.1492417	.007365	-20.26	0.000	-.1636769	-.1348066
_cons	.272101	.0144362	18.85	0.000	.2438066	.3003954
/athrho	-.3623361	.0782835	-4.63	0.000	-.515769	-.2089031
rho	-.34727	.0688428			-.4744279	-.2059164

```
Likelihood-ratio test of rho=0:     chi2(1) =  22.9216    Prob > chi2 = 0.0000
. replace P=0
(5813 real changes made)
. predict P0, pmarg1
. replace P=1
(10000 real changes made)
. predict P1, pmarg1
.
. g mx=P1-P0
.
. su mx
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mx	10000	.1381713	.0786861	2.98e-06	.2365768

The maximum likelihood models involving joint normality that we considered generally per-

formed fairly well from the standpoint of delivering estimates of program impact in the near neighborhood of the truth. However, it must be stressed that the simulated examples in STATA do-files 6.1.do and 6.2.do involved error terms in the outcome and participation equations that were jointly normal. Recall that the system of equations is

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$
$$\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

and

$$\begin{aligned} P_i^* &= \beta_1 - \gamma_0 - \gamma_1 \cdot x_i - \gamma_2 \cdot z_{1i} - \gamma_3 \cdot z_{2i} - \gamma_4 \cdot \mu_i - \epsilon_i^C \\ &= \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \end{aligned}$$

We thus have the error terms

$$\xi_i^Y = \beta_3 \cdot \mu_i + \epsilon_i^Y$$
$$\xi_i^C = -\gamma_4 \cdot \mu_i - \epsilon_i^C$$

In the simulated example μ_i, ϵ_i^Y and ϵ_i^C were drawn from the normal distribution. Since linear sums of normally distributed random variables are themselves normally distributed ξ_i^Y and ξ_i^C are normally distributed as well. Moreover, since μ_i appears in both ξ_i^Y and ξ_i^C they are correlated. In other words, the error terms in our simulation samples were indeed jointly normally distributed.

However, there is no real reason to believe that the error terms are necessarily jointly normally distributed in real world samples. In fact, in many applications is seems somewhat implausible. Suppose, for instance, the program is a weight control initiative and the outcome is a binary indicator of whether an individual is overweight or not. It is easy to imagine that at least a substantial minority of the population falls into extreme categories such as these, regardless of their observed characteristics such as age, race, income, etc., such as:

- Those who will always elect to enroll in the program and will always avoid being overweight (implying a very large positive value of ξ^C but a very large negative value for ξ^Y);
- Those who will always elect to enroll in such programs but never avoid being overweight (implying a very large positive value for both ξ^C and ξ^Y);
- Those who will never enroll in such a program but always be overweight (implying a large negative value for ξ^C and a large positive value for ξ^Y).

The point is that in the real world it is likely that for many applications at least a significant minority of the population or any representative sample from it is characterized by pronounced unobserved behavioral tendencies regarding program participation and the outcome of interest.

However, this is exactly the sort of behavior that the normal distribution is not particularly suitable for modelling: the normal distribution tends to be a poor framework for capturing "extreme" behavior. For instance, under the normal distribution roughly 68 percent of the population (or any reasonable sized representative sample from it) has a value for a normally distributed characteristic within one standard deviation of the mean, while over 99 percent have a value for that characteristic within three standard deviations of the mean.[25] It is not really a distribution

[25] This is sometimes referred to as the "68-95-99.7" rule. If W is normally distributed with mean μ and standard deviation σ,

$$Pr\left(\mu - \sigma \leq W \leq \mu + \sigma\right) \approx 0.683$$
$$Pr\left(\mu - 2 \cdot \sigma \leq W \leq \mu + 2 \cdot \sigma\right) \approx 0.955$$
$$Pr\left(\mu - 3 \cdot \sigma \leq W \leq \mu + 3 \cdot \sigma\right) \approx 0.997$$

where $\approx$ means "equals approximately".

for capturing behavioral tendencies that are extreme (for instance, in the sense of involving large positive or negative error values that overwhelm any role for observed characteristics) but not rare in the population.

To make this a bit more concrete, in Figure 6.8 we plot a typical bivariate normal density. Specifically, we plot the joint probabilities for various values of two random variables, ξ^Y and ξ^C, that we assume together follow a bivariate normal distribution. Notice that this distribution is has a sharply defined summit. This summit occurs at the mean values for ξ^Y and ξ^C. The height of the surface (along the vertical axis labelled $Pr(\xi^Y, \xi^C)$) represents the joint probability of the combination of any particular combination of values of ξ^Y and ξ^C along the horizontal axes occurring. Notice that the joint probability of combinations of ξ^Y and ξ^C falls quickly as we move away from the summit of the probability surface at the means of ξ^Y and ξ^C. What this means basically is that values away from the means of ξ^Y and ξ^C have a rapidly declining probability of occurring, ruling out "extreme" behavior for even a significant minority of the population. Since this distribution only has one peak (i.e. it is "unimodal") it also rules out distinct concentrations of extreme types of individuals. This distribution simply is not a good vehicle for capturing the likely profile of unobserved types that might emerge in the context of the just-discussed weight control initiative.

A better representation of a population with different pronounced unobserved types is found in

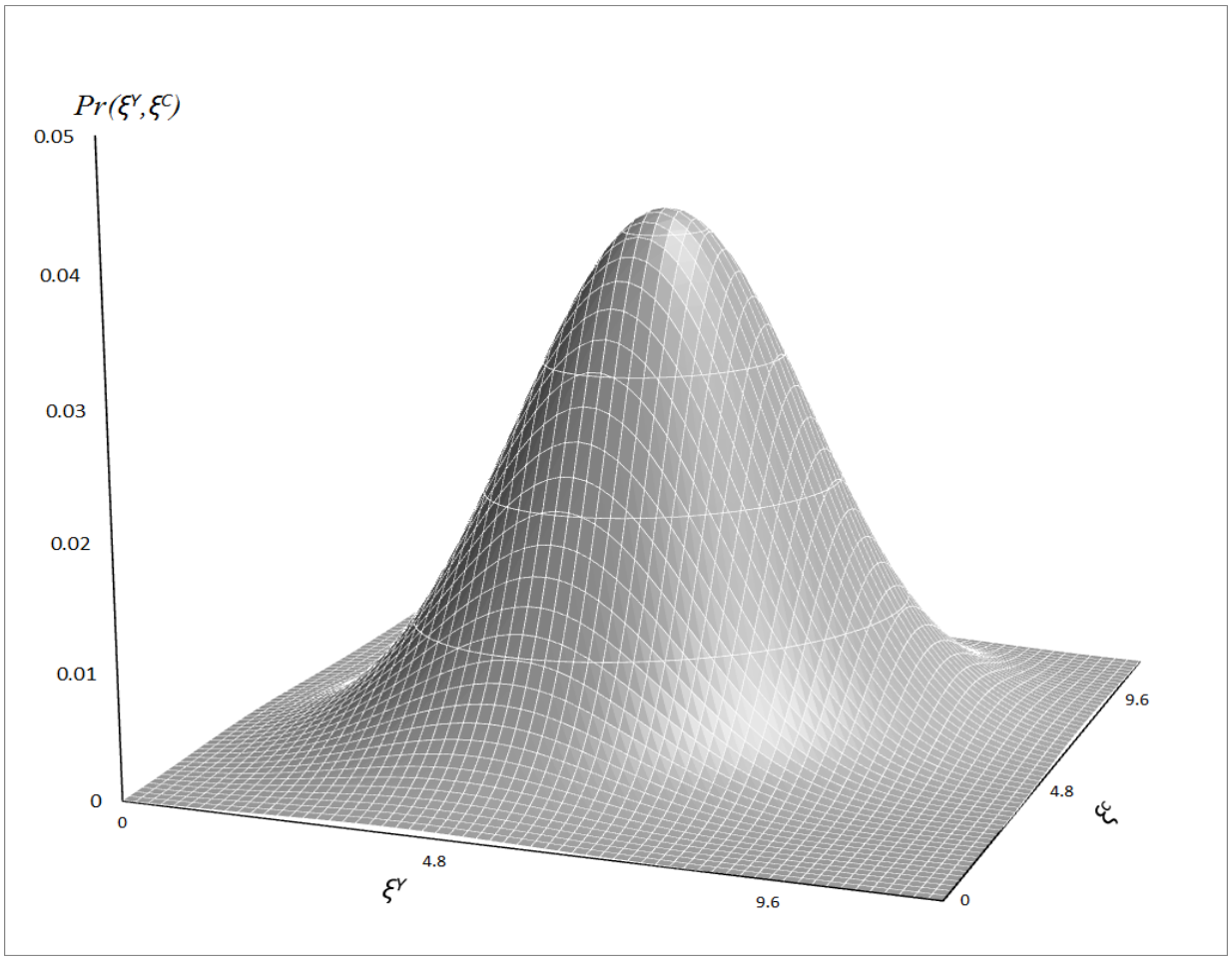


Figure 6.8: A Bivariate Normal Density

Figure 6.9. This once again simply plots the joint probabilities for various values of two random variables, ξ^Y and ξ^C, but in this case the surface has multiple peaks. In other words, it is "multimodal". In this Figure, there are four distinct peaks. Once can think of each peak as representing one pronounced "type" within the population.

A recent paper (Guilkey and Lance 2014) considers the performance of various estimators of program impact when the outcome and program participation are both binary. Specifically, the authors perform Monte Carlo experiments under which they apply these various estimators of program impact (some of which you have just learned about) to the behavioral framework

$$Y_i^* = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

and

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

This is, of course, Case 3 from our discussion. Each Monte Carlo experiment involves a particular combination of features for the sample stemming from this behavioral model: the number of observations, the commonness Y and P (e.g. the percentage of the sample that experiences the outcome $Y = 1$ or participates in the program), the explanatory strength of the instruments $\{z_1, z_2\}$, the distribution of the error terms $\{\xi^Y, \xi^C\}$, etc. For the last of these characteristics they consider

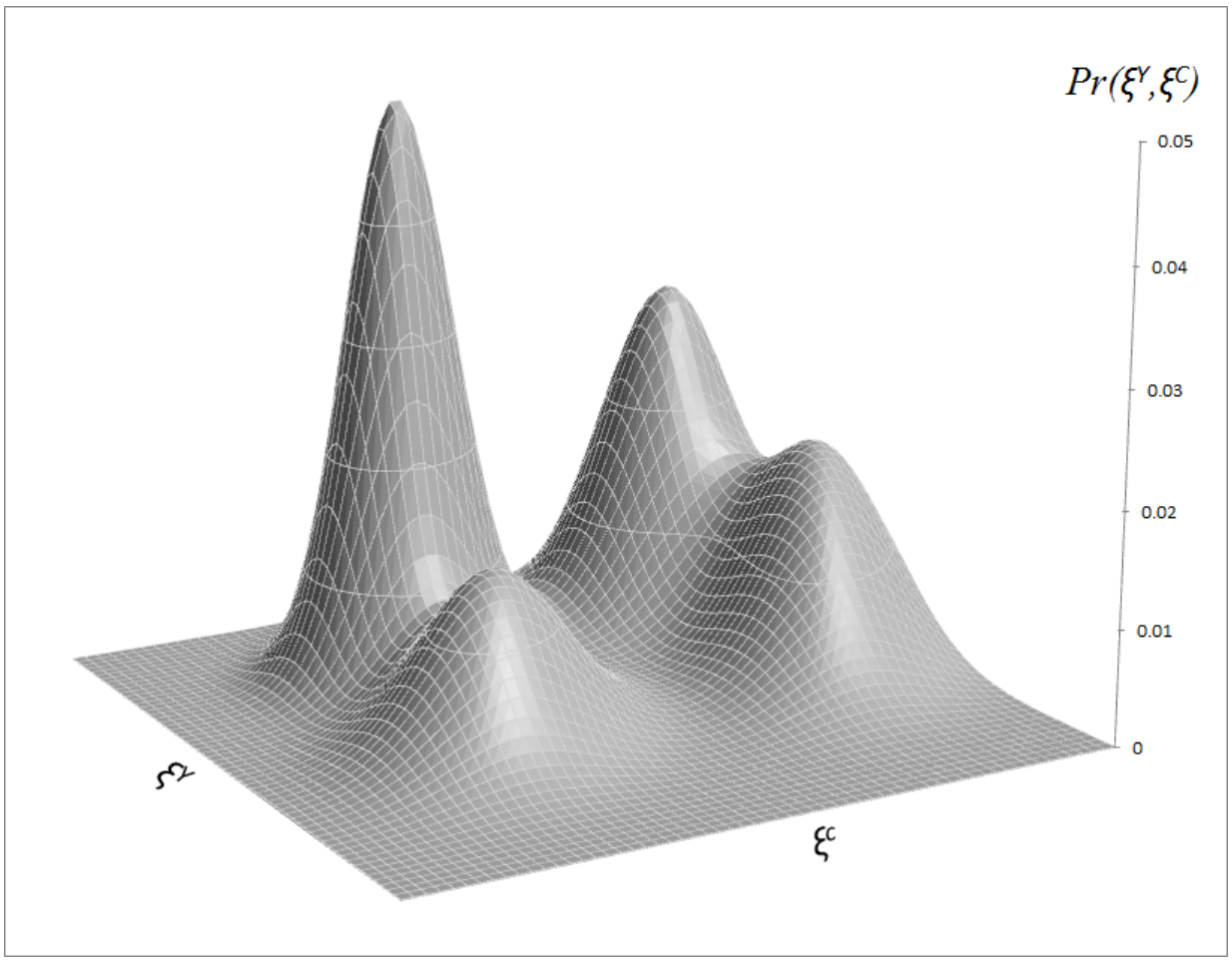


Figure 6.9: A Multimodal Density: Different Unobserved Types

cases where $\{\xi^Y, \xi^C\}$ are jointly normally distributed and cases where they are not. For each combination of characteristics they drew either 500 or 1,000 samples (called "Monte Carlo replications") satisfying those characteristics and estimated program impact with each of the impact evaluation estimators considered. The mean absolute deviation of the estimate of program impact from true impact is then calculated for each impact evaluation estimator across the replications for that experiment.

This design thus allows the authors to consider, for instance, how the relative performance of the various impact evaluation estimators differs between the cases where the error terms are jointly normal and where they are not jointly normal. Guilkey and Lance (2014) find that models that rely on an assumption of joint normality perform relatively well when errors are indeed jointly normal, but their relative performance suffers when they are not. Put differently, when you adopt an impact evaluation estimator that assumes joint normality, you really are betting that the error terms are indeed jointly normal.

This finding is not that shocking and in fact has many antecedents (e.g. Mroz 1999 or even the debates about the appropriate manner to model health care in the course of the RAND Health Insurance Experiment). The impetus to explore semi- and non-parametric estimators in other contexts was partly driven by a sense that more parametric models might not be very robust to violations of their assumptions. This is a sentence worth dissecting. By parametric models we typically mean models that involve somehow a specific distributional assumption, such as normality. Semi-parametric models are models that involve few or weak distributional assumptions. Non-parametric models eschew specific distributional assumptions altogether.

When we say that parametric models might not be very robust to the violations of their assumptions, we mean that model performance (in terms, for instance, of mean absolute deviation or some other measure of the accuracy of estimates in terms of true population values) might be poor when the sample did not arise from the assumptions inherent in the distribution assumed. For instance, the normal distribution involves a skewness of 0 and kurtosis of 3. If the true population data generating process behind the variable for which we assume a normal distribution involves a distribution with a skewness that is actually more than zero, for instance, then the sample from that population has values of that variable that in their distribution violate a key assumption behind the normal distribution. The basic fear then is that the performance of estimators that involve specific distributional assumptions such as normality might not be very good when those assumptions are violated by the data.

Interestingly, Guilkey and Lance (2014) find that a semi-nonparametric estimator of program impact performs relatively well in their experiments regardless of the actual specific distribution of $\{\xi^Y, \xi^C\}$. This flexible estimator was first proposed by Heckman and Singer (1984) in the context of hazard models but adopted to joint estimation of equations (such as the outcome and participation equations in the present discussion) by work such as Guilkey and Mroz (1992).[26]

To get some idea of how this works, we begin with the system

$$Y_i^* = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \mu_i^Y + \epsilon_i^Y$$

and

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \mu_i^C + \epsilon_i^C$$

where the μs and ϵs are not observed. ϵ_i^Y and ϵ_i^C are independently distributed, but μ^Y and μ^C might be correlated. For concreteness, we retain the assumption that Y_i and P_i are binary in nature.

[26] Mroz (1999) examines the performance of this estimator in a slightly different context through a series of similar Monte Carlo experiments.

The next step involves making some distributional assumption for the errors ϵ_i^Y and ϵ_i^C. Since they are independently distributed, it actually isn't essential to make the same distributional assumption (though it might look odd if we decided to model one binary model as probit and the other as logit). In this chapter we have thus far focused on the assumption of normality, so for variety we will instead assume that the ϵs are each the difference of two Type-I Extreme value distributed random variables. As we have learned, the cumulative density for such variables is given by the logistic function.

Specifically, individual i's probability for their observed outcome Y_i and program participation P_i is

$$\begin{aligned}
&Pr\left(Y_i, P_i|\beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3, \mu_i^Y, \mu_i^C\right) \\
&= \left(\frac{exp\left(\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \mu_i^Y\right)}{1 + exp\left(\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \mu_i^Y\right)}\right)^{Y_i} \\
&\cdot \left(\frac{1}{1 + exp\left(\beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \mu_i^Y\right)}\right)^{(1-Y_i)} \\
&\cdot \left(\frac{exp\left(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \mu_i^C\right)}{1 + exp\left(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \mu_i^C\right)}\right)^{P_i} \\
&\cdot \left(\frac{1}{1 + exp\left(\delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \mu_i^C\right)}\right)^{(1-P_i)}
\end{aligned}$$

Notice that this conditions on the unobservables μ_i^Y and μ_i^C in the same sense that it does the unknown population parameters $\{\beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3\}$. The reason is that we cannot observe the μs and control for them like we can observables like x_i, z_{1i} and z_{2i}. However, we cannot exactly estimate the μs like we can the βs or δs since the μs are individual-level unobservables, but we have not assumed either that we have longitudinal data or that the μs are necessarily fixed with respect to time (both of which are required for the dummy variable version of the fixed-effects estimator, for instance).

The way that this semi-parametric estimator deals with the μs is to estimate a discrete approximation to their true joint distribution and then integrates out with respect to them. This is another sentence that merits dissection. First, the semi-parametric estimator involves a discrete approximation to their joint distribution, meaning that while the μs might be able to take on an infinite combination of values and hence have a smooth, continuous joint distribution, this estimator discretely approximates it with a few mass points and probabilities of those mass points occurring (in other words, it assumes that they can take on only a finite number of combinations in values, in the hope that that will provide a decent approximation of the truth). The advantage of this approach is that no particular assumption is made about the joint distribution of the μs. Rather, the data are allowed to dictate an approximation as part of the estimation process. We then "integrate out" over the μs by allowing transforming every individual's probability for their observed $\{Y_i, P_i\}$ into a weighted sum based on the various discrete values that the μs can take on, with the weights being the probabilities associated with those various values of μ.

Specifically, the individual's contribution is now

$$Pr\left(Y_i, P_i|\beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3, \mu_2^Y, ..., \mu_K^Y, \mu_2^C, ..., \mu_K^C, \theta_2, ..., \theta_K\right)$$

$$= \sum_{k=1}^{K} \pi_k \cdot Pr\left(Y_i, P_i | \beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3, \mu^Y = \mu_k^Y, \mu^C = \mu_k^C\right)$$

where

$$\pi_k = Pr\left(\mu^Y = \mu_k^Y, \mu^C = \mu_k^C\right) = \frac{exp\left(\theta_k\right)}{1 + \sum_{j=2}^{K} exp\left(\theta_j\right)}$$

for $k = 2, ..., K$ and

$$\pi_1 = Pr\left(\mu^Y = 0, \mu^C = 0\right) = \frac{1}{1 + \sum_{j=2}^{K} exp\left(\theta_j\right)}$$

The normalization that μs at one of the mass points equal 0 (i.e. that $\mu_1^Y = \mu_1^C = 0$) is standard and necessary to identify separately the constant terms β_0 and δ_0. If we have a sample of N individuals along the lines we have considered this far, the likelihood function is then a product of the individual probabilities for the observed outcomes Y_i and P_i:

$$L\left(\beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3, \mu_2^Y, ..., \mu_K^Y, \mu_2^C, ..., \mu_K^C, \theta_2, ..., \theta_K\right)$$

$$= \prod_{i=1}^{N} Pr\left(Y_i, P_i | \beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3, \mu_2^Y, ..., \mu_K^Y, \mu_2^C, ..., \mu_K^C, \theta_2, ..., \theta_K\right)$$

which is maximized with respect to $\{\beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2, \delta_3, \mu_2^Y, ..., \mu_K^Y, \mu_2^C, ..., \mu_K^C, \theta_2, ..., \theta_K\}$.

The number of mass points K is also determined by the data. A variety criteria have been proposed for determining the number of mass points K (see, for instance, Mroz 1999) but most frankly amount to the following: keep adding mass points and re-maximizing the likelihood until adding further mass points fails to improve the maximized likelihood value significantly.[27] Surprisingly, this model often performs quite well even with few mass points. Why this is so is not clear.

The main advantage of this semi-parametric estimator, and the likely key to its success, is its flexibility. By imposing no particular assumption about the joint distribution of the μs, it can accommodate all kinds of possibilities, including the kinds of extreme behavior discussed earlier (it just implies large positive or negative mass point values). Currently, this estimator is not supported in STATA (though one of the authors of this manual is working on a command).

Nonetheless, we present estimates from this semi-parametric estimator applied to the present numerical example. Specifically, we jointly estimate the outcome and program participation equations with three points of support. The basic parameter estimates are presented in Table 6.1.[28] First, from the table title, this semi-parametric estimator is often referred to as the "Discrete Factor Model". Table 6.1 presents results from both the outcome and program participation equations. One set of mass points is normalized to zero (in order to separately identify the mass points from the constant). The estimates of the second and third mass points are positive for the outcome equation (at 1.45849 and 3.37042, respectively) while both for the program equation are negative.

[27] The differences between approaches generally involve the method of determining whether the likelihood has improved "significantly" or not, since some minuscule improvement is nearly guaranteed with the addition of more mass points.

[28] As mentioned, this model is not supported in STATA. The results presented are from a Fortran program that estimates this model. Since this is not STATA output, the results are placed in tables.

Table 6.1. The Discrete Factor Model				
	Outcome Equation		Participation Equation	
	Estimate	Std. Error	Estimate	Std. Error
Constant	-1.33676	1.6496	2.98458	0.8503
x	1.11966	0.2198	-1.01545	0.3523
z_1			0.30962	0.1130
z_2			-0.36407	0.1326
P	1.18058	0.4660		
Mass Point 1	0.00000	(Fixed)	0.00000	(Fixed)
Mass Point 2	1.45849	0.9816	-2.97451	1.3692
Mass Point 3	3.37042	2.9642	-5.12749	3.1763

In Table 6.2 we present the estimated probabilities associated with the mass points. Finally, the estimated program impact is .1543151, which is not far from the truth. Note that the errors are in fact jointly normally distributed. Models (such as the bi-probit) that assume joint normality (i.e. are based on the correct distributional assumption) should exhibit somewhat superior performance. Yet the discrete factor model estimate of program impact is remarkably close to true impact. The results from Guilkey and Lance (2014) suggest that the model would likely far outperform models based on an assumption of joint normality when errors are not, in fact, jointly normal.

Table 6.2. Mass Point Probabilities			
	Estimate	Std. Error	Probability
Mass Point 1	0.00000	(Fixed)	0.31095844
Mass Point 2	0.61258	0.8048	0.57377612
Mass Point 3	-0.99242	1.3839	0.11526545

6.1.3 Testing

We now briefly discuss testing in the instrumental variables setting.[29] Essentially, these are tests of whether the instrumental variables model is correctly specified (e.g. whether the instruments are appropriately excluded from the equation of interest, whether they really explain the endogenous variable, etc.). As such they are typically referred to as **specification tests**.[30] There are basically three types of tests that we consider:

1. Tests of the endogeneity of the allegedly endogenous variables;
2. Tests of the "strength" of the instrument(s) as determinants of variation in the endogenous variable;
3. Tests of whether the instruments indeed satisfy the properties of instruments.

We briefly discuss each type of test. Our focus is on the linear instrumental setting (because it is in the linear setting that the most work has been done on testing) but also discuss non-linear tests where applicable.

[29] A much-consulted resource for this general topic during one of the author's graduate school days was Dow (2001), the memory of which likely heavily informs elements of the discussions to follow. Apologies in advance to Dr. Dow if there are any phrases herein for which the author's memory of Dow (2001) from graduate school days was precise: any failure of citation is unintentional. The reader is advised to seek out and read Dow (2001) for a more thorough treatment of this subject.

[30] Specification tests have been developed for other sorts of models as well.

We begin with testing for the most basic assumption motivating the instrumental variables model (as well as almost every other quasi-experimental estimator of program impact): that program participation is endogenous. Once again, we use the system from the first subsection to organize our thinking:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y \end{aligned}$$

and

$$\begin{aligned} P_i^* &= \beta_1 - \gamma_0 - \gamma_1 \cdot x_i - \gamma_2 \cdot z_{1i} - \gamma_3 \cdot z_{2i} - \gamma_4 \cdot \mu_i - \epsilon_i^C \\ &= \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C \end{aligned}$$

Tests of endogeneity tend to assume that the instruments (z_{1i} and z_{2i}) are validly excluded and significant predictors of the potentially endogenous variable. In other words, they tend to assume that instrumental variables methods such as two-stage least squares will consistently estimate β_1.

An obvious possibility from the last Chapter is a Hausman-type test. As explained in Chapter 5, Hausman tests compare a consistent (i.e. consistent whether program participation P_i is endogenous or not) but less efficient estimator with a more efficient estimator that would not be consistent if P_i was indeed endogenous. For instance, two-stage least squares is a consistent estimator of β_1 regardless of whether P_i is endogenous or not (the consistency of two-stage least squares does not depend on P_i and ξ_i^Y actually being correlated). Of course, it is only consistent if the exclusion restrictions are valid and the excluded instruments are significant predictors of z_{1i} and z_{2i}, hence the statement in the preceding paragraph that tests of the endogeneity of P_i tend to assume that the assumed instrumental variables specification is correct, at least so far as the excluded instruments are concerned. Two-stage least squares is less efficient than simple regression of Y_i on P_i and x_i, because two-stage least squares involves additional sampling variation from the "first-stage" regression. If P_i is exogenous (i.e. independent or mean independent of ξ_i^Y) both two-stage least squares and the simple regression of Y_i on P_i and x_i should yield consistent, and hence presumably similar, estimates of β_1. However, if P_i and ξ_i^Y *are* correlated, then there should be large differences in the estimates of β_1 provided by two-stage least squares (which would be consistent under this circumstance) and regression of Y_i on P_i and x_i (which would not be consistent).

The intuition behind the test is perhaps best understood by referring back to the simplistic Hausman statistic that we discussed in Chapter 5:

$$\frac{\left(\hat{\beta}_1^C - \hat{\beta}_1^E\right)^2}{\hat{var}\left(\hat{\beta}_1^C\right) - \hat{var}\left(\hat{\beta}_1^E\right)}$$

where $\hat{\beta}_1^C$ and $\hat{\beta}_1^E$ are the estimates of β_1 emerging, respectively, from the consistent and efficient models. $\hat{var}(\hat{\beta}_1^C)$ and $\hat{var}(\hat{\beta}_1^E)$ are the respective variance estimates from the two models. If P_i is not endogenous (i.e. not correlated with ξ_i^Y) then $\hat{\beta}_1^C$ and $\hat{\beta}_1^E$ should be similar because both are consistent and the test statistic, which is based on their difference, should be small, leading us to accept a null hypothesis (H^0) of P_i being exogenous. If the estimates of β_1 are very different, this is evidence that simple regression is potentially inconsistent due to endogeneity. The test statistic would have a large value and lead us to reject the null hypothesis in favor of the alternative hypothesis (H^a). The differences in the estimates of β_1 are normalized by the differences in the variances of the two estimates to correct for differences in the estimates that one might expect simply due to the superior efficiency of the efficient estimator.

Remember, the validity of this test rests entirely on the assumption that two-stage least squares is consistent regardless of whether P_i is endogenous or not. That assumption in turn rests on the

instruments being correctly specified. In our context, this means that z_{1i} and z_{2i} are legitimately excluded from the outcome equation and uncorrelated with ξ_i^Y.

An example of this based on the "linear" circumstances of STATA do-file 6.1.do is offered in Outputs 6.32-6.34. Once again, conducting a Hausman test in STATA involves in essence constructing the components of the test and then conducting it with the `hausman` command. Output 6.32 presents results from estimation of the consistent model and the storing of the results, and Output 6.33 does the same for the efficient model. Finally, Output 6.34 presents results from the Hausman test itself. The test statistic value is 35.51, which leads us to reject the null hypothesis that two-stage least squares and the simple regression model provide the same estimates. Notice that, unlike the simplistic test statistic where we considered only β_1 (the coefficient on P_i) in Output 6.34 the coefficients on both P_i and x_i are tested.

STATA Output 6.32 (6.1.do)

```
. * A "consistent" model
. ivregress 2sls Y x (P=x z1 z2)
Instrumental variables (2SLS) regression              Number of obs =   10000
                                                      Wald chi2(2)  = 9487.14
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.4727
                                                      Root MSE      =  3.6031
------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    2.108621   .3202348     6.58   0.000     1.480972    2.736269
           x |    2.001747    .042218    47.41   0.000     1.919001    2.084493
       _cons |    1.892261   .1892769    10.00   0.000     1.521285    2.263237
------------------------------------------------------------------------------
Instrumented:  P
Instruments:   x z1 z2
.
. estimates store c1
```

STATA Output 6.33 (6.1.do)

```
* An "efficient" model
. reg Y P x
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) = 4968.55
       Model |  122730.14      2  61365.0699            Prob > F      =  0.0000
    Residual |  123469.862  9997  12.3506914            R-squared     =  0.4985
-------------+------------------------------           Adj R-squared =  0.4984
       Total |  246200.001  9999  24.6224624            Root MSE      =  3.5144
------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    .263087   .0813785     3.23   0.001     .1035687    .4226052
           x |    1.781695   .0200693    88.78   0.000     1.742355    1.821035
       _cons |    2.963129   .0588625    50.34   0.000     2.847747    3.078512
------------------------------------------------------------------------------
.
. estimates store e1
```

There are a few variations on this approach. For instance, one could estimate the first-stage program participation equation by ordinary least squares, obtain the predicted participation $\hat{P}_i$ and then regress Y_i on P_i, $\hat{P}_i$ and x_i. A significance test of the estimated coefficient on $\hat{P}_i$ is a kind of endogeneity test: it will be significant only if P_i is endogenous, in which case the estimated coefficient on $\hat{P}_i$ represents the difference between the coefficient on P_i and true program impact β_1. Another version of this test uses the first stage predicted residuals instead of $\hat{P}_i$, but otherwise follows the same logic. Finally, one could simply regress Y_i on P_i, z_{1i}, z_{2i} and x_i. Assuming that the instruments z_{1i} and z_{2i} are validly excluded, a test on the significance of these instruments is an endogeneity test. Notice that every one of these tests still requires the assumption that the instruments z_{1i} and z_{2i} are indeed legitimately excludable from the output equation and uncorrelated with ξ_i^Y. A big caveat about these tests is that they often have fairly low power to reject the null hypothesis of exogeneity. What that means is that failing this test and rejecting the null hypothesis of exogeneity probably really means something, while accepting the null hypothesis of exogeneity may not be fully reassuring that P_i really is exogenous.

STATA Output 6.34 (6.1.do)

```
. * Hausman test
. hausman c1 e1, equations(1:1)
                 ---- Coefficients ----
              |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
              |       c1           e1         Difference          S.E.
--------------+----------------------------------------------------------------
            P |    2.108621     .263087        1.845534        .3097222
            x |    2.001747    1.781695        .2200516        .0371427
------------------------------------------------------------------------------
                    b = consistent under Ho and Ha; obtained from ivregress
         B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       35.51
                Prob>chi2 =      0.0000
```

In the limited dependent variable setting the assumption of joint normality frequently provides a natural test of correlation because these methods typically generate an estimate of the correlation ρ. These were generally discussed with the estimation results for these models presented in the last subsection. For instance, in the bivariate probit model ρ is directly estimated as part of the full information maximum likelihood estimation process. These tests of course assume that joint normality is an appropriate assumption regarding the joint distribution of $\{\xi_i^Y, \xi_i^C\}$.

We now briefly discuss tests of instrument strength. As we mentioned earlier at the conclusion of the discussion of linear instrumental variables, a series of prominent papers consider the performance of linear instrumental when endogenous variables are only weakly correlated with the instruments (i.e., in the context of our example to this point, when the variables z_{1i} and z_{2i} are only weakly correlated with P_i). The general conclusion is that instrumental variables performs quite poorly, providing potentially wildly biased and inconsistent estimates of program impact, with such weak instruments.

Unfortunately, at this stage in the evolution of the literature on this subject, we can offer more warning than guidance. Clear, widely agreed upon guidelines (for instance an acceptable threshold level of significance for the instruments) have not really been established. For a linear first stage, some have suggested an F-statistic threshold of 10 for a test of the joint significance of

the instruments. Casual, unpublished experiments by one of the authors of this manual for a first stage nonlinear binary model (e.g. logit regression of program participation P_i on the observed exogenous characteristics x and the instruments z_{1i} and z_{2i}) and a binary outcome variable have suggested a threshold of generally reliable performance with a joint significance χ^2 statistic of perhaps 35-50. However, there is no real consensus on this and the figures suggested may be quite misleading outside of the context of the particular simulations from which they arose.

STATA Output 6.35 (6.1.do)

```
. * First stage of ``manual´´ two-stage least squares,
. * over-identified case
. * instrument significance test
. reg P x z1 z2
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  3,  9996) = 1333.14
       Model |  695.526508     3  231.842169            Prob > F      =  0.0000
    Residual |  1738.37659  9996  .173907222            R-squared     =  0.2858
-------------+------------------------------           Adj R-squared =  0.2856
       Total |   2433.9031  9999  .243414651            Root MSE      =  .41702

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.119577   .0020847   -57.36   0.000    -.1236635   -.1154906
          z1 |   .0365565   .0020887    17.50   0.000     .0324622    .0406508
          z2 |  -.0428673   .0020938   -20.47   0.000    -.0469716   -.0387629
       _cons |   .5791768   .0041704   138.88   0.000     .5710019    .5873517
------------------------------------------------------------------------------

.
. test z1 z2
 ( 1)  z1 = 0
 ( 2)  z2 = 0

       F(  2,  9996) =  363.97
            Prob > F =    0.0000
```

We present examples of such tests in Outputs 6.35 and 6.36, which are based on the example in STATA do-file 6.1.do. Output 6.35 provides the F-test for the first stage linear probability model while Output 6.36 provides the χ^2 statistic appropriate when the first stage is estimated by probit or logit. In either case, the process is as simple as running the model and then typing

```
test z1 z2
```

As the reader can see, the test statistic values (of 363.97 in Output 6.35 and 631.92 in Output 6.36) are quite large. While there are no hard and fast guidelines for first stage instrument strength sufficient to allay concerns about weak instruments, the authors would be unconcerned in the face of these test statistic values.

Finally, we turn to tests of whether the instruments satisfy some the most important requirements for an instrument: a valid instrument must be statistically unrelated to the error term ξ_i^Y. An immediate implication of this is that it must be legitimately excludable from the outcome equation (if not it would by exclusion be relegated to the error term ξ_i^Y and hence correlated with it). One important limitation that holds for tests of instrument validity:

Such tests can only ever be performed with over-identified models.

In other words, the models must have more identification than is technically required for the model to be estimable. Following discussions in the first subsection, in the linear setting this means

that instrument validity can be tested only when the number of instruments/exclusion restrictions exceeds the number of endogenous variables.

STATA Output 6.36 (6.1.do)

```
. * First stage of ``manual´´ two-stage least squares,
. * two instrument (over-identified) case
. * instrument significance test
. logit P x z1 z2

Iteration 0:   log likelihood = -6798.6892
Iteration 1:   log likelihood = -5124.7194
Iteration 2:   log likelihood = -5107.5952
Iteration 3:   log likelihood = -5107.5649
Iteration 4:   log likelihood = -5107.5649

Logistic regression                               Number of obs   =      10000
                                                  LR chi2(3)      =    3382.25
                                                  Prob > chi2     =     0.0000
Log likelihood = -5107.5649                       Pseudo R2       =     0.2487

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.7020947   .0161289   -43.53   0.000    -.7337067   -.6704827
          z1 |   .2137315   .0125744    17.00   0.000     .1890861     .238377
          z2 |  -.2508408   .0126878   -19.77   0.000    -.2757084   -.2259732
       _cons |   .4584433   .0247126    18.55   0.000     .4100074    .5068792
------------------------------------------------------------------------------

.
. test z1 z2

 ( 1)  [P]z1 = 0
 ( 2)  [P]z2 = 0

           chi2(  2) =  631.92
         Prob > chi2 =    0.0000
```

However, in the limited dependent variable setting non-linearity could be the basis for over-identification. In other words, a bivariate probit model with one endogenous variable and one instrument is still overidentified because the model is nonlinear, and that non-linearity provides identification in and of itself (whether such identification is confidence inspiring is another issue altogether). Because these tests can only ever be performed in the context of overidentified models, they are sometimes referred to as **overidentification tests**.

There are basically two types of overidentification test. One type assumes that one of the sources of identification is correct, and then assesses the validity of the other source of identification by examining whether the results (in our context, program impact estimates) generated with the latter differ from those generated with only the former. For instance, if you have two instruments, you test the validity of one of them on the assumption that the other is validly excluded from the equation of interest and uncorrelated with the error term from that equation. The advantage of that test is that it is immediately clear which source of identification is invalid. The disadvantage is that this conclusion depends entirely on the assumption that one source of identification is valid. The other type remains agnostic about which of the instruments is valid. Of course, if the test statistic indicates instrument invalidity, it is not clear whether that finding is being driven by all of the instruments (e.g. they are all invalid) or just some subset of them.

We begin with the former type of identification test (i.e. the type that assumes that one of the instruments/exclusion restrictions is correct). One natural possibility would be a type of Hausman test. If we assume that one of the instruments (z_{1i}, both to fix ideas and because we must assume that one instrument is valid for this test) is valid we could then compare a consistent estimator

(using just z_{1i} as an instrument) with an efficient estimator (using both z_{1i} and z_{2i} as instruments). The latter is more efficient because of the greater capability to explain variation in the endogenous variable P_i but will be inconsistent if z_{2i} is correlated with ξ_i^Y. For instance, under two-stage least squares, first stage predicted program participation $\hat{P}_i$ will be correlated with ξ_i^Y if z_{2i} and ξ_i^Y are correlated, rendering the two-stage least squares estimator with both instruments inconsistent.

STATA Output 6.37 (6.1.do)

```
. * A "consistent" model
. ivregress 2sls Y x (P=x z1 ), first
First-stage regressions
```

Number of obs	=	10000			
F(2, 9997)	=	1718.26			
Prob > F	=	0.0000			
R-squared	=	0.2558			
Adj R-squared	=	0.2557			
Root MSE	=	0.4257			

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.1192699	.0021278	-56.05	0.000	-.1234408	-.115099
z1	.0367036	.002132	17.22	0.000	.0325245	.0408826
_cons	.5798419	.0042566	136.22	0.000	.571498	.5881857

Instrumental variables (2SLS) regression

Number of obs	=	10000	
Wald chi2(2)	=	9333.40	
Prob > chi2	=	0.0000	
R-squared	=	0.4652	
Root MSE	=	3.6287	

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.361004	.4951864	4.77	0.000	1.390457	3.331552
x	2.03184	.061767	32.90	0.000	1.910779	2.152901
_cons	1.745816	.2896135	6.03	0.000	1.178184	2.313448

```
Instrumented:  P
Instruments:   x z1
.
. estimates store c2
```

We illustrate this tests in Outputs 6.37-6.39, which use the framework of the simulation example in STATA do-file 6.1.do to perform this test. Output 6.37 provides estimates from the consistent model that relies on z_{1i} alone as an instrument. Output 6.38 does the same for the (comparatively) efficient model that relies on both z_{1i} and z_{2i} as instruments.

In Output 6.39 the results of the Hausman test are reported. Now we are presented with a relatively small test statistic value of 0.45, with an accompanying p-value of 0.7999. We would thus accept the null hypothesis that z_{2i} is a legitimate instrument (which is not a surprising conclusion since it is a legitimate instrument by construction per the setup in STATA do-file 6.1.do).

We now move on to tests that do not rely on the assumption that some particular subset of the instruments are valid. One classic test involves estimating the model

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \xi_i^Y$$

and

$$P_i^* = \delta_0 + \delta_1 \cdot x_i + \delta_2 \cdot z_{1i} + \delta_3 \cdot z_{2i} + \xi_i^C$$

by two-stage least squares. The predicted residuals $\hat{\xi}_i^Y$ are then calculated from the fitted model and regressed on x_i, z_{1i} and z_{2i}. $N \cdot R^2$, where N is the sample size and R^2 is the coefficient of determination from the regression of $\hat{\xi}_i^Y$ on x_i, z_{1i} and z_{2i}, follows a χ^2 distribution with degrees of freedom equal to the degree of over-identification (i.e. the number of instruments employed beyond what was required to achieve exact identification, which is 1 in this case) if all of the instruments are valid. A large test statistic implies that at least one instrument is invalidly excluded. This is sometimes referred to as a Lagrange Multiplier (or LM) test.

STATA Output 6.38 (6.1.do)

```
. * An "efficient" model
.
. ivregress 2sls Y x (P=x z1 z2), first
First-stage regressions
```

```
                                                Number of obs   =      10000
                                                F(   3,   9996) =    1333.14
                                                Prob > F        =     0.0000
                                                R-squared       =     0.2858
                                                Adj R-squared   =     0.2856
                                                Root MSE        =     0.4170
```

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.119577	.0020847	-57.36	0.000	-.1236635	-.1154906
z1	.0365565	.0020887	17.50	0.000	.0324622	.0406508
z2	-.0428673	.0020938	-20.47	0.000	-.0469716	-.0387629
_cons	.5791768	.0041704	138.88	0.000	.5710019	.5873517

```
Instrumental variables (2SLS) regression              Number of obs =   10000
                                                      Wald chi2(2)  = 9487.14
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.4727
                                                      Root MSE      =  3.6031
```

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.108621	.3202348	6.58	0.000	1.480972	2.736269
x	2.001747	.042218	47.41	0.000	1.919001	2.084493
_cons	1.892261	.1892769	10.00	0.000	1.521285	2.263237

```
Instrumented:  P
Instruments:   x z1 z2
.
. estimates store e2
```

The generalized method of moments tradition has yielded a popular test of over-identification based on something known as Hansen's J statistic. First, let us recall our simplified generalized method of moments framework. We have the following moment conditions for our overidentified model:

$$g_1 = \sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i \right)$$

$$g_2 = \sum_{i=1}^{N} z_{1i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i \right)$$

$$g_3 = \sum_{i=1}^{N} x_i \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right)$$

$$g_4 = \sum_{i=1}^{N} z_{2i} \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot P_i - \hat{\beta}_2 \cdot x_i\right)$$

In this over-identified model there are no estimate values $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$ such that $g_1 = g_2 = g_3 = g_4 = 0$. Instead, the goal of generalized method of moments estimation is to find estimates $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$ such that g_1, g_2, g_3 and g_4 collectively are as close to zero as possible. We gave a very simplified example of the actual objective function minimized in generalized method of moments estimation:

$$Q = \begin{bmatrix} g_1 & g_2 & g_3 & g_4 \end{bmatrix} \begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}$$

$$= w_1 \cdot (g_1)^2 + w_2 \cdot (g_2)^2 + w_3 \cdot (g_3)^2 + w_4 \cdot (g_4)^2$$

In practice, generalized method of moments estimation employs more complex "optimal" weighting matrices, but this gives the basic idea of the generalized method of moments objective function.

STATA Output 6.39 (6.1.do)

```
. * Hausman test
. hausman c2 e2, equations(1:1)
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       c2           e2       Difference          S.E.
-------------+----------------------------------------------------------------
           P |    2.361004     2.108621        .2523834        .3777026
           x |     2.03184     2.001747        .0300928        .0450867
------------------------------------------------------------------------------
                    b = consistent under Ho and Ha; obtained from ivregress
     B = inconsistent under Ha, efficient under Ho; obtained from ivregress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =         0.45
                Prob>chi2 =       0.7999
```

The Hansen J-statistic is basically just $N \cdot Q$, which follows a χ^2 distribution with degrees of freedom equal to the number of moment conditions (4 in our example) minus the number of parameters to be estimated (3 in our example). The null hypothesis (H_0) is that the expected value of the moment conditions is equal to zero, which we would reject in the face of a large J-statistic. One intuitive explanation of Hansen's J test is that if the instruments are valid, it should be possible to get the moment conditions close to zero. Failure to do so is evidence that at least one of them is not valid. However, one must be cautious with Hansen-J tests because other forms of model mis-specification could potentially lead to a large test statistic. In Output 6.40 we provide an example of a Hansen's J test. The test statistic value is, at .461503, not particularly large, with a correspondingly large p-value of 0.4969, which would lead us to accept the null hypothesis that the instruments are valid.

STATA Output 6.40 (6.1.do)

```
. * Gmm estimation
.
. ivregress gmm Y x (P=x z1 z2), first
First-stage regressions
```

Number of obs	=	10000			
F(3, 9996)	=	2308.62			
Prob > F	=	0.0000			
R-squared	=	0.2858			
Adj R-squared	=	0.2856			
Root MSE	=	0.4170			

P	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.119577	.0016949	-70.55	0.000	-.1228994	-.1162547
z1	.0365565	.0020593	17.75	0.000	.0325199	.0405931
z2	-.0428673	.0020369	-21.05	0.000	-.04686	-.0388746
_cons	.5791768	.0041754	138.71	0.000	.5709922	.5873614

Instrumental variables (GMM) regression

Number of obs	=	10000
Wald chi2(2)	=	9351.62
Prob > chi2	=	0.0000
R-squared	=	0.4727
Root MSE	=	3.6031

GMM weight matrix: Robust

Y	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.108944	.324678	6.50	0.000	1.472587	2.745301
x	2.001951	.0430581	46.49	0.000	1.917559	2.086344
_cons	1.893205	.1920943	9.86	0.000	1.516708	2.269703

```
Instrumented:  P
Instruments:   x z1 z2
.
. * The Hansen test
.
. estat overid
  Test of overidentifying restriction:
  Hansen´s J chi2(1) = .461503 (p = 0.4969)
```

STATA Output 6.41 (6.2.do)

```
. biprobit (Y= x P z1 z2) (P=x z1 z2)
Fitting comparison equation 1:
Iteration 0:   log likelihood = -5741.0029
Iteration 1:   log likelihood = -4153.3347
Iteration 2:   log likelihood = -4113.6203
Iteration 3:   log likelihood = -4113.4501
Iteration 4:   log likelihood = -4113.4501
Fitting comparison equation 2:
Iteration 0:   log likelihood = -6798.6892
Iteration 1:   log likelihood = -5106.9076
Iteration 2:   log likelihood =  -5102.731
Iteration 3:   log likelihood = -5102.7308
Comparison:    log likelihood = -9216.1809
Fitting full model:
Iteration 0:   log likelihood = -9216.1809
Iteration 1:   log likelihood = -9214.3282
Iteration 2:   log likelihood = -9213.1495
Iteration 3:   log likelihood = -9213.1285
Iteration 4:   log likelihood = -9213.1283
Seemingly unrelated bivariate probit              Number of obs   =      10000
                                                  Wald chi2(7)    =    4471.91
Log likelihood = -9213.1283                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Y            |
           x |   .5583344   .0194147    28.76   0.000     .5202822    .5963866
           P |   .6259885   .2395513     2.61   0.009     .1564766     1.0955
          z1 |  -.0004073   .0122222    -0.03   0.973    -.0243623    .0235478
          z2 |   .0021148   .0135981     0.16   0.876     -.024537    .0287667
       _cons |   .5037425   .1703472     2.96   0.003     .1698681    .8376169
-------------+----------------------------------------------------------------
P            |
           x |  -.4155645   .0088435   -46.99   0.000    -.4328975   -.3982316
          z1 |   .1264503   .0073257    17.26   0.000     .1120923    .1408083
          z2 |  -.1493738   .0074034   -20.18   0.000    -.1638841   -.1348635
       _cons |   .2720356   .0144469    18.83   0.000     .2437202    .3003509
-------------+----------------------------------------------------------------
     /athrho |  -.3784434   .1631063    -2.32   0.020    -.6981259   -.0587609
-------------+----------------------------------------------------------------
         rho |  -.3613549   .1418083                     -.6031769   -.0586934
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0:     chi2(1) =  6.10514    Prob > chi2 = 0.0135
.
. test [Y]z1 [Y]z2
 ( 1)  [Y]z1 = 0
 ( 2)  [Y]z2 = 0
           chi2(  2) =    0.03
         Prob > chi2 =    0.9844
```

Finally, in instances where both the outcome Y and the program participation variable P are binary in nature, a sort of rough over-identification test is sometimes performed in the context of joint estimation of the participation and outcome equations via full information maximum likelihood estimation. Essentially, the test involves including all of the instruments as right-hand side explanatory variables in *both* equations. The idea behind this is that program impact should already be identified by non-linearity (although this can be a shaky foundation the appropriateness

of which often depends on making the right assumption about the joint distribution of $\{\xi_i^Y, \xi_i^C\}$[31]) and the instruments should therefore serve simply as controls for themselves in the output equation (as opposed to their role in the endogeneity test we just mentioned in the linear instrumental variables setting).

A test of their significance in the outcome equation is thus a test of their legitimate exclusion from the outcome equation. Output 6.41 reports results from estimation of the bi-probit model using the simulation in STATA do-file 6.2.do as a departure point. In this case, the instruments are included in both equations. Their z-statistics and p-values in the outcome equation (with a z-statistic of -0.03 and p-value of 0.973 in the case of z_{1i} and a z-statistic of 0.16 and p-value of 0.876 in the case of z_{1i}) certainly do not suggest that individually they play a significant role in the determination of the outcome Y_i. Following bi-probit estimation their joint significance in the outcome equation is tested. The resulting χ^2 statistic is 0.03 with an accompanying p-value of 0.9844 suggests that they are not collectively significant direct determinants of Y_i either.

6.1.4 Natural and Purposeful Social Experiments

We briefly discuss instrumental variables and experiments. The basic challenge we face in this manual is that program participation is not randomly determined. In this and the last chapter, we have considered quasi-experimental estimators that address the possibility that there are unobserved determinants of the outcome of interest with which program participation is correlated. If this is the case, then simple estimators of program impact, such as regression of the outcome on program participation and its other observed determinants (x_i in the running example of this chapter) is biased and inconsistent.

Essentially, an instrument is a variable that allows us to concentrate on only some portion of the variation in program participation not associated with the unobserved determinants of the outcome. By concentrating on this "channel" of variation in program participation we are able to estimate program impact without worrying about potential correlation of overall impact with such unobservables.

In this sense, an instrument can be viewed as a kind of experiment. It provides a channel of experimental variation in program participation in the sense that that channel is not associated with unobserved determinants of the outcome.

Indeed, instrumental variables estimation often appeals to the basic concepts and terminology of experiments. One important instance are cases where the natural course of human events gives rise to particular institutional circumstances that generate a random channel of variation in program participation. Such instances are frequently referred to as **natural experiments**.

John Snow's cholera experiment can be viewed as a natural experiment. The outcome of interest was cholera cases. The "program" was exposure to tainted water. Program participation was thus driven by myriad individual decisions (which pubs to visit, water sources to use outside of the home, etc.). These decisions were driven by the observed and unobserved characteristics of London residents, many of which may also have influenced whether they developed cholera. This made it impossible to argue that exposure to tainted water per se caused cholera: one could have argued that exposure to tainted water was simply a proxy for other factors that actually drove cholera infection.

[31] We have already seen an example where "getting the joint distribution right" was not be enough to insure a reasonable estimate of program impact: in that earlier example the errors were assumed to be and were in reality jointly normal, but identification by non-linearity did not work very well when there was insufficient variation in the observable x.

Snow focused on an extremely unusual institutional arrangement in the form of the happenstance assignment of households to the two water companies and the sudden divergence in their water sources in the late 1840s. A household's water company could thus be viewed as an instrument for their exposure to tainted water: out of all of the possible sources of variation in tainted water, the household's water company provided a random channel of variation in exposure to tainted water.

Instrumental variables is also a popular impact estimation option in purposeful social experiments, particularly when compliance with experimental assignment is not complete. The "Oregon Experiment" from Chapter 3 is an example of this. In that case, 70,000 qualified individuals applied for roughly 10,000 insurance policies under Oregon Health Plan Standard, a Medicaid program for adults who are low-income, uninsured, "able-bodied" and not eligible for other public insurance in Oregon. Of them, 30,000 were randomly selected to apply. However, many of the selected did not ultimately complete the enrollment process of their own volition.

This creates real problems for the purposes of studying the impact of health insurance on health care demand or health outcomes. Failure to apply by some was without doubt associated with many observed and unobserved determinants of health care demand and, ultimately, health. This means that eventual insurance status was also associated with these characteristics, returning the experiment's investigators to the original circumstances (the potential endogeneity of insurance status as a determinant of health and health care demand) that prompted the experiment in the first place.

One solution to this is to use the random selection to apply as an instrument for health insurance. The theory behind this is that the attempt at purposeful randomization in this experiment still provides a channel of random variation in health insurance status that should not have a direct effect on health or health care demand and should not be correlated with any unobservables that influence health or health care demand. In other words, it would seem to satisfy the classic requirements for an instrument.

6.2 Local Average Treatment Effects

We concluded the last section by discussing explicitly conceptual links between instrumental variables estimation and experiments, advancing the idea that instruments can be viewed as representing experiments. Further, it was suggested that random assignment can serve as an instrument, a particularly attractive fall-back position in purposeful randomized social experiments where there is not complete compliance with randomized experimental assignment to participation status.

Unfortunately, however, there is a potential wrinkle to this insurance policy, as well as to the more general interpretation of linear instrumental variables in the context of experiments: it turns out that linear instrumental variables estimates program impact only for those who comply with their assignment per the experiment that the instrument represents. This section is concerned with this complication, which has been a central topic of the recent literature regarding instrumental variables and led some to question its value as a program impact estimator (indeed, even more broadly as a causal estimator).

Our overarching goal is to estimate average program impact across some population, typically using the information provided by a random sample from that population. Until this point in this chapter we have assumed that program impact was constant across the population. The notion that instrument variables recovers an estimate of average program impact only for the subpopulation that complies with the program participation status implied by their assigned value for the instrument would not necessarily be problematic in this setting: since program impact was constant across the

population, an estimator that identified average program impact for a subgroup of the population might still effectively provide information about average impact at the population level since the average impact for the subgroup in question would be the same as that for the population. In other words, if program impact was constant then capturing impact only for those who comply with their experimental assignment would not necessarily be a problem since their impact would be the same as that of non-compliers.

Thus, the notion that instrumental variables recovers estimates of average program impact consistent for those who comply with the program participation status implied by their assigned value of the instrument is only a problem if program impacts are heterogeneous. Heterogeneous impact means that program impact varies from individual to individual, and average program impacts might differ between different types of individuals. Then, one has to consider the possibility that those who comply with their experimental participation assignment have different average program impacts from those who do not. In this case instrumental variables estimators of program impact might provide extremely misleading estimates of average program impact at the population level. The estimate of program impact that we do recover from instrumental variables in this case is commonly referred to as a **Local Average Treatment Effect** (often referred to simply with the acronym **LATE**, a convention that we adhere to for parsimony's sake for the remainder of the chapter). The term "local" distinguishes the average program impact involved from the "global" average impact that would apply to the entire population.

It is worth delving into what "compliance" means where an instrument is concerned. When we say that instruments represent experiments, this implies that the various values that the instruments take on represent experimental assignments. Thus, when we speak of compliance in the instrumental variables context, it might be useful to think in terms of a binary instrument. One value of the instrument (say 0) might indicate assignment to non-participant status, while the other value (1) might indicate assignment to participant status. A useful way to think about compliance is that compliers are those whose participation status always reflects their participation assignment per their instrument value.

It would perhaps be useful to begin approaching this issue a bit more formally. For the purpose of thinking about what exactly linear instrumental variables identifies when there is program impact heterogeneity and varying experimental assignment compliance across individuals, it is useful to think of each individual (either from the population of interest or in any representative sample from that population) as falling into one of three categories with regard to a potential instrument z:

1. **Always Takers:** These are people who will always participate in a program, regardless of the value of z assigned to them;

2. **Never Takers:** These are people who will never participate, regardless of the value of z assigned to them;

3. **Compliers:** These are the people who will participate if their assigned value of z indicates that they should do so, and not participate if their assigned value of z indicates that they should not do so.

One could conceive of a fourth group: those who would always select the opposite of the participation status indicated by their value of z. In other words, if their assigned value of z suggest that they should participate, they will not do so. However, if their assigned value of z suggests that they should participate, then they will elect not to participate. This possibility is typically ruled out by what is usually referred to as the **monotonicity assumption**.

Suppose now that:

- We have an instrument z that can take on only the values 1 (indicating that the individual should participate) and 0 (indicating that the individual should not participate);
- Only a subset of the population of interest (or only a subsample of any random sample from the population) changes their participation status from non-participation ($P = 0$) to participation ($P = 1$) as the value of the instrument z changes from 0 to 1;
- Program impact varies between individuals (i.e. there is program impact heterogeneity).

When we use z as the sole instrument, the resulting estimate of program impact consistently estimates program impact only for those who are compliers with respect to z. In other words, instrumental variables estimation based on the instrument z provides an estimate of average program impact that applies only to the compliers with respect to z. In effect, the instrument z allows us to identify a channel of experimental variation in program participation P associated with or meaningful for only one group: the compliers.

It is important to remember that complier status is instrument-specific. For instance, if there were two instruments, z_1 and z_2, the subpopulations of compliers for the two instruments might be completely different (or they might overlap heavily). Generally speaking, different instruments will yield different local average treatment effects (since each potential instrument might influence different types of individuals in terms of their program participation decision or, put a little bit differently, each implies a different group of compliers). As we will discuss below, there are a few things that can be learned about the complier population.

Let us outline a model. To our knowledge, there is no consensus approach for motivating this issue from a theoretical (i.e. behavioral) standpoint. We therefore adopt our familiar potential outcome approach to this issue. To begin with, suppose that we have the potential outcome equations

$$Y_i^0 = \beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$
$$Y_i^1 = \beta_0 + \beta_{1i} + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

We have an observed (x_i) and unobserved (μ_i) individual-level characteristic determining potential outcomes as well as a purely random, idiosyncratic unobserved component ϵ_i^Y. We assume that the three are independently distributed. Program impact for individual i is

$$\begin{aligned} & Y_i^1 - Y_i^0 \\ &= \beta_0 + \beta_{1i} + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y \\ &\quad - \left(\beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &= \beta_{1i} \end{aligned}$$

Notice that program impact is individual-specific. In other words, this model builds in potential for heterogeneous program impact.

A regression specification can be derived in much the same fashion as with behavioral models in preceding chapters. The observed outcome is

$$\begin{aligned} & Y_i = P_i \cdot Y_i^1 + (1 - P_i) \cdot Y_i^0 \\ &= P_i \cdot \left(\beta_0 + \beta_{1i} + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &\quad + (1 - P_i) \cdot \left(\beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \end{aligned}$$

$$= \beta_0 + \beta_{1i} \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

Notice that the unobservable μ_i appears in this equation, and would hence be an element of the regression residual.

The cost of program participation is given by

$$C_i = \gamma_0 + \gamma_1 \cdot x_i + \gamma_{2i} \cdot z_i + \gamma_4 \cdot \mu_i + \epsilon_i^C$$

We now focus on one instrument, z_i. For the purposes of this discussion, we will generally assume that this instrument is binary. We will also assume that $\gamma_{2i} < 0$. This means that an individual becomes more likely to participate as the value of z_i switches from 0 to 1 because as it does so their cost of participation falls by $\gamma_{2i} \cdot z_i$ (it falls because $\gamma_{2i} < 0$).

Individual i will choose to participate (i.e. their value for the program participation indicator P, P_i, equals 1) if

$$Y_i^1 - Y_i^0 - C_i > 0$$

or

$$\beta_{1i} - \gamma_0 - \gamma_1 \cdot x_i - \gamma_{2i} \cdot z_i - \gamma_4 \cdot \mu_i - \epsilon_i^C > 0$$

where we assume that ϵ_i^C is uncorrelated with both ϵ_i^Y and z_i. In this framework, LATE would typically involve some correlation between β_{1i} and γ_{2i} The motivation for this assumption is that those most responsive to the instrument (via a large negative value to γ_{2i}) also have a program impact that deviates from the overall population average impact.

Let us now consider a numerical example based on this system. This example is captured in the STATA do-file 6.3.do. The potential outcome and cost equations are parameterized as follows for 50,000 individuals:

$$Y_i^1 = 2 + \omega_i + 1.5 \cdot x_i + \mu_i + \epsilon_i^Y$$
$$Y_i^0 = 2 + 1.5 \cdot x_i + \mu_i + \epsilon_i^Y$$
$$C_i = 5.5 + 2 \cdot x - 1.4 \cdot \omega_i \cdot z_i + \mu_i + \epsilon_i^C$$

where x and μ are independently normally distributed with mean 0 and variance 4 (i.e. x, $\mu \sim$ N(0,4))) and the ϵs are independently normally distributed with mean 0 and variance 9 (i.e. ϵs $\sim$ N(0,9)). The instrument z equals 1 if a normally distributed random variable with mean 0 exceeds 0, and equals 0 otherwise. Finally, the variable ω_i is distributed uniformly on the interval $[0, 4]$ (this means that ω can take on any value from 0 to 4, with all possible values in that interval having equal probability of occurring).

Before proceeding to the results, it is worth considering the role that ω_i plays in this system. Program impact is

$$\begin{aligned} Y_i^1 - Y_i^0 &= 2 + \omega_i + 1.5 \cdot x_i + \mu_i + \epsilon_i^Y \\ &\quad - \left(2 + 1.5 \cdot x_i + \mu_i + \epsilon_i^Y\right) \\ &= \omega_i \end{aligned}$$

Program impact is thus individual specific. Individual i participates if

$$Y_i^1 - Y_i^0 - C_i > 0$$

or

$$\omega_i - \left(5.5 + 2 \cdot x - 1.4 \cdot \omega_i \cdot z_i + \mu_i + \epsilon_i^C\right) > 0$$

Clearly, the instrument z_i will influence the participation decision, but the degree to which it does will depend on ω_i, which can take on values as small as 0 (in which case the instrument z plays no roles in determining participation) or as large as 4. Since program impact is ω_i, this means that individuals with larger program impact will have greater responsiveness to the instrument z as a determinant of their participation decision.

Output 6.42 provides the overall participation pattern. 36.12 percent participate in the program. This is somewhat below the participation rates in many examples to this point in the manual.

STATA Output 6.42 (6.3.do)

```
. * Basic summary statistics: participation
. tabulate P
```

P	Freq.	Percent	Cum.
0	31,942	63.88	63.88
1	18,058	36.12	100.00
Total	50,000	100.00	

Output 6.43 provides summary statistics by participation status. Non-participants have much higher average values for the potential outcomes, but also much higher costs. Interestingly, the difference in the potential outcomes is just about 1.76 for non-participants and about 2.4 for participants. In other words, participants have higher program impacts. Of course, this may also reflect the fact that those with higher program impact also tend to have lower costs of participation via the same mechanism: ω. Participants have lower values for the observed and unobserved, respectively, background characteristics x and μ. Given the positive role they play in shaping potential outcomes, this explains much of the difference in average potential outcomes $\{Y^0, Y^1\}$ between participants and non-participants.

The difference in the average value of the unobserved characteristic μ between participants and non-participants hints at potential problems estimating program impact by simple means (such as comparison of average outcomes Y between participants and non-participants or straightforward regression of Y on P and x). The reason should be a familiar one by now: program participation is associated with the unobservable μ. Therefore, when one attempts to capture average differences in the outcome Y between participants and non-participants, either by simple comparison of averages or straightforward regression, program participation plays two empirical roles: a control for the experience of program participation (which is what we want it to do) and proxy for the unobserved characteristic μ (which we do not want it to do). Its muddled empirical role yields a biased and inconsistent estimate of program impact that combines both the true effect with a contribution that reflects program participation P's burden of serving as a proxy for the unobservable μ.

Output 6.44 presents correlations among some of the key variables in this simulation. First, and most importantly, program participation P is highly correlated with the unobservable μ. It is also highly correlated with the instrument z, which provides some re-assuring *prima facie* evidence that, at the least, weak instruments will not be a concern. Moreover, the instrument z is not really correlated with the error term from the outcome equation, ϵ^Y (with a meager correlation of -0.0024). Program participation is highly correlated with ω (`omega` in Output 6.44).

STATA Output 6.43 (6.3.do)

```
. * Basic summary statistics: variable means
.
. by P, sort: summarize Y y1 y0 c x* z* mu epsilon*
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	31942	3.504195	4.264767	-16.55925	21.04068
y1	31942	5.269855	4.525998	-16.37425	24.27966
y0	31942	3.504195	4.264767	-16.55925	21.04068
c	31942	7.351672	3.856453	.045818	28.47001
x0	31942	1	0	1	1
x	31942	.765381	1.732334	-5.846226	8.543729
z	31942	.4285893	.494882	0	1
mu	31942	.3818189	1.944444	-7.467243	8.587901
epsilony	31942	-.0256958	2.987694	-11.88791	12.87173
epsilonc	31942	.8952693	2.772225	-8.871226	12.05819

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	18058	1.652806	4.473777	-17.68306	18.638
y1	18058	1.652806	4.473777	-17.68306	18.638
y0	18058	-.7560726	4.14057	-19.3104	16.24554
c	18058	-1.672156	3.215136	-21.42513	3.920629
x0	18058	1	0	1	1
x	18058	-1.379782	1.668691	-8.141914	5.449922
z	18058	.6299147	.4828406	0	1
mu	18058	-.6756291	1.926847	-8.096637	6.72061
epsilony	18058	-.0107699	2.999512	-11.36862	12.5643
epsilonc	18058	-1.564312	2.751916	-12.37837	9.243759

STATA Output 6.44 (6.3.do)

```
. * Correlations among unobservables
.
. correlate P z* mu epsilony epsilonc omega
(obs=50000)
```

	P	z	mu	epsilony	epsilonc	omega
P	1.0000					
z	0.1934	1.0000				
mu	-0.2535	0.0068	1.0000			
epsilony	0.0024	-0.0022	0.0018	1.0000		
epsilonc	-0.3929	0.0051	-0.0065	-0.0025	1.0000	
omega	0.2674	-0.0083	0.0034	-0.0004	0.0010	1.0000

Output 6.45 presents an estimate of average program impact at the population level from our simulated sample of 50,000. The estimate is, unsurprisingly, just about 2 at 1.997965. This is unsurprising because average program impact is simply the average of

$$Y_i^1 - Y_i^0 = \omega_i$$

However, we have assumed that ω is a uniformly distributed random variable on the interval

spanning 0 to 4. The expected value of such a variable is 2. We have not explicitly presented the estimate of average program impact in the simulation examples to this point. We do so now simply because below we will be examining average program impact across the subpopulations of importance for LATE.

STATA Output 6.45 (6.3.do)

```
. * Average program impact
.
. generate TE=y1-y0
. summarize TE
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          TE |     50000    1.997965     1.15535   .0000782   3.999935
```

Output 6.46 presents estimates from straightforward regression of Y on P and x. The estimate of program impact is, at 1.110648, just over half the true average impact of 2. This is the consequence of the correlation between P and μ on display. The direction of the bias is exactly what the omitted variables formulas worked out in earlier chapters suggested: because μ and Y should be positively correlated but μ and P are negatively correlated, the bias term is negative, suggesting that the bias to the estimate of program impact from straightforward regression should be downward.

STATA Output 6.46 (6.3.do)

```
. * Cross sectional regression
.
. regress Y x P
      Source |       SS       df       MS              Number of obs =   50000
-------------+------------------------------           F(  2, 49997) =11983.07
       Model |  318162.39     2  159081.195             Prob > F      =  0.0000
    Residual |  663735.141 49997  13.2754993            R-squared     =  0.3240
-------------+------------------------------           Adj R-squared =  0.3240
       Total |  981897.531 49999  19.6383434            Root MSE      =  3.6436

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.380798   .0095312   144.87   0.000     1.362116    1.399479
           P |    1.110648   .0396082    28.04   0.000     1.033015    1.188281
       _cons |    2.447358   .0216525   113.03   0.000     2.404919    2.489798
------------------------------------------------------------------------------
```

In Output 6.47 we present estimates of average program impact from compliers, never-takers and always-takers. Since this is a simulated example, we parameterized the system and thus have the luxury of easily determining the subpopulation among these three to which a particular observation belongs. Doing so basically involves a hypothetical calculation. First, set the instrument equal to 0 for every individual in the sample. Then determine whether they would participate given that instrument value, the value we attached to every parameter of the system, and the individual's draw for every other variable (e.g. x_i, μ_i, ϵ_i^Y, etc.). Let the resulting participation outcome be $P_i^{z=0}$. Then set the instrument equal to 1 for each individual in the sample and determine their participation outcome under this regime. We will call in $P_i^{z=1}$.

The never-takers are the individuals for whom

$$P_i^{z=0} = P_i^{z=1} = 0$$

In other words, the never-takers are those who never elect to participation, regardless of whether the instrument z equals 0 or 1. Similarly, the always-takers are those for whom

$$P_i^{z=0} = P_i^{z=1} = 1$$

The always-takers thus always participate, regardless of instrument value. The compliers are those for whom

$$P_i^{z=0} = 0$$

and

$$P_i^{z=1} = 1$$

In other words, changing the instrument z's value from 0 to 1 induces the compliers to switch from non-participation to participation in the program. There is of course another hypothetical possibility, that

$$P_i^{z=0} = 1$$

and

$$P_i^{z=1} = 0$$

In other words, there is the hypothetical possibility that some would switch from participant to non-participant as the instrument z switches from 0 to 1. It is this possibility that is commonly ruled out by the monotonicity assumption.

STATA Output 6.47 (6.3.do)

```
. * The average treatment effect for compliers
. summarize TE if comp==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          TE |       9642    2.705054     .913936   .0256362    3.99993
. * The average treatment effect for the never takers
. summarize TE if never==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          TE |      27130     1.59954    1.080428   .0000782   3.999935
. * The average treatment effect for the always takers
. summarize TE if always==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          TE |      13228    2.299711     1.12321   .0001178   3.999846
```

From Output 6.47 we can see that average program impact differs between these three subpopulations. Compliers (for whom the variable `comp` in Output 6.47 equals 1.) have the highest estimated average program impact at 2.705054. The lowest estimated average program impact at 1.59954 is among the never-takers (indicated by `never`). Note as well that the three subpopulations contain, collectively, 9,642 + 27,130 + 13,228 = 50,000 observations. This means that there is no one for whom

$$P_i^{z=0} = 1$$

and

$$P_i^{z=1} = 0$$

In other words, this simulated example satisfies the monotonicity assumption.

Finally, in Output 6.48 we present results from two-stage least squares estimation of program impact. The estimate of program impact is 2.742388. Notice that this is almost exactly the same as the estimate of program impact for compliers of 2.705054 that we estimated directly. It is not close to the program impact estimates for the never- and always-takers. This is the essence of LATE: linear instrumental variables has yielded an estimate of program impact that reflects the program impact of compliers.

STATA Output 6.48 (6.3.do)

```
. ivregress 2sls Y x (P=x z), first
First-stage regressions
```

```
                                          Number of obs   =      50000
                                          F(   2,  49997) =   10972.17
                                          Prob > F        =     0.0000
                                          R-squared       =     0.3050
                                          Adj R-squared   =     0.3050
                                          Root MSE        =     0.4004
```

P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	-.1244896	.0008972	-138.76	0.000	-.1262481	-.1227312
z	.1886522	.0035817	52.67	0.000	.181632	.1956725
_cons	.2654226	.002536	104.66	0.000	.2604521	.2703932

```
Instrumental variables (2SLS) regression      Number of obs =   50000
                                              Wald chi2(2)  =22663.92
                                              Prob > chi2   =  0.0000
                                              R-squared     =  0.3011
                                              Root MSE      =  3.7048
```

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
P	2.742388	.1756519	15.61	0.000	2.398116	3.086659
x	1.58349	.0233447	67.83	0.000	1.537735	1.629245
_cons	1.859938	.0653687	28.45	0.000	1.731817	1.988058

```
Instrumented:  P
Instruments:   x z
```

The possibility that an instrumental variable estimate reflects LATE has enormous implications. If the estimate has a LATE interpretation, then the instrumental variables estimator is simply not yielding an estimate of average program impact that is relevant to the entire population for which we wish to learn program impact. Rather, it consistently estimates average program impact only for the complier population: those whose participation behavior responds to changes in the value of the instrument.

This is a potential enormous complication, and raises reasonable questions about the usefulness of instrumental variables as an estimator of program impact. There are a couple of things to bear in mind, however. First, LATE is at least a consistent estimate of program impact for *some* subpopulations. Second, the possibility of LATE in some sense strengthens the argument for over-identified models: to the extent that different subpopulations respond to different specific

instruments z, the more instruments included the potentially broader the subpopulation for which the instrumental variables estimator consistently estimates average program impact.

Finally, while we cannot know which of the three subpopulations (never-takers, always-takers and compliers) a particular observation belongs to, we can characterize some of the features of the complier subpopulation as a whole.[32] To begin with, we can estimate the proportion of the sample that belongs to the complier subpopulation. The size of the complier subpopulation is simply the sample average of

$$Pr\left(P_i = 1 | z_i = 1\right) - Pr\left(P_i = 1 | z_i = 0\right)$$

In other words, it is the average marginal effect of z from the first stage (the average change in the probability of participation as the instrument z switches from 0 to 1 in value).

STATA Output 6.49 (6.3.do)

```
. tabulate comp
       comp |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     40,358       80.72       80.72
          1 |      9,642       19.28      100.00
------------+-----------------------------------
      Total |     50,000      100.00

. generate zt=z

. logit P x z
Iteration 0:   log likelihood = -32704.135
Iteration 1:   log likelihood =  -23801.44
Iteration 2:   log likelihood = -23502.054
Iteration 3:   log likelihood = -23500.356
Iteration 4:   log likelihood = -23500.356
Logistic regression                               Number of obs   =      50000
                                                  LR chi2(2)      =   18407.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -23500.356                       Pseudo R2       =     0.2814
------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.8094979   .0080578  -100.46   0.000    -.8252909   -.7937049
           z |   1.198323    .023642    50.69   0.000     1.151986    1.244661
       _cons |  -1.469398   .0184647   -79.58   0.000    -1.505588   -1.433207
------------------------------------------------------------------------------

. replace z=0
(25065 real changes made)

. predict p0
(option pr assumed; Pr(P))

. replace z=1
(50000 real changes made)

. predict p1
(option pr assumed; Pr(P))

. generate mx=p1-p0

. su mx
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mx |     50000    .1885606    .0861325   .0005275    .290929
```

Recall that we calculated that 9,642 of the 50,000 observations, or about 19.3 percent, were

[32] For those who wish to move a bit beyond the coverage of this topic below, we recommend the excellent discussion in Angrist and Pischke (2009), Section 4.4.4.

compliers. In Output 6.49 we estimate the proportion of the sample who are compliers. First, we actually tabulate `comp` to re-confirm that around 19.3 percent of the sample are compliers. Next we regress P on x and z by logit regression, and use the fitted model (i.e. the model with the parameter estimates from the regression) to compute the marginal effect of the instrument z). We then average the marginal effect across the population to arrive at the estimate that 18.85606 percent of the sample are compliers, which is extremely close to the actual value of 19.3 percent.

Thus, even with just the information that we would normally actually observe (i.e. $\{Y_i, P_i, x_i, z_i\}$) we would have been able to estimate that just under one in five in the sample are compliers. This is a low but in no way atypical figure. It is common to find that the compliers are a distinct minority of the population and of any representative sample from that population.

It is also possible to learn something about the average characteristics of the compliers. Since the average characteristics of the population as a whole are easily estimated (simply compute the sample average for each characteristic across a random sample from that population) it is thus possible to gain insights into the ways in which compliers differ from the typical individual in the population of interest.

A variety of approaches (i.e. formulas) have been proposed for this. We will focus on the "kappa weighting" estimator from Abadie (2003), under which

$$E\left(x_i|comp_i = 1\right) = \frac{E\left(\kappa_i \cdot x_i\right)}{E\left(\kappa_i\right)}$$

where

$$\kappa_i = 1 - \frac{P_i \cdot (1 - z_i)}{1 - Pr\left(z_i = 1|x_i\right)} - \frac{(1 - P_i) \cdot z_i}{Pr\left(z_i = 1|x_i\right)}$$

This formula thus allows us to estimate the expected value of x_i given that the individual is a complier (i.e. $comp_i = 1$, where *comp* is a binary indicator that equals 1 if the individual is a complier and 0 otherwise along the lines of the variable `comp` from the various Outputs in this section).

Output 6.50 reports the average of x across the entire sample of 50,000 individuals as well as for the 9,642 who are compliers. The compliers have a lower estimated average value for x (-.2576623 against -.0093662). We are able to recover directly an estimate of the average of x because this is a simulation of our own creation (hence we could determine whether each individual in the sample is a complier or not). With a real world sample we would observe only $\{Y_i, P_i, x_i, z_i\}$ and thus would need to estimate the average value of x with something like the kappa weighting estimator, since it relies only on this more limited set of variables.

STATA Output 6.50 (6.3.do)

```
. su x
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           x |      50000   -.0093662    1.996121  -8.141914   8.543729

.
. su x if comp==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           x |       9642   -.2576623    1.443206  -5.658031   5.449922
```

Output 6.51 presents the computation of the the kappa weighting estimate of the average of x among compliers. The approach taken is to construct κ_i (referred to simply as k in the STATA code), compute $\kappa_i \cdot x_i$ (kx in the STATA code), calculate the sample average for each and divide. The estimate of the average of x among compliers is -.30396967, which is not far off of the true sample value of -.2576623.

STATA Output 6.51 (6.3.do)

```
. logit z x
Iteration 0:   log likelihood =  -34657.19
Iteration 1:   log likelihood = -34656.368
Iteration 2:   log likelihood = -34656.368
Logistic regression                                     Number of obs   =      50000
                                                        LR chi2(1)      =       1.64
                                                        Prob > chi2     =     0.1999
Log likelihood = -34656.368                             Pseudo R2       =     0.0000

------------------------------------------------------------------------------
           z |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .0057439   .0044811     1.28   0.200    -.0030389    .0145267
       _cons |    .005254   .0089445     0.59   0.557     -.012277     .022785
------------------------------------------------------------------------------

.
. predict pz
(option pr assumed; Pr(z))
.
. g k=1-(P*(1-z))/(1-pz) - ((1-P)*z)/pz
.
. g kx=k*x
.
. su k
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           k |     50000    .1886487    .9784497  -1.027593          1

.
. loc a=r(mean)
.
. su kx
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          kx |     50000   -.0573435    1.979768  -7.841552   8.543729

.
. loc b=r(mean)
.
. di `b´/`a´
-.30396967
```

We conclude by noting in passing that nonlinear instrumental variables models often technically do not suffer from concerns regarding LATE. The reason is that these models typically rely on distributional assumptions of the sort we considered in the last section, and the nonlinearity associated with those distributional assumptions provides identification beyond what instruments provide. For instance, in a bivariate probit model the assumed joint normality of the error terms from the outcome and program participation latent variable equations is a source of identification separate from that generated by the instruments, which are the source of the LATE concern. Thus the estimate of average program impact generated by such models is not necessarily subject to a LATE interpretation. Of course, the nonlinearity is the source of this additional identification, and

as we have seen that might not constitute very reliable identification.

6.3 Regression Discontinuity Designs

We now briefly discuss an approach to program impact evaluation that could have served as the basis for a short chapter of its own, but is presented here because perhaps the most widely applicable version of it can be naturally viewed as a species of instrumental variables. The departure point for this estimation tradition is the reality that some programs impose participation criteria in the form of eligibility thresholds. In particular, these eligibility thresholds are generally defined with respect to the characteristics of the prospective participants. For instance, some programs might allow only those below a poverty threshold to enroll.

The eligibility threshold can then become a point at which the probability of participation might experience a "discontinuity". Specifically, on the ineligible side of the threshold are individuals who are not eligible to enroll because of their characteristics. They are then either unable or unwilling to participate in the program due to the burden of ineligibility per program's eligibility rules (and the difference between these two cases is of intellectual interest). Compared with those who are just on the ineligible side of the threshold, those whose traits place them at the threshold of eligibility suddenly have the constraints imposed by these eligibility criteria removed. This can generate a rather sudden spike in the probability of program participation compared with those who are just barely ineligible.

Let us consider the simplest case. Suppose that there is a program in which everyone would participate if they could. Suppose as well that individuals have one observed characteristic x. A program decides to allow individuals to participate if their observed characteristic x is below an eligibility threshold e. In other words, the individual is eligible to participate if

$$x \leq e$$

while he or she is ineligible if

$$x > e$$

Suppose as well that the eligibility criteria is strict in the sense that no one for whom $x > e$ will be allowed to participate.

This circumstance is referred to as a **sharp discontinuity** and is illustrated in Figure 6.10. This Figure graphs the probability of participation $Pr(P = 1|x)$ against the observed characteristic x. Notice that this probability suddenly falls from 1 to 0 at the threshold. This is classically referred to as a sharp discontinuity because participation is essentially a completely deterministic function of the characteristic x. In other words, whether one is a participant or not is basically completely determined by which side of the eligibility threshold their value for x places them.

Let us suppose that participation in the program actually has some positive impact (we will refer to that impact as I) on an outcome Y. Further, we will assume that the outcome depends at least in part on the observable x as well.[33] To fix ideas, let us assume that x has a positive effect on Y. Figure 6.11 graphs the outcome Y against the observable x.

The relationship between Y and x experiences a sudden discontinuity at e. The reason for this is that participation falls suddenly and dramatically immediately above $x = e$. In other words, the discontinuity (i.e. break) in the relationship between Y and x at $x = e$ is driven completely by the discontinuity in program participation. It reflects the dynamics of program participation with respect to x.

[33]For the purposes of the immediate point being made, the second assumption is simply for intuitive appeal and our ability to recover an estimate of program impact does not depend on it.

The discontinuity/break in the relationship between Y and x is an artifact of the discontinuity/break in program participation P at $x = e$. Thus, the discontinuity in Figure 6.11 is more or less driven by program participation. Therefore, it is revealing program impact.

One possible regression specification that could tease out this program impact is

$$Y = \beta_0 + \beta_1 \cdot I\left(x \leq e\right) + \beta_2 \cdot x + \epsilon$$

where $I(\cdot)$ is an indicator function that equals 1 if the logical condition it references (in this case $x \leq e$) holds and 0 otherwise. The term

$$\beta_1 \cdot I\left(x \leq e\right)$$

basically captures the possibility of a shift in the constant term between the eligible and the ineligible. Then, the effective constant term for ineligible individuals is β_0 while that for the eligible individuals is

$$\beta_0 + \beta_1$$

Program impact is then β_1.

Notice that the regression specification

$$Y = \beta_0 + \beta_1 \cdot I\left(x \leq e\right) + \beta_2 \cdot x + \epsilon$$

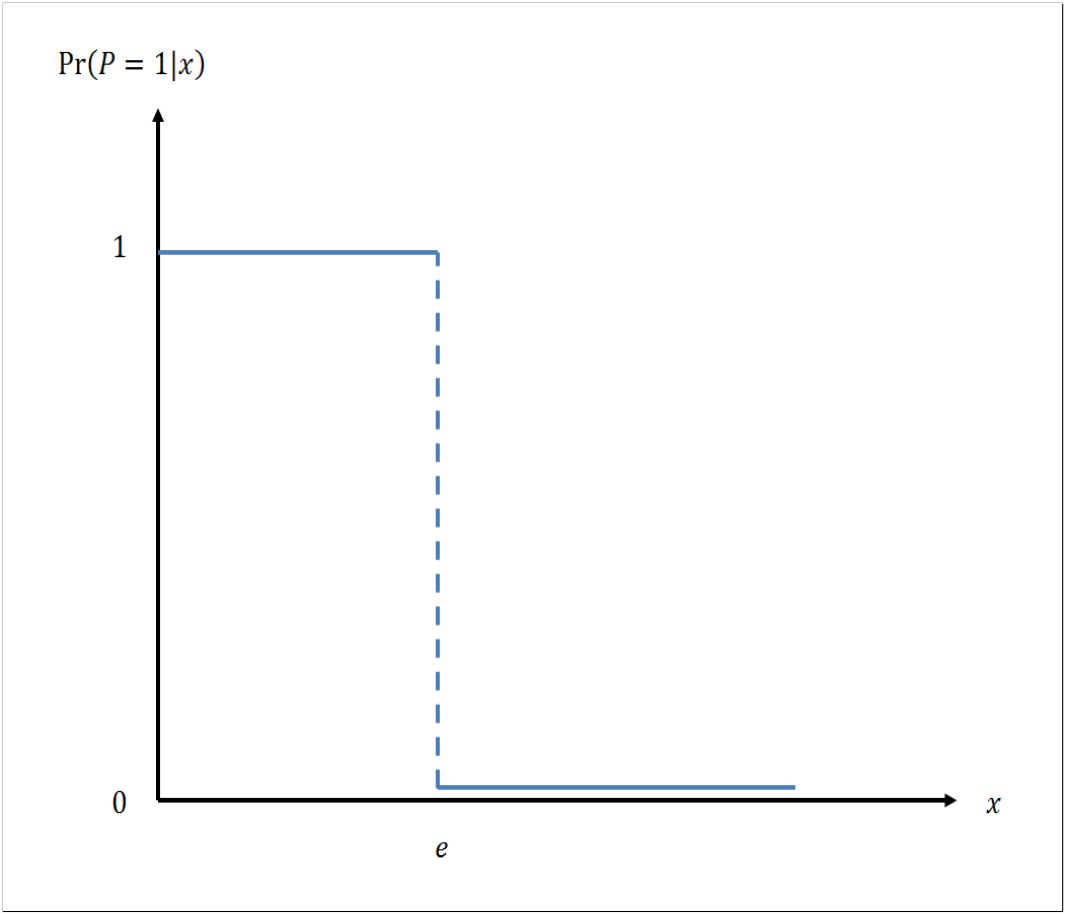


Figure 6.10: A "Sharp" Discontinuity Design

is for all intents and purposes identical to

$$Y = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot x + \epsilon$$

where P is an indicator of program participation that equals 1 if the individual participates and 0 otherwise. The reason is that an individual's value for P depends wholly on whether their observed characteristic x is above or below the threshold e. In other words,

$$P = I\left(x \leq e\right)$$

This highlights the critical feature of the sharp design: program participation is a mechanical function of the observed characteristic x.

This means that the program participation decision rule we have discussed to this point is not really meaningful. Specifically, our typical participation decision is based on whether

$$Y^1 - Y^0 - C > 0$$

Now, participation is not really a decision for those on the ineligible side of the threshold: those who are ineligible cannot enroll no matter what.

One detail we have not tied down is what insures that all eligible individuals do enroll. To guarantee this, we typically need to add the condition

$$Y^1 - Y^0 - C > 0$$

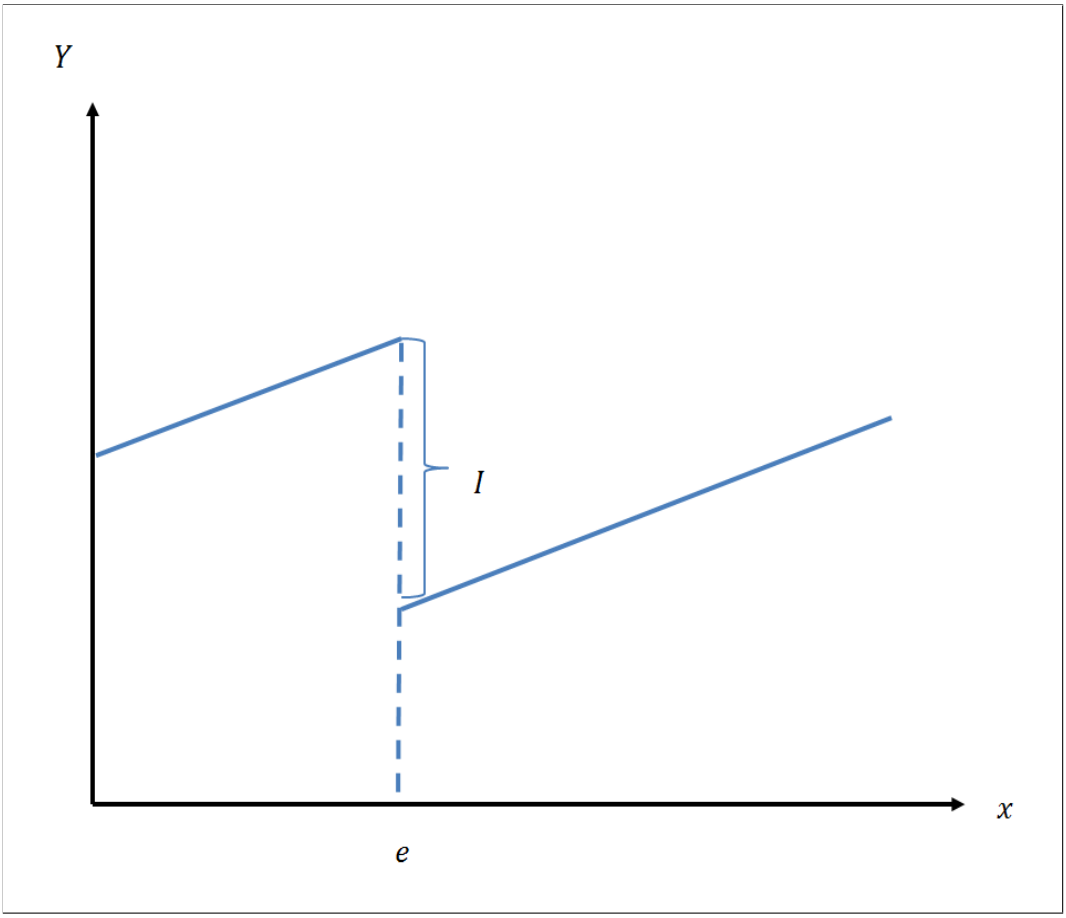


Figure 6.11: A "Sharp" Discontinuity Design Treatment Effect

At a minimum, this must hold for all individuals who are eligible (i.e. for all individuals for whom $x \leq e$) although many probably think of this as holding at all values for x (so that a pure sharp design obtains regardless of the value for the threshold e).

The upshot of this is that from a regression standpoint the strict sharp design is in some sense a classic selection on observables circumstance. To explore this briefly before moving on to a framework more directly relevant to instrumental variables, let us consider a behavioral model.

Let us assume again that we have a sample of N individuals from a population indexed by i. Suppose that potential outcomes are

$$Y_i^0 = \beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

Notice that program impact is simply

$$Y_i^1 - Y_i^0 = \beta_1$$

In other words, for now we assume constant program impact.

Let us consider eligibility and the cost of participation. To begin with, an individual is eligible to participate if $x_i \leq e$. They are ineligible if $x_i > e$. It is still useful to specify a cost function because even eligible individuals could face some cost associated with participation. Let us assume that the cost of participation takes on the simple form

$$C_i = \gamma_0 + \gamma_1 \cdot (1 - I(x_i \leq e)) + \gamma_2 \cdot x_i + \gamma_3 \cdot \mu_i + \epsilon_i^C$$

Notice that the individual's cost of participation changes by γ_1 if they are ineligible. In other words, being eligible removes a cost of participation of γ_1. Let us also define

$$\tilde{C}_i = \gamma_0 + \gamma_2 \cdot x_i + \gamma_3 \cdot \mu_i + \epsilon_i^C$$

This is simply the cost of participation without considering eligibility.

The individual participates if

$$Y_i^1 - Y_i^0 - C_i > 0$$

A classic sharp design as presented can be introduced with two assumptions. First, we must assume that

$$Y_i^1 - Y_i^0 - \tilde{C}_i > 0$$

for all individuals i (or at least for the eligible). This insures that all individuals would elect to participate were they eligible to do so. Second, the term γ_1 must be so large that it overwhelms every other determinant of cost (x_i, μ_i and ϵ_i^C) and the program return β_1 as a determinant of the participation decision: if the individual is ineligible, they face the additional cost γ_1, which makes participation impossible.

Notice that this means that there is no real reason to believe that the unobservable μ_i is correlated with program participation P_i. All individuals would participate if given the opportunity to do so and what separates participants and non-participants is simply eligibility.

Once again, a regression specification is straightforward. We have

$$Y_i = P_i \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right)$$

$$+ (1 - P_i) \cdot \left(\beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y\right)$$

$$= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

Note that we could substitute eligibility $I(x_i \leq e)$ for P in this equation.

We briefly discuss a numerical example. Specifically, in STATA do-file 6.4.do we simulate 10,000 observations from the model:

$$Y_i^1 = 13 + 2.5 \cdot x_i + \mu_i + \epsilon_i^Y$$
$$Y_i^0 = 2 + 2.5 \cdot x_i + \mu_i + \epsilon_i^Y$$

where x, μ and ϵ^Y are all normally distributed with mean 0 and variance 1 (i.e. x, μ, $\epsilon^Y \sim$ N(0,1)). Individual i is eligible to participate if $x_i \leq 0$. The cost of participation is

$$C_i = 1 + 100 \cdot (1 - I(x_i \leq 0)) - 1.5 \cdot x_i - \mu + \epsilon_i^C$$

where ϵ_i^C is normally distributed with mean 0 and variance 1 (i.e. $\epsilon^C \sim$ N(0,1)). $I(x_i \leq 0)$ is an indicator variable that equals 1 if $x_i \leq 0$ and 0 otherwise. Notice that the costs triggered by ineligibility are, at 100, large in magnitude compared with the rest of the model. It would be useful to keep in mind cost without considering eligibility:

$$\tilde{C}_i = 1 - 1.5 \cdot x_i - \mu + \epsilon_i^C$$

This will allow us to consider the individual's intentions in the absence of eligibility criteria.

We present some basic summary statistics in Outputs 6.52 through 6.55. In Output 6.52 we tabulate and cross tabulate the program participation and eligibility status (denoted by `elig`). Roughly half participate and are eligible. However, the major take-away from this Output is that program participation and eligibility are the same (as the sharp design would predict).

STATA Output 6.52 (6.4.do)

```
. * Basic summary statistics: participation and eligibility
. tabulate P
```

P	Freq.	Percent	Cum.
0	5,037	50.37	50.37
1	4,963	49.63	100.00
Total	10,000	100.00	

```
.
. tabulate elig
```

elig	Freq.	Percent	Cum.
0	5,037	50.37	50.37
1	4,963	49.63	100.00
Total	10,000	100.00	

```
.
. tabulate P elig
```

P \ elig	0	1	Total
0	5,037	0	5,037
1	0	4,963	4,963
Total	5,037	4,963	10,000

Output 6.53 presents the sample averages of the key variables by program participation status (which is the same as saying that it does so by eligibility status). Perhaps the most important thing to note is that there is no overlap in the values of x between non-participants/ineligibles and participants/eligibles. As Angrist and Pischke (2009) highlight (with reference to Imbens and Lemieux (2008)), this reflects one respect with which regression discontinuity (or at least the sharp discontinuity variant) differs from other quasi-experimental methods: it depends on there being no value for x that occurs among participants and non-participants. While the mean of x differs between the two groups, those for μ, ϵ^Y (represented by `ey`) and ϵ^C (represented by `ec`) are basically the same.

Costs net of eligibility threshold considerations $\tilde{C}$ (represented by `Ct`) differ between the two groups, but this reflects mainly the differences in the average values for x between the two groups (since the means for μ and ϵ^C are basically the same across them). Notice the average cost for non-participants is very high (at 99.818) compared with that for participants. This reflects the ineligibility cost penalty.

STATA Output 6.53 (6.4.do)

```
. by P, sort: summarize x mu Y y1 y0 C  Ct elig ey ec
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
x	5037	.7921362	.6011398	.0000284	3.641588
mu	5037	-.0020822	.988782	-3.328795	4.167702
Y	5037	3.962622	2.043871	-2.493304	11.74773
y1	5037	14.96262	2.043871	8.506696	22.74773
y0	5037	3.962622	2.043871	-2.493304	11.74773
C	5037	99.818	1.652332	93.41628	104.9598
Ct	5037	-.5780662	1.831878	-8.036539	4.823392
elig	5037	0	0	0	0
ey	5037	-.0156363	1.00283	-3.579602	3.630332
ec	5037	.004124	.994509	-3.786709	3.449271

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
x	4963	-.8051544	.604685	-3.918931	-.000185
mu	4963	-.0009141	1.000649	-3.379955	3.25495
Y	4963	10.98765	2.067781	2.205629	17.54821
y1	4963	10.98765	2.067781	2.205629	17.54821
y0	4963	-.0123504	2.067781	-8.794371	6.548206
C	4963	2.21441	1.662647	-3.398403	8.797449
Ct	4963	2.616987	1.844013	-3.230353	10.12673
elig	4963	1	0	1	1
ey	4963	.0014498	1.007571	-3.204001	3.915892
ec	4963	.0057641	.9885451	-3.794222	3.700645

Output 6.54 presents the mean of net return

$$Y_i^1 - Y_i^0 - \tilde{C}_i$$

across the sample. This is represented in Output 6.54 by `ret`. Notice that it is never negative. This is necessary to insure that everyone would elect to participate if eligible (which is the most strict way to pose the strict sharp discontinuity story). When we present numerical examples we

typically offer some language to the effect that the values for the parameters used for simulation are more or less randomly chosen. That is not really the case in this instance: we had to set program impact (represented by `ATE` in Output 6.54) to 11 to insure that this condition held. Even if we had decided to pursue the minimalist approach to generating a sharp discontinuity and insure only that

$$Y_i^1 - Y_i^0 - \tilde{C}_i > 0$$

for the eligible (i.e. those for whom $x_i \leq 0$) it would still have required setting program impact to some level high enough to insure even that. This highlights the sort of strong behavioral conditions required to give rise to a pure sharp discontinuity.

STATA Output 6.54 (6.4.do)

```
. * Cost and ATE
.
. summarize ret ATE
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         ret |      10000    9.992361    2.435115   .8732748   19.03654
         ATE |      10000          11    1.99e-07         11         11
```

Output 6.55 presents some key correlations. Program participation and eligibility status are perfectly correlated (as one would expect). μ is uncorrelated with program participation status. This reflects the fact that participation is completely driven by the eligibility condition, removing any scope for sorting of participants and non-participants by μ. x is highly negatively correlated with both program participation and eligibility. This is an artifact of the fact that an individual participates (i.e. is eligible) only if x falls below some threshold.

STATA Output 6.55 (6.4.do)

```
. * Key correlations
.
. corr P elig x mu
(obs=10000)
             |        P     elig        x       mu
-------------+------------------------------------
           P |   1.0000
        elig |   1.0000   1.0000
           x |  -0.7981  -0.7981   1.0000
          mu |   0.0006   0.0006  -0.0032   1.0000
```

Outputs 6.56 and 6.57 present results from regression of Y on participation P and eligibility status, respectively. There are two major things to note from these results. First, the results of the two regressions are identical, which is a reflection of the fact that participation and eligibility status are identical. Second, the estimate of program impact is, at 7.025028, a big underestimate from the standpoint of true ATE. This is because of a key omitted variable: x. x influences both Y and eligibility/participation.

The low estimate of program impact thus reflects classic omitted variable bias (though we offer the usual caveat that, as presented, this does not prove bias). It would seem that, quite likely, this

could be dealt with simply by controlling for x (after all, the lack of correlation between P and μ rules out selection on unobservables).

STATA Output 6.56 (6.4.do)

```
. * basic regression of Y on P
.
. reg Y P
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  1,  9998) =29191.89
       Model |  123370.777     1  123370.777           Prob > F      =  0.0000
    Residual |  42253.546   9998  4.22619984           R-squared     =  0.7449
-------------+------------------------------           Adj R-squared =  0.7449
       Total |  165624.323   9999  16.5640887           Root MSE      =  2.0558

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    7.025028   .0411166   170.86   0.000     6.944431    7.105624
       _cons |    3.962622    .028966   136.80   0.000     3.905843    4.019401
------------------------------------------------------------------------------
```

STATA Output 6.57 (6.4.do)

```
. * basic regression of Y on eligibility
.
. reg Y elig
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  1,  9998) =29191.89
       Model |  123370.777     1  123370.777           Prob > F      =  0.0000
    Residual |  42253.546   9998  4.22619984           R-squared     =  0.7449
-------------+------------------------------           Adj R-squared =  0.7449
       Total |  165624.323   9999  16.5640887           Root MSE      =  2.0558

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        elig |    7.025028   .0411166   170.86   0.000     6.944431    7.105624
       _cons |    3.962622    .028966   136.80   0.000     3.905843    4.019401
------------------------------------------------------------------------------
```

In Output 6.58 we include x as a regressor. The resulting estimate of program impact is 11.00097, which as it happens is very close to the true value of 11. This is the simplest possible sharp discontinuity regression.

We can now see from this exercise that strict sharp discontinuity circumstances are in essence cases of selection on observables. A key to this that deserves highlighting is that the eligibility threshold had nothing to do with the unobservable μ. Had that not been the case (i.e. had the threshold been determined in such a fashion that the average value for μ was somehow different between those who are eligible and those who are not) then the simple regression controlling for the observable x would not have yielded an unbiased estimate of program impact since P would have been correlated with the residual via μ.

We illustrate this sharp discontinuity example graphically in Figures 6.12 and 6.13. Figure 6.12 graphs both actual observed Y (in light blue dots) and predicted Y based on the estimated

model from Output 6.58 (red dots). Notice that the predicted regression line experiences a clear discontinuity at $x = 0$. This reflects the shift in participation status at the eligibility threshold (which is the same: $x = 0$). The shift is program impact, as illustrated in Figure 6.13.

Before moving on, we note that, in practice, the strictly linear control for x in our sharp discontinuity regression might not be very safe. The reason is that the underlying relationship between x and $\{Y^0, Y^1\}$ might not be linear. If this is the case a sharp discontinuity specification that offers just one linear control for x might yield very misleading estimates of program impact. The reason for this is that with only a linear control for x when the actual relationship between x and $\{Y^0, Y^1\}$ is nonlinear it can be difficult to differentiate the effect of the program from this non-linearity.

To illustrate, consider the following trivial extension of the numerical example:

$$Y_i^1 = 13 + 2.5 \cdot x_i + 2 \cdot x_i^2 - 2 \cdot x_i^3 + \mu_i + \epsilon_i^Y$$

$$Y_i^0 = 2 + 2.5 \cdot x_i + 2 \cdot x_i^2 - 2 \cdot x_i^3 + \mu_i + \epsilon_i^Y$$

The major innovation is that we have introduced scope for non-linearity to the relationship between x and $\{Y^0, Y^1\}$ by introducing terms for x squared (x^2) and x cubed (x^3). This extension effectively implies the behavioral framework

$$Y_i^0 = \beta_0 + \beta_2 \cdot x_i + \beta_4 \cdot x_i^2 + \beta_5 \cdot x_i^3 + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

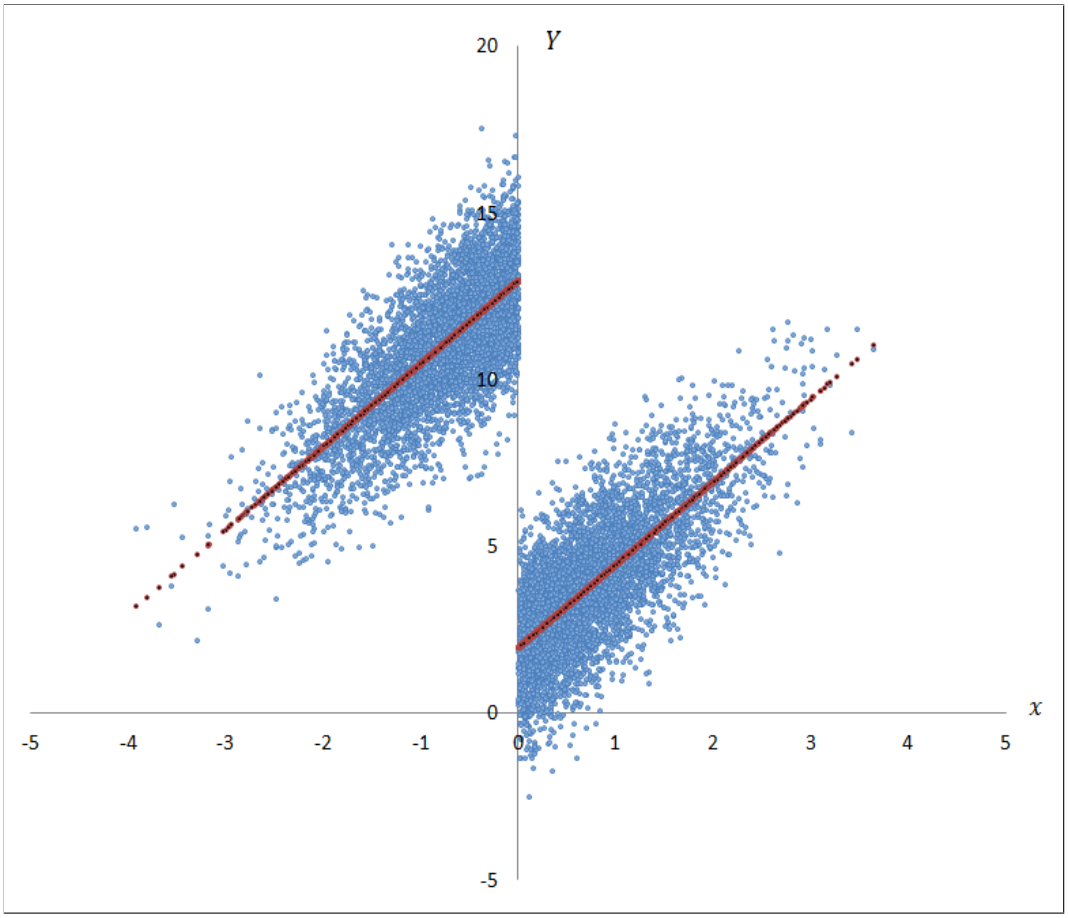


Figure 6.12: Numerical Example: A "Sharp" Discontinuity

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_4 \cdot x_i^2 + \beta_5 \cdot x_i^3 + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

The framework now allows for a more complicated, non-linear relationship between x and the potential outcomes $\{Y^1, Y^0\}$.

This in turn gives rise to the regression specification

$$\begin{aligned} Y_i &= P_i \cdot Y_i^1 + (1 - P_i)\, Y_i^0 \\ &= P_i \cdot \left(\beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_4 \cdot x_i^2 + \beta_5 \cdot x_i^3 + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &+ (1 - P_i) \cdot \left(\beta_0 + \beta_2 \cdot x_i + \beta_4 \cdot x_i^2 + \beta_5 \cdot x_i^3 + \beta_3 \cdot \mu_i + \epsilon_i^Y\right) \\ &= \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_4 \cdot x_i^2 + \beta_5 \cdot x_i^3 + \beta_3 \cdot \mu_i + \epsilon_i^Y \end{aligned}$$

The most straightforward regression model emerging from this extended framework thus explicitly recognizes the non-linearity of the relationship between Y and x. In other words, the original regression specification

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

no longer truly captures the full, true functional form of the relationship between x and Y.

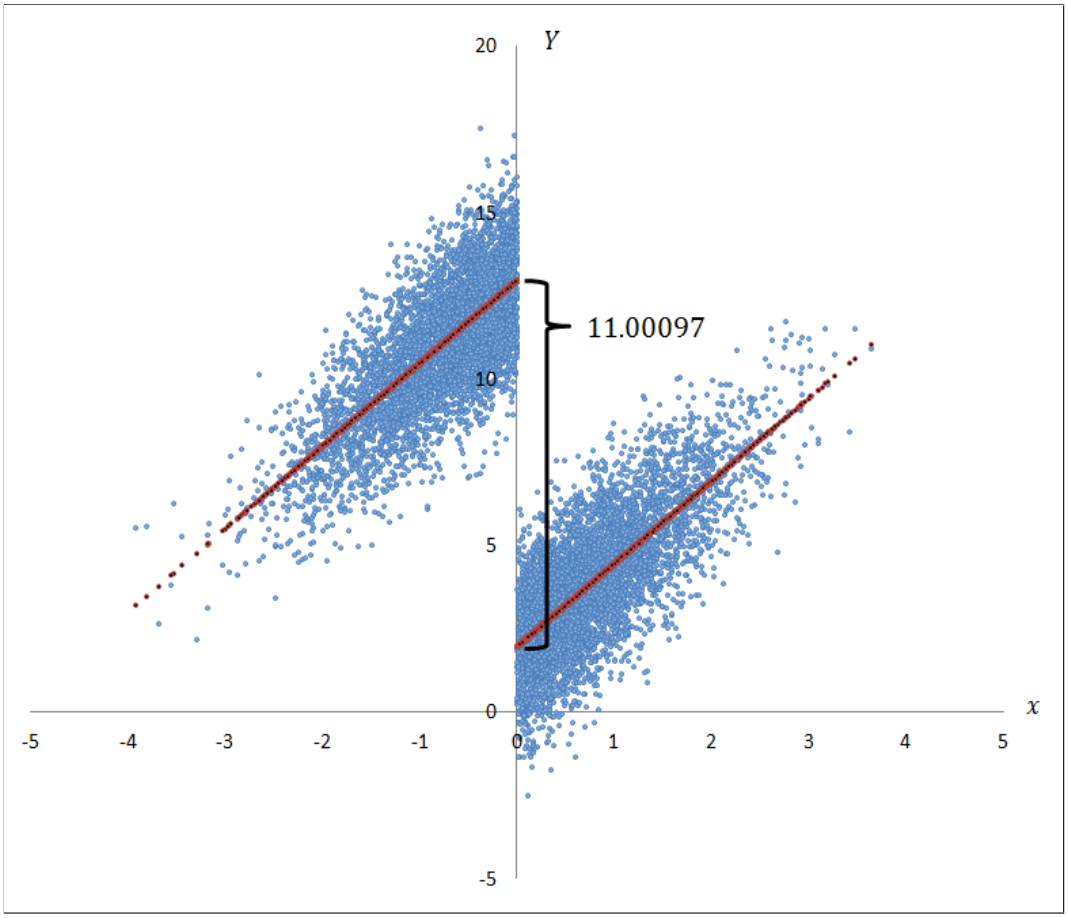


Figure 6.13: Numerical Example: Program Impact

STATA Output 6.58 (6.4.do)

```
. * basic regression of Y on P and x
.
. reg Y P x
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) =36948.63
       Model |  145888.211     2  72944.1055           Prob > F      =  0.0000
    Residual |  19736.1122  9997  1.97420348           R-squared     =  0.8808
-------------+------------------------------           Adj R-squared =  0.8808
       Total |  165624.323  9999  16.5640887           Root MSE      =  1.4051

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    11.00097   .0466443   235.85   0.000     10.90954     11.0924
           x |     2.48918   .0233073   106.80   0.000     2.443493    2.534867
       _cons |    1.990853   .0270704    73.54   0.000     1.937789    2.043916
------------------------------------------------------------------------------
```

In Outputs 6.59 and 6.60 we present results from two sharp discontinuity regressions with a sample simulated under this extension (these were estimated through STATA do-file 6.5.do, which contains the simulation per this extension to allow for a nonlinear relationship between x and Y). In Output 6.59, which repeats the same sort of estimation exercise as in Output 6.58, we now recover an estimate of program impact of 2.361814 , which is far below the true value of 11.

STATA Output 6.59 (6.5.do)

```
. * basic regression of Y on P and x
.
. reg Y P x
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) =11699.65
       Model |  628287.172     2  314143.586           Prob > F      =  0.0000
    Residual |   268426.16  9997  26.8506712           R-squared     =  0.7007
-------------+------------------------------           Adj R-squared =  0.7006
       Total |  896713.332  9999  89.6803013           Root MSE      =  5.1818

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    2.361814   .1720206    13.73   0.000     2.024619    2.699009
           x |   -6.947736   .0859556   -80.83   0.000    -7.116226   -6.779245
       _cons |     8.29658   .0998336    83.10   0.000     8.100886    8.492274
------------------------------------------------------------------------------
```

In Output 6.60 we present results from the richer approach of regressing Y on P (or one could have used the eligibility status since program participation and eligibility status are still identical) and x, x^2 and x^3. The estimate of program impact is now 11.00064, which is very close to the true value of 11.

The discrepancy in program impact estimates between Outputs 6.59 and 6.60 reflects the failure to capture the nonlinearity of the relationship between Y and x with the specification in Output 6.59. In that instance the program participation variable was forced in essence to capture some of

that non-linearity as well as program impact. With the inclusion of terms (x^2 and x^3) that explicitly capture such nonlinearity, program participation/eligibility is in some sense freed to capture impact. That said, even our extended example is restrictive (in the sense that the non-linear relationship between x and Y^1 is the same as that between x and Y^0), so real world applications might require even richer specifications. See Angrist and Pischke (2009) for one of the clearest discussions of appropriately capturing such non-linearity that we have seen.

STATA Output 6.60 (6.5.do)

```
. * basic regression of Y on P and x, x2, x3
.
. reg Y P x x2 x3
      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  4,  9995) =       .
       Model |  876977.746     4  219244.436           Prob > F      =  0.0000
    Residual |  19735.5868  9995  1.97454595           R-squared     =  0.9780
-------------+------------------------------           Adj R-squared =  0.9780
       Total |  896713.332  9999  89.6803013           Root MSE      =  1.4052

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   11.00064   .0555559    198.01   0.000     10.89174    11.10954
           x |   2.488684   .0402076     61.90   0.000     2.409869    2.567499
          x2 |   1.994864   .0099724    200.04   0.000     1.975316    2.014412
          x3 |  -1.999895   .0069435   -288.02   0.000    -2.013506   -1.986284
       _cons |   1.996159   .0323807     61.65   0.000     1.932686    2.059631
------------------------------------------------------------------------------
```

We included controls for x^2 and x^3 because we controlled the design of the data generating process and hence knew exactly the features of the non-linear relationship between x and Y. With real world samples we really would not know the nature (e.g. functional form) of the relationship between the x and Y. It is therefore advisable to include much richer terms to capture possible non-linearity: not only x, x^2 and x^3 but also x^4, x^5, x^6 and so on. We are unaware of any established best practice for this, but it seems reasonable that two "stopping criteria" for determining the order of polynomial term (i.e. the R in x^R) up to which to control is to add terms until further polynomial terms are no longer significant (with some overshooting past the last significant result to be more confident that they are indeed no longer significant) and until the estimate of program impact does not seem to change appreciably with the inclusion of further polynomial terms.

We now extend this basic idea to the arena of instrumental variables estimation. To begin with, we return to the original behavioral framework for the sharp discontinuity design, including the potential outcome equations

$$Y_i^0 = \beta_0 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

$$Y_i^1 = \beta_0 + \beta_1 + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

We also retain the eligibility threshold $x_i \leq e$ and the cost function

$$C_i = \gamma_0 + \gamma_1 \cdot (1 - I(x_i \leq e)) + \gamma_2 \cdot x_i + \gamma_3 \cdot \mu_i + \epsilon_i^C$$

In other words, to this point we retain exactly the framework that guided the discussion of sharp discontinuity.

We now depart from the sharp discontinuity story in two respects:

- We no longer assume that $Y_i^1 - Y_i^0 - \tilde{C}_i > 0$ for all individuals in our sample, nor do we assume that it must hold for eligible individuals. This implies that some eligible individuals might not choose to participate and some ineligible individuals would not have participated even if they had been eligible;
- We no longer assume that γ_1 is so large as to overwhelm any role that x, μ or ϵ^C might play in shaping the enrollment decision. This implies that some who are ineligible might participate anyway.

These two small modifications essentially shatter the clean and clear circumstances of the sharp discontinuity case.

There are several immediate implications to these extensions. First, it is no longer the case that the eligibility threshold perfectly sorts participants and non-participants: indeed there could be both participants and non-participants on both sides of the threshold. An immediate implication of this is that eligibility and program participation are not the same thing:

$$P \neq I\left(x \leq e\right)$$

The fact that there could be participants and non-participants on either side of the threshold has led to this new circumstance being labelled **fuzzy discontinuity**.

Second, to the extent that γ_1 is large enough that it still has some teeth (by which we mean that it still can determine the participation decision of at least some individuals) we would still expect some kind of discontinuity in the probability of program participation at the eligibility threshold (since the additional costs γ_1 suddenly kick in at the threshold). In particular, we now face the possibility of individuals for whom

$$Y_i^1 - Y_i^0 - \tilde{C}_i > 0$$

while

$$Y_i^1 - Y_i^0 - \left(\tilde{C}_i + \gamma\right) \leq 0$$

In other words, whether these individuals participate depends on whether they are eligible and hence face the cost γ: if they do not face them they will participate and if they do they won't do so. Hence, the imposition of the cost γ starting at the threshold can cause a discrete fall in the probability of program participation.

Finally, since there is now scope for μ to play a role in the program participation decision, it is no longer clear that we are still in the realm where estimation of program impact can rely on selection on observables methods. We likely cannot rely on methods essentially straight out of Chapter 4.

Suppose, however, that the eligibility threshold is unrelated to the unobservable μ. This behavioral framework gives rise to the regression model

$$Y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot x_i + \beta_3 \cdot \mu_i + \epsilon_i^Y$$

The residual for this is

$$\beta_3 \cdot \mu_i + \epsilon_i^Y$$

Notice that eligibility status $I(x \leq e)$ does not appear directly in the regression specification, would have a influence on Y_i only through P_i (since this is the only thing that becoming eligible might shift) and is unrelated to the residual. In other words, it influences the outcome only through the endogenous variable P_i and is not correlated with the true regression residual. In other words, eligibility $I(x_i \leq e)$ *is essentially an instrument* for P_i.

Let us consider another numerical example. Specifically, in STATA do-file 6.6.do we simulate 1,000,000 observations for the model

$$Y_i^1 = 5 - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y$$

$$Y_i^0 = 1 - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y$$

$$C_i = 2 + x_i + 5 \cdot (1 - I(x_i \leq .3)) + 3 \cdot \mu + \epsilon_i^C$$

where x and μ are drawn from a standard normal distribution (i.e. x, $\mu \sim$ N(0,1)), ϵ^Y is drawn from a normal distribution with mean 0 and variance 9 (e.g. $\epsilon^Y \sim$ N(0,9)) and ϵ^C is drawn from a normal distribution with mean 0 and variance 16 ($\epsilon^C \sim$ N(0,16)). The large sample size was designed to insure that results did not reflect any small sample quirk in light of the fact that the instrumental variables estimator is consistent.

In Output 6.61 we present patterns to participation and eligibility. First, note that around 51 percent participate while roughly 62 percent are eligible to do so. Around 21 percent of the ineligible participate. The figure for the eligible is 69.83741. This does not prove that the eligibility threshold influences participation (since the threshold itself depends on x which also influences the participation decision). However, it certainly can be viewed as *prima facie* evidence consistent with this possibility.

STATA Output 6.61 (6.6.do)

```
. * Basic summary statistics: participation and eligibility
. summarize P elig
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           P |   1000000     .512429    .4998457          0          1
        elig |   1000000      .61749    .4860003          0          1
.
. by elig, sort: su P
-------------------------------------------------------------------------------
-> elig = 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           P |    382510    .2122559    .4089056          0          1
-------------------------------------------------------------------------------
-> elig = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           P |    617490    .6983741    .4589641          0          1
```

In Output 6.62 we present the means of key variables by program participation status. To begin with, note that the means of x and μ differ between the participant and non-participant subsamples. The latter represents a distinct departure from the sharp discontinuity setting. This introduces the prospect that program participation is associated with the unobservable. In other words, it introduces the prospect that program participation is endogenous. Note as well that, unlike the sharp discontinuity setting, there is heavy overlap in the range of values for x found among participants and non-participants. Perhaps unsurprisingly, the non-participants face much higher average costs of participation. Finally, non-participants are less likely to have been eligible to participate in the program.

In Output 6.63 we present correlations for key variables. Eligibility and program participation are now heavily correlated (with a correlation of .4727) but not perfectly so. The observed characteristics x is heavily correlated with program participation, but so is the unobservable μ. Note, however, that eligibility (our candidate instrument) is essentially uncorrelated with μ with a correlation of -0.0014. The upshot of these correlations is that we likely have an endogeneity problem (were we simply to regress Y on P and x) but we also have a good candidate instrument in eligibility status, which is highly correlated with the endogenous variable (P) but not with the unobserved determinant of Y, μ.

STATA Output 6.62 (6.6.do)

```
. * Summary statistics
.
. by P, sort: summarize x mu Y y1 y0 C  elig ey ec
```

-> P = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
x	487571	.4302715	.9478949	-4.468497	4.690036
mu	487571	.3980109	.9256133	-3.870993	4.8915
Y	487571	2.926242	7.741386	-32.62632	42.46924
y1	487571	6.926242	7.741386	-28.62632	46.46924
y0	487571	2.926242	7.741386	-32.62632	42.46924
C	487571	8.837613	3.621684	4.000004	31.57618
elig	487571	.3819977	.4858765	0	1
ey	487571	.000709	3.001664	-15.50635	13.23711
ec	487571	2.123298	3.45043	-11.76793	18.73824

-> P = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
x	512429	-.4089781	.8695881	-4.9255	4.071929
mu	512429	-.3779881	.9169414	-4.885721	3.985869
Y	512429	3.174202	7.546185	-30.73956	39.41773
y1	512429	3.174202	7.546185	-30.73956	39.41773
y0	512429	-.8257979	7.546185	-34.73956	35.41773
C	512429	-.7850499	3.47768	-21.03528	3.999999
elig	512429	.8415585	.3651548	0	1
ey	512429	.0021629	3.001856	-13.64473	13.53007
ec	512429	-2.034315	3.385695	-21.70841	11.8032

STATA Output 6.63 (6.6.do)

```
. * Key correlations
.
. corr P elig x mu
(obs=1000000)
```

	P	elig	x	mu
P	1.0000			
elig	0.4727	1.0000		
x	-0.4192	-0.7850	1.0000	
mu	-0.3881	-0.0014	0.0013	1.0000

In Output 6.64 we present estimates from regression of Y on P and x. To assess these results, one must recognize that the true average treatment effect is

$$Y_i^1 - Y_i^0$$

$$5 - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y$$

$$-\left(1 - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y\right)$$

$$= 4$$

Against this, the estimate of program impact from the regression (-2.577132) is truly dreadful.

STATA Output 6.64 (6.6.do)

```
. reg Y P x
      Source |       SS       df       MS              Number of obs = 1000000
-------------+------------------------------           F(  2,999997) =95527.00
       Model |  9370250.38     2  4685125.19            Prob > F      =  0.0000
    Residual |  49044888.3999997  49.0450354            R-squared     =  0.1604
-------------+------------------------------           Adj R-squared =  0.1604
       Total |  58415138.6999999  58.4151971           Root MSE      =  7.0032

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |  -2.577132   .0154319  -167.00   0.000    -2.607378   -2.546886
           x |  -3.366212   .0077076  -436.74   0.000    -3.381318   -3.351105
       _cons |   4.374627   .0105636   414.12   0.000     4.353923    4.395331
------------------------------------------------------------------------------
```

In Output 6.65 we provide two-stage least squares estimates of program impact using eligibility as an instrument. First, in terms of first stage predictive power note that eligibility (`elig` in the Output) is highly significant. The first stage explanatory power standard seems easily met. The second stage estimate of program impact is 3.975063, which is very close to the true value of 4.

Finally, we conduct two interesting graphical exercises. First, we performed a logit regression of program participation P on x and $I(x \leq .3)$ (in other words, the eligibility indicator). We use the estimated model to predict participation. In Figure 6.14 we provide a scatter plot of participation (light blue) and predicted participation (crimson) against x. Notice that there are participants and non-participants at each value for x. More interestingly, there is a clear discontinuity to the probability of participation conditional on x, and it is at $x = .3$ (though that is admittedly difficult to determine precisely given the detail on the graph; we assure the reader that the discontinuity is indeed at $x = .3$). This is the discrete drop in the probability of participation given the imposition of the eligibility cost as soon as $x > .3$.

We repeat this exercise, only this time the focus is on μ. Hence, we perform logit regression of P on μ and $I(x \leq .3)$. In Figure 6.15 we scatter plot program participation (light blue) and predicted program participation (crimson) against μ. The pattern this time exhibits no evidence for a discontinuity in the probability of participation, but the plot of the predicted probabilities against μ clearly indicates that there appear to be two curves governing this relationship, one associated with a consistently higher predicted probability of participation than the other. The "wedge" separating these two curves is eligibility. But the important takeaway is that there is no evidence for a discontinuity with respect to the unobservable μ.

Up to this point in the discussion of the fuzzy design, we have assumed a constant program impact across the sample. However, in the context of heterogeneous program impact, the fuzzy RD (or at the least the instrumental variables implementation of it) yields an estimate with a local average interpretation. In this setting the compliers are those who would switch their participation status depending on whether they faced the eligibility cost.

To explore this, we introduce a trivial extension of our fuzzy design numerical example. Specifically, we now extend one of the potential outcome equations as follows:

$$Y_i^1 = 5 + w_i - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y$$

where w is distributed uniformly on the interval [0,15]. We simulate 1,000,000 observations for this extension in STATA do-file 6.7.do.

Before proceeding, note that program impact is now

$$\begin{aligned} & Y_i^1 - Y_i^0 \\ & = 5 + w_i - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y \\ & - \left(1 - 2 \cdot x_i + 7 \cdot \mu + \epsilon_i^Y\right) \\ & = 4 + w_i \end{aligned}$$

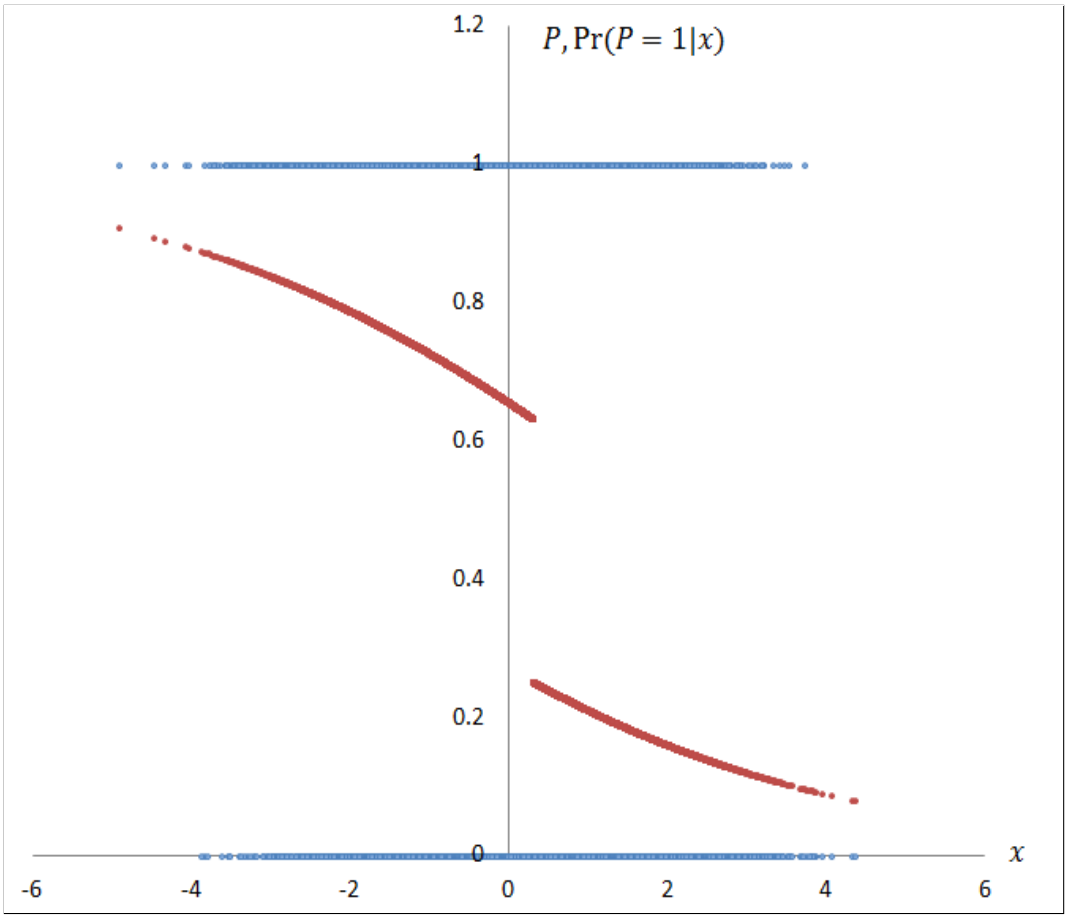


Figure 6.14: Discontinuity in the Probability of Participation: x

In other words, the program impact varies from individual to individual.

We skip the standard Outputs involving basic statistics and instead begin with Output 6.66, which presents the average treatment effect across the entire sample, as well as program impact only among compliers. The average treatment effect is 11.49778 while average program impact among compliers (i.e. the local average treatment effect, at least with respect to the instrument $I(x \leq .3)$ is 8.348064. That LATE is less than ATE is not a big surprise: the compliers are those responsive (in terms of switching their program participation decision) to the eligibility cost. Those who were closer to indifference between participating and not doing so are more likely to allow a factor such as the eligibility cost to drive their decision. Around 17.7 percent of the sample are compliers.

Output 6.67 presents results from simple regression of Y on P and x. The estimate of program impact is, at 6.961003, far from the true average treatment effect across the sample of 11.49778. The two-stage least squares instrumental variable (with eligibility $I(x \leq .3)$ as the instrument) estimate is presented in Output 6.68. The estimate of program impact from this two-stage estimation is 8.452753.

This clearly differs from the average treatment effect across the sample, but is very close to the LATE value of 8.348064. We can thus see that, in the presence of heterogeneous treatment effects, the LATE interpretation of the instrumental variables case carries over to the instrumental variables fuzzy discontinuity setting.

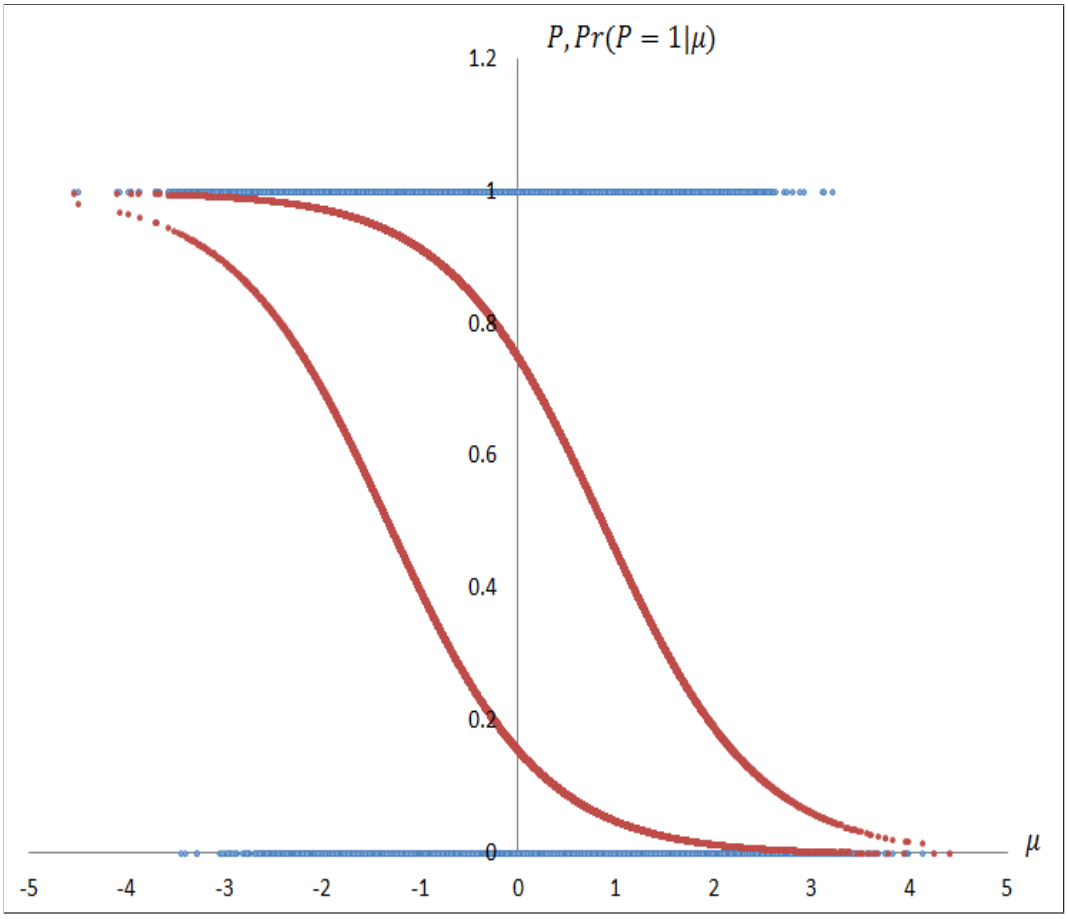


Figure 6.15: No Discontinuity in the Probability of Participation: μ

STATA Output 6.65 (6.6.do)

```
. ivreg Y x (P=elig x), first
First-stage regressions
------------------------

      Source |       SS       df       MS              Number of obs = 1000000
-------------+------------------------------           F(  2,999997) =       .
       Model |  57325.1164     2  28662.5582            Prob > F      =  0.0000
    Residual |  192520.404999997  .192520981            R-squared     =  0.2294
-------------+------------------------------           Adj R-squared =  0.2294
       Total |  249845.52999999   .24984577            Root MSE      =  .43877

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.062661   .0007077   -88.55   0.000     -.064048    -.061274
        elig |   .3848326   .0014572   264.08   0.000     .3819765    .3876888
       _cons |   .2748122   .0010012   274.48   0.000     .2728499    .2767746
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression
      Source |       SS       df       MS              Number of obs = 1000000
-------------+------------------------------           F(  2,999997) =70954.09
       Model |  528699.804     2  264349.902            Prob > F      =  0.0000
    Residual |  57886438.8999997  57.8866125            R-squared     =  0.0091
-------------+------------------------------           Adj R-squared =  0.0090
       Total |  58415138.6999999  58.4151971            Root MSE      =  7.6083

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |   3.975063   .0656616    60.54   0.000     3.846369    4.103757
           x |  -1.994455    .015709  -126.96   0.000    -2.025244   -1.963666
       _cons |   1.016797   .0344993    29.47   0.000     .9491791    1.084414
------------------------------------------------------------------------------
Instrumented:  P
Instruments:   x elig
------------------------------------------------------------------------------
```

STATA Output 6.66 (6.7.do)

```
. * ATE, whole sample
.
. summarize ATE
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         ATE |   1000000    11.49778    4.327637   4.000031   18.99995
.
. * ATE, compliers
.
. summarize ATE if complier==1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         ATE |    177378    8.348064    3.162492   4.000098   18.99975
```

The two-stage least squares specification presented has involved simply linear terms in x. In practice, the first and second stage specifications often involve additional polynomial terms in x, $\{x, x^2, x^3, ..., x^R\}$, and in the first stage specification, interactions between $I(x \leq e)$ and those

polynomials, $\{x \cdot I, x^2 \cdot I, x^3 \cdot I, ..., x^R \cdot I\}$ (where we abbreviate $I(x \leq e)$ as I). These sort of polynomial terms are introduced for a very similar reason to that which led to their introduction in the sharp discontinuity case. In our example we got away with not doing so because the underlying relationships between x and the potential outcomes and cost was linear.

STATA Output 6.67 (6.7.do)

```
. reg Y P x
      Source |       SS       df       MS              Number of obs = 1000000
-------------+------------------------------           F(  2,999997) =95547.58
       Model |  14080954.1     2  7040477.06           Prob > F      =  0.0000
    Residual |  73685339.5999997  73.6855605           R-squared     =  0.1604
-------------+------------------------------           Adj R-squared =  0.1604
       Total |  87766293.6999999  87.7663814           Root MSE      =   8.584

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    6.961003   .0246269   282.66   0.000     6.912735    7.009271
           x |   -2.145149   .0089427  -239.88   0.000    -2.162676   -2.127621
       _cons |    5.420296   .0224751   241.17   0.000     5.376245    5.464346
------------------------------------------------------------------------------
```

STATA Output 6.68 (6.7.do)

```
. ivreg Y x (P=elig x), first
First-stage regressions

      Source |       SS       df       MS              Number of obs = 1000000
-------------+------------------------------           F(  2,999997) =59555.65
       Model |  14056.3422     2  7028.17109           Prob > F      =  0.0000
    Residual |  118009.796999997   .11801015           R-squared     =  0.1064
-------------+------------------------------           Adj R-squared =  0.1064
       Total |  132066.138999999   .13206627           Root MSE      =  .34353

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.027981   .0005541   -50.50   0.000     -.029067   -.0268951
        elig |    .1960948   .0011409   171.87   0.000     .1938587     .198331
       _cons |    .7223344   .0007839   921.49   0.000     .7207981    .7238708
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs = 1000000
-------------+------------------------------           F(  2,999997) =57080.49
       Model |  13810587.1     2  6905293.53           Prob > F      =  0.0000
    Residual |  73955706.5999997  73.9559284           R-squared     =  0.1574
-------------+------------------------------           Adj R-squared =  0.1574
       Total |  87766293.6999999  87.7663814           Root MSE      =  8.5998

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           P |    8.452753   .1456513    58.03   0.000     8.167282    8.738225
           x |   -1.991898    .017255  -115.44   0.000    -2.025717   -1.958079
       _cons |    4.162098   .1231483    33.80   0.000     3.920732    4.403465
------------------------------------------------------------------------------
Instrumented:  P
Instruments:   x elig
------------------------------------------------------------------------------
```

6.4 Some Closing Thoughts

In this chapter we have discussed instrumental variables, a very slick and potentially powerful method of program impact evaluation that allows us to use a variable (the instrument) to capture a random channel of program participation, which is otherwise not random at all. This has proven to be an extremely popular method of impact evaluation (and causal modelling more generally).

It is not one without its perils, however. Perhaps the most important disadvantage of this approach is that it requires that the evaluator model participation, at least in part. Crucially, one must assume that there is a variable (or, more specifically, at least one variable) that influences program participation but otherwise has no effect on the outcome of interest and is not related to the unobserved determinants of the outcome of interest. It can be very difficult to persuasively defend assumptions along this line, and they cannot always be tested.

Beyond this central issue, the performance of the estimator can be rather fragile. It can yield very misleading impact estimates with small samples, instruments that only weakly influence participation, etc. In the realm of limited dependent outcomes of interest, there can be a trade-off between more easily estimated models requiring restrictive distributional assumptions and more flexible and potentially robust approaches that can comparatively be a nightmare to estimate (and are not available in most standard commercial statistical analysis packages).

There is a potentially serious interpretational issue with the instrumental variables strategy. Even if all of the aforementioned requirements for successful instrumental variables estimation are met, it can yield estimates of program impact that, while consistent, are so only for the types of individuals who would allow their program participation decision to be shaped by the instrument. In other words, it does not necessarily yield an estimate of the sort of true population average program impact that might be of interest. While it is true that some important things can be learned about the population whose average treatment effect is being recovered (the so-called "compliers") this still does not necessarily tell us anything about how their impact relates to that for, for instance, the average person in society.

Chapter 7

Final Thoughts

We have now reached the end of our journey through the subject of program impact evaluation methods. This has hopefully left the reader with a richer, more nuanced understanding of alternative program impact evaluation methods.

The fundamental identification problem of program impact evaluation is that we cannot observe the value of an outcome of interest for any given individual both when they participate (Y^1, in the potential outcome terms we have utilized throughout these chapters) and when they do not participate (Y^0) in a program. We can only observe one of these for a particular individual. This makes it impossible to estimate program impact at the individual level.

A natural response is to shift focus from program impact at the individual level to average impact at the population level. In principle, estimation of average impact would entail, essentially, comparison of outcomes across participants with those across non-participants. Specific methods might involve taking the difference in average outcomes across representative samples of participants and non-participants or regression of the outcome of interest on an indicator of program participation.

However, such measures of differences in outcomes between participants and non-participants might not reflect the impact of program participation. Under normal real world circumstances, individuals decide whether to participate in a program. When they do so, that decision is typically influenced by their background characteristics, as well as the environmental circumstances and constraints confronting them. This means that these sorts of factors help to sort individuals into participant and non-participant sub-populations. We would therefore expect that the distribution of these factors would differ between these two groups. Even when participation is a program-level decision (as in when program officials decide which communities will and will not receive the treatment), the characteristics of the unit of participation typically inform that decision.

When such non-random assignment to program participation status occurs, simple methods of impact evaluation (such as the aforementioned comparison of average outcomes between participants and non-participants or regression of an outcome of interest on an indicator of program participation) will not yield unbiased estimates of program impact. The reason is simple: these methods basically simply compare outcomes between participants and non-participants. However, when participants and non-participants differ on average in their characteristics, it becomes impossible to say whether any differences in outcomes between the two groups is due to program participation or differences in their other characteristics that might also generate differences in outcomes.

The solution to this potential challenge can be looked at in two different but related ways. First, one could view the challenge of program impact evaluation as being one of finding samples of

program participants and non-participants that are sufficiently comparable in their characteristics that one could argue that any differences in average outcomes between the two must be driven by the experience of program participation.

Second, one could view the challenge of program impact evaluation as one of finding a random channel of variation in program participation out of overall variation, which might not be random. Variation driven by this random channel should then effectively identify comparable (in the sense of reflecting similar background characteristics) participants and non-participants or differences in outcomes between the two that reflect only variation in the experience of program participation. To one degree or another, every impact evaluation methodology considered in this manual implicitly uses one or both of these angles as a departure point for developing an estimation approach that yields unbiased and/or consistent estimates of program impact.

7.1 A Brief Recap

We have covered four major traditions of program impact evaluation methods designed to recover estimates of impact that reflect the causal impact of program participation on an outcome of interest. It is perhaps useful to recap briefly the methods, summarizing and hence contrasting them in one place.

7.1.1 Randomized Control Trials

Randomized control trials (i.e. RCTs) involve deliberate randomization of program participation. This approach has great appeal: it relies on a simple mechanism (randomization) for generating estimates of program impact. However, there are limitations. Randomization can be tricky in practice and cannot always be tested. To the extent that the latter is true, the distinction between RCTs and quasi-experimental methods becomes blurred: both essentially rely on assumptions in order to interpret program impact estimates as the causal impact of program participation on an outcome. Moreover, many interesting programs cannot be evaluated (or, more broadly, many interesting causal questions cannot be addressed) by them because randomization is not always feasible and even when it is some simple potential parameters of interest (such as median program impact) cannot be identified. More broadly, there will always be lingering doubt that humans will ever passively accept their experimental assignment.

7.1.2 Selection on Observables Models

Selection on observable methods for estimating program impact essentially assume that we can observe all of the factors (background characteristics, constraints, environmental circumstances, etc.) that influence both the program participation decision and the outcome of interest. The two major branches of this approach are regression and matching, though as we have seen they are not really altogether distinct (e.g. regression can be seen as a kind of matching estimator). In principle, both of these estimation approaches are simple, but in practice there appears to be considerable methodological disagreement about the specifics of implementation (though it is unclear how often these differences are important in practice). In general, the identifying assumption of these models (that we can observe everything that influences both participation and the outcome) is a very strong one. Nonetheless, this is a very popular approach to impact evaluation.

7.1.3 Within Models

Within methods assume that any factors that influence program participation and the outcome of interest, and which we cannot observe, are somehow fixed. The classic example involves longitudinal data and assumes that such factors are fixed over time, but the basic approach has sometimes been applied in a strictly cross-sectional context (e.g. assuming that the unobserved confounding factor is constant across communities). This assumption, that the unobserved confounder is fixed, is the key identifying assumption of these models. This is indeed a potentially powerful identification mechanism. For instance, with longitudinal data, every individual in some sense becomes a control for themselves. However, the assumption that all unobserved confounding factors are so fixed is a strong one. Moreover, this estimation approach can actually worsen some types of bias (i.e. measurement error bias) and cannot be applied to many modelling circumstances of potential interest (e.g. the limited dependent variable options are, well, limited).

7.1.4 Instrumental Variables

Instrumental variables methods rely on instruments which are correlated with random channels of program participation, out of overall variation in participation (which might not be random). A valid instrument needs to be correlated with program participation, have no independent role in determining the outcome of interest and be uncorrelated with any unobserved determinants of program participation. The latter two assumptions can generally be tested only in the case where there are multiple instruments per endogenous variable (i.e. the over-identified case). Even when these assumptions regarding the instrument(s) are met interpretation can be complicated by the possibility of local average treatment effects. Specifically, in the local average treatment effects case instrumental variables generates a consistent estimate of program impact only for the subpopulation whose program participation is responsive to variation in the instrument. Thus, in this instance instrumental variables does consistently identify program impact for some subpopulation about which some can be learned (e.g. their proportion of the population and basic characteristics), but it remains unclear how a local average treatment effect might relate to, say, the average program impact across the population. Compared with within models, there are far more options for instrumental variables estimation with limited dependent variables, though one needs to be mindful of the additional assumptions many such models entail. One important application of instrumental variables is to the "fuzzy" regression discontinuity design, whereby some eligibility criteria generates a discontinuity in the probability of participation and eligibility status can serve as an instrument, though estimates generated by this application can have a local average interpretation as well.

7.2 The Future

Although we have focused on these four central traditions for estimation of program impact, this taxonomy is not necessarily set in stone. For one thing, hybrid approaches are quite popular (for instance, fixed effects instrumental variable approaches whereby some instrument is available for the change in program participation over time). However, there are several alternative approaches that may become increasingly prominent in years to come.

Perhaps chief among these are "Deeply Structural Models". These are often applied to longitudinal data with a focus on dynamic decision making, in which case the term "Dynamic Programming" is often applied. Whatever the case, the key feature of this approach is that it involves the estimation of a model of how the individual makes decisions that shape the outcome. Once the model parameters are estimated, one could simulate behavioral choices under all sorts of new constraints

and possibilities, including programs or policies that did not exist for the estimating sample. A good example is Gilleskie (1998), who used a fitted dynamic programming model to simulate the expected behavioral and health implications of various potential (but as of her writing untried) national health care or insurance reform schemes.

Because they require potentially extensive assumptions regarding the behavioral process behind individual decision making (i.e. the nature of welfare/utility functions, budget constraint functions, health production functions, expectations formation processes and dynamic constraint and production functions in the case of dynamic models, etc.) some have expressed skepticism regarding these models. Their basic criticism is that models laden with so many assumptions are unlikely to provide reliable inferences regarding program impacts. However, some attempts at validating these models (e.g. Todd and Wolpin 2006) have quieted these criticisms to some degree.

Perhaps the biggest historical obstacles to the implementation of these models have been their computational intensity and, particularly for dynamic programming models, a paucity of the sort of longitudinal data most ideal as an estimation platform for these models. Both constraints are easing as computers grow ever faster and the data infrastructure in higher and lower income societies grows ever richer. As a result, these may become more popular as program impact evaluation tools in years to come.

Finally, the emergence of non-traditional information sources such as social media may create demand and scope for new impact evaluation methodologies. For instance, an innovative recent paper (Kearney and Levine 2014) used information from "old" media sources (e.g. Nielsen ratings) but also "new" social media (e.g. Twitter) to assess the impact of MTV's[1] program *16 and Pregnant* on teen childbearing. Evaluations such as these highlight the emerging possibilities offered by the information generated by new media, including social media such as Twitter and Facebook (and their international counterparts, such as Weibo). As the availability of these information sources grows, there may well be need for methodological innovation.

7.3 Further Reading

This manual has provided an overview of methods for program impact evaluation that strove not to presume much prior knowledge by the reader. As a result, many relationships and results the background derivation of which is simply skipped in other treatments of this topic were explicitly derived in this manual. This could be a grueling process (e.g. the explicit derivation of many nonlinear instrumental variables models). Hopefully, the end result is that the reader has a better grasp of the topic because they understand the foundation of so many important relationships and results on which the discussion of impact evaluation methods turns.

Many readers might want to continue their journey learning about the methods covered in this manual. Moreover, having completed this manual the reader is better positioned to participate in a discussion of these methods that is a step up the complexity ladder from that in this manual (which sought to convey a solid foundation in the essential properties and basic behavior of these impact evaluation methods). For instance, the estimation of the variance of estimates and relative efficiency are not topics on which we have long dwelt in this manual.

Clearly, the references at the end of the manual offer a starting point. However, among them we would particularly recommend Angrist and Pischke (2009) as a next step. It too provides a review of program impact evaluation methods (though one could argue that they more explicitly broaden the focus to causal modelling in general). The topics covered in this manual overlap considerably with

[1]MTV, or Music Television, is an American cable network originally based on music videos but now more focused on programming, especially so-called "reality" programming, aimed at teens.

those in Angrist and Pischke (2009). The difference is that their discussion is at the next level (for instance, it presumes understanding of many things explicitly explained and derived in this manual). Despite this, it is still quite approachable by the standards of this literature. Moreover, Angrist and Pischke (2009) is more editorial in nature, offering more normative (but in the case of both authors very informed!) judgments about the tradeoffs between different estimation alternatives. Developing some sense of opinion about these methods is important, and Angrist and Pischke do a very good job introducing the reader to their world view without in any sense allowing it to distort the discussion.

7.4 Impact Evaluation Meets Philosophy

At this point, some degree of cynicism on the reader's part might be justified. A major implicit message of this manual would seem to be that there is no "Gold Standard" method of impact evaluation. All of the methods discussed involve assumptions, not all of which are testable, that allow one to interpret the estimates generated by them as program impact (or, more broadly, as reflecting a causal relationship). Some present inherent limitations in terms of the parameters that can be estimated.

In the absence of a Gold Standard, it can be difficult to know what method is preferred. This is why careful impact evaluation work is so important. One must have a good sense of the institutional and environmental framework in which a program operates, as well as a good understanding of the design and procedures of the program itself and the types of populations motivated to participate and why they would be. This allows the evaluator to have an informed sense of what assumptions are (probably) reasonable, and hence which impact evaluation methods might be preferred and how much weight to assign to the estimates generated by them. It is true that even then assumptions (or, at the least, certainly *untestable* assumptions) are glorified opinions, but they will at least be informed opinions.

In closing, as a philosophical issue, it is important to remember that program impact evaluation is not a creature like theoretical mathematics, with its appeal to the clear, beautiful (and occasionally merciless) prospect of uncovering an absolute truth. Program impact evaluation is a statistical exercise. And one can never prove anything statistically: all that can ever be generated is evidence. Even the impact estimates generated by a "perfect" RCT still do not represent proof of anything. Rather, those estimates constitute evidence. And evidence does not necessarily reveal truth: there are many instances in the human story where an eventually revealed truth has proven surprising in light of the available evidence to that point. Indeed, mere evidence often most readily lends itself simply to the formation of opinion. It is therefore altogether natural that there should also be scope for opinion about the most appropriate or credible manner in which to build evidence. Taken against this fundamental backdrop, the complexities of the tradeoffs between impact evaluation methods, as well as the pattern of ebb and flow in the popularity of alternative approaches, seem more inevitable and less intrinsically intellectually concerning.

References

Abadie, A. (2003) "Semiparametric Instrumental Variable Estimation Of Treatment Response Models." *Journal of Econometrics* 113(2): 231-263.

Abadie, A., D. Drukker, J. Herr, G. Imbens. (2004) "Implementing Matching Estimators For Average Treatment Effects In STATA." *The STATA Journal* 4(3): 290-311.

Angeles, G., D. Guilkey, T. Mroz. (1998). "Purposive Program Placement And The Estimation Of Family Planning Program Effects In Tanzania." *Journal of the American Statistical Association* 93(443): 884-899.

Angrist, J. (1990) "Lifetime Earnings And The Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records." *American Economic Review* 80(3): 313-336.

Angrist, J. (2004) "Treatment Effect Heterogeneity In Theory And Practice." *The Economic Journal* 114(494): C52-C83.

Angrist, J., E. Bettinger, E. Bloom, E. King, M. Kremer. (2002) "Vouchers For Private Schooling In Colombia: Evidence From A Randomized Natural Experiment." *American Economic Review* 92(5): 1535-1558.

Angrist, J., J. Hahn. (2004) "When To Control For Covariates? Panel Asymptotics For Estimates Of Treatment Effects." *The Review of Economics and Statistics* 86(1): 58-72.

Angrist, J., G. Imbens. (1991) "Sources Of Identifying Information In Evaluation Models." NBER Technical Working Paper #117.

Angrist, J., G. Imbens, D. Rubin. (1996) "Identification Of Causal Effects Using Instrumental Variables." with comments and rejoinder, *Journal of the American Statistical Association* 91(434): 444-455.

Angrist, J., A. Krueger. (1991) "Does Compulsory Schooling Attendance Affect Schooling And Earnings?" *The Quarterly Journal of Economics* 106(4): 979-1014.

Angrist, J., A. Krueger. (1999) "Empirical Strategies In Labor Economics." in *Handbook Of Labor Economics Vol. 3A*, O. Ashenfelter, D. Card (eds.) North Holland: Amsterdam.

Angrist, J., A. Krueger. (2001) " Instrumental Variables And The Search For Identification." *Journal of Economic Perspectives* 15(4): 69-85.

Angrist, J., V. Lavy. (1999) "Using Maimonides' Rule To Estimate The Effect Of Class Size On Student Achievement." *The Quarterly Journal of Economics* 114(2): 533-575.

Angrist, J., J. Pischke. (2009) *Mostly Harmless Econometrics* Princeton University Press, Princeton.

Ashenfelter, O., D. Card (1985) "Using The Longitudinal Structure Of Earnings To Estimate The Effect Of Training Programs." *The Review of Economics and Statistics* 67(4): 648-660.

Ashraf, N., O. Bandiera, K. Jack. (2013) "No Margin, No Mission? A Field Experiment On Incentives For Public Service Delivery." Available at http://www.povertyactionlab.org/publication/no-margin-no-mission-field-experiment-incentives-pro-social-tasks

Austin P. (2008) "Assessing Balance In Measured Baseline Covariates When Using Many-To-One Matching On The Propensity-Score." *Pharmacoepidemiology and Drug Safety* 17(12): 1218-1225.

Austin, P. (2009) "The Relative Ability Of Different Propensity Score Methods To Balance Measured Covariates Between Treated And Untreated Subjects In Observational Studies." *Medical Decision Making* 29(6): 661-677.

Austin, P. (2011) "An Introduction To Propensity Score Methods For Reducing The Effects Of Confounding In Observational Studies." *Multivariate Behavioral Research* 46(3): 399-424.

Baicker, K., S. Taubman, H. Allen, M. Bernstein, J. Gruber, J. Newhouse, E. Schneider, B. Wright, A. Zaslavsky, A. Finkelstein. (2013) "The Oregon Experiment: Effects Of Medicaid On Clinical Outcomes." *The New England Journal of Medicine* 368(18): 1713-1722.

Banerjee, A., S. Cole, E. Duflo, L. Linden. (2007) "Remedying Education: Evidence From Two Randomized Experiments In India." *The Quarterly Journal of Economics* 122(3): 1235-1264.

Battistin, E., E. Rettore. (2003) "Another Look At The Regression Discontinuity Design." Working Paper.

Becker, S., A. Ichino. (2002) "Estimation Of Average Treatment Effects Based On Propensity Scores." *The STATA Journal* 2(4): 358-377.

Behrman, J., J. Hoddinott. (2000) "An Evaluation Of The Impact Of PROGRESA On Pre-School Child Height." International Food Policy Research Institute.

Behrman, J., P. Segupta, P. Todd. (2000) "The Impact Of PROGRESA On Achievement Test Scores In The First Year." Final report, International Food Policy Research Institute.

Behrman, J., P. Segupta, P. Todd. (2001) "Progressing Through Progresa: An Impact Assessment Of A School Subsidy Experiment." Penn Institute for Economic Research Working Paper 01-033.

Bertand, J., G. Escudero. (2002) *Compendium of Indicators For Evaluating Reproductive Health Programs* MEASURE Evaluation Manual Series, No. 6.

Bertrand, M., S. Mullainathan. (2004) "Are Emily And Greg More Employable Than Lakisha And Jamal? A Field Experiment On Labor Market Discrimination." *The American Economic Review* 94(4): 991-1013.

Bertrand, M., Duflo, E., S. Mullainathan. (2004) "How Much Should We Trust Difference-in-Differences Estimates?" *The Quarterly Journal of Economics* 119(1): 249-275.

Bingham, P., N. Verlander, M. Cheal. (2004) "John Snow, William Farr And The 1849 Outbreak Of Cholera That Affected London: A Reworking Of The Data Highlights The Importance Of The Water Supply." *Public Health* 118(6): 387-394.

Blundell, R., M. Costas Dias. (2002) "Alternative Approaches To Evaluation In Empirical Microeconomics." *Portuguese Economic Journal* 1(2): 91-115.

Bollen, K., D. Guilkey, T. Mroz. (1995) "Binary Outcomes And Endogenous Explanatory Variables: Tests And Solutions With An Application To The Demand For Contraceptive Use In Tunisia." *Demography* 32(1): 111-131.

Bonnal, L., D. Fougere, A. Serandon. (1997) "Evaluating The Impact Of French Employment Policies On Individual Labour Market Histories." *The Review of Economic Studies* 64(4): 683-713.

Bound, J., D. Jaeger, R. Baker. (1995) "Problems With Instrumental Variables Estimation When The Correlation Between The Instruments And The Endogenous Regressors Is Weak." *Journal of the American Statistical Association* 90(430): 443-450.

Buckley, J., Y. Shang. (2003) "Estimating Policy and Program Effects with Observational Data: The "Difference-in-Differences"Estimator." *Practical Assessment, Research & Evaluation* 8(24). (Electronic journal.)

Buddelmeyer, H., E. Skoufias. (2003) "An Evaluation Of The Performance Of Regression Discontinuity Design On Progresa."IZA Discussion Paper #827.

Cain, G., D. Wissoker. (1990) "A Reanalysis Of Marital Stability In The Seattle-Denver Income Maintenance Experiment." *American Journal of Sociology* 95(5): 1235-1260.

Cameron, A., P. Trivedi. (1998) *Regression Analysis Of Count Data* Cambridge University Press: Cambridge.

Campbell, D., J. Stanley. (1966) *Experimental And Quasi-Experimental Designs For Research* Houghton Mifflin: Boston.

Card, D. (1999) "The Causal Effect Of Education On Earnings."in *Handbook of Labor Economics, Vol. 3A*, O. Ashenfelter, D. Card (eds.) North Holland: Amsterdam.

Cullen, J., B. Jacob, S. Levitt. (2006) "The Effect Of School Choice On Participants: Evidence From Randomized Lotteries." *Econometrica* 74(5): 1191-1230.

Deaton, A. (1997) *The Analysis of Household Surveys: A Microeconometric Approach To Development Policy* World Bank/The Johns Hopkins University Press.

Deaton, A. (2010) "Instruments, Randomization, And Learning About Development." *The Journal of Economic Literature* 48(2): 424-455.

Dow, W. (2001) "A Guide To Specification Tests For Two-Stage Least Squares." Working Paper, Department of Health Policy and Planning, UNC-Chapel Hill.

Dow, W., E. Norton. (2003) "Choosing Between And Interpreting The Heckit And Two-Part Models For Corner Solutions." *Health Services & Outcomes Research Methodology* 4(1): 5-18.

Duan, N., W. Manning, C. Morris, J. Newhouse. (1983) "A Comparison Of Alternative Models For The Demand For Medical Care." *Journal of Business and Economic Statistics* 1(2): 115-126.

Duan, N., W. Manning, C. Morris, J. Newhouse. (1984) "Choosing Between The Sample Selection And Multi-part Model." *Journal of Business and Economic Statistics* 2(3): 283-289.

Duan, N., W. Manning, C. Morris, J. Newhouse. (1985) "Comments On Selectivity Bias." in *Advances in Health Economics and Health Services Research, Vol. 6*, R. Schleffer, L. Rossiter (eds.) JAI Press: Greenwhich, CT.

Duflo, E., P. Dupas, M. Kremer. (2012) "Education, HIV And Early Fertility: Experimental Evidence From Kenya." Available at http://www.povertyactionlab.org/publication/education-hiv-and-early-fertility-experimental-evidence-kenya

Fisher, R. (1935) *The Design of Experiments* Oliver and Boyd: London.

Gertler, P. (2000) "Final Report: The Impact Of PROGRESA On Health." International Food Policy Research Institute.

Gertler, P., S. Boyce. (2001) "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico." Working paper, UC Berkeley.

Gertler, P, J. Molyneaux. (1994). "How Economic Development And Family Planning Programs Combined To Reduce Indonesian Fertility." *Demography* 31(1): 33-63.

Gilleskie, D. (1998) "A Dynamic Stochastic Model Of Medical Care Use And Work Absence." *Econometrica* 66(1): 1-45.

Granger, C. (1969) "Investigating Causal Relations By Econometric Models And Cross-Spectral Methods." *Econometrica* 37(3): 424-438.

Greene, W. (2000) *Econometric Analysis (4th Edition)* Prentice Hall: Upper Saddle River, NJ.

Greene, W. (2004) "Fixed Effects And Bias Due To The Incidental Parameters Problem In The Tobit Model." *Econometric Reviews* 23(2): 125-147.

Griffeths, W., R. Hill, G. Judge. (1993) *Learning and Practicing Econometrics* John Wiley & Sons: New York.

Guilkey, D., P. Lance. (2014) "Program Impact Estimation With Binary Outcome Variables: Monte Carlo Results For Alternative Estimators And Empirical Examples." in *Festschrift in Honor of Peter Schmidt*, R. Sickles, W. Horrace (eds.) Springer-Verlag.

Guilkey, D., T. Mroz. (1992) "Discrete Factor Approximations For Use In Simultaneous Equation Models With Both Continuous And Discrete Endogenous Variables." Working Paper, University of North Carolina at Chapel Hill.

Haas, J., T. Brownie. (2001) "Iron Deficiency And Reduced Work Capacity: A Critical Review Of The Research To Determine A Causal Relationship." *The Journal of Nutrition* 131(2): 676S-690S.

Hahn, J. (1998) "On The Role Of The Propensity Score In Efficient Semiparametric Estimation Of Average Treatment Effects." *Econometrica* 66(2): 315-331.

Hausman, J., D. Wise (eds). (1985) *Social Experimentation* The University of Chicago Press: Chicago.

Hay, J., R. Olsen. (1984) "Let Them Eat Cake: A Note On Comparing Alternative Models Of The Demand For Health Care." *The Journal of Business and Economic Statistics* 2(3): 279-282.

Heckman, J., B. Singer. (1984) "A Method For Minimizing The Impact Of Distributional Assumptions In Econometric Models For Duration Data.'" *Econometrica* 52(2): 271-320.

Heckman, J., B. Honore. (1990) "The Empirical Content Of The Roy Model." *Econometrica* 58(5): 1121-1149.

Heckman, J., J. Hotz. (1989) "Choosing Among Alternative Non-experimental Methods For Estimating The Impact Of Social Programs: The Case Of Manpower Training." *Journal of the American Statistical Association* 84(408): 862-880.

Heckman, J., H. Ichimura, J. Smith, P. Todd. (1996) "Sources Of Selection Bias In Evaluating Social Programs: An Interpretation Of Conventional Measures And Evidence On The Effectiveness Of Matching As A Program Evaluation Method." *Proceedings of the National Academy of Sciences* 93(23): 13416-13420.

Heckman, J., H. Ichimura, P. Todd. (1997a) "Matching As An Econometric Evaluation Estimator: Evidence From A Job Training Programme." *Review of Economic Studies* 64(4): 605-654.

Heckman, J., H. Ichimura, J. Smith, P. Todd. (1997b) "Characterizing Selection Bias Using Experimental Data." Working Paper.

Heckman, J., R. LaLonde, J. Smith. (1999) "The Economics And Econometrics Of Active Labor Market Programs."in *Handbook of Labor Economics Volume III*, O. Ashenfelter, D. Card (eds.) Elsevier: Amsterdam, Lausanne, New York, Oxford, Shannon, Singapore, Tokyo.

Heckman, J., L. Lochner, C. Taber. (1998) "General Equilibrium Treatment Effects: A Study Of Tuition Policy."NBER Working paper #6426.

Heckman, J., J. Smith. (1993) "Assessing The Case For Randomized Evaluation Of Social Programs."in *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*, K. Jensen, P. Madsen (eds) Ministry of Labour: Copenhagen.

Heckman, J., J. Smith. (1995) "Assessing The Case For Social Experiments." *Journal of Economic Perspectives* 9(2): 85-110.

Hill, A. (1965) "The Environment And Disease: Association Or Causation?" *Proceedings of the Royal Society of Medicine* 58(5): 295-300.

Hirano, K., G. Imbens, G. Ridder. (2003) "Efficient Estimation Of Average Treatment Effects Using The Estimated Propensity Score." *Econometrica* 71(4): 1161-1189.

Hoddinott, J., E. Skoufias, R. Washburn. (2000) "The Impact Of PROGRESA On Consumption: A Final Report."International Food Policy Research Institute.

Hsiao, Cheng (1986) *Analysis Of Panel Data. Econometrics Society Monographs* Cambridge University Press: New York.

Hutchinson, P., P. Lance, D. Guilkey, M. Shahjahan, S. Haque. (2006) "Analyzing The Cost-Effectiveness Of A National Health Communication Programs: The Smiling Sun Program In Bangladesh." *Journal of Health Communication* 11(Supplement 2):91-121.

Imbens, G. (2004) "Nonparametric Estimation Of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1): 4-29.

Imbens, G., J. Angrist. (1994) "Identification And Estimation Of Local Average Treatment Effects." *Econometrica* 62(2): 467-75.

Imbens, G., J. Angrist. (1999) "Comment On: "Instrumental Variables: A Study Of Implicit Behavioral Assumptions Used In Making Program Evaluations", by J. Heckman." *Journal of Human Resources* 34(4): 823-827.

Imbens, G., T. Lemieux. (2008) "Regression Discontinuity Designs: A Guide To Practice." *Journal of Econometrics* 142(2): 615-35.

Imbens, G., R. Rosenbaum. (forthcoming) "Randomization Inference With An Instrumental Variable." *Journal of the Royal Statistical Society, Series B.*

Kearney, M., P. Levine. (2014) "Media Influences On Social Outcomes: The Impact Of MTV's 16 And Pregnant On Teen Childbearing." NBER Working Paper 19795.

LaLonde, R. (1986) "Evaluating The Econometric Evaluations Of Training Programs With Experimental Data." *American Economic Review* 76(4): 604-620.

Lalonde, R. (1995) "The Promise Of Public Sector-Sponsored Training Programs." *The Journal of Economic Perspectives* 9(2): 149-168.

Leuven E., B. Sianesi. (2003) "`psmatch2`: Stata Module To Perform Full Mahalanobis And Propensity Score Matching, Common Support Graphing, And Covariate Imbalance Testing." http://idea s.repec.org/c/boc/bocode/s432001.html

Levitt, S., J. List. (2011) "Was There Really A Hawthorne Effect At The Hawthorne Plant? An Analysis Of The Original Illumination Experiments." *American Economic Journal: Applied Economics* 3(1): 224-238.

Lewbel, A., Y. Dong, T. Yang. (2012) "Comparing Features Of Convenient Estimators For Binary Choice Models With Endogenous Regressors." *Canadian Journal of Economics* 45(3): 809-829.

Lucas, R. (1976) "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1(1): 19-46.

Maddala, G. (1983) *Limited-Dependent And Qualitative Variables in Econometrics. Econometric Society Monographs* Cambridge University Press: New York.

Maddala, G. (1985) "A Survey Of The Literature On Selectivity Bias As It Pertains To Health Care Markets." in *Advances in Health Economics and Health Services Research, Vol. 6*, R. Schleffer, L. Rossiter (eds.) JAI Press: Greenwhich, CT, 3-17.

Manski, C. (1995) *Identification Problems In The Social Sciences* Harvard University Press: Cambridge, London.

Manski, C. (1996) "Learning About Treatment Effects From Experiments With Random Assignment To Treatments." *Journal of Human Resources* 31(4): 709-733.

Manski, C., I. Garfinkel (eds). (1992) *Evaluating Welfare And Training Programs* Harvard University Press: Cambridge.

McFadden, D. (2001) "Statistical Tools For Economists." Available at http://elsa.berkeley.edu/users/mcfadden/e240a_sp01/.

Meyer, B. (1995) "Natural And Quasi-Experiments In Economics." *Journal of Business & Economic Statistics* 13(2): 151-161.

Miguel, E., M. Kremer. (2004) "Worms: Identifying Impacts On Education And Health In The Presence Of Treatment Externalities." *Econometrica* 72(1):159-217. (Also see NBER Working Paper #8481.)

Miller, G. (2010) "Contraception As Development? New Evidence From Family Planning In Columbia." *The Economic Journal* 120(545): 709-736.

Millimet, D. (2001) "What Is The Difference Between 'Endogeneity' And 'Sample Selection Bias'."Stata Resources & Support FAQ. (Available at http://www.stata.com/support/faqs/stat/bias.html)

Mobarak, A., M. Rosenzweig. (2013) "Informal Risk Sharing, Index Insurance And Risk Taking In Developing Countries. " *American Economic Review* 103(3): 375-380.

Moffit, R. (2004) "The Role Of Randomized Field Trials In Social Science Research." *American Behavioral Scientist* 47(5): 506-540.

Mroz, T. (1999) "Discrete Factor Approximations In Simultaneous Equation Models: Estimating The Impact Of A Dummy Endogenous Variable On A Continuous Outcome." *Journal of Econometrics* 92(2): 233-274.

Newey, W. (1987) "Efficient Estimation Of Limited Dependent Variable Models With Endogenous Explanatory Variables." *Journal of Econometrics* 36(3): 231-250.

Newhouse, J. (1993) *Free for All? Lessons From The RAND Insurance Experiment* Harvard University Press: Camridge, London.

Neyman, J., E. Scott. (1948) "Consistent Estimates Based On Partially Consistent Observations." *Econometrica* 16(1): 1-32.

Orr, L. (1998) *Social Experiments: Evaluating Public Programs With Experimental Methods* Sage Publications, Thousand Oaks.

Orr, L., H. Bloom, S. Bell, W. Lin, G. Cave, F. Doolittle. (1995) *The National JTPA Study: Impacts, Benefits, And Costs Of Title II-A* Abt Associates: Bethesda.

Parker, S., E. Skoufias. (2000) "Final Report: The Impact Of PORGRESA On Work, Leisure And Time Allocation."International Food Policy Research Institute.

Pianto, D., S. Soares. (2004) " Use Of Survey Design For The Evaluation Of Social Programs: The PNAD And PETI." Area ANPEC: 6 Economia Social.

Pitt, M., M. Rosenzweig, D. Gibbons. (1993) "The Determinants and Consequences of the Placement of Government Programs in Indonesia." *World Bank Economic Review* 7(3): 319-348.

Rivers, D., Q. Vuong. (1988) "Limited Information Estimators And Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics* 39(3): 347-366.

Rodgers, J., W. Nicewander, L. Toothaker. (1984) "Linearly Independent, Orthogonal And Uncorrelated Variables." *The American Statistician* 38(2): 133-134.

Rosenbaum, P., D. Rubin. (1984) "Reducing Bias In Observational Studies Using Subclassification On The Propensity Score." *Journal of the American Statistical Association* 79(387): 516-524.

Rosenzweig, M., K. Wolpin. (1986) "Evaluating The Effects Of Optimally Distributed Programs." *American Economic Review* 76(3): 470-482.

Roy, A. (1951) "Some Thoughts On The Distribution Of Earnings." *Oxford Economic Papers (New Series).* 3(2): 135-146.

Schultz, T. (2000a) "School Subsidies For The Poor: Evaluating A Mexican Strategy For Reducing Poverty."International Food Policy Research Institute.

Schultz, T. (2000b) "Final Report: The Impact Of PROGRESA On School Enrollments." International Food Policy Research Institute.

Schultz, T. (2000c) "Impact of PROGRESA On School Attendance Rates In The Sampled Population."International Food Policy Research Institute.

Schultz, T. (2001) "School Subsidies For The Poor: Evaluating The Mexican PROGRESA Poverty Program."Yale Economic Growth Center Discussion Paper No. 834.

Smith, J., P. Todd. (2001) "Reconciling Conflicting Evidence On The Performance Of Propensity Score Matching Methods." *American Economic Review* 91(2): 112-119.

Snow, J. (1855) *On The Mode Of Communication Of Cholera* London: John Churchill, New Burlington Street, England.

Staiger, D., J. Stock. (1997) "Instrumental Variables Regression With Weak Instruments." *Econometrica* 65(3): 557-586.

Stock, J., M. Yogo. (2005) "Testing For Weak Instruments In Linear IV Regression.", in *Identification And Inference For Econometric Models: Essays In Honor Of Thomas Rothenberg* D. Andrews and J. Stock (eds) Cambridge University Press.

Teruel, G., B. Davis. (2000) "Final Report: An Evaluation Of The Impact Of PROGRESA Cash Payments On Private Inter-Household Transfers."International Food Policy Research Institute.

Terza, J., A. Basu, P. Rathouz. (2008) "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity In Health Econometric Modelling." *Journal of Health Economics* 27(3): 531-543.

Thomas, D., E. Frankenberg, J. Friedman, J. Habicht, M. Hakimi, Jaswadi, N. Jones, C. McKelvey, G. Pelto, B. Sikoki, T. Seeman, J. Smith, C. Sumantri, W. Suriastini, S. Wilopo. (2003) "Iron Deficiency And The Well-Being Of Older Adults: Early Results From A Randomized Nutrition Intervention."Working Paper. UCLA.

Todd, P., K. Wolpin. (2006) "Assessing The Impact Of A School Subsidy Program In Mexico: Using A Social Experiment To Validate A Dynamic Behavioral Model Of Child Schooling And

Fertility." *American Economic Review* 96(5): 1384-1417.

Wooldridge, J. (2001) *Econometric Analysis of Cross Section and Panel Data* MIT Press.

MEASURE Evaluation
Carolina Population Center
The University of North Carolina at Chapel Hill, CB 3446
Chapel Hill, NC 27516 USA
www.cpc.unc.edu/measure

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation