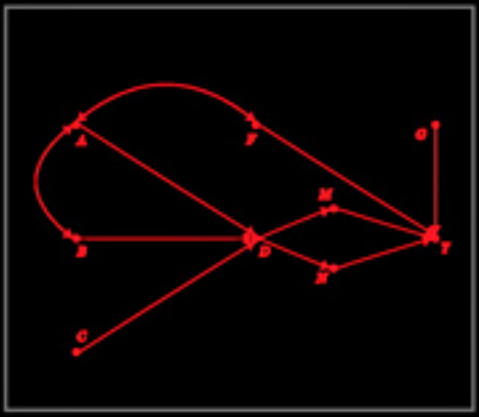
ANALYTICAL METHODS FOR SOCIAL RESEARCH



Counterfactuals and Causal Inference

Methods and Principles for Social Research

STEPHEN L. MORGAN
CHRISTOPHER WINSHIP

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521856157

This page intentionally left blank

COUNTERFACTUALS AND CAUSAL INFERENCE

Did mandatory busing programs in the 1970s increase the school achievement of disadvantaged minority youth? Does obtaining a college degree increase an individual's labor market earnings? Did the use of a butterfly ballot in some Florida counties in the 2000 presidential election cost Al Gore votes? Simple cause-and-effect questions such as these are the motivation for much empirical work in the social sciences. In this book, the counterfactual model of causality for observational data analysis is presented, and methods for causal effect estimation are demonstrated using examples from sociology, political science, and economics.

Stephen L. Morgan is Associate Professor of Sociology and the Director of the Center for the Study of Inequality at Cornell University. His previous publications include *On the Edge of Commitment: Educational Attainment and Race in the United States* (2005).

Christopher Winship is Diker–Tishman Professor of Sociology at Harvard University. For the past twelve years he has served as editor of *Sociological Methods and Research*. He has published widely in a variety of journals and edited volumes.

Analytical Methods for Social Research

Analytical Methods for Social Research presents texts on empirical and formal methods for the social sciences. Volumes in the series address both the theoretical underpinnings of analytical techniques as well as their application in social research. Some series volumes are broad in scope, cutting across a number of disciplines. Others focus mainly on methodological applications within specific fields such as political science, sociology, demography, and public health. The series serves a mix of students and researchers in the social sciences and statistics.

Other Titles in the Series:

Event History Modeling: A Guide for Social Scientists, by Janet M. Box-Steffensmeier and Bradford S. Jones

Ecological Inference: New Methodological Strategies, edited by Gary King, Ori Rosen, and Martin A. Tanner

Spatial Models of Parliamentary Voting, by Keith T. Poole

Essential Mathematics for Political and Social Research, by Jeff Gill

Political Game Theory: An Introduction, by Nolan McCarty and Adam Meirowitz

Data Analysis Using Regression and Multilevel/Hierarchical Models, by Andrew Gelman and Jennifer Hill

Counterfactuals and Causal Inference

Methods and Principles for Social Research

STEPHEN L. MORGAN

Cornell University

CHRISTOPHER WINSHIP

Harvard University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521856157

First published in print format 2007

ISBN-13 978-0-511-34926-3 eBook (EBL)
ISBN-10 0-511-34926-2 eBook (EBL)

ISBN-13 978-0-521-85615-7 hardback
ISBN-10 0-521-85615-9 hardback

To my wife, Sydney, my son, Vinny, and my daughter, Beatrix
– Steve Morgan

To my wife, Nancy, and my sons, David and Michael
– Chris Winship

Contents

Acknowledgments

Without yet knowing it, we began to write this book in 1997 when collaborating on a paper for the 1999 volume of the *Annual Review of Sociology*, titled "The Estimation of Causal Effects from Observational Data." We benefited from many helpful comments in the preparation of that manuscript, and we were pleased that many of our colleagues found it to be a useful introduction to a literature that we were, at the time, still working to understand ourselves. Since then, considerable progress in the potential outcomes and counterfactual modeling literature has been achieved, which led us into long discussions of the utility of writing a more comprehensive introduction. In the end, our motivation to learn even more of the literature was the decisive factor.

We thank Richard Berk, Felix Elwert, Jeremy Freese, Glenn Firebaugh, George Farkas, Andrew Gelman, Gary King, Trond Petersen, David Weakliem, and Kim Weeden for reading some or all of the penultimate draft of the book. We also thank the anonymous reviewer recruited by Cambridge University Press. The insightful comments of all of these readers helped tremendously. We also thank our students at Cornell and Harvard, from whom we have learned much in the course of learning and then presenting this material to them. Their comments and questions were more valuable than they are probably aware.

Finally, we thank Kelly Andronicos and Jenny Todd at Cornell University for assistance with the preparation of the manuscript, as well as Larry Wu and Ed Parsons at Cambridge University Press, Project Manager Peter Katsirubas at Aptara, Inc., and Victoria Danahy at In Other Words.

Part 1: Counterfactual Causality and Empirical Research in the Social Sciences

Chapter 1

Introduction

Did mandatory busing programs in the 1970s increase the school achievement of disadvantaged minority youth? If so, how much of a gain was achieved? Does obtaining a college degree increase an individual's labor market earnings? If so, is this particular effect large relative to the earnings gains that could be achieved only through on-the-job training? Did the use of a butterfly ballot in some Florida counties in the 2000 presidential election cost Al Gore votes? If so, was the number of miscast votes sufficiently large to have altered the election outcome?

At their core, these types of questions are simple cause-and-effect questions of the form, Does X cause Y? If X causes Y, how large is the effect of X on Y? Is the size of this effect large relative to the effects of other causes of Y?

Simple cause-and-effect questions are the motivation for much empirical work in the social sciences, even though definitive answers to cause-and-effect questions may not always be possible to formulate given the constraints that social scientists face in collecting data. Even so, there is reason for optimism about our current and future abilities to effectively address cause-and-effect questions. In the past three decades, a counterfactual model of causality has been developed, and a unified framework for the prosecution of causal questions is now available. With this book, we aim to convince more social scientists to apply this model to the core empirical questions of the social sciences.

In this introductory chapter, we provide a skeletal precis of the main features of the counterfactual model. We then offer a brief and selective history of causal analysis in quantitatively oriented observational social science. We develop some background on the examples that we will draw on throughout the book, concluding with an introduction to graphical causal models that also provides a roadmap to the remaining chapters.

1.1 The Counterfactual Model for Observational Data Analysis

With its origins in early work on experimental design by Neyman (1990 [1923], 1935), Fisher (1935), Cochran and Cox (1950), Kempthorne (1952), and Cox (1958), the counterfactual model for causal analysis of observational data was formalized in a series of papers by Donald Rubin (1974, 1977, 1978, 1980a, 1981, 1986, 1990). In the statistics tradition, the model is often referred to as the potential outcomes framework, with reference to potential yields from Neyman's work in agricultural statistics (see Gelman and Meng 2004; Rubin 2005). The counterfactual model also has roots in the economics literature (Roy 1951; Quandt 1972), with important subsequent work by James Heckman (see Heckman 1974, 1978, 1979, 1989, 1992, 2000), Charles Manski (1995, 2003), and others. Here, the model is also frequently referred to as the potential outcomes framework. The model is now dominant in both statistics and economics, and it is being used with increasing frequency in sociology, psychology, and political science.

A counterfactual account of causation also exists in philosophy, which began with the seminal 1973 article of David Lewis, titled "Causation."[1] It is related to the counterfactual model for observational data analysis that we will present in this book, but the philosophical version, as implied by the title of Lewis' original article, aims to be a general model of causality. As noted by the philosopher James Woodward in his 2003 book, *Making Things Happen: A Theory of Causal Explanation*, the counterfactual approach to causality championed by Lewis and his students has not been influenced to any substantial degree by the potential outcomes version of counterfactual modeling that we will present in this book. However, Woodward attempts to bring the potential outcomes literature into dialogue with philosophical models of causality, in part by augmenting the important recent work of the computer scientist Judea Pearl. We will also use Pearl's work extensively in our presentation, drawing on his 2000 book, *Causality: Models, Reasoning, and Inference*. We will discuss the broader philosophical literature in Chapters 8 and 10, as it does have some implications for social science practice and the pursuit of explanation more generally.

[1] In this tradition, causality is defined with reference to counterfactual dependence (or, as is sometimes written, the "ancestral" to counterfactual dependence). Accordingly, and at the risk of a great deal of oversimplification, the counterfactual account in philosophy maintains that it is proper to declare that, for events c and e, c causes e if (1) c and e both occur and (2) if c had not occurred and all else remained the same, then e would not have occurred. The primary challenge of the approach is to define the counterfactual scenario in which c does not occur (which Lewis did by imagining a limited "divergence miracle" that prevents c from occurring in a closest possible hypothetical world where all else is the same except that c does not occur). The approach differs substantially from the regularity-based theories of causality that dominated metaphysics through the 1960s, based on relations of entailment from covering law models. For a recent collection of essays in philosophy on counterfactuals and causation, see Collins, Hall, and Paul (2004).

The core of the counterfactual model for observational data analysis is simple. Suppose that each individual in a population of interest can be exposed to two alternative states of a cause. Each state is characterized by a distinct set of conditions, exposure to which potentially affects an outcome of interest, such as labor market earnings or scores on a standardized mathematics test. If the outcome is earnings, the population of interest could be adults between the ages of 30 and 50, and the two states could be whether or not an individual has obtained a college degree. Alternatively, if the outcome is a mathematics test score, the population of interest could be high school seniors, and the two states could be whether or not a student has taken a course in trigonometry. In the counterfactual tradition, these alternative causal states are referred to as alternative treatments. When only two treatments are considered, they are referred to as treatment and control. Throughout this book, we will conform to this convention.

The key assumption of the counterfactual framework is that each individual in the population of interest has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time. For example, for the causal effect of having a college degree rather than only a high school degree on subsequent earnings, adults who have completed high school degrees have theoretical what-if earnings under the state "have a college degree," and adults who have completed college degrees have theoretical what-if earnings under the state "have only a high school degree." These what-if potential outcomes are counterfactual.

Formalizing this conceptualization for a two-state treatment, the potential outcomes of each individual are defined as the true values of the outcome of interest that would result from exposure to the alternative causal states. The potential outcomes of each individual i are y_i^1 and y_i^0, where the superscript 1 signifies the treatment state and the superscript 0 signifies the control state. Because both y_i^1 and y_i^0 exist in theory for each individual, an individual-level causal effect can be defined as some contrast between y_i^1 and y_i^0, usually the simple difference $y_i^1 - y_i^0$. Because it is impossible to observe both y_i^1 and y_i^0 for any individual, causal effects cannot be observed or directly calculated at the individual level.[2]

By necessity, a researcher must analyze an observed outcome variable Y that takes on values y_i for each individual i that are equal to y_i^1 for those in the treatment state and y_i^0 for those in the control state. We usually refer to those in the treatment state as the treatment group and those in the control state as the control group.[3] Accordingly, y_i^0 is an unobservable counterfactual

[2]The only generally effective strategy for estimating individual-level causal effects is a crossover design, in which individuals are exposed to two alternative treatments in succession and with enough time elapsed in between exposures such that the effects of the cause have had time to dissipate (see Rothman and Greenland 1998). Obviously, such a design can be attempted only when a researcher has control over the allocation of the treatments and only when the treatment effects are sufficiently ephemeral. These conditions rarely exist for the causal questions that concern social scientists.

[3]We assume that, for observational data analysis, an underlying causal exposure mechanism exists in the population, and thus the distribution of individuals across the treatment and

outcome for each individual i in the treatment group, and y_i^1 is an unobservable counterfactual outcome for each individual i in the control group.

In the counterfactual modeling tradition, attention is focused on estimating various average causal effects, by analysis of the values y_i, for groups of individuals defined by specific characteristics. To do so effectively, the process by which individuals of different types are exposed to the cause of interest must be modeled. Doing so involves introducing defendable assumptions that allow for the estimation of the average unobservable counterfactual values for specific groups of individuals. If the assumptions are defendable, and a suitable method for constructing an average contrast from the data is chosen, then an average difference in the values of y_i can be given a causal interpretation.

1.2 Causal Analysis and Observational Social Science

The challenges of using observational data to justify causal claims are considerable. In this section, we present a selective history of the literature on these challenges, focusing on the varied history of the usage of experimental language in observational social science. We will also consider the growth of survey research and the shift toward outcome-equation-based motivations of causal analysis that led to the widespread usage of regression estimators. Many useful discussions of these developments exist, and our presentation here is not meant to be complete.[4] We review only the literature that is relevant for explaining the connections between the counterfactual model and other traditions of quantitatively oriented analysis that are of interest to us here. We return to these issues again in Chapters 8 and 10.

1.2.1 Experimental Language in Observational Social Science

Although the word experiment has a very broad definition, in the social sciences it is most closely associated with randomized experimental designs, such as the double-blind clinical trials that have revolutionized the biomedical sciences and the routine small-scale experiments that psychology professors perform on

control states exists independently of the observation and sampling process. Accordingly, the treatment and control groups exist in the population, even though we typically observe only samples of them in the observed data. We will not require that the labels "treatment group" and "control group" refer only to the observed treatment and control groups.

[4] For a more complete synthesis of the literature on causality in observational social science, see, for sociology, Berk (1988, 2004), Bollen (1989), Goldthorpe (2000), Lieberson (1985), Lieberson and Lynn (2002), Marini and Singer (1988), Singer and Marini (1987), Sobel (1995, 1996, 2000), and Smith (1990, 2003). For economics, see Angrist and Krueger (1999), Heckman (2000, 2005), Moffitt (2003), Pratt and Schlaifer (1984), and Rosenzweig and Wolpin (2000). For political science, see Brady and Collier (2004), King, Keohane, and Verba (1994), and Mahoney and Goertz (2006).

their own students.[5] Randomized experiments have their origins in the work of statistician Ronald A. Fisher during the 1920s, which then diffused throughout various research communities via his widely read 1935 book, *The Design of Experiments*.

Statisticians David Cox and Nancy Reid (2000) offer a definition of an experiment that focuses on the investigator's deliberate control and that allows for a clear juxtaposition with an observational study:

> The word *experiment* is used in a quite precise sense to mean an investigation where the system under study is under the control of the investigator. This means that the individuals or material investigated, the nature of the treatments or manipulations under study and the measurement procedures used are all selected, in their important features at least, by the investigator.
>
> By contrast in an observational study some of these features, and in particular the allocation of individuals to treatment groups, are outside the investigator's control. (Cox and Reid 2000:1)

We will maintain this basic distinction throughout this book. We will argue in this section that the counterfactual model of causality that we introduced in the last section is valuable precisely because it helps researchers to stipulate assumptions, evaluate alternative data analysis techniques, and think carefully about the process of causal exposure. Its success is a direct result of its language of potential outcomes, which permits the analyst to conceptualize observational studies as if they were experimental designs controlled by someone other than the researcher – quite often, the subjects of the research. In this section, we offer a brief discussion of other important attempts to use experimental language in observational social science and that succeeded to varying degrees.

Samuel A. Stouffer, the sociologist and pioneering public opinion survey analyst, argued that "the progress of social science depends on the development of limited theories – of considerable but still limited generality – from which prediction can be made to new concrete instances" (Stouffer 1962[1948]:5). Stouffer argued that, when testing alternative ideas, "it is essential that we always keep in mind the model of a controlled experiment, even if in practice we may have to deviate from an ideal model" (Stouffer 1950:356). He followed this practice over his career, from his 1930 dissertation that compared experimental with case study methods of investigating attitudes, to his leadership of the team that produced *The American Soldier* during World War II (see Stouffer 1949), and in his 1955 classic *Communism, Conformity, and Civil Liberties.*

On his death, and in celebration of a posthumous collection of his essays, Stouffer was praised for his career of survey research and attendant explanatory success. The demographer Philip Hauser noted that Stouffer "had a hand

[5] The *Oxford English Dictionary* provides the scientific definition of experiment: "An action or operation undertaken in order to discover something unknown, to test a hypothesis, or establish or illustrate some known truth" and also provides source references from as early as 1362.

in major developments in virtually every aspect of the sample survey – sampling procedures, problem definition, questionnaire design, field and operating procedures, and analytic methods" (Hauser 1962:333). Arnold Rose (1962:720) declared, "Probably no sociologist was so ingenious in manipulating data statistically to determine whether one hypothesis or another could be considered as verified." And Herbert Hyman portrayed his method of tabular analysis in charming detail:

> While the vitality with which he attacked a table had to be observed in action, the characteristic strategy he employed was so calculating that one can sense it from reading the many printed examples. . . . Multivariate analysis for him was almost a way of life. Starting with a simple cross-tabulation, the relationship observed was elaborated by the introduction of a third variable or test factor, leading to a clarification of the original relationship. . . . But there was a special flavor to the way Sam handled it. With him, the love of a table was undying. Three variables weren't enough. Four, five, six, even seven variables were introduced, until that simple thing of beauty, that original little table, became one of those monstrous creatures at the first sight of which a timid student would fall out of love with our profession forever. (Hyman 1962:324-5)

Stouffer's method was to conceive of the experiment that he wished he could have conducted and then to work backwards by stratifying a sample of the population of interest into subgroups until he felt comfortable that the remaining differences in the outcome could no longer be easily attributed to systematic differences within the subgroups. He never lost sight of the population of interest, and he appears to have always regarded his straightforward conclusions as the best among plausible answers. Thus, as he said, "Though we cannot always design neat experiments when we want to, we can at least keep the experimental model in front of our eyes and behave cautiously" (Stouffer 1950:359).

Not all attempts to incorporate experimental language into observational social science were as well received. Most notably in sociology, F. Stuart Chapin had earlier argued explicitly for an experimental orientation to nearly all of sociological research, but while turning the definition of an experiment in a direction that agitated others. For Chapin, a valid experiment did not require that the researcher obtain control over the treatment to be evaluated, only that observation of a causal process be conducted in controlled conditions (see Chapin 1932, 1947). He thus considered what he called "ex post facto experiments" to be the solution to the inferential problems of the social sciences, and he advocated matching designs to select subsets of seemingly equivalent individuals from those who were and were not exposed to the treatment of interest. In so doing, however, he proposed to ignore the incomparable, unmatched individuals, thereby losing sight of the population that Stouffer the survey analyst always kept in the foreground.

Chapin thereby ran afoul of emergent techniques of statistical inference, and he suffered attacks from his natural allies in quantitative analysis. The

statistician Oscar Kempthorne, whose 1952 book *The Design and Analysis of Experiments* would later became a classic, dismissed Chapin's work completely. In a review of Chapin's 1947 book, *Experimental Designs in Sociological Research*, Kempthorne wrote:

> The usage of the word "experimental design" is well established by now to mean a plan for performing a comparative experiment. This implies that various treatments are actually applied by the investigator and are not just treatments that happened to have been applied to particular units for some reason, known or unknown, before the "experiment" was planned. This condition rules out practically all of the experiments and experimental designs discussed by the author. (Kempthorne 1948:491)

Chapin's colleagues in sociology were often just as unforgiving. Nathan Keyfitz (1948:260), for example, chastised Chapin for ignoring the population of interest and accused him of using terms such as "experimental design" merely to "lend the support of their prestige."

In spite of the backlash against Chapin, in the end he has a recognizable legacy in observational data analysis. The matching techniques he advocated will be discussed later in Chapter 4. They have been reborn in the new literature, in part because the population of interest has been brought back to the foreground. But there is an even more direct legacy. Many of Chapin's so-called experiments were soon taken up, elaborated, and analyzed by the psychologist Donald T. Campbell and his colleagues under the milder and more general name of "quasi-experiments."[6]

The first widely read presentation of the Campbell's perspective emerged in 1963 (see Campbell and Stanley 1966[1963]), in which quasi-experiments were discussed alongside randomized and fully controlled experimental trials, with an evaluation of their relative strengths and weaknesses in alternative settings. In the subsequent decade, Campbell's work with his colleagues moved closer toward observational research, culminating in the volume by Cook and Campbell (1979), *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, wherein a whole menu of quasi-experiments was described and analyzed: from the sort of *ex post* case-control matching studies advocated by Chapin (but relabelled more generally as nonequivalent group designs) to novel proposals for regression discontinuity and interrupted time series designs (which we will discuss later in Chapter 9). For Cook and Campbell, the term quasi-experiment refers to "experiments that have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons

[6] In his first publication on quasi-experiments, Campbell (1957) aligned himself with Stouffer's perspective on the utility of experimental language, and in particular Stouffer (1950). Chapin is treated roughly by Campbell and Stanley (1963:70), even though his *ex post facto* design is identified as "one of the most extended efforts toward quasi-experimental design."

from which treatment-caused change is inferred" (Cook and Campbell 1979:6).[7] And, rather than advocate for a reorientation of a whole discipline as Chapin had, they pitched the approach as a guide for field studies, especially program evaluation studies of controlled interventions. Nonetheless, the ideas were widely influential throughout the social sciences, as they succeeded in bringing a tamed experimental language to the foreground in a way that permitted broad assessments of the strengths and weaknesses of alternative study designs and data analysis techniques.

1.2.2 "The Age of Regression"

Even though the quasi-experiment tradition swept through the program evaluation community and gained many readers elsewhere, it lost out in both sociology and economics to equation-based motivations of observational data analysis, under the influence of a new generation of econometricians, demographers, and survey researchers who developed structural equation and path-model techniques. Many of the key methodological advances took place in the field of economics, as discussed by Goldberger (1972) and Heckman (2000), even though the biologist Sewall Wright (1925, 1934) is credited with the early development of some of the specific techniques.

In the 1960s, structural equation models spread quickly from economics throughout the social sciences, moving first to sociology via Hubert Blalock and Otis Dudley Duncan, each of whom is usually credited with introducing the techniques, respectively, via Blalock's 1964 book *Causal Inferences in Nonexperimental Research* and Duncan's 1966 article, "Path Analysis: Sociological Examples," which was published as the lead article in that year's *American Journal of Sociology*. In both presentations, caution is stressed. Blalock discusses carefully the differences between randomized experiments and observational survey research. Duncan states explicitly in his abstract that "Path analysis focuses on the problem of interpretation and does not purport to be a method for discovering causes," and he concludes his article with a long quotation from Sewall Wright attesting to the same point.

A confluence of developments then pushed structural equations toward widespread usage and then basic regression modeling toward near complete dominance of observational research in some areas of social science. In sociology, the most important impetus was the immediate substantive payoff to the techniques. *The American Occupational Structure*, which Duncan cowrote with Peter Blau and published in 1967, transformed scholarship on social stratification, offering new decompositions of the putative causal effects of parental background and individuals' own characteristics on later educational and occupational

[7] Notice that Cook and Campbell's definition of quasi-experiments here is, in fact, consistent with the definition of an experiment laid out by Cox and Reid, which we cited earlier. For that definition of an experiment, control is essential but randomization is not. The text of Cook and Campbell (1979) equivocates somewhat on these issues, but it is clear that their intent is to discuss controlled experiments in which randomization is not feasible and that they then label quasi-experiments.

attainment. Their book transformed a core subfield of the discipline of sociology, leading to major theoretical and methodological redirections of many existing lines of scholarship.[8]

In part because of this success, it appears undeniable that Blalock and Duncan became, for a time, less cautious. Blalock had already shown a predilection toward slippage. When introducing regression equations in his 1964 book, specified as $Y_i = a + bX_i + e_i$, where X is the causal variable of interest and Y is the outcome variable of interest, Blalock then states correctly and clearly:

> What if there existed a major determinant of Y, not explicitly contained in the regression equation, which was in fact correlated with some of the independent variables X_i? Clearly, it would be contributing to the error term in a manner so as to make the errors systematically related to these particular X_i. If we were in a position to bring this unknown variable into the regression equation, we would find that at least some of the regression coefficients (slopes) would be changed. This is obviously an unsatisfactory state of affairs, making it nearly impossible to state accurate scientific generalizations. (Blalock 1964:47)

But Blalock ends his book with a set of numbered conclusions, among which can be found a different characterization of the same issue. Instead, he implies that the goal of causal inference should not be sacrificed even when these sorts of assumptions are dubious:

> We shall assume that error terms are uncorrelated with each other and with any of the independent variables in a given equation . . . . In nonexperimental studies involving nonisolated systems, this kind of assumption is likely to be unrealistic. This means that disturbing influences must be explicitly brought into the model. But at some point one must stop and make the simplifying assumption that variables left out do not produce confounding influences. Otherwise, causal inferences cannot be made. (Blalock 1964:176)

Blalock then elevates regression models to high scientific status: "In causal analyses our aim is to focus on causal laws as represented by regression equations and their coefficients" (Blalock 1964:177). And he then offers the practical advice that "The method for making causal inferences may be applied to models based on a priori reasoning, or it may be used in exploratory fashion to arrive at models which give closer and closer approximations to the data" (Blalock 1964:179).

Not only are these conclusions unclear – Should the exploration-augmented model still be regarded as a causal model? – they misrepresent the first 171 pages of Blalock's own book, in which he stressed the importance of assumptions grounded in substantive theory and offered repeated discussion of the differences

[8] For example, compare the methods (and substantive motivations) in Sewell (1964), with its nonparametric table standardization techniques, to Sewell, Haller, and Portes (1969), with its path model of the entire stratification process.

between regression equations embedded in recursive path models and the sorts of randomized experiments that often yield more easily defendable causal inferences. They also misrepresent the closing pages of his book, in which he returns with caution to argue that a researcher should remain flexible, report inferences from multiple models, and give an accounting of exploratory outcomes.

Duncan's record is less obviously equivocal, as he never failed to mention that assumptions about causal relationships must be grounded in theory and cannot be revealed by data. Yet, as Abbott (2001[1998]:115) notes, "Duncan was explicit in [*The American Occupational Structure*] about the extreme assumptions necessary for the analysis, but repeatedly urged the reader to bear with him while he tried something out to see what could be learned." What Duncan learned transformed the field, and it was thus hard to ignore the potential power of the techniques to move the literature.

Duncan's 1975 methodological text, *Introduction to Structural Equation Models*, is appropriately restrained, with many fine discussions that echo the caution in the abstract of his 1966 article. Yet he encourages widespread application of regression techniques to estimate causal effects, and at times he leaves the impression that researchers should just get on with it as he did in the *The American Occupational Structure*. For example, in his Chapter 8, titled "Specification Error," Duncan notes that "it would require no elaborate sophistry to show that we will never have the 'right' model in any absolute sense" (Duncan 1975:101). But he then continues:

> As the term will be used here, analysis of specification error relates to a rhetorical strategy in which we suggest a model as the "true" one for sake of argument, determine how our working model [the model that has been estimated] differs from it and what the consequences of the difference(s) are, and thereby get some sense of how important the mistakes we will inevitably make may be. Sometimes it is possible to secure genuine comfort by this route. (Duncan 1975:101-2)

As is widely known, Duncan later criticized the widespread usage of regression analysis and structural equation modeling more generally, both in his 1984 book *Notes on Social Measurement: Historical and Critical* and in private communication in which he reminded many inside and outside of sociology of his long-standing cautionary perspective (see Xie 2006).

Finally, the emergent ease with which regression models could be estimated with new computing power was important as well. No longer would Stouffer have needed to concentrate on a seven-way cross tabulation. His descendants could instead estimate and then interpret only a few estimated regression slopes, rather than attempt to make sense of the hundred or so cells that Stouffer often generated by subdivision of the sample. Aage Sørensen has given the most memorable indictment of the consequences of this revolution in computing power:

> With the advent of the high-speed computer, we certainly could study the relationships among many more variables than before.

> More importantly, we could compute precise quantitative measures of the strength of these relationships. The revolution in quantitative sociology was a revolution in statistical productivity. Social scientists could now calculate almost everything with little manual labor and in very short periods of time. Unfortunately, the sociological workers involved in this revolution lost control of their ability to see the relationship between theory and evidence. Sociologists became alienated from their sociological species being. (Sørensen 1998:241)

As this quotation intimates, enthusiasm for regression approaches to causal inference had declined dramatically by the mid-1990s. Naive usage of regression modeling was blamed for nearly all the ills of sociology, everything from stripping temporality, context, and the valuation of case study methodologies from the mainstream (see Abbott 2001 for a collections of essays), the suppression of attention to explanatory mechanisms (see Hedström 2005 and Goldthorpe 2001), the denial of causal complexity (see Ragin 1987, 2000), and the destruction of mathematical sociology (Sørensen 1998).

It is unfair to lay so much at the feet of least squares formulas, and we will argue later that regression can be put to work quite sensibly in the pursuit of causal questions. However, the critique of practice is largely on target. For causal analysis, the rise of regression led to a focus on equations for outcomes, rather than careful thinking about how the data in hand differ from what would have been generated by the ideal experiments one might wish to have conducted. This sacrifice of attention to experimental thinking might have been reasonable if the outcome-equation tradition had led researchers to specify and then carefully investigate the plausibility of alternative explanatory mechanisms that generate the outcomes of the equations. But, instead, it seems that researchers all too often chose not to develop fully articulated mechanisms that generate outcomes and instead chose to simply act as if the regression equations somehow mimic appreciably well (by a process not amenable to much analysis) the experiments that researchers might otherwise have wished to undertake.

The counterfactual model for observational data analysis has achieved success in the past two decades in the social sciences because it brings experimental language back into observational data analysis. But it does so in the way that Stouffer used it: as a framework in which to ask carefully constructed "what-if" questions that lay bare the limitations of observational data and the need to clearly articulate assumptions grounded in theory that is believable.

1.3 Types of Examples Used Throughout the Book

In this section, we offer background on the main substantive examples that we will draw on throughout the book when discussing the methods and approach abstractly and then when demonstrating particular empirical analysis strategies.

1.3.1 Broad Examples from Sociology, Economics, and Political Science

We first outline three prominent classic examples that, in spite of their distinct disciplinary origins, are related to each other: (1) the causal effects of family background and mental ability on educational attainment, (2) the causal effects of educational attainment and mental ability on earnings, and (3) the causal effects of family background, educational attainment, and earnings on political participation. These examples are classic and wide ranging, having been developed, respectively, in the formative years of observational data analysis in sociology, economics, and political science.

The Causal Effects of Family Background and Intelligence on Educational Attainment

In the status attainment tradition in sociology, as pioneered by Blau and Duncan (1967), family background and mental ability are considered to be ultimate causes of educational attainment. This claim is grounded on the purported existence of a specific causal mechanism that relates individuals' expectations and aspirations for the future to the social contexts that generate them. This particular explanation is most often identified with the Wisconsin model of status attainment, which was based on early analyses of the Wisconsin Longitudinal Survey (see Sewell, Haller, and Portes 1969; Sewell, Haller, and Ohlendorf 1970).

According to the original Wisconsin model, the joint effects of high school students' family backgrounds and mental abilities on their eventual educational attainments can be completely explained by the expectations that others hold of them. In particular, significant others – parents, teachers, and peers – define expectations based on students' family background and observable academic performance. Students then internalize the expectations crafted by their significant others. In the process, the expectations become individuals' own aspirations, which then compel achievement motivation.

The implicit theory of the Wisconsin model maintains that students are compelled to follow their own aspirations. Accordingly, the model is powerfully simple, as it implies that significant others can increase high school students' future educational attainments merely by increasing their own expectations of them.[9] Critics of this status attainment perspective argued that structural constraints embedded in the opportunity structure of society should be at the center of all models of educational attainment, and hence that concepts such as aspirations and expectations offer little or no explanatory power. Pierre Bourdieu (1973) dismissed all work that asserts that associations between aspirations and attainments are causal. Rather, for Bourdieu, the unequal opportunity structures of society "determine aspirations by determining the extent to which they can be satisfied" (Bourdieu 1973:83). And, as such, aspirations have no autonomous explanatory power because they are nothing other than alternative indicators of structural opportunities and resulting attainment.

[9] See Hauser, Warren, Huang, and Carter (2000) for the latest update of the original model.

The Causal Effects of Educational Attainment and Mental Ability on Earnings

The economic theory of human capital maintains that education has a causal effect on the subsequent labor market earnings of individuals. The theory presupposes that educational training provides skills that increase the potential productivity of workers. Because productivity is prized in the labor market, firms are willing to pay educated workers more.

These claims are largely accepted within economics, but considerable debate remains over the size of the causal effect of education. In reflecting on the first edition of his book, *Human Capital*, which was published in 1964, Gary Becker wrote nearly 30 years later:

> Education and training are the most important investments in human capital. My book showed, and so have many other studies since then, that high school and college education in the United States greatly raise a person's income, even after netting out direct and indirect costs of schooling, and after adjusting for the better family backgrounds and greater abilities of more educated people. Similar evidence is now available for many points in time from over one hundred countries with different cultures and economic systems. (Becker 1993[1964]:17)

The complication, hinted at in this quotation, is that economists also accept that mental ability enhances productivity as well. Thus, because those with relatively high ability are assumed to be more likely to obtain higher educational degrees, the highly educated are presumed to have higher innate ability and higher natural rates of productivity. As a result, some portion of the purported causal effect of education on earnings may instead reflect innate ability rather than any productivity-enhancing skills provided by educational institutions (see Willis and Rosen 1979). The degree of "ability bias" in standard estimates of the causal effect of education on earnings has remained one of the largest causal controversies in the social sciences since the 1970s (see Card 1999).

The Causal Effects of Family Background, Educational Attainment, and Earnings on Political Participation

The socioeconomic status model of political participation asserts that education, occupational attainment, and income predict strongly most measures of political participation (see Verba and Nie 1972). Critics of this model maintain instead that political interests and engagement determine political participation, and these are merely correlated with the main dimensions of socioeconomic status.[10]

[10]This interest model of participation has an equally long lineage. Lazarsfeld, Berelson, and Gaudet (1955[1948]:157) write that, in their local sample, "the difference in deliberate non-voting between people with more or less education can be completely accounted for by the notion of interest."

In other words, those who have a predilection to participate in politics are likely to show commitment to other institutions, such as the educational system.

Verba, Schlozman, and Brady (1995) later elaborated the socioeconomic status model, focusing on the contingent causal processes that they argue generate patterns of participation through the resources conferred by socioeconomic position. They claim:

> ... interest, information, efficacy, and partisan intensity provide the desire, knowledge, and self-assurance that impel people to be engaged by politics. But time, money, and skills provide the wherewithal without which engagement is meaningless. It is not sufficient to know and care about politics. If wishes were resources, then beggars would participate. (Verba et al. 1995:355-6)

They reach this conclusion through a series of regression models that predict political participation. They use temporal order to establish causal order, and they then claim to eliminate alternative theories that emphasize political interests and engagement by showing that these variables have relatively weak predictive power in their models.

Moreover, they identify education as the single strongest cause of political participation. Beyond generating the crucial resources of time, money, and civic skills, education shapes preadult experiences and transmits differences in family background (see Verba et al. 1995, Figure 15.1). Education emerges as the most powerful cause of engagement because it has the largest net association with measures of political participation.

Nie, Junn, and Stehlik-Barry (1996) then built on the models of Verba and his colleagues, specifying in detail the causal pathways linking education to political participation. For this work, the effects of education, family income, and occupational prominence (again, the three basic dimensions of socioeconomic status) on voting frequency are mediated by verbal proficiency, organizational membership, and social network centrality. Nie et al. (1996:76) note that these variables "almost fully explain the original bivariate relationship between education and frequency of voting."

Each of these first three examples, as noted earlier, is concerned with relationships that unfold over the lifecourse of the majority of individuals in most industrialized societies. As such, these examples encompass some of the most important substantive scholarship in sociology, economics, and political science. At the same time, however, they pose some fundamental challenges for causal analysis: measurement complications and potential nonmanipulability of the causes of interest. Each of these deserves some comment before the narrower and less complicated examples that follow are introduced.

First, the purported causal variables in these models are highly abstract and internally heterogeneous. Consider the political science example. Political participation takes many forms, from volunteer work to financial giving and voting. Each of these, in turn, is itself heterogeneous, given that individuals can contribute episodically and vote in only some elections. Furthermore,

family background and socioeconomic status include at least three underlying dimensions: family income, parental education, and occupational position. But other dimensions of advantage, such as wealth and family structure, must also be considered, as these are thought to be determinants of both an individual's educational attainment and also the resources that supposedly enable political participation.[11]

Scholars who pursue analysis of these causal effects must therefore devote substantial energy to the development of measurement scales. Although very important to consider, in this book we will not discuss measurement issues so that we can focus closely on causal effect estimation strategies. But, of course, it should always be remembered that, in the absence of agreement on issues of how to measure a cause, few causal controversies can be resolved, no matter what estimation strategy seems best to adopt.

Second, each of these examples concerns causal effects for individual characteristics that are not easily manipulable through external intervention. Or, more to the point, even when they are manipulable, any such induced variation may differ fundamentally from the naturally occurring (or socially determined) variation with which the models are most directly concerned. For example, family background could be manipulated by somehow convincing a sample of middle-class and working-class parents to exchange their children at particular well-chosen ages, but the subsequent outcomes of this induced variation may not correspond to the family background differences that the original models attempt to use as explanatory differences.

As we will discuss later, whether nonmanipulability of a cause presents a challenge to an observational data analyst is a topic of continuing debate in the methodological and philosophical literature. We will discuss this complication at several points in this book, including a section in the concluding chapter. But, given that the measurement and manipulability concerns of the three broad examples of this section present challenges at some level, we also draw on more narrow examples throughout the book, as we discuss in the next section. For these more recent and more narrow examples, measurement is generally less controversial and potential manipulability is more plausible (and in some cases is completely straightforward).

1.3.2 Narrow and Specific Examples

Throughout the book, we will introduce recent specific examples, most of which can be considered more narrow causal effects that are closely related to the broad causal relationships represented in the three examples presented in the last section. These examples will include, for example, the causal effect of education on mental ability, the causal effect of military service on earnings, and the causal effect of felon disenfranchisement on election outcomes. To give a sense of the general characteristics of these narrower examples, we describe in

[11] Moreover, education as a cause is somewhat ungainly as well. For economists who wish to study the effects of learned skills on labor market earnings, simple variables measuring years of education obtained are oversimplified representations of human capital.

the remainder of this section four examples that we will use at multiple points throughout the book: (1) the causal effect of Catholic schooling on learning, (2) the causal effect of school vouchers on learning, (3) the causal effect of manpower training on earnings, and (4) the causal effect of alternative voting technology on valid voting.

The Causal Effect of Catholic Schooling on Learning

James S. Coleman and his colleagues presented evidence that Catholic schools are more effective than public schools in teaching mathematics and reading to equivalent students (see Coleman and Hoffer 1987; Coleman, Hoffer, and Kilgore 1982; Hoffer, Greeley, and Coleman 1985). Their findings were challenged vigorously by other researchers who argued that public school students and Catholic school students are insufficiently comparable, even after adjustments for family background and measured motivation to learn (see Alexander and Pallas 1983, 1985; Murnane, Newstead, and Olsen 1985; Noell 1982; Willms 1985; see Morgan 2001 for a summary of the debate). Although the challenges were wide ranging, the most compelling argument raised (and that was foreseen by Coleman and his colleagues) was that students who are most likely to benefit from Catholic schooling are more likely to enroll in Catholic schools net of all observable characteristics. Thus, self-selection on the causal effect itself may generate a mistakenly large apparent Catholic school effect. If students instead were assigned randomly to Catholic and public schools, both types of schools would be shown to be equally effective on average.

To address the possibility that self-selection dynamics create an illusory Catholic school effect, a later wave of studies then assessed whether or not naturally occurring experiments were available that could be used to more effectively estimate the Catholic school effect. Using a variety of variables that predict Catholic school attendance (e.g., share of the local population that is Catholic) and putting forth arguments for why these variables do not directly determine achievement, Evans and Schwab (1995), Hoxby (1996), and Neal (1997) generated support for Coleman's original conclusions.

The Causal Effect of School Vouchers on Learning

In response to a perceived crisis in public education in the United States, policymakers have introduced publicly funded school choice programs into some metropolitan areas in an effort to increase competition among schools on the assumption that competition will improve school performance and resulting student achievement (see Chubb and Moe 1990; see also Fuller and Elmore 1996). Although these school choice programs differ by school district, the prototypical design is the following. A set number of $3000 tuition vouchers redeemable at private schools are made available to students resident in the public school district, and all parents are encouraged to apply for one of these vouchers. The vouchers are then randomly assigned among those who apply. Students who

receive a voucher remain eligible to enroll in the public school to which their residence status entitles them. But they can choose to enroll in a private school. If they choose to do so, they hand over their $3000 voucher and pay any required top-up fees to meet the private school tuition.

The causal effects of interest resulting from these programs are numerous. Typically, evaluators are interested in the achievement differences between those who attend private schools using vouchers and other suitable comparison groups. Most commonly, the comparison group is the group of voucher applicants who lost out in the lottery and ended up in public schools (see Howell and Peterson 2002; Hoxby 2003; Ladd 2002; Neal 2002). And, even though these sorts of comparisons may seem entirely straightforward, the published literature shows that considerable controversy surrounds how best to estimate these effects, especially given the real-world complexity that confronts the implementation of randomization schemes (see Krueger and Zhu 2004; Peterson and Howell 2004).

For this example, other effects are of interest as well. A researcher might wish to know how the achievement of students who applied for vouchers but did not receive them changed in comparison with those who never applied for vouchers in the first place (as this would be crucial for understanding how the self-selecting group of voucher applicants may differ from other public school students). More broadly, a researcher might wish to know the expected achievement gain that would be observed for a public school student who was randomly assigned a voucher irrespective of the application process. This would necessitate altering the voucher assignment mechanism, and thus it has not been an object of research. Finally, the market competition justification for creating these school choice policies implies that the achievement differences of primary interest are those among public school students who attend voucher-threatened public schools (i.e., public schools that feel as if they are in competition with private schools but that did not feel as if they were in competition with private schools before the voucher program was introduced).

The Causal Effect of Manpower Training on Earnings

The United States federal government has supported manpower training programs for economically disadvantaged citizens for decades (see LaLonde 1995). Through a series of legislative renewals, these programs have evolved substantially, and program evaluations have become an important area of applied work in labor and public economics.

The services provided to trainees differ and include classroom-based vocational education, remedial high school instruction leading to a general equivalency degree, and on-the-job training (or retraining) for those program participants who have substantial prior work experience. Moreover, the types of individuals served by these programs are heterogeneous, including ex-felons, welfare recipients, and workers displaced from jobs by foreign competition. Accordingly, the causal effects of interest are heterogeneous, varying with individual characteristics and the particular form of training provided.

Even so, some common challenges have emerged across most program evaluations. Ashenfelter (1978) discovered what has become known as "Ashenfelter's dip," concluding after his analysis of training program data that

> ... all of the trainee groups suffered unpredicted earnings declines in the year prior to training.... This suggests that simple before and after comparisons of trainee earnings may be seriously misleading evidence. (Ashenfelter 1978:55)

Because trainees tend to have experienced a downward spiral in earnings just before receiving training, the wages of trainees would rise to some degree even in the absence of any training. Ashenfelter and Card (1985) then pursued models of these "mean reversion" dynamics, demonstrating that the size of treatment effect estimates is a function of alternative assumptions about pretraining earnings trajectories. They called for the construction of randomized field trials to improve program evaluation.

LaLonde (1986) then used results from program outcomes for the National Supported Work (NSW) Demonstration, a program from the mid-1970s that randomly assigned subjects to alternative treatment conditions. LaLonde argued that most of the econometric techniques used for similar program evaluations failed to match the experimental estimates generated by the NSW data. Since LaLonde's 1986 paper, econometricians have continued to refine procedures for evaluating both experimental and nonexperimental data from training programs, focusing in detail on how to model the training selection mechanism (see Heckman, LaLonde, and Smith 1999; Manski and Garfinkel 1992; Smith and Todd 2005).

The Causal Effect of Alternative Voting Technology on Valid Voting

For specific causal effects embedded in the larger political participation debates, we could focus on particular decision points – the effect of education on campaign contributions, net of income, and so on. However, the politics literature is appealing in another respect: outcomes in the form of actual votes cast and subsequent election victories. These generate finely articulated counterfactual scenarios.

In the contested 2000 presidential election in the United States, considerable attention focused on the effect of voting technology on the election outcome in Florida. Wand et al. (2001) published a refined version of their analysis that spread like wildfire on the Internet in the week following the presidential election. They asserted that

> ... the butterfly ballot used in Palm Beach County, Florida, in the 2000 presidential election caused more than 2,000 Democratic voters to vote by mistake for Reform candidate Pat Buchanan, a number larger than George W. Bush's certified margin of victory in Florida. (Wand et al. 2001:793)

Reflecting on efforts to recount votes undertaken by various media outlets, Wand and his colleagues identify the crucial contribution of their analysis:

> Our analysis answers a counterfactual question about voter intentions that such investigations [by media outlets of votes cast] cannot resolve. The inspections may clarify the number of voters who marked their ballot in support of the various candidates, but the inspections cannot tell us how many voters marked their ballot for a candidate they did not intend to choose. (Wand et al. 2001:804)

Herron and Sekhon (2003) then examined invalid votes that resulted from overvotes (i.e., voting for more than one candidate), arguing that such overvotes further hurt Gore's vote tally in two crucial Florida counties. Finally, Mebane (2004) then considered statewide voting patterns, arguing that if voters' intentions had not been thwarted by technology, Gore would have won the Florida presidential election by 30,000 votes. One particularly interesting feature of this example is that the precise causal effect of voting technology on votes is not of interest, only the extent to which such causal effects aggregate to produce an election outcome inconsistent with the preferences of those who voted.

1.4 Observational Data and Random-Sample Surveys

When we discuss methods and examples throughout this book, we will usually assume that the data have been generated by a relatively large random-sample survey. We will also assume that the proportion and pattern of individuals who are exposed to the cause are fixed in the population by whatever process generates causal exposure.

We rely on the random-sample perspective because we feel it is the most natural framing of these methods for the typical social scientist, even though many of the classic applications and early methodological pieces in this literature do not reference random-sample surveys. For the examples just summarized, the first three have been examined primarily with random-sample survey data, but many of the others have not. Some, such as the manpower training example, depart substantially from this sort of setup, as the study subjects for the treatment in that example are a nonrandom and heterogeneous collection of welfare recipients, ex-felons, and displaced workers.[12]

[12] Partly for this reason, some of the recent literature (e.g., Imbens 2004) has made careful distinctions between the sample average treatment effect (SATE) and the population average treatment effect (PATE). In this book, we will focus most of our attention on the PATE (and other conditional PATEs). We will generally write under the implicit assumption that a well-defined population exists (generally a superpopulation with explicit characteristics) and that the available data are a random sample from this population. However, much of our treatment of these topics could be rewritten without the large random-sample perspective and focusing only on the average treatment effect within the sample in hand. Many articles in this tradition of analysis adopt this alternative starting point (especially those relevant for small-scale studies in epidemiology and biostatistics for which the "sample" is generated in such a way

Pinning down the exact consequences of the data generation and sampling scheme of each application is important for developing estimates of the expected variability of a causal effect estimate. We will therefore generally modify the random-sampling background when discussing what is known about the expected variability of the alternative estimators we will present. However, we focus more in this book on parameter identification than on the expected variability of an estimator in a finite sample, as we discuss in the next section.

In fact, as the reader will notice in subsequent chapters, we often assume that the sample is infinite. This preposterous assumption is useful for presentation purposes because it simplifies matters greatly; we can then assume that sampling error is zero and assert, for example, that the sample mean of an observed variable is equal to the population expectation of that variable. But this assumption is also an indirect note of caution: It is meant to appear preposterous and unreasonable in order to reinforce the point that the consequences of sampling error must always be considered in any empirical analysis.[13]

Moreover, we will also assume for our presentation that the variables in the data are measured without error. This perfect measurement assumption is, of course, also entirely unreasonable. But it is commonly invoked in discussions of causality and in many, if not most, other methodological pieces. We will indicate in various places throughout the book when random measurement error is especially problematic for the methods that we present. We leave it as self-evident that nonrandom measurement error can be debilitating for all methods.

1.5 Identification and Statistical Inference

In the social sciences, identification and statistical inference are usually considered separately. In his 1995 book, *Identification Problems in the Social Sciences*, the economist Charles Manski writes:

> ... it is useful to separate the inferential problem into statistical and identification components. Studies of identification seek to characterize the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations. Identification problems cannot be solved by gathering more of the same kind of data. (Manski 1995:4)

He continues:

> Empirical research must, of course, contend with statistical issues as well as with identification problems. Nevertheless, the two types of

that a formal connection to a well-defined population is impossible). We discuss these issues in substantial detail in Chapter 2, especially in the appendix on alternative population models.

[13] Because we will in these cases assume that the sample is infinite, we must then also assume that the population is infinite. This assumption entails adoption of the superpopulation perspective from statistics (wherein the finite population from which the sample is drawn is regarded as one realization of a stochastic superpopulation). Even so, and as we will explain in Chapter 2, we will not clutter the text of the book by making fine distinctions between the observable finite population and its more encompassing superpopulation.

> inferential difficulties are sufficiently distinct for it to be fruitful to study them separately. The study of identification logically comes first. Negative identification findings imply that statistical inference is fruitless: it makes no sense to try to use a sample of finite size to infer something that could not be learned even if a sample of infinite size were available. Positive identification findings imply that one should go on to study the feasibility of statistical inference. (Manski 1995:5)

In contrast, in his 2002 book, *Observational Studies*, the statistician Paul Rosenbaum sees the distinction between identification and statistical inference as considerably less helpful:

> The idea of identification draws a bright red line between two types of problems. Is this red line useful? ... In principle, in a problem that is formally not identified, there may be quite a bit of information about β, perhaps enough for some particular practical decision .... Arguably, a bright red line relating assumptions to asymptotics is less interesting than an exact confidence interval describing what has been learned from the evidence actually at hand. (Rosenbaum 2002:185–6)

Rosenbaum's objection to the bright red line of identification is issued in the context of analyzing a particular type of estimator – an instrumental variable estimator – that can offer an estimate of a formally identified parameter that is so noisy in a dataset of any finite size that one cannot possibly learn anything from the estimate. However, an alternative estimator – usually a least squares regression estimator in this context – that does not formally identify a parameter because it remains asymptotically biased even in an infinite sample may nonetheless provide sufficiently systematic information so as to remain useful, especially if one has a sense from other aspects of the analysis of the likely direction and size of the bias.

We accept Rosenbaum's perspective; it is undeniable that an empirical researcher who forsakes all concern with statistical inference could be led astray by considering only estimates that are formally identified. But, for this book, our perspective is closer to that of Manski, and we focus on identification problems almost exclusively. Nonetheless, where we feel it is important, we will offer discussions of the relative efficiency of estimators, such as for matching estimators and instrumental variable estimators. And we will discuss the utility of comparing alternative estimators based on the criterion of mean-squared error. Our primary goal, however, remains the clear presentation of material that can help researchers to determine what assumptions must be maintained in order to identify causal effects, as well as the selection of an appropriate technique that can be used to estimate an identified causal effect from a sample of sufficient size under whatever assumptions are justified.

1.6 Causal Graphs as an Introduction to the Remainder of the Book

After introducing the main pieces of the counterfactual model in Chapter 2, we will then present conditioning techniques for causal effect estimation in Part 2 of the book. In Chapter 3, we will present a basic conditioning framework using causal diagrams. Then, in Chapters 4 and 5, we will explain matching and regression estimators, showing how they are complementary variants of a more general conditioning approach.

In Part 3 of the book, we will then make the transition from "easy" to "hard" instances of causal effect estimation, for which simple conditioning will not suffice because relevant variables that determine causal exposure are not observed. After presenting the general predicament in Chapter 6, we will then offer Chapters 7 through 9 on instrumental variable techniques, mechanism-based estimation of causal effects, and the usage of over-time data to estimate causal effects.

Finally, in Chapter 10 we will provide a summary of some of the objections that others have developed against the counterfactual model. And we will conclude the book with a broad discussion of the complementary modes of causal inquiry that comprise causal effect estimation in observational social science.

In part because the detailed table of contents already gives an accurate accounting of the material that we will present in the remaining chapters, we will not provide a set of detailed chapter summaries here. Instead, we will conclude this introductory chapter with three causal diagrams and the causal effect estimation strategies that they suggest. These graphs allow us to foreshadow many of the specific causal effect estimation strategies that we will present later.

Because the remainder of the material in this chapter will be reintroduced and more fully explained later (primarily in Chapters 3, 6, and 8), it can be skipped now without consequence. However, our experience in teaching this material suggests that many readers may benefit from a quick graphical introduction to the basic estimation techniques before considering the details of the counterfactual framework for observational data analysis.

Graphical Representations of Causal Relationships

Judea Pearl (2000) has developed a general set of rules for representing causal relationships with graph theory. We will provide a more complete introduction to Pearl's graph-theoretic modeling of causal relationships in Chapter 3, but for now we use the most intuitive pieces of his graphical apparatus with only minimal explanation. That these graphs are readily interpretable and provide insight with little introduction is testament to the clarity of Pearl's contribution to causal analysis.

Consider the causal relationships depicted in the graph in Figure 1.1 and suppose that these relationships are derived from a set of theoretical propositions that have achieved consensus in the relevant scholarly community. For this graph, each node represents an observable random variable. Each directed edge

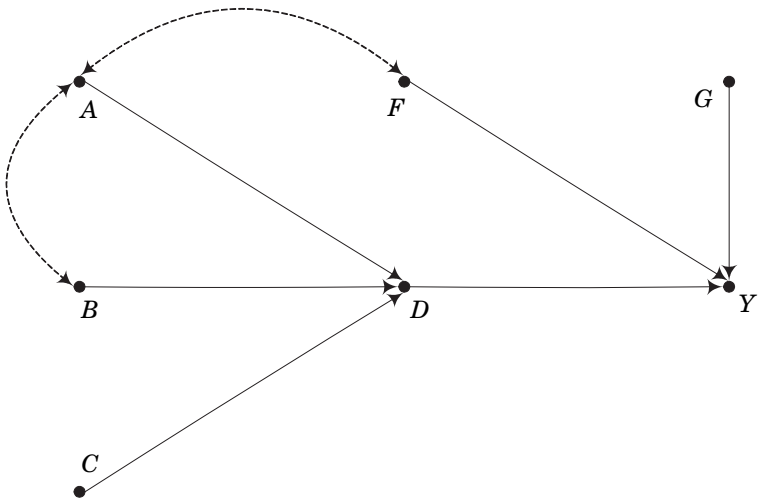


Figure 1.1: A causal diagram in which back-door paths from D to Y can be blocked by observable variables and C is an instrumental variable for D.

(i.e., single-headed arrow) from one node to another signifies that the variable at the origin of the directed edge causes the variable at the terminus of the directed edge. Each curved and dashed bidirected edge (i.e., double-headed arrow) signifies the existence of common unobserved nodes that cause both terminal nodes. Bidirected edges represent common causes only, not mere correlations with unknown sources and not relationships of direct causation between the two variables that they connect.

Now, suppose that the causal variable of primary interest is D and that the causal effect that we wish to estimate is the effect of D on Y. The question to consider is the following: Given the structure of causal relationships represented in the graph, which variables must we observe and then use in a data analysis routine to estimate the size of the causal effect of D on Y?

Before answering this question, consider some of the finer points of the graph. In Pearl's framework, the causal variable D has a probability distribution. The causal effects emanating from the variables A, B, and C are explicitly represented in the graph by directed edges, but the relative sizes of these effects are not represented in the graph. Other causes of D that are unrelated to A, B, and C are left implicit, as it is merely asserted in Pearl's framework that D has a probability distribution net of the systematic effects of A, B, and C on D.[14]

[14]There is considerable controversy over how to interpret these implicit causes. For some, the assertion of their existence is tantamount to asserting that causality is fundamentally probabilistic. For others, these implicit causes merely represent causes unrelated to the systematic causes of interest. Under this interpretation, causality can still be considered a structural, deterministic relation. The latter position is closest to the position of Pearl (2000; see sections 1.4 and 7.5).

The outcome variable, Y, is likewise caused by F, G, and D, but there are other implicit causes that are unrelated to F, G, and D that give Y its probability distribution.

This graph is not a full causal model in Pearl's framework because some second-order causes of D and Y create supplemental dependence between the observable variables in the graph.[15] These common causes are represented in the graph by bidirected edges. In particular, A and B share some common causes that cannot be more finely specified by the state of knowledge in the field. Likewise, A and F share some common causes that also cannot be more finely specified by the state of knowledge in the field.

The Three Basic Strategies to Estimate Causal Effects

Three basic strategies for estimating causal effects will be covered in this book. First, one can condition on variables (with procedures such as stratification, matching, weighting, or regression) that block all back-door paths from the causal variable to the outcome variable. Second, one can use exogenous variation in an appropriate instrumental variable to isolate covariation in the causal and outcome variables. Third, one can establish an isolated and exhaustive mechanism that relates the causal variable to the outcome variable and then calculate the causal effect as it propagates through the mechanism.

Consider the graph in Figure 1.1 and the opportunities it presents to estimate the causal effect of D on Y with the conditioning estimation strategy. First note that there are two back-door paths from D to Y in the graph that generate a supplemental noncausal association between D and Y: (1) D to A to F to Y and (2) D to B to A to F to Y.[16] Both of these back-door paths can be blocked in order to eliminate the supplemental noncausal association between D and Y by observing and then conditioning on A and B or by observing and then conditioning on F. These two conditioning strategies are general in the sense that they will succeed in producing consistent causal effect estimates of the effect of D on Y under a variety of conditioning techniques and in the presence of nonlinear effects. They are minimally sufficient in the sense that one can observe and then condition on any subset of the observed variables in $\{A, B, C, F, G\}$ as long as the subset includes either $\{A, B\}$ or $\{F\}$.[17]

[15] Pearl would refer to this graph as a semi-Markovian causal diagram rather than a fully Markovian causal model (see Pearl 2000, Section 5.2).

[16] As we note later in Chapter 3 when more formally defining back-door paths, the two paths labeled "back-door paths" in the main text here may represent many back-door paths because the bidirected edges may represent more than one common cause of the variables they point to. Even so, the conclusions stated in the main text are unaffected by this possibility because the minimally sufficient conditioning strategies apply to all such additional back-door paths as well.

[17] For the graph in Figure 1.1, one cannot effectively estimate the causal effect of D on Y by simply conditioning only on A. We explain this more completely in Chapter 3, where we introduce the concept of a collider variable. The basic idea is that conditioning only on A, which is a collider, creates dependence between B and F within the strata of A. As a result, conditioning only on A fails to block all back-door paths from D to Y.

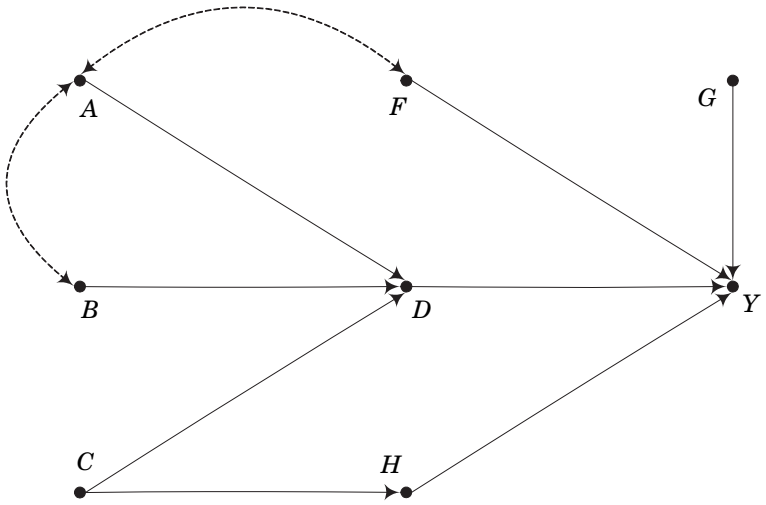


Figure 1.2: A causal diagram in which C is no longer an instrumental variable for D.

Now, consider the second estimation strategy, which is to use an instrumental variable for D to estimate the effect of D on Y. This strategy is completely different from the conditioning strategy just summarized. The goal is not to block back-door paths from the causal variable to the outcome variable but rather to use a localized exogenous shock to both the causal variable and the outcome variable in order to estimate indirectly the relationship between the two. For the graph in Figure 1.1, the variable C is a valid instrument for D because it causes D but does not have an effect on Y except though its effect on D. As a result, one can estimate consistently the causal effect of D on Y by taking the ratio of the relationships between C and Y and between C and D.[18] For this estimation strategy, A, B, F, and G do not need to be observed if the only interest of a researcher is the causal effect of D on Y.

To further consider the differences between these first two strategies, now consider the alternative graph presented in Figure 1.2. There are five possible strategies for estimating the causal effect of D on Y for this graph, and they differ from those for the set of causal relationships in Figure 1.1 because a third back-door path is now present: D to C to H to Y. For the first four strategies, all back-door paths can be blocked by conditioning on $\{A, B, C\}$, $\{A, B, H\}$,

[18] Although all other claims in this section hold for all distributions of the random variables and all types of nonlinearity of causal relationships, one must assume for IV estimation what Pearl labels a linearity assumption. What this assumption means depends on the assumed distribution of the variables. It would be satisfied if the causal effect of C on D is linear and the causal effect of D on Y is linear. Both of these would be true, for example, if both C and D were binary variables and Y were an interval-scaled variable, and this is the most common scenario we will consider in this book.

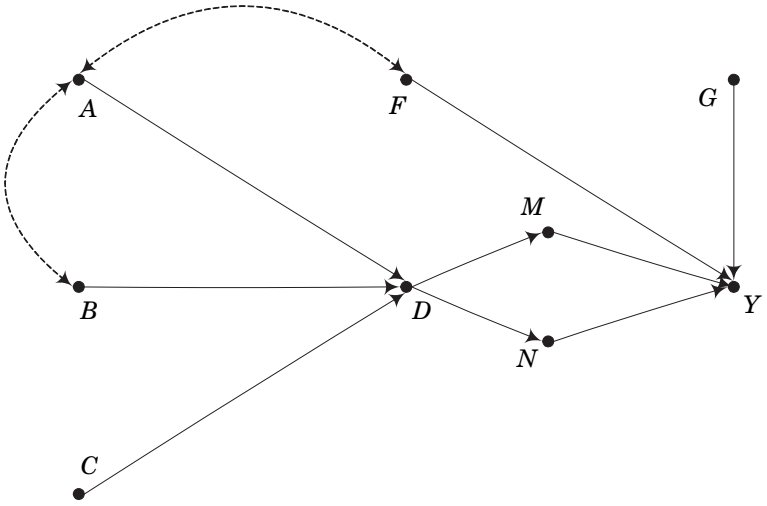


Figure 1.3: A causal diagram in which M and N represent an isolated and exhaustive mechanism for the causal effect of D on Y.

$\{F, C\}$, or $\{F, H\}$. For the fifth strategy, the causal effect can be estimated by conditioning on H and then using C as an instrumental variable for D.

Finally, to see how the third mechanistic estimation strategy can be used effectively, consider the alternative graph presented in Figure 1.3. For this graph, four feasible strategies are available as well. The same three strategies proposed for the graph in Figure 1.1 can be used. But, because the mediating variables M and N completely account for the causal effect of D on Y, and because M and N are not determined by anything other than D, the causal effect of D on Y can also be calculated by estimation of the causal effect of D on M and N and then subsequently the causal effects of M and N on Y. And, because this strategy is available, if the goal is to obtain the causal effect of D on Y, then the variables A, B, C, F, and G can be ignored.[19]

In an ideal scenario, all three of these forms of causal effect estimation could be used to obtain estimates, and all three would generate equivalent estimates (subject to the expected variation produced by a finite sample from a population). If a causal effect estimate generated by conditioning on variables that block all back-door paths is similar to a causal effect estimate generated by a valid instrumental variable estimator, then each estimate is bolstered.[20] Better

[19]Note that, for the graph in Figure 1.3, both M and N must be observed. If, instead, only M were observed, then this mechanistic estimation strategy will not identify the full causal effect of D on Y. However, if M and N are isolated from each other, as they are in Figure 1.3, the portion of the causal effect that passes through M or N can be identified in the absence of observation of the other. We discuss these issues in detail in Chapters 6 and 8.

[20]As we discuss in detail in Chapter 7, estimates generated by conditioning techniques and by valid instrumental variables will rarely be equivalent when individual-level heterogeneity of the causal effect is present (even in an infinite sample).

yet, if a mechanism-based strategy then generates a third equivalent estimate, all three causal effect estimates would be even more convincing. And, in this case, an elaborated explanation of how the causal effect comes about is also available, as a researcher could then describe how the causal effect is propagated through the intermediate mechanistic variables M and N.

The foregoing skeletal presentation of causal effect estimation is, of course, inherently misleading. Rarely does a state of knowledge prevail in a field that allows a researcher to specify causes as cleanly as in the causal diagrams in these figures. Accordingly, estimating causal effects is a great deal more challenging.

Nonetheless, beyond introducing the basic estimation techniques, these simple graphs convey two important sets of points that we will emphasize throughout the book. First, there is often more than one way to estimate a causal effect, and simple rules such as "control for all other causes of the outcome variable" can be poor guides for practice. For example, for the graph in Figure 1.1, there are two completely different and plausible conditioning strategies: either condition on F or on A and B. The strategy to "control for all other causes of the outcome variable" is misleading because (1) it suggests that one should condition on G as well, which is unnecessary if all one wants to obtain is the causal effect of D on Y and (2) it does not suggest that one can estimate the causal effect of D on Y by conditioning on a subset of the variables that cause the causal variable of interest. In this case, one can estimate the causal effect of D on Y without conditioning on any of the other causes of Y, but instead by conditioning on the variables that cause D. Even so, this last conditioning strategy should not be taken too far. One need not condition on C when also conditioning on both A and B. Not only is this unnecessary (just as for G with the other conditioning strategy), in doing so one fails to use C in its most useful way: as an instrumental variable that can be used to consistently estimate the causal effect of D on Y, ignoring completely A, B, F, and G.

Second, the methods we will present, as we believe is the case with all estimation strategies in the social sciences, are not well suited to discovering the causes of outcomes and then comprehensively estimating the relative effects of all alternative causes. The way in which we have presented these graphs is telling on this point. Consider again the question that we posed after introducing the graph in Figure 1.1. We asked a simpler version of the following question: Given the structure of causal relationships that relate A, B, C, D, F, G, and Y to each other (represented by presupposed edges that signify causal effects of unknown magnitude), which variables must we observe and then use in a data analysis routine to estimate the size of the causal effect of D on Y? This sort of constrained question (i.e., beginning with the conditional "given" clause) is quite a bit from different from seeking to answer the more general question: What are the causes of Y? The methods that we will present are not irrelevant to this broader question, but they are designed to answer simpler subordinate questions.

Consider Figure 1.1 again. If we had estimated the effect of D on Y by observing only A, B, D, and Y and then conditioning on A and B, and if we

then found that D had a trivially small effect on Y, we would then want to observe both F and G and think further about whether what we considered to be common causes of both A and F might be known and observable after all. However, if we did not have a theory and its associated state of knowledge that suggested that F and G have causal effects on Y (i.e., and instead thought that D was the only systematic cause of Y), then determining that D has a small to nonexistent effect on Y would not help us to find any of the other causes of Y that may be important.

The limited nature of the methods that we will present implies two important features of causal effect estimation from the perspective of counterfactual modeling. To offer a precise and defendable causal effect estimate, a well-specified theory is needed to justify assumptions about underlying causal relationships. And, if theory is poorly specified, or divergent theories exist in the relevant scholarly community that support alternative assumptions about underlying causal relationships, then alternative causal effect estimates may be considered valid conditional on the validity of alternative maintained assumptions. We discuss these issues in depth in the concluding section of the book, after presenting the framework and the methods that generate estimates that must then be placed in their proper context.

Chapter 2

The Counterfactual Model

In this chapter, we introduce the foundational components of the counterfactual model of causality, which is also known as the potential outcome model. We first discuss causal states and the relationship between potential and observed outcome variables. Then we introduce average causal effects and discuss the assumption of causal effect stability, which is maintained in most applications of the counterfactual model. We conclude with a discussion of simple estimation techniques, in which we demonstrate the importance of considering the relationship between the potential outcomes and the process of causal exposure.

2.1 Causal States and Potential Outcomes

For a binary cause, the counterfactual framework presupposes the existence of two well-defined causal states to which all members of the population of interest could be exposed.[1] These two states are usually labeled treatment and control. When a many-valued cause is analyzed, the convention is to refer to the alternative states as alternative treatments.

Consider the examples introduced in Section 1.3. Some of these examples have well-defined states, and others do not. The manpower training example is completely straightforward, and the two states are whether an individual is enrolled in a training program or not. The Catholic school example is similar. Here, the alternative states are "Catholic school" and "public school." The only complication with these examples is the possibility of inherent differences across training programs and Catholic schools. If any such treatment-site heterogeneity exists, then stratified analyses may be necessary, perhaps by regions

[1] We justify the importance of carefully defining the boundaries of the population of interest when presenting average causal effects later in this chapter. As we note there, we also provide an appendix to this chapter, in which we explain the general superpopulation model that we will adopt when the boundaries of the population can be clearly defined and when we have the good fortune of having a large random sample from the population.

of the country, size of the program, or whatever other dimension suggests that variability of the causal states deserves explicit modeling.[2]

Other examples have less clearly defined causal states. Consider the classic political participation line of inquiry. For the relationship between socioeconomic status and political participation, there are many underlying well-defined causal effects, such as the effect of having obtained at least a college degree on the frequency of voting in local elections and the effect of having a family income greater than some cutoff value on the amount of money donated to political campaigns. Well-defined causal states exist for these narrow causal effects, but it is not clear at all that well-defined causal states exist for the broad and internally differentiated concepts of socioeconomic status and political participation.

Finally, consider a related political science example. Beyond the voting technology effect discussed in Subsection 1.3.2 on the outcome of the 2000 presidential election, a broader set of question has been asked. To what extent do restrictions on who can vote determine who wins elections? A recent and highly publicized variant of this question is this: What is the effect on election outcomes of laws that forbid individuals with felony convictions from voting?[3] Uggen and Manza (2002) make the straightforward claim that the 2000 presidential election would have gone in favor of Al Gore if felons and ex-felons had been permitted to vote:

> Although the outcome of the extraordinarily close 2000 presidential election could have been altered by a large number of factors, it would almost certainly have been reversed had voting rights been extended to any category of disenfranchised felons. (Uggen and Manza 2002:792)

Uggen and Manza (2002) then note an important limitation of their conclusion:

> ... our counterfactual examples rely upon a ceteris paribus assumption – that nothing else about the candidates or election would change save the voting rights of felons and ex-felons. (Uggen and Manza 2002:795)

When thinking about this important qualification, one might surmise that a possible world in which felons had the right to vote would probably also be a world in which the issues (and probably candidates) of the election would be very different. Thus, the most challenging definitional issue here is not who counts as a felon or whether or not an individual is disenfranchised, but rather how well the alternative causal states can be characterized.

As this example illustrates, it is important that the "what would have been" nature of the conditionals that define the causal states of interest be carefully

[2]Hong and Raudenbush (2006) provide a careful analysis of retention policies in U.S. primary education, implementing this type of treatment-site stratification based on the average level of retention in different schools.

[3]Behrens, Uggen, and Manza (2003), Manza and Uggen (2004), and Uggen, Behrens, and Manza (2005) give historical perspective on this question.

laid out. When a *ceteris paribus* assumption is relied on to rule out other contrasts that are nearly certain to occur at the same time, the posited causal states are open to the charge that they are too metaphysical to justify the pursuit of causal analysis.[4]

Given the existence of well-defined causal states, causal inference in the counterfactual tradition proceeds by stipulating the existence of potential outcome random variables that are defined over all individuals in the population of interest. For a binary cause, we will denote potential outcome random variables as Y^1 and Y^0.[5] We will also adopt the notational convention from statistics in which realized values for random variables are denoted by lowercase letters. Accordingly, y_i^1 is the potential outcome in the treatment state for individual i, and y_i^0 is the potential outcome in the control state for individual i. The individual-level causal effect of the treatment is then defined as

$$\delta_i = y_i^1 - y_i^0. \tag{2.1}$$

Individual-level causal effects can be defined in ways other than as a linear difference in the potential outcomes. For example, the individual-level causal effect could be defined instead as the ratio of one individual-level potential outcome to another, as in y_i^1/y_i^0. In some applications, there may be advantages to these sorts of alternative definitions at the individual level, but the overwhelming majority of the literature represents individual-level causal effects as linear differences, as in Equation (2.1).[6]

[4]This may well be the case with the felon disenfranchisement example, but this is a matter for scholars in political sociology and criminology to debate. Even if the charge sticks, this particular line of research is nonetheless still an important contribution to the empirical literature on how changing laws to allow felons and ex-felons to vote could have potential effects on election outcomes.

[5]There is a wide variety of notation in the potential outcome and counterfactuals literature, and we have adopted the notation that we feel is the easiest to grasp. However, we should note that Equation (2.1) and its elements are often written as one of the following alternatives:

$$\begin{aligned} \Delta_i &= Y_{1i} - Y_{0i}, \\ \delta_i &= Y_i^t - Y_i^c, \\ \tau_i &= y_i(1) - y_i(0), \end{aligned}$$

and variants thereof. We use the right-hand superscript to denote the potential treatment state of the corresponding potential outcome variable, but other authors use the right-hand subscript or parenthetical notation. We also use numerical values to refer to the treatment states, but other authors (including us, see Morgan 2001, Winship and Morgan 1999, and Winship and Sobel 2004) use values such as t and c for the treatment and control states, respectively. There is also variation in the usage of uppercase and lowercase letters. We do not claim that everyone will agree that our notation is the easiest to grasp, and it is certainly not as general as, for example, the parenthetic notation. But it does seem to have proven itself in our own classes, offering the right balance between specificity and compactness.

[6]Moreover, the individual-level causal effect could be defined as the difference between the expectations of individual-specific random variables, as in $E[Y_i^1] - E[Y_i^0]$, where $E[.]$ is the expectation operator from probability theory (see, for a clear example of this alternative setup, King et al. 1994:76-82). In thinking about individuals self-selecting into alternative treatment states, it can be useful to set up the treatment effects in this way. In many applications, individuals are thought to consider potential outcomes with some recognition of

2.2 Treatment Groups and Observed Outcomes

For a binary cause with two causal states and associated potential outcome variables Y^1 and Y^0, the convention in the counterfactuals literature is to define a causal exposure variable, D, which takes on two values: D is equal to 1 for members of population who are exposed to the treatment state and equal to 0 for members of the population who are exposed to the control state. Exposure to the alternative causal states is determined by a particular process, typically an individuals's decision to enter one state or another, an outside actor's decision to allocate individuals to one state or another, a planned random allocation carried out by an investigator, or some combination of these alternatives.

By convention, those who are exposed to the treatment state are referred to as the treatment group whereas those who are exposed to the control state are referred to as the control group. Because D is defined as a population-level random variable (at least in most cases in observational data analysis), the treatment group and control group exist in the population as well as the observed data. Throughout this book, we will use this standard terminology, referring to treatment and control groups when discussing those who are exposed to alternative states of a binary cause. If more than two causal states are of interest, then we will shift to the semantics of alternative treatments and corresponding treatment groups, thereby discarding the baseline labels of control state and control group. Despite our adoption of this convention, we could rewrite all that follows referring to members of the population as what they are: those who are exposed to alternative causal states.

When we refer to individuals in the observed treatment and control groups, we will again adopt the notational convention from statistics in which realized values for random variables are denoted by lowercase letters. Accordingly, the random variable D takes on values of $d_i = 1$ for each individual i who is an observed member of the treatment group and $d_i = 0$ for each individual i who is an observed member of the control group.

Given these definitions of Y^1, Y^0, and D (as well as their realizations y_i^1, y_i^0, d_i), we can now define the observed outcome variable Y in terms of them. We can observe values for a variable Y as $y_i = y_i^1$ for individuals with $d_i = 1$ and as $y_i = y_i^0$ for individuals with $d_i = 0$. The observable outcome variable Y is therefore defined as

$$\begin{aligned} Y &= Y^1 \text{ if } D = 1, \\ Y &= Y^0 \text{ if } D = 0. \end{aligned}$$

the inherent uncertainty of their beliefs, which may properly reflect true variability in their potential outcomes. But, when data for which a potential outcome is necessarily observed for any individual as a scalar value (via an observational outcome variable, defined later) are analyzed, this individual-level, random-variable definition is largely redundant. Accordingly, we will denote individual-level potential outcomes as values such as y_i^1 and y_i^0, regarding these as realizations of population-level random variables Y^1 and Y^0 while recognizing, at least implicitly, that they could also be regarded as realizations of individual-specific random variables Y_i^1 and Y_i^0.

Table 2.1: The Fundamental Problem of Causal Inference

Group	Y^1	Y^0
Treatment group $(D = 1)$	Observable as Y	Counterfactual
Control group $(D = 0)$	Counterfactual	Observable as Y

This paired definition is often written compactly as

$$Y = DY^1 + (1 - D)Y^0. \tag{2.2}$$

In words, one can never observe the potential outcome under the treatment state for those observed in the control state, and one can never observe the potential outcome under the control state for those observed in the treatment state. This impossibility implies that one can never calculate individual-level causal effects.

Holland (1986) describes this challenge as the fundamental problem of causal inference in his widely read introduction to the counterfactual model. Table 2.1 depicts the problem. Causal effects are defined within rows, which refer to groups of individuals in the treatment state or in the control state. However, only the diagonal of the table is observable, thereby rendering impossible the direct calculation of individual-level causal effects merely by means of observation and then subtraction.[7]

As shown clearly in Equation (2.2), the outcome variable Y, even if we could enumerate all of its individual-level values y_i in the population, reveals only half of the information contained in the underlying potential outcome variables. Individuals contribute outcome information only from the treatment state in which they are observed. This is another way of thinking about Holland's fundamental problem of causal inference. The outcome variables we must analyze – labor market earnings, test scores, and so on – contain only a portion of the information that would allow us to directly calculate causal effects for all individuals.

2.3 The Average Treatment Effect

Because it is typically impossible to calculate individual-level causal effects, we focus attention on the estimation of aggregated causal effects, usually alternative

[7] As Table 2.1 shows, we are more comfortable than some writers in using the label "counterfactual" when discussing potential outcomes. Rubin (2005), for example, avoids the term counterfactual, under the argument that potential outcomes become counterfactual only after treatment assignment has occurred. Thus no potential outcome is ever *ex ante* counterfactual. We agree, of course. But, because our focus is on observational data analysis, we find the counterfactual label useful for characterizing potential outcomes that are rendered unobservable *ex post* to the treatment assignment/selection mechanism.

average causal effects. With $E[.]$ denoting the expectation operator from probability theory, the average treatment effect in the population is

$$\begin{aligned} E[\delta] &= E[Y^1 - Y^0] \\ &= E[Y^1] - E[Y^0]. \end{aligned} \tag{2.3}$$

The second line of Equation (2.3) follows from the linearity of the expectation operator: The expectation of a difference is equal to the difference of the two expectations.[8]

For Equation (2.3), the expectation is defined with reference to the population of interest. For the political science examples in Chapter 1, the population could be "all eligible voters" or "all eligible voters in Florida." For other examples, such as the manpower training example, the population would be defined similarly as "all adults eligible for training," and eligibility would need to be defined carefully. Thus, to define average causal effects and then interpret estimates of them, it is crucial that researchers clearly define the characteristics of the individuals in the assumed population of interest.[9]

Note also that the subscripting on i for δ_i has been dropped for Equation (2.3). Even so, δ is not necessarily constant in the population, as it is a random variable just like Y^1 and Y^0. We can drop the subscript i in this equation because the causal effect of a randomly selected individual from the population is equal to the average causal effect across individuals in the population. We will at times throughout this book reintroduce redundant subscripting on i in order to reinforce the inherent individual-level heterogeneity of the potential outcomes and the causal effects they define, but we will be clear when we are doing so.

Consider the Catholic school example from Subsection 1.3.2 that demonstrates the relationship between observed and potential outcomes and how these are related to typical estimation of the average causal effect in Equation (2.3). For the Catholic school effect on learning, the potential outcome under the treatment, y_i^1, is the what-if achievement outcome of individual i if he or she were enrolled in a Catholic school. The potential outcome under the control, y_i^0, is the what-if achievement outcome of individual i if he or she were enrolled in a public school. Accordingly, the individual-level causal effect, δ_i, is the what-if difference in achievement that could be calculated if we could simultaneously educate individual i in both a Catholic school and a public school.[10] The average

[8] However, more deeply, it also follows from the assumption that the causal effect is defined as a linear difference at the individual level, which allows the application of expectations in this simple way to characterize population-level average effects.

[9] And, regardless of the characterization of the features of the population, we will assume throughout this book that the population is a realization of an infinite superpopulation. We discuss our decision to adopt this underlying population model in an appendix to this chapter. Although not essential to understanding most of the material in this book, some readers may find it helpful to read that appendix now in order to understand how these definitional issues are typically settled in this literature.

[10] However, it is a bit more complex than this. Now that we have introduced a real-world scenario, other assumptions must also be invoked, notably the stable unit treatment value assumption, introduced and explained in the next section.

causal effect, $E[\delta]$, is then the mean value among all students in the population of these what-if differences in test scores. In general, the average causal effect is equal to the expected value of the what-if difference in test scores for a randomly selected student from the population.

2.4 The Stable Unit Treatment Value Assumption

In most applications, the counterfactual model retains its transparency through the maintenance of a very simple but strong assumption known as the stable unit treatment value assumption or SUTVA (see Rubin 1980b, 1986). In economics, this is sometimes referred to as a no-macro-effect or partial equilibrium assumption (see Heckman 2000, 2005 and Garfinkel, Manski, and Michalopoulos 1992 for the history of these ideas and Manski and Garfinkel 1992 for examples). SUTVA, as implied by its name, is a basic assumption of causal effect stability that requires that the potential outcomes of individuals be unaffected by potential changes in the treatment exposures of other individuals. In the words of Rubin (1986:961), who developed the term,

> SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive.

Consider the idealized example in Table 2.2, in which SUTVA is violated because the treatment effect varies with treatment assignment patterns. For the idealized example, there are three randomly drawn subjects from a population of interest, and the study is designed such that at least one of the three study subjects must receive the treatment and at least one must receive the control. The first column of the table gives the six possible treatment assignment patterns.The first row of Table 2.2 presents all three ways to assign one individual to the treatment and the other two to the control, as well the potential outcomes for each of the three subjects. Subtraction within the last column shows that the individual-level causal effect is 2 for all three individuals. The second row of Table 2.2 presents all three ways to assign two individuals to the treatment and one to the control. As shown in the last column of the row, the individual-level causal effects implied by the potential outcomes are now 1 instead of 2. Thus, for this idealized example, the underlying causal effects are a function of the treatment assignment patterns, such that the treatment is less effective when more individuals are assigned to it. For SUTVA to hold, the potential outcomes would need to be identical for both rows of the table.

This type of treatment effect dilution is only one way in which SUTVA can be violated. More generally, suppose that $\mathbf{d}$ is an $N \times 1$ vector of treatment indicator variables for N individuals (analogous to the treatment assignment vectors in the first column of Table 2.2), and define potential outcomes generally as functions of the vector $\mathbf{d}$. The outcome for individual i under the

Table 2.2: A Hypothetical Example in Which SUTVA is Violated

Treatment assignment patterns			Potential outcomes	
$\begin{bmatrix} d_1 = 1 \\ d_2 = 0 \\ d_3 = 0 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 0 \\ d_2 = 1 \\ d_3 = 0 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 0 \\ d_2 = 0 \\ d_3 = 1 \end{bmatrix}$	$y_1^1 = 3$ $y_2^1 = 3$ $y_3^1 = 3$	$y_1^0 = 1$ $y_2^0 = 1$ $y_3^0 = 1$
$\begin{bmatrix} d_1 = 1 \\ d_2 = 1 \\ d_3 = 0 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 0 \\ d_2 = 1 \\ d_3 = 1 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 1 \\ d_2 = 0 \\ d_3 = 1 \end{bmatrix}$	$y_1^1 = 2$ $y_2^1 = 2$ $y_3^1 = 2$	$y_1^0 = 1$ $y_2^0 = 1$ $y_3^0 = 1$

treatment is $y_i^1(\mathbf{d})$, and the outcome for individual i under the control is $y_i^0(\mathbf{d})$. Accordingly, the individual-level causal effect for individual i is $\delta_i(\mathbf{d})$. SUTVA is what allows us to write $y_i^1 = y_i^1(\mathbf{d})$ and $y_i^0 = y_i^0(\mathbf{d})$ and, as a result, assert that individual-level causal effects δ_i exist that are independent of the assignment process itself.[11]

Sometimes it is argued that SUTVA is so restrictive that we need an alternative conception of causality for the social sciences. We agree that SUTVA is very sobering. However, our position is that SUTVA reveals the limitations of observational data and the perils of immodest causal modeling rather than the limitations of the counterfactual model itself. Rather than consider SUTVA as overly restrictive, researchers should always reflect on the plausibility of SUTVA in each application and use such reflection to motivate a clear discussion of the meaning and scope of a causal effect estimate.

Consider the example of the Catholic school effect again. For SUTVA to hold, the effectiveness of Catholic schooling cannot be a function of the number (and/or composition) of students who enter the Catholic school sector. For a variety of reasons – endogenous peer effects, capacity constraints, and so on – most school effects researchers would probably expect that the Catholic school effect would change if large numbers of public school students entered the Catholic school sector. As a result, because there are good theoretical reasons to believe that macro effects would emerge if Catholic school enrollments ballooned, it may be that researchers can estimate the causal effect of Catholic schooling only for those who would typically choose to attend Catholic schools, but also subject to the constraint that the proportion of students educated in Catholic schools remain relatively constant. Accordingly, it may be impossible to determine from any data that could be collected what the Catholic school effect on achievement would be under a new distribution of students across school sectors that would result from a large and effective policy intervention.

[11] In other words, if SUTVA is violated, then Equation (2.1) must be written in its most general form as $\delta_i(\mathbf{d}) = y_i^1(\mathbf{d}) - y_i^0(\mathbf{d})$. In this case, individual-level treatment effects could be different for every possible configuration of treatment exposures.

As a result, the implications of research on the Catholic school effect for research on school voucher programs (see Subsection 1.3.2) may be quite limited, and this has not been clearly enough recognized by some (see Howell and Peterson 2002, Chapter 6).

Consider also the manpower training example introduced in Subsection 1.3.2. Here, the suitability of SUTVA may depend on the particular training program. For small training programs situated in large labor markets, the structure of wage offers to retrained workers may be entirely unaffected by the existence of the training program. However, for a sizable training program in a small labor market, it is possible that the wages on offer to retrained workers would be a function of the way in which the price of labor in the local labor market responds to the movement of trainees in and out of the program (as might be the case in a small company town after the company has just gone out of business and a training program is established). As a result, SUTVA may be reasonable only for a subset of the training sites for which data have been collected.

Finally, consider SUTVA in the context of an example that we will not consider in much detail in this book: the evaluation of the effectiveness of mandatory school desegregation plans in the 1970s on the subsequent achievement of black students. Gathering together the results of a decade of research, Crain and Mahard (1983) conducted a meta-analysis of 93 studies of the desegregation effect on achievement. They argued that the evidence suggests an increase of .3 standard deviations in the test scores of black students across all studies.[12] It seems undeniable that SUTVA is violated for this example, as the effect of moving from one school to another must be a function of relative shifts in racial composition across schools. Breaking the analysis into subsets of cities where the compositional shifts were similar could yield average treatment effect estimates that can be more clearly interpreted. In this case, SUTVA would be abandoned in the collection of all desegregation events, but it could then be maintained for some groups (perhaps in cities where the compositional shift was relatively small).

In general, if SUTVA is maintained but there is some doubt about its validity, then certain types of marginal effect estimates can usually still be defended. The idea here would be to state that the estimates of average causal effects hold only for what-if movements of a very small number of individuals from one hypothetical treatment state to another. If more extensive what-if contrasts are of interest, such as would be induced by a widespread intervention, then SUTVA would need to be dropped and variation of the causal effect as a function of

[12] As reviewed by Schofield (1995) and noted in Clotfelter (2004), most scholars now accept that the evidence suggests that black students who were bused to predominantly white schools experienced small positive reading gains but no substantial mathematics gains. Cook and Evans (2000:792) conclude that "... it is unlikely that efforts at integrating schools have been an important part of the convergence in academic performance [between whites and blacks], at least since the early 1970s" (see also Armor 1995; Rossell, Armor, and Walberg 2002). Even so, others have argued that the focus on test score gains has obscured some of the true effectiveness of desegregation. In a review of these longer-term effects, Wells and Crain (1994:552) conclude that "interracial contact in elementary and secondary school can help blacks overcome perpetual segregation."

treatment assignment patterns would need to be modeled explicitly. This sort of modeling can be very challenging and generally requires a full model of causal effect exposure that is grounded on a believable theoretical model that sustains subtle predictions about alternative patterns of individual behavior. But it is not impossible, and it represents a frontier of research in many well-established causal controversies (see Heckman 2005, Sobel 2006).

2.5 Treatment Assignment and Observational Studies

A researcher who wishes to estimate the effect of a treatment that he or she can control on an outcome of interest typically designs an experiment in which subjects are randomly assigned to alternative treatment and control groups. Other types of experiments are possible, as we described earlier in Chapter 1, but randomized experiments are the most common research design when researchers have control over the assignment of the treatment.

After randomization of the treatment, the experiment is run and the values of the observed outcome, y_i, are recorded for those in the treatment group and for those in the control group. The mean difference in the observed outcomes across the two groups is then anointed the estimated average causal effect, and discussion (and any ensuing debate) then moves on to the particular features of the experimental protocol and the degree to which the pool of study participants reflects the population of interest for which one would wish to know the average treatment effect.

Consider this randomization research design with reference to the underlying potential outcomes defined earlier. For randomized experiments, the treatment indicator variable D is forced by design to be independent of the potential outcome variables Y^1 and Y^0. (However, for any single experiment with a finite set of subjects, the values of d_i will be related to the values of y_i^1 and y_i^0 because of chance variability.) Knowing whether or not a subject is assigned to the treatment group in a randomized experiment yields no information whatsoever about a subject's what-if outcome under the treatment state, y_i^1, or, equivalently, about a subject's what-if outcome under the control state, y_i^0. Treatment status is therefore independent of the potential outcomes, and the treatment assignment mechanism is said to be ignorable.[13] This independence assumption is usually written as

$$(Y^0, Y^1) \perp\!\!\!\perp D, \tag{2.4}$$

where the symbol $\perp\!\!\!\perp$ denotes independence and where the parentheses enclosing

[13]Ignorability holds in the weaker situation in which S is a set of observed variables that completely characterize treatment assignment patterns and in which $(Y^0, Y^1) \perp\!\!\!\perp D \mid S$. Thus treatment assignment is ignorable when the potential outcomes are independent of D, conditional on S. We will offer a more complete discussion of ignorability in the next three chapters.

Y^0 and Y^1 stipulate that D must be jointly independent of all functions of the potential outcomes (such as δ). For a properly run randomized experiment, learning the treatment to which a subject has been exposed gives no information whatsoever about the size of the treatment effect.

At first exposure, this way of thinking about randomized experiments and potential outcomes can be confusing. The independence relationships represented by Equation (2.4) seem to imply that even a well-designed randomized experiment cannot tell us about the causal effect of the treatment on the outcome of interest. But, of course, this is not so, as Equation (2.4) does not imply that D is independent of Y. If individuals are randomly assigned to both the treatment and the control states, and individual causal effects are nonzero, then the definition of the outcome variable, $Y = DY^1 + (1-D)Y^0$ in Equation (2.2), ensures that Y and D will be dependent.

Now consider the additional challenges posed by observational data analysis. It is the challenges to causal inference that are the defining features of an observational study according to Rosenbaum (2002:vii):

> An *observational study* is an empiric investigation of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects.

This definition is consistent with the Cox and Reid definition quoted in Chapter 1 (see page 7).

Observational data analysis in the counterfactual tradition is thus defined by a lack of control over the treatment [and, often more narrowly by the infeasibility of randomization designs that allow for the straightforward maintenance of the independence assumption in Equation (2.4)]. An observational researcher, hoping to estimate a causal effect, begins with observed data in the form of values $\{y_i, d_i\}_i^N$ for an observed outcome variable, Y, and a treatment status variable, D. To determine the causal of effect of D on Y, the first step in analysis is to investigate the treatment selection mechanism. Notice the switch in language from assignment to selection. Because observational data analysis is defined as empirical inquiry in which the researcher does not have the capacity to assign individuals to treatments (or, as Rosenbaum states equivalently, to assign treatments to individuals), researchers must instead investigate how individuals end up in alternative treatment states.

And herein lies the challenge of much scholarship in the social sciences. Although some of the process by which individuals select alternative treatments can be examined empirically, a full accounting of treatment selection is sometimes impossible (e.g., if subjects are motivated to select on the causal effect itself and a researcher does not have a valid measure of their expectations). As much as this challenge may be depressing to a dispassionate policy designer/evaluator, this predicament should not be depressing for social scientists in general. On the contrary, our existential justification rests on the pervasive need to deduce theoretically from a set of basic principles or infer from experience and knowledge of related studies the set of defendable assumptions about

the missing components of the treatment selection mechanism. Only through such effort can it be determined whether causal analysis can proceed or whether further data collection and preliminary theoretical analysis are necessary.

2.6 Average Causal Effects and Naive Estimation

As described in prior sections of this chapter, the fundamental problem of causal inference requires that we focus on non-individual-level causal effects, maintaining assumptions about treatment assignment and treatment stability that will allow us to give causal interpretations to differences in average values of observed outcomes. In the remainder of this chapter, we define average treatment effects of varying sorts and then lay out the complications of estimating them. In particular, we consider how average treatment effects vary across those who receive the treatment and those who do not.

2.6.1 Conditional Average Treatment Effects

The average causal effect, known as the average treatment effect in the counterfactual tradition, was defined in Equation (2.3) as $E[\delta] = E[Y^1 - Y^0]$. This average causal effect is the most common subject of investigation in the social sciences, and it is the causal effect that is closest to the sorts of effects investigated in the three broad foundational examples introduced in Chapter 1: the effects of family background and mental ability on educational attainment, the effects of educational attainment and mental ability on earnings, and the effects of socioeconomic status on political participation. More narrowly defined average causal effects are of interest as well in virtually all of the other examples introduced in Chapter 1.

Two conditional average treatment effects are of particular interest. The average treatment effect for those who typically take the treatment is

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1], \end{aligned} \tag{2.5}$$

and the average treatment effect for those who typically do not take the treatment is

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0], \end{aligned} \tag{2.6}$$

where, as for the average treatment effect in Equation (2.3), the second line of each definition follows from the linearity of the expectation operator. These two conditional average causal effects are often referred to by the acronyms ATT and ATC, which signify the average treatment effect for the treated and the average treatment effect for the controls, respectively.

Consider the examples again. For the Catholic school example, the average treatment effect for the treated is the average effect of Catholic schooling on the achievement of those who typically attend Catholic schools rather than across all students who could potentially attend Catholic schools. The difference between the average treatment effect and the average treatment effect for the treated can also be understood with reference to individuals. From this perspective, the average treatment effect in Equation (2.3) is the expected what-if difference in achievement that would be observed if we could educate a randomly selected student in both a public school and a Catholic school. In contrast, the average treatment effect for the treated in Equation (2.5) is the expected what-if difference in achievement that would be observed if we could educate a randomly selected Catholic school student in both a public school and a Catholic school.

For this example, the average treatment effect among the treated is a theoretically important quantity, for if there is no Catholic school effect for Catholic school students, then most reasonable theoretical arguments would maintain that it is unlikely that there would be a Catholic school effect for students who typically attend public schools (at least after adjustments for observable differences between Catholic and public school students). And, if policy interest were focused on whether or not Catholic schooling is beneficial for Catholic school students (and thus whether public support of transportation to Catholic schools is a benevolent government expenditure, etc.), then the Catholic school effect for Catholic school students is the only quantity we would want to estimate. The treatment effect for the untreated would be of interest as well if the goal of analysis is ultimately to determine the effect of a potential policy intervention, such as a new school voucher program, designed to move more students out of public schools and into Catholic schools. In fact, an even narrower conditional treatment effect might be of interest: $E[\delta|D = 0, \text{CurrentSchool} = \text{Failing}]$, where of course the definition of being currently educated in a failing school would have to be clearly specified.

The manpower training example is similar, in that the subject of first investigation is surely the treatment effect for the treated (as discussed in detail in Heckman et al. 1999). If a cost-benefit analysis of a program is desired, then a comparison of the aggregate net benefits for the treated to the overall costs of the program to the funders is needed. The treatment effect for other potential enrollees in the treatment program could be of interest as well, but this effect is secondary (and may be impossible to estimate for groups of individuals completely unlike those who have enrolled in the program in the past).

The butterfly ballot example is somewhat different as well. Here, the treatment effect of interest is bound by a narrow question that was shaped by media attention. The investigators were interested only in what actually happened in the 2000 election, and they focused very narrowly on whether the effect of having had a butterfly ballot rather than an optical scan ballot caused some individuals to miscast their votes. And, in fact, they were most interested in narrow subsets of the treated, for whom specific assumptions were more easily asserted and defended (e.g., those who voted for Democrats in all other races on the ballot but who voted for Pat Buchanan or Al Gore for president). In this

case, the treatment effect for the untreated, and hence the all-encompassing average treatment effect, was of little interest to the investigators (or to the contestants and the media).

As these examples demonstrate, more specific average causal effects (or more general properties of the distribution of causal effects) are often of greater interest than simply the average causal effect in the population. In this book, we will focus mostly on the three types of average causal effects represented by Equations (2.3), (2.5), and (2.6), as well as simple conditional variants of them. But, especially when presenting instrumental variable estimators later and discussing general heterogeneity issues, we will also focus on more narrowly defined causal effects. Heckman (2000), Manski (1995), and Rosenbaum (2002) all give full discussions of the variety of causal effects that may be relevant for different types of applications, such as quantiles of the distribution of individual-level causal effects in subpopulations of interest and the probability that the individual-level causal effect is greater than zero among the treated (see also Heckman, Smith, and Clements 1997).

2.6.2 Naive Estimation of Average Treatment Effects

Suppose again that randomization of the treatment is infeasible and thus that only an observational study is possible. Instead, an autonomous fixed treatment selection regime prevails, where π is the proportion of the population of interest that takes the treatment instead of the control. In this scenario, the value of π is fixed in the population by the behavior of individuals, and it is unknown. Suppose further that we have observed survey data from a relatively large random sample of the population of interest.

Because we are now shifting from the population to data generated from a random sample of the population, we must use appropriate notation to distinguish sample-based quantities from the population-based quantities that we have considered until now. For the sample expectation of a quantity in a sample of size N, we will use a subscript on the expectation operator, as in $E_N[.]$. With this notation, $E_N[d_i]$ is the sample mean of the dummy treatment variable, $E_N[y_i|d_i = 1]$ is the sample mean of the outcome for those observed in the treatment group, and $E_N[y_i|d_i = 0]$ is the sample mean of the outcome for those observed in the control group.[14] The naive estimator of the average causal effect is then defined as

$$\hat{\delta}_{\text{NAIVE}} \equiv E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0], \tag{2.7}$$

which is simply the difference in the sample means of the observed outcome variable Y for the observed treatment and control groups.

In observational studies, the naive estimator rarely yields a consistent estimate of the average treatment effect because it converges to a contrast,

[14] In other words, the subscript N serves the same basic notational function as an overbar on y_i, as in $\bar{y}_i$. We use this sub-N notation, as it allows for greater clarity in aligning sample and population-level conditional expectations for subsequent expressions.

$E[Y|D = 1] - E[Y|D = 0]$, that is not equivalent to (and usually not equal to) any of the average causal effects defined earlier. To see why, decompose the average treatment effect in Equation (2.3) as

$$\begin{aligned} E[\delta] = &\{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &- \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\}. \end{aligned} \tag{2.8}$$

The average treatment effect is then a function of five unknowns: the proportion of the population that is assigned to (or self-selects into) the treatment along with four conditional expectations of the potential outcomes. Without introducing additional assumptions, we can consistently estimate with observational data from a random sample of the population only three of the five unknowns on the right-hand side of Equation (2.8), as we now show.

We know that, for a very large random sample, the mean of realized values for the dummy treatment variable D would be equal to the true proportion of the population that would be assigned to (or would select into) the treatment. More precisely, we know that the sample mean of the values d_i converges in probability to π, which we write as

$$E_N[d_i] \xrightarrow{p} \pi. \tag{2.9}$$

Although the notation of Equation (2.9) may appear unfamiliar, the claim is that, as the sample size N increases, the sample mean of the values d_i approaches the true value of π, which we assume is a fixed population parameter equal exactly to $E[D]$. Thus, the notation $\xrightarrow{p}$ denotes convergence in probability for a sequence of estimates over a set of samples where the sample size N is increasing to infinity.[15] We can offer similar claims about two other unknowns in Equation (2.8):

$$E_N[y_i|d_i = 1] \xrightarrow{p} E[Y^1|D = 1], \tag{2.10}$$

$$E_N[y_i|d_i = 0] \xrightarrow{p} E[Y^0|D = 0], \tag{2.11}$$

which indicate that the sample mean of the observed outcome in the treatment group converges to the true average outcome under the treatment state for those in the treatment group (and analogously for the control group and control state).

Unfortunately, however, there is no assumption-free way to effectively estimate the two remaining unknowns in Equation (2.8): $E[Y^1|D = 0]$ and $E[Y^0|D = 1]$. These are counterfactual conditional expectations: the average outcome under the treatment for those in the control group and the average outcome under the control for those in the treatment group. Without further assumptions, no estimated quantity based on observed data from a random sample of the population of interest would converge to the true values for these unknown counterfactual conditional expectations. For the Catholic school example, these are the average achievement of public school students if they had instead been

[15] Again, see our appendix to this chapter on our assumed superpopulation model.

educated in Catholic schools and the average achievement of Catholic school students if they had instead been educated in public schools.

2.6.3 Expected Bias of the Naive Estimator

In the last subsection, we noted that the naive estimator $\hat{\delta}_{\text{NAIVE}}$, which is defined as $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$, converges to $E[Y^1|D = 1] - E[Y^0|D = 0]$. In this subsection, we show why this contrast can be uninformative about the causal effect of interest in an observational study by analyzing the expected bias in the naive estimator as an estimator of the average treatment effect.[16] Consider the following rearrangement of the decomposition in Equation (2.8):

$$\begin{aligned} E[Y^1|D = 1] - E[Y^0|D = 0] \quad &= \quad E[\delta] \\ &\quad + \{E[Y^0|D = 1] - E[Y^0|D = 0]\} \\ &\quad + (1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\}. \end{aligned} \tag{2.12}$$

The naive estimator converges to the left-hand side of this equation, and thus the right-hand side shows both the true average treatment effect, $E[\delta]$, plus the expectations of two potential sources of expected bias in the naive estimator.[17] The first source of potential bias, $\{E[Y^0|D = 1] - E[Y^0|D = 0]\}$, is a *baseline bias* equal to the difference in the average outcome in the absence of the treatment between those in the treatment group and those in the control group. The second source of potential bias, $\{(1 - \pi)E[\delta|D = 1] - E[\delta|D = 0]\}$, is a *differential treatment effect bias* equal to the expected difference in the treatment effect between those in the treatment and those in the control group (multiplied by the proportion of the population under the fixed treatment selection regime that does not select into the treatment).

To clarify this decomposition of the bias of the naive estimator, consider a substantive example – the effect of education on an individual's mental ability. Assume that the treatment is college attendance. After administering a test to a group of young adults, we find that individuals who have attended college score higher than individuals who have not attended college. There are three possible reasons that we might observe this finding. First, attending college might make individuals smarter on average. This effect is the average treatment effect, represented by $E[\delta]$ in Equation (2.12). Second, individuals who

[16] An important point of this literature is that the bias of an estimator is a function of what is being estimated. Because there are many causal effects that can be estimated, general statements about the bias of particular estimators are always conditional on a clear indication of the causal parameter of interest.

[17] The referenced rearrangement is simply a matter of algebra. Let $E[\delta] = e$, $E[Y^1|D = 1] = a$, $E[Y^1|D = 0] = b$, $E[Y^0|D = 1] = c$, and $E[Y^0|D = 0] = d$ so that Equation (2.8) can be written more compactly as $e = \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\}$. Rearranging this expression as $a - d = e + a - b - \pi a + \pi b + \pi c - \pi d$ then simplifies to $a - d = e + \{c - d\} + \{(1 - \pi)[(a - c) - (b - d)]\}$. Substituting for a, b, c, d, and e then yields Equation (2.12).

Table 2.3: An Example of the Bias of the Naive Estimator

Group	$E[Y^1\|.]$	$E[Y^0\|.]$
Treatment group ($D = 1$)	10	6
Control group ($D = 0$)	8	5

attend college might have been smarter in the first place. This source of bias is the baseline difference represented by $E[Y^0|D = 1] - E[Y^0|D = 0]$. Third, the mental ability of those who attend college may increase more than would the mental ability of those who did not attend college if they had instead attended college. This source of bias is the differential effect of the treatment, represented by $E[\delta|D = 1] - E[\delta|D = 0]$.

To further clarify the last term in the decomposition, consider the alternative hypothetical example depicted in Table 2.3. Suppose, for context, that the potential outcomes are now some form of labor market outcome, and that the treatment is whether or not an individual has obtained a college degree. Suppose further that 30 percent of the population obtains college degrees, such that π is equal to .3. As shown on the main diagonal of Table 2.3, the average (or expected) potential outcome under the treatment is 10 for those in the treatment group, and the average (or expected) potential outcome under the control for those in the control group is 5. Now, consider the off-diagonal elements of the table, which represent the counterfactual average potential outcomes. According to these values, those who have college degrees would have done better in the labor market than those without college degrees in the counterfactual state in which they did not in fact obtain college degrees (i.e., on average they would have received 6 instead of 5). Likewise, those who do not obtain college degrees would not have done as well as those who did obtain college degrees in the counterfactual state in which they did in fact obtain college degrees (i.e., on average they would have received only 8 instead of 10). Accordingly, the average treatment effect for the treated is 4, whereas the average treatment effect for the untreated is only 3. Finally, if the proportion of the population that completes college is .3, then the average treatment effect is 3.3, which is equal to $.3(10 - 6) + (1 - .3)(8 - 5)$.

Consider now the bias in the naive estimator. For this example, the naive estimator, as defined in Equation (2.7), would be equal to 5, on average, across repeated samples from the population (i.e., because $E[Y^1|D = 1] - E[Y^0|D = 0] = 10 - 5$). Thus, over repeated samples, the naive estimator would be upwardly biased for the average treatment effect (i.e., yielding 5 rather than 3.3), the average treatment effect for the treated (i.e., yielding 5 rather than 4), and the average treatment effect for the untreated (i.e., yielding 5 rather than 3). Equation (2.12) gives the components of the total expected bias of 1.7 for the naive estimator as an estimate of the average treatment effect. The term $\{E[Y^0|D = 1] - E[Y^0|D = 0]\}$, which we labeled the expected baseline bias, is

$6 - 5 = 1$. The term $(1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\}$, which is the expected differential treatment effect bias, is $(1 - .3)(4 - 3) = .7$.[18]

2.6.4 Estimating Causal Effects Under Maintained Assumptions About Potential Outcomes

What assumptions suffice to enable unbiased and consistent estimation of the average treatment effect with the naive estimator? There are two basic classes of assumptions: (1) assumptions about potential outcomes for subsets of the population defined by treatment status and (2) assumptions about the treatment assignment/selection process in relation to the potential outcomes. These two types of assumptions are variants of each other, and each may have a particular advantage in motivating analysis in a particular application.

In this section, we discuss only the first type of assumption, as it suffices for the present examination of the fallibility of the naive estimator. And our point in introducing these assumptions is simply to explain in one final way why the naive estimator will fail in most social science applications to generate an unbiased and consistent estimate of the average causal effect when randomization of the treatment is infeasible.

Consider the following two assumptions:

$$\text{Assumption 1:} \quad E[Y^1|D = 1] = E[Y^1|D = 0], \tag{2.13}$$

$$\text{Assumption 2:} \quad E[Y^0|D = 1] = E[Y^0|D = 0]. \tag{2.14}$$

If one asserts these two equalities and then substitutes into Equation (2.8), the number of unknowns is reduced from the original five parameters to the three parameters that we know from Equations (2.9)–(2.11) can be consistently estimated with data generated from a random sample of the population. If both Assumptions 1 and 2 are maintained, then the average treatment effect, the average treatment effect for the treated, and the average treatment effect for the untreated in Equations (2.3), (2.5), and (2.6), respectively, are all equal. And the naive estimator is consistent for all of them.

When would Assumptions 1 and 2 in Equations (2.13) and (2.14) be reasonable? Clearly, if the independence of potential outcomes, as expressed in Equation (2.4), is valid because the treatment has been randomly assigned, then Assumptions 1 and 2 in Equations (2.13) and (2.14) are implied. But, for observational data analysis, for which random assignment is infeasible, these assumptions would rarely be justified.

Consider the Catholic school example introduced in Subsection 1.3.2. If one were willing to assume that those who choose to attend Catholic schools

[18] In general, the amount of this expected differential treatment effect bias declines as more of the population is characterized by the treatment effect for the treated than by the treatment effect for the untreated (i.e., as π approaches 1).

do so for completely random reasons, then these two assumptions could be asserted. But we know from the applied literature that this characterization of treatment selection is false. Nonetheless, one might be able to assert instead a weaker narrative to warrant these two assumptions. One could maintain that students and their parents make enrollment decisions based on tastes for an education with a religious foundation and that this taste is unrelated to the two potential outcomes, such that those with a taste for the religious foundations of education would not necessarily benefit more from actually being educated in a Catholic school than in other schools. This possibility also seems unlikely, in part because it implies that those with a distaste for a religious education do not attend Catholic schools and it seems reasonable to assume that they would perform substantially worse in a Catholic school than the typical student who does attend a Catholic school.

Thus, at least for the Catholic school example, there seems no way to justify the naive estimator as an unbiased and consistent estimator of the average treatment effect (or of the average treatment effect for the treated and the average treatment effect for the untreated). We encourage the reader to consider all of the examples presented in the first chapter, and we suspect that all will agree that Assumptions 1 and 2 in Equations (2.13) and (2.14) cannot be sustained for any of them.

But it is important to recognize that assumptions such as these can (and should) be evaluated separately. Consider the two relevant cases for Assumptions 1 and 2:

1. If Assumption 1 is true but Assumption 2 is not, then $E[Y^1|D = 1] = E[Y^1|D = 0]$ whereas $E[Y^0|D = 1] \neq E[Y^0|D = 0]$. In this case, the naive estimator remains biased and inconsistent for the average treatment effect, but it is now unbiased and consistent for the average treatment effect for the untreated. This result is true because of the same sort of substitution we noted earlier. We know that the naive estimator $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$ converges to $E[Y^1|D = 1] - E[Y^0|D = 0]$. If Assumption 1 is true, then one can substitute $E[Y^1|D = 0]$ for $E[Y^1|D = 1]$. Then, one can state that the naive estimator converges to the contrast $E[Y^1|D = 0] - E[Y^0|D = 0]$ when Assumption 1 is true. This contrast is defined in Equation (2.6) as the average treatment effect for the untreated.

2. If Assumption 2 is true but Assumption 1 is not, then $E[Y^0|D = 1] = E[Y^0|D = 0]$ whereas $E[Y^1|D = 1] \neq E[Y^1|D = 0]$. The opposite result to the prior case follows. One can substitute $E[Y^0|D = 1]$ for $E[Y^0|D = 0]$ in the contrast $E[Y^1|D = 1] - E[Y^0|D = 0]$. Then, one can state that the naive estimator converges to the contrast $E[Y^1|D = 1] - E[Y^0|D = 1]$ when Assumption 2 is true. This contrast is defined in Equation (2.5) as the average treatment effect for the treated.

Considering the validity of Assumptions 1 and 2 separately shows that the naive estimator may be biased and inconsistent for the average treatment effect and yet may be unbiased and consistent for either the average treatment

effect for the treated or the average treatment effect for the untreated. These possibilities can be important in practice. For some applications, it may be the case that we have good theoretical reason to believe that (1) Assumption 2 is valid because those in the treatment group would, on average, do no better or no worse under the control than those in the control group, and (2) Assumption 1 is invalid because those in the control group would not do nearly as well under the treatment as those in the treatment group. Under this scenario, the naive estimator will deliver an unbiased and consistent estimate of the average treatment effect for the treated, even though it is still biased and inconsistent for both the average treatment effect for the untreated and the unconditional average treatment effect.

Now, return to the case in which neither Assumption 1 nor Assumption 2 is true. If the naive estimator is therefore biased and inconsistent for the typical average causal effect of interest, what can be done? The first recourse is to attempt to partition the sample into subgroups within which assumptions such as Assumptions 1 and/or 2 can be defended. The strategy amounts to conditioning on one or more variables that identify such strata and then asserting that the naive estimator is unbiased and consistent within these strata for one of the average treatment effects. One can then average estimates from these strata in a reasonable way to generate the average causal effect estimate of interest. We turn, in the next part of the book, to the two major conditioning strategies – matching and regression analysis – for estimating average causal effects when the naive estimator is biased and inconsistent.

2.7 Conclusions

In this chapter, we have introduced the main components of the counterfactual model of causality, also known as the potential outcome model. To motivate the presentation of matching and regression in the next part of the book, we first reintroduce causal graphs and the notational framework for modeling the treatment assignment mechanism in the next chapter. Although we will then show that matching and regression share many connections, we also aim to demonstrate that they are typically motivated in two entirely different ways, as, in the first case, an attempt to balance the variables that predict treatment assignment/selection and as, in the second case, an attempt to condition on all other relevant direct causes of the outcome. The causal graphs show this connection clearly, and hence we begin by showing how conditioning strategies represent an attempt to eliminate all net associations between the causal variable and the outcome variable that are produced by back-door paths that confound the causal effect of interest.

Appendix A: Population and Data Generation Models

In the counterfactual tradition, no single agreed-on way to define the population exists. In a recent piece, for example, Rubin (2005:323) introduces the primary elements of the potential outcome model without taking any particular position on the nature of the population, writing that "'summary' causal effects can also be defined at the level of collections of units, such as the mean unit-level causal effect for all units." As a result, a variety of possible population-based (and "collection"-based) definitions of potential outcomes, treatment assignment patterns, and observed outcomes can be used. In this appendix, we explain the choice of population model that we will use throughout the book (and implicitly, unless otherwise specified).

Because we introduce populations, samples, and convergence claims in this chapter, we have placed this appendix here. Nonetheless, because we have not yet introduced models of causal exposure, some of the fine points in the following discussion may well appear confusing (notably how "nature" performs randomized experiments behind our backs). For readers who wish to have a full understanding of the implicit superpopulation model we will adopt, we recommend a quick reading of this appendix now and then a second more careful reading after completing Chapters 3 and 4.

Our Implicit Superpopulation Model

The most expedient population and data generation model to adopt is one in which the population is regarded as a realization of an infinite superpopulation. This setup is the standard perspective in mathematical statistics, in which random variables are assumed to exist with fixed moments for an uncountable and unspecified universe of events. For example, a coin can be flipped an infinite number of times, but it is always a Bernoulli distributed random variable for which the expectation of a fair coin is equal to .5 for both heads and tails. For this example, the universe of events is infinite because the coin can be flipped forever.

Many presentations of the potential outcome framework adopt this basic setup, presumably following Rubin (1977) and Rosenbaum and Rubin (1983b, 1985a). For a binary cause, potential outcomes Y^1 and Y^0 are implicitly assumed to have expectations $E[Y^1]$ and $E[Y^0]$ in an infinite superpopulation. Individual realizations of Y^1 and Y^0 are then denoted y_i^1 and y_i^0. These realizations are usually regarded as fixed characteristics of each individual i.

This perspective is tantamount to assuming a population machine that spawns individuals forever (i.e., the analog to a coin that can be flipped forever). Each individual is born as a set of random draws from the distributions of Y^1, Y^0, and additional variables collectively denoted by S. These realized values y^1, y^0, and s are then given individual identifiers i, which then become y_i^1, y_i^0, and s_i.

The challenge of causal inference is that nature also performs randomized experiments in the superpopulation. In particular, nature randomizes a causal variable D within strata defined by the values of S and then sets the value of Y as y_i equal to y_i^1 or y_i^0, depending on the treatment state that is assigned to each individual. If nature assigns an individual to the state $D = 1$, nature then sets y_i equal to y_i^1. If nature assigns an individual to the state $D = 0$, nature then sets y_i equal to y_i^0. The differential probability of being assigned to $D = 1$ instead of $D = 0$ may be a function in S, depending on the experiment that nature has decided to conduct (see Chapters 3 and 4). Most important, nature then deceives us by throwing away y_i^1 and y_i^0 and giving us only y_i.

In our examples, a researcher typically obtains data from a random sample of size N from a population, which is in the form of a dataset $\{y_i, d_i, s_i\}_{i=1}^N$. The sample that generates these data is drawn from a finite population that is itself only one realization of a theoretical superpopulation. Based on this set-up, the joint probability distribution in the sample $\Pr_N(Y, D, S)$ must converge in probability to the true joint probability distribution in the superpopulation $\Pr(Y, D, S)$ as the sample size approaches infinity. The main task for analysis is to model the relationship between D and S that nature has generated in order use observed data on Y to estimate causal effects defined by Y^1 and Y^0.

Because of its expediency, we will usually write with this superpopulation model in the background, even though the notions of infinite superpopulations and sequences of sample sizes approaching infinity are manifestly unrealistic. We leave the population and data generation model largely in the background in the main text, so as not to distract the reader from the central goals of our book.

Alternative Perspectives

There are two main alternative models of the population that we could adopt. The first, which is consistent with the most common starting point of the survey sampling literature (e.g., Kish 1965), is one in which the finite population is recognized as such but treated as so large that is convenient to regard it as infinite. Here, values of a sample statistic (such as a sample mean) are said to equal population values in expectation, but now the expectation is taken over repeated samples from the population (see Thompson 2002 for an up-to-date accounting of this perspective). Were we to adopt this perspective, rather than our superpopulation model, much of what we write would be the same. However, this perspective tends to restrict attention to large survey populations (such as all members of the U.S. population older than 18) and makes it cumbersome to discuss some of the estimators we will consider (e.g., in Chapter 4, where we will sometimes define causal effects only across the common support of some random variables, thereby necessitating a redefinition of the target population).

The second alternative is almost certainly much less familiar to many empirical social scientists but is a common approach within the counterfactual causality literature. It is used often when no clearly defined population exists from which the data can be said to be a random sample (such as when a collection

of data of some form is available and an analyst wishes to estimate the causal effect for those appearing in the data). In this situation, a dataset exists as a collection of individuals, and the observed individuals are assumed to have fixed potential outcomes y_i^1 and y_i^0. The fixed potential outcomes have average values for those in the study, but these average values are not typically defined with reference to a population-level expectation. Instead, analysis proceeds by comparison of the average values of y_i for those in the treatment and control groups with all other possible average values that could have emerged under all possible permutations of treatment assignment. This perspective then leads to a form of randomization inference, which has connections to exact statistical tests of null hypotheses most commonly associated with Fisher (1935). As Rosenbaum (2002) shows, many of the results we present in this book can be expressed in this framework (see also Rubin 1990, 1991). But the combinatoric apparatus required for doing so can be cumbersome (and at times requires constraints, such as homogeneity of treatment effects, that are restrictive). Nonetheless, because the randomization inference perspective has some distinct advantages in some situations, we will refer to it at several points throughout the book. And we strongly recommend that readers consult Rosenbaum (2002) if the data under consideration arise from a sample that has no straightforward and systematic connection to a well-defined population. In this case, sample average treatment effects may be the only well-defined causal effects, and, if so, then the randomization inference tradition is a clear choice.

Appendix B: Extension of the Framework to Many-Valued Treatments

In this chapter, we have focused discussion mostly on binary causal variables, conceptualized as dichotomous variables that indicate whether individuals are observed in treatment and control states. As we show here, the counterfactual framework can be used to analyze causal variables with more than two categories.

Potential and Observed Outcomes for Many-Valued Treatments

Consider the more general setup, in which we replace the two-valued causal exposure variable, D, and the two potential outcomes Y^1 and Y^0 with (1) a set of J treatment states, (2) a corresponding set of J causal exposure dummy variables, $\{Dj\}_{j=1}^J$, and (3) a corresponding set of J potential outcome random variables, $\{Y^{Dj}\}_{j=1}^J$. Each individual receives only one treatment, which we denote Dj^*. Accordingly, the observed outcome variable for individual i, y_i, is then equal to $y_i^{Dj^*}$. For the other $J-1$ treatments, the potential outcomes of individual i exist in theory as $J-1$ other potential outcomes y_i^{Dj} for $j \neq j^*$, but they are counterfactual.

Consider the fundamental problem of causal inference for many-value treatments presented in Table 2.4 (which is simply an expansion of Table 2.1 to

Table 2.4: The Fundamental Problem of Causal Inference for Many-Valued Treatments

Group	Y^{D1}	Y^{D2}	$\cdots$	Y^{DJ}
Takes $D1$	Observable as Y	Counterfactual	$\cdots$	Counterfactual
Takes $D2$	Counterfactual	Observable as Y		Counterfactual
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Takes DJ	Counterfactual	Counterfactual	$\cdots$	Observable as Y

many-valued treatments). Groups exposed to alternative treatments are represented by rows with, for example, those who take treatment $D2$ in the second row. For a binary treatment, we showed earlier that the observed variable Y contains exactly half of the information contained in the underlying potential outcome random variables. In general, for a treatment with J values, Table 2.4 shows that the observed outcome variable Y contains only $1/J$ of the total amount of information contained in the underlying potential outcome random variables. Thus, the proportion of unknown and inherently unobservable information increases as the number of treatment values, J, increases.

For an experimentalist, this decline in the relative amount of information in Y is relatively unproblematic. Consider an example in which a researcher wishes to know the relative effectiveness of three pain relievers for curing headaches. The four treatments are "Take nothing," "Take aspirin," "Take ibuprofen," and "Take acetaminophen." Suppose that the researcher rules out an observational study, in part because individuals have constrained choices (i.e., pregnant women may take acetaminophen but cannot take ibuprofen; many individuals take a daily aspirin for general health reasons). Instead, she gains access to a large pool of subjects not currently taking any medication and not prevented from taking any of the three medicines.[19] She divides the pool randomly into four groups, and the drug trial is run. Assuming all individuals follow the experimental protocol, at the end of the data collection period the researcher calculates the mean length and severity of headaches for each of the four groups.

Even though three quarters of the cells in a 4×4 observability table analogous to Table 2.4 are counterfactual, she can effectively estimate the relative effectiveness of each of the drugs in comparison with each other and in comparison with the take-nothing control group. Subject to random error, contrasts such as $E_N[y_i|\text{Take aspirin}] - E_N[y_i|\text{Take ibuprofen}]$ reveal all of the average treatment effects of interest. The experimental design allows her to ignore the counterfactual cells in the observability table by assumption. In other words, she can assume that the average counterfactual value of Y^{Aspirin} for those who

[19] Note that, in selecting this group, she has adopted a definition of the population of interest that does not include those who (1) take one of these pain relievers regularly for another reason and (2) do not have a reason to refuse to take one of the pain relievers.

Table 2.5: The Observability Table for Estimating how Education Increases Earnings

Education	Y^{HS}	Y^{AA}	Y^{BA}	Y^{MA}
Obtains HS	Observable as Y	Counterfactual	Counterfactual	Counterfactual
Obtains AA	Counterfactual	Observable as Y	Counterfactual	Counterfactual
Obtains BA	Counterfactual	Counterfactual	Observable as Y	Counterfactual
Obtains MA	Counterfactual	Counterfactual	Counterfactual	Observable as Y

took nothing, took ibuprofen, and took acetaminophen (i.e., $E[Y^{\text{Aspirin}}|\text{Take nothing}]$, $E[Y^{\text{Aspirin}}|\text{Take ibuprofen}]$, and $E[Y^{\text{Aspirin}}|\text{Take acetaminophen}]$) can all be assumed to be equal to the average observable value of Y for those who take the treatment aspirin, $E[Y|\text{Take aspirin}]$. She can therefore compare sample analogs of the expectations in the cells of the diagonal of the observability table, and she does not have to build contrasts within its rows. Accordingly, for this type of example, comparing the effects of multiple treatments with each other is no more complicated than the bivariate case, except insofar as one nonetheless has more treatments to assign and resulting causal effect estimates to calculate.

Now consider a variant on the education-earnings example from the first chapter. Suppose that a researcher hopes to estimate the causal effect of different educational degrees on labor market earnings, and further that only four degrees are under consideration: a high school degree (HS), an associate's degree (AA), a bachelor's degree (BA), and a master's degree (MA). For this problem, we therefore have four dummy treatment variables corresponding to each of the treatment states: HS, AA, BA, and MA. Table 2.5 has the same structure as Table 2.4. Unlike the pain reliever example, random assignment to the four treatments is impossible. Consider the most important causal effect of interest for policy purposes, $E[Y^{\text{BA}} - Y^{\text{HS}}]$, which is the average effect of obtaining a bachelor's degree instead of a high school degree.

Suppose that an analyst has survey data on a set of middle-aged individuals for whom earnings at the most recent job and highest educational degree is recorded. To estimate this effect without asserting any further assumptions, the researcher would need to be able to consistently estimate population-level analogs to the expectations of all of the cells of Table 2.5 in columns 1 and 3, including six counterfactual cells off of the diagonal of the table. The goal would be to formulate consistent estimates of $E[Y^{\text{BA}} - Y^{\text{HS}}]$ for all four groups of differentially educated adults. To obtain a consistent estimate of $E[Y^{\text{BA}} - Y^{\text{HS}}]$, the researcher would need to be able to consistently estimate $E[Y^{\text{BA}} - Y^{\text{HS}}|HS = 1]$, $E[Y^{\text{BA}} - Y^{\text{HS}}|AA = 1]$, $E[Y^{\text{BA}} - Y^{\text{HS}}|BA = 1]$, and $E[Y^{\text{BA}} - Y^{\text{HS}}|MA = 1]$, after which these estimates would be averaged across the distribution of educational attainment. Notice that this requires the consistent estimation of some doubly counterfactual contrasts, such as the effect on earnings of shifting from

a high school degree to a bachelor's degree for those who are observed with a master's degree. The researcher might boldly assert that the wages of all high school graduates are, on average, equal to what all individuals would obtain in the labor market if they instead had high school degrees. But this is very likely to be a mistaken assumption if it is the case that those who carry on to higher levels of education would have been judged more productive workers by employers even if they had not attained more than high school degrees.

As this example shows, a many-valued treatment creates substantial additional burden on an analyst when randomization is infeasible. For any two-treatment comparison, one must find some way to estimate a corresponding $2(J-1)$ counterfactual conditional expectations, because treatment contrasts exist for individuals in the population whose observed treatments place them far from the diagonal of the observability table.

If estimating all of these counterfactual average outcomes is impossible, analysis can still proceed in a more limited fashion. One might simply define the parameter of interest very narrowly, such as the average causal effect of a bachelor's degree only for those who typically attain high school degrees: $E[Y^{\text{BA}} - Y^{\text{HS}}|\text{HS} = 1]$. In this case, the causal effect of attaining a bachelor's degree for those who typically attain degrees other than a high school degree are of no interest for the analyst.

Alternatively, there may be reasonable assumptions that one can invoke to simplify the complications of estimating all possible counterfactual averages. For this example, many theories of the relationship between education and earnings suggest that, for each individual i, $y_i^{\text{HS}} \leq y_i^{\text{AA}} \leq y_i^{\text{BA}} \leq y_i^{\text{MA}}$. In other words, earnings never decrease as one obtains a higher educational degree. Asserting this assumption (i.e., taking a theoretical position that implies it) may allow one to ignore some cells of the observability table that are furthest from the direct comparison one hopes to estimate.

Other Aspects of the Counterfactual Model for Many-Valued Treatments

Aside from the expansion of the number of causal states, and thus also treatment indicator variables and corresponding potential outcome variables, all other features of the counterfactual model remain essentially the same. SUTVA must still be maintained, and, if it is unreasonable, then more general methods must again be used to model treatment effects that may vary with patterns of treatment assignment. Modeling treatment selection remains the same, even though the added complexity of having to model movement into and out of multiple potential treatment states can be taxing. And the same sources of bias in standard estimators must be considered, only here again the complexity can be considerable when there are multiple states beneath each contrast of interest.

To avoid all of this complexity, one temptation is to assume that treatment effects are linear additive in an ordered set of treatment states. For the effect of education on earnings, a researcher might instead choose to move forward under the assumption that the effect of education on earnings is linear additive

in the years of education attained. For this example, the empirical literature has demonstrated that this is a particularly poor idea. For the years in which educational degrees are typically conferred, individuals appear to receive an extra boost in earnings. When later discussing the estimation of treatment effects using linear regression for many-valued treatments, we will discuss a piece by Angrist and Krueger (1999) that shows very clearly how far off the mark these methods can be when motivated by unreasonable linearity and additivity assumptions.

Part 2: Estimating Causal Effects by Conditioning

Chapter 3

Causal Graphs, Identification, and Models of Causal Exposure

In this chapter, we present the basic conditioning strategy for the identification and estimation of causal effects. After introducing a methodology for building causal graphs, we present what has become known as the back-door criterion for sufficient conditioning to identify a causal effect. We then present models of causal exposure, introducing the treatment assignment and treatment selection literature from statistics and econometrics. We then return to the back-door criterion and discuss the two basic motivations of conditioning – balancing determinants of the cause of interest and adjusting for other causes of the outcome. We conclude with a discussion of the identification and estimation of conditional average causal effects by conditioning.

3.1 Causal Graphs and Conditioning as Back-Door Identification

In his 2000 book titled *Causality: Models, Reasoning, and Inference*, Judea Pearl lays out a powerful and extensive graphical theory of causality. Here, we present and use only the most basic elements of his theory. To the reader familiar with traditional linear path models, much of this material will look familiar. There are, however, important and subtle differences between traditional path models and Pearl's usage of directed acyclic graphs (DAGs).

Pearl's work provides a language and a framework for thinking about causality that differs from the potential outcome perspective presented in the last chapter. Beyond the alternative terminology and notation, Pearl (2000, Section 7.3) proves that the fundamental concepts underlying the potential outcome model and his more recent perspective are equivalent. In some cases, causal

statements in the potential outcome framework can be represented concisely by a causal graph. But it can be awkward to represent many of the complications created by causal effect heterogeneity. Accordingly, in this section we suppress potential outcome random variables and use only observed outcome variables.[1] Furthermore, we implicitly focus on only the unconditional average treatment effect first, although we will return to a discussion of the estimation of conditional average treatment effects in the final section of the chapter.

Even though we must suppress some of the very useful generality of the potential outcome framework, Pearl has shown that graphs nonetheless provide a direct and powerful way of thinking about causal systems of variables and the identification strategies that can be used to estimate the effects within them. Thus, some of the advantage of the framework is precisely that it permits suppression of what could be a dizzying amount of notation to reference each potential causal state in a system of equations. In this sense, Pearl's perspective is a reaffirmation of the utility of graphical models in general, and its appeal to us is similar to the general appeal of path models, which have retained their adherents in spite of some of their known limitations.

For our purposes in this chapter, Pearl's work is important for three different reasons. First, his framework is completely nonparametric, and as a result it is usually unnecessary to specify the nature of the functional dependence of an outcome Y on the variables that cause it. Thus, $X \rightarrow Y$ simply implies that X causes Y, without specifying whether the effect is linear, quadratic, or some other highly nonlinear function in the values of X. This generality allows for a theory of causality without side assumptions about functional form, such as the linearity assumptions that became the Achilles' heel of traditional path models. Second, Pearl's approach shows clearly the critical importance of what he labels "collider" variables, which are specific types of endogenous variables that must be treated with caution. Finally, Pearl shows that there are three basic methods for identifying a causal effect: conditioning on variables that block all back-door paths, conditioning on variables that allow for estimation by a mechanism, and estimating a causal effect by an instrumental variable that is an exogenous shock to to the cause. Each of these identification strategies was already introduced briefly in Section 1.6. Here, we provide the foundations of his approach and then offer a more detailed presentation of the conditioning strategy for estimating causal effects by invoking Pearl's back-door criterion. In Chapter 6, we then return to the framework to discuss the estimation of causal effects more generally when important variables are unobserved. There, we will then more fully present the front-door and instrumental variable approaches.

3.1.1 Basic Elements

As with standard path models, the basic goal of drawing a causal system as a DAG is to represent explicitly all causes of the outcome of interest. As we

[1] Nonetheless, we still maintain that the observed outcome variable is generated by a process of causal exposure to alternative causal states with their attendant potential outcome variables. See Pearl (2000, Chapter 7) for the formal connections between potential outcomes and causal graphs.

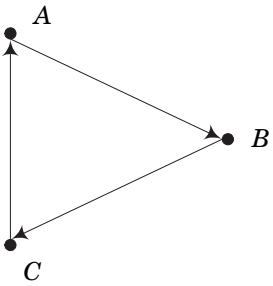


Figure 3.1: A directed graph that includes a cycle.

discussed earlier in Section 1.6, each node of a causal graph represents a random variable and is labeled by a letter such as A, B, or C. Nodes that are represented by a solid circle • are observed random variables, whereas nodes that are represented by a hollow circle ∘ are unobserved random variables. Causes are represented by directed edges → (i.e., single-headed arrows), such that an edge from one node to another signifies that the variable at the origin of the directed edge causes the variable at the terminus.[2]

Unlike standard path models, a DAG does not permit a representation of simultaneous causation. Only directed edges are permissible, and thus direct causation can run in only one direction, as in $X \rightarrow Y$. Furthermore, a DAG is defined to be an acyclic graph. Accordingly, no directed paths emanating from a causal variable also terminate at the same causal variable. Figure 3.1 presents a graph that includes a cycle, and as a result it is not a DAG, even though it includes only directed edges.

Under some circumstances it is useful to use a curved and dashed bidirected edge (as in Figures 1.1–1.3 earlier) as a shorthand device to indicate that two variables are mutually dependent on one or more (typically unobserved) common causes. In this shorthand, the two graphs presented in panels (a) and (b) of Figure 3.2 are equivalent. When this shorthand representation is used, the resulting graph is no longer a DAG by its formal definition. But, because the bidirected edge is a mere shorthand semantic substitution, the graph can be treated usually as if it were a DAG.[3] Such shorthand can be helpful in suppressing a complex set of background causal relationships that are irrelevant to the empirical analysis at hand. Nonetheless, these bidirected edges should not be interpreted in any way other than as we have just stated. They are not

[2]In Pearl's framework, each random variable is assumed to have an implicit probability distribution net of the causal effects represented by the directed edges. This position is equivalent to assuming that background causes of each variable exist that are independent of the causes explicitly represented in the graph by directed edges.

[3]Pearl would refer to such a graph as a semi-Markovian causal diagram rather than a fully Markovian causal model, but he would nonetheless treat it as if it were a full DAG when considering the identification of causal effects that are represented by directed edges in the graph (see Pearl 2000, Section 5.2).

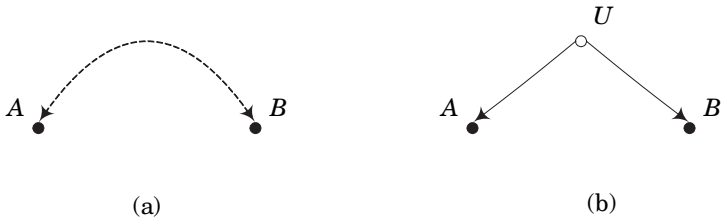


Figure 3.2: Two representations of the joint dependence of A and B on common causes.

indicators of mere correlations between the variables that they connect, and they do not signify that either of the two variables has a direct cause on the other one. Rather, they represent an unspecified set of common causes of the two variables that they connect.

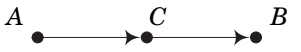
Figure 3.3 presents the three basic patterns of causal relationships that would be observed for any three variables that are related to each other: a chain of mediation, a fork of mutual dependence, and an inverted fork of mutual causation. Pearl's analysis of the first two types of relationship is conventional. For the graph in panel (a), A affects B through A's causal effect on C and C's causal effect on B. This type of a causal chain renders the variables A and B unconditionally associated. For the graph in panel (b), A and B are both caused by C. Here, A and B are also unconditionally associated, but now it is because they mutually depend on C.[4]

For the third graph in panel (c), A and B are again connected by a pathway through C. But now A and B are both causes of C. Pearl labels C a "collider" variable. Formally, a variable is a collider along a particular path if it has two arrows running into it. Figuratively, the causal effects of A and B "collide" with each other at C. Collider variables are common in social science applications: Any endogenous variable that has two or more causes is a collider along some path.

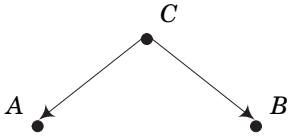
A path that is connected by a collider variable does not generate an unconditional association between the variables that cause the collider variable. For the mutual causation graph in panel (c) of Figure 3.3, the pathway between A and B through C does not generate an unconditional association between A and B. As a result, if nothing is known about the value that C takes on, then knowing the value that A takes on yields no information about the value that B takes on. Pearl's language is quite helpful here. The path $A \rightarrow C \leftarrow B$ does not generate an association between A and B because the collider variable C "blocks" the possible causal effects of A and B on each other.

Even though collider variables do not generate unconditional associations between the variables that determine them, we will show in the next subsection

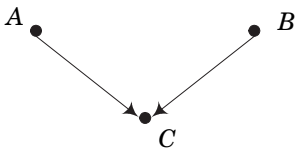
[4]The unconditional associations between A and B for both graphs mean that knowing the value that A takes on gives one some information on the likely value that B takes on. This unconditional association between A and B, however, is completely indirect, as neither A nor B has a direct causal effect on each other.



(a) Mediation



(b) Mutual dependence



(c) Mutual causation

Figure 3.3: Basic patterns of causal relationships for three variables.

that incautious handling of colliders can create conditional dependence that can sabotage a causal analysis. The importance of considering collider variables is a key insight of Pearl's framework, and it is closely related to the familiar concerns of selecting on the dependent variable and conditioning on an endogenous variable. But, to understand these complications, we first must introduce basic conditioning techniques in the context of graphical models.

3.1.2 Conditioning on Observable Variables

One of the most common modeling strategies to prosecute causal questions in quantitative research is to analyze a putative causal relationship within groups defined by one or more variables. Whether referred to as subgroup analysis, subclassification, stratification, or tabular decomposition, the usual motivation is to analyze the data after conditioning on membership in groups identified by values of a variable that is thought to be related to both the causal variable and the outcome variable.

From a graphical perspective, the result of such a modeling strategy is to generate simplified subgraphs for each subgroup or stratum of the data that correspond to each value of the conditioning variable. This procedure is analogous to disconnecting the conditioning variable from all other variables that it points

to in the original graph and rewriting the graph as many times as there are values for the conditioning variable. In the mutual dependence graph in panel (b) of Figure 3.3, conditioning on C results in separate graphs for each value c of C; in each of these subgraphs, A and B are disconnected. The reasoning here should be obvious: If analysis is carried out for a group in which all individuals have a particular value for the variable C, then the variable C is constant within the group and cannot therefore be associated with A or B. Thus, from a graphical perspective, conditioning is a means of transforming one graph into a simpler set of component graphs where fewer causes are represented.

As a technique for estimating a causal effect, conditioning is a very powerful and very general strategy. But one very important qualification must be noted: Conditioning on a collider variable does not simplify the original graph but rather adds complications by creating new associations. To see why, reconsider the mutual causation graph in panel (c) of Figure 3.3, where A and B are unrelated to each other but where both cause the collider variable C. In this case, conditioning on C will induce a relationship between A and B for at least one subgroup defined by the values of C.

The reasoning here is not intuitive, but it can be conveyed by a simple example with the mutual causation graph in panel (c) of Figure 3.3. Suppose that the population of interest is a set of applicants to a particular selective college and that C indicates whether applicants are admitted or rejected (i.e., $C = 1$ for admitted applicants and $C = 0$ for rejected applicants). Admissions decisions at this hypothetical college are determined entirely by two characteristics of students that are known to be independent within the population of applicants: SAT scores and a general rating of motivation based on an interview. These two factors are represented by A and B in panel (c) of Figure 3.3. Even though SAT score and motivation are unrelated among applicants in general, they are not unrelated when the population is divided into admitted and rejected applicants. Among admitted applicants, those with the highest SAT scores are on average the least motivated, and those with the lowest SAT scores are on average the most motivated. Thus, the college's sorting of applicants generates a pool of admitted students within which SAT and motivation are negatively related.[5]

This example is depicted in Figure 3.4 for 250 simulated applicants to this hypothetical college. For this set of applicants, SAT and motivation have a very small positive correlation of .035.[6] Offers of admission are then determined by the sum of SAT and motivation and granted to the top 15 percent of applicants

[5] A negative correlation will emerge for rejected students as well if (1) SAT scores and motivation have similarly shaped distributions and (2) both contribute equally to admissions decisions. As these conditions are altered, other patterns can emerge for rejected students, such as if admissions decisions are a nonlinear function of SAT and motivation.

[6] The values for SAT and motivation are 250 independent draws from standard normal variables. The draws result in an SAT variable with mean of .007 and a standard deviation of 1.01 as well as a motivation variable with mean of −.053 and a standard deviation of 1.02. Although the correlation between SAT and motivation is a small positive value for this simulation, we could drive the correlation arbitrarily close to 0 by increasing the number of applicants for the simulation.

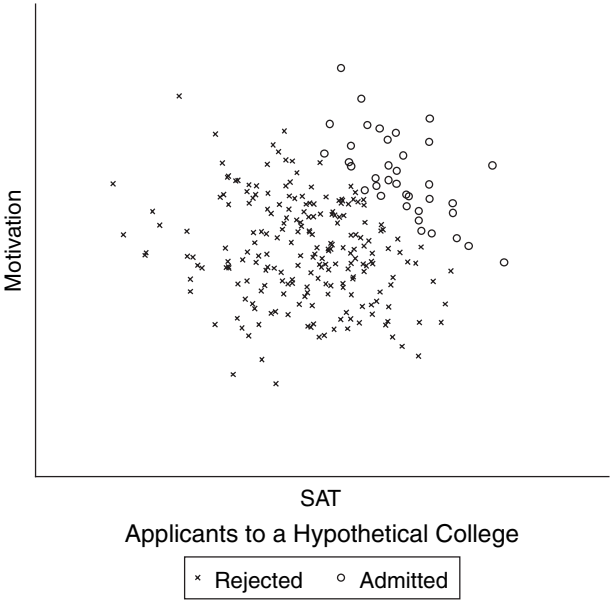


Figure 3.4: Simulation of conditional dependence within values of a collider variable.

(as shown in the upper right-hand portion of Figure 3.4).[7] Among admitted applicants, the correlation between SAT and motivation is −.641 whereas among rejected applicants the correlation between SAT and motivation is −.232. Thus, within values of the collider (the admissions decision), SAT and motivation are negatively related.

As Pearl documents comprehensively with a wide range of examples, this is a very general feature of causal relationships and is present in many real-world applications. In the next section, we show that care must be taken when attempting to estimate a causal effect by conditioning because conditioning on a collider variable can spoil an analysis.

3.1.3 Point Identification by Conditioning on Variables that Satisfy the Back-Door Criterion

Pearl elaborates three different approaches to identifying causal effects, which we already introduced in Section 1.6 as (1) conditioning on variables that block all back-door paths from the causal variable to the outcome variable, (2) using exogenous variation in an appropriate instrumental variable to isolate covariation in the causal variable and the outcome variable, and (3) establishing an

[7] Admission is offered to the 37 of 250 students (14.8 percent) whose sum of SAT and motivation is greater than or equal to 1.5.

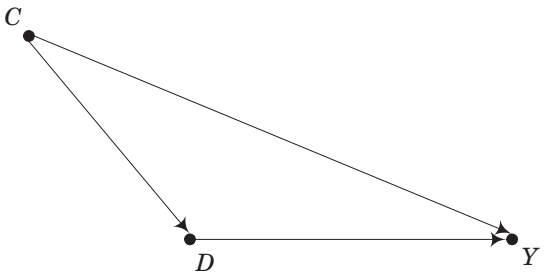


Figure 3.5: A causal diagram in which the effect of D on Y is confounded by C.

isolated and exhaustive mechanism (or set of mechanisms) that intercepts the effect of the causal variable on the outcome variable and then calculating the causal effect as it propagates through the mechanism. In this subsection, we consider the first of these strategies, which motivates the basic matching and regression techniques that we will present in the next two chapters.

Perhaps the most general concern of a researcher seeking to estimate a causal effect is that the causal variable D and the outcome variable Y are mutually dependent on a common third variable C. This common but simple scenario is represented by the DAG in Figure 3.5. In this case, the total association between D and Y represents the genuine causal effect of D on Y and the common dependence of D and Y on C. In this case, it is often said that the causal effect of D on Y is confounded by C. In particular, the presence of the causal effects $C \rightarrow D$ and $C \rightarrow Y$ confound the causal effect $D \rightarrow Y$.

For this example, the causal effect of D on Y can be consistently estimated by conditioning on C. We will explain this claim more formally and more generally in the remainder of this section. For now, consider conditioning only as a data analysis procedure in order to understand the end point of the discussion that follows. As an operational data analysis routine, the effect of D on Y can be estimated by conditioning on C in two steps: (1) calculate the association between D and Y for each subgroup with C equal to c and then (2) average these c-specific associations over the marginal distribution of the values c that the variable C takes on. The resulting weighted average is the causal effect of D on Y in Pearl's framework, which would be labeled the average treatment effect in the counterfactual causality literature. In this sense, conditioning on C identifies the causal effect of D on Y in this example.

Conditioning is a powerful strategy for estimating causal effects, and it is both successful and completely transparent for simple examples such as this one. But it is a much more complicated procedure in general than is suggested by our short accounting of this example. The complications arise when colliders are present, and Pearl has explained systematically how to resolve these complications.

Before considering a complex example, consider a more formal analysis of the example in Figure 3.5. Pearl characterizes the confounding created by both $C \rightarrow D$ and $C \rightarrow Y$ in a novel way, using the language of back-door paths. For Pearl, a path is any any sequence of edges pointing in any direction that connects one variable to another. A back-door path is then defined as a path between any causally ordered sequence of two variables that includes a directed edge $\rightarrow$ that points to the first variable. For the DAG in Figure 3.5, there are two paths that connect D and Y: $D \rightarrow Y$ and $D \leftarrow C \rightarrow Y$. For the causally ordered pair of variables D and Y, the path $D \leftarrow C \rightarrow Y$ is a back-door path because it includes a directed edge pointing to D. Likewise, the path $D \rightarrow Y$ is not a back-door path because it does not include a directed edge pointing to D.

The problem with back-door paths is that they may contribute to the association between D and Y. As a result, the observed association between D and Y may not consistently estimate the causal effect of D on Y. In Pearl's language, the observed association between D and Y does not identify the causal effect because the total association between D and Y is an unknown composite of the true causal effect $D \rightarrow Y$ and the back-door path $D \leftarrow C \rightarrow Y$.

With this language, Pearl then develops what he labels the "back-door criterion" for determining whether or not conditioning on a given set of observed variables will identify the causal effect of interest. If one or more back-door paths connects the causal variable to the outcome variable, Pearl shows that the causal effect is identified by conditioning on a set of variables Z if and only if all back-door paths between the causal variable and the outcome variable are blocked after conditioning on Z. He then proves that all back-door paths are blocked by Z if and only if each back-door path

1. contains a chain of mediation $A \rightarrow C \rightarrow B$, where the middle variable C is in Z, or

2. contains a fork of mutual dependence $A \leftarrow C \rightarrow B$, where the middle variable C is in Z, or

3. contains an inverted fork of mutual causation $A \rightarrow C \leftarrow B$, where the middle variable C and all of C's descendants are *not* in Z.[8]

Conditions 1 and 2 of the back-door criterion should be clear as stated; they imply that back-door associations between the causal variable and the outcome variable can be eliminated by conditioning on observed variables that block each back-door path. Condition 3, however, is quite different and is not intuitive. It states instead that the set of conditioning variables Z cannot include collider

[8]This claim is a combination of Pearl's definition of d-separation (Pearl 2000:16-17) and his definition of the back-door criterion (Pearl 2000:79). The back-door criterion is interpretable only when the causal effect of interest is specified as a component of a graph that is representable as a causal model (or at least a component of locally Markovian causal model, in which the underspecified causal relations are irrelevant to an evaluation of the back-door criterion for the particular causal effect under consideration). See Pearl's causal Markov condition for the existence of a causal model (Pearl 2000, Section 1.4.2, Theorem 1.4.1).

variables that lie along back-door paths.[9] We will explain the importance of condition 3, as well as the underlying rationale for it, in the examples that follow.[10]

First, return one last time to the simple example in Figure 3.5. Here, there is a single back-door path, which is a fork of mutual dependence where C causes both D and Y. By Pearl's back-door criterion, conditioning on C blocks $D \leftarrow C \rightarrow Y$ because C is the middle variable in a fork of mutual dependence. As a result, C satisfies Pearl's back-door criterion, and the causal effect of D on Y is identified by conditioning on C.

Consider now a more complex example, which involves the do-not-condition-on-colliders condition 3 of the back-door criterion. A common but poorly justified practice in the social sciences is to salvage a regression model from suspected omitted-variable bias by adjusting for an endogenous variable that can be represented as a proxy for the omitted variable that is unobserved. In many cases, this strategy will fail because the endogenous variable is usually a collider.

To set up this example, suppose that an analyst is confronted with a basic DAG, similar to the one in Figure 3.5, in which the causal effect of D on Y is confounded by a variable such as C. But suppose now that the confounder variable is unobserved, and thus cannot be conditioned on. When in this situation, researchers often argue that the effects of the unobserved confounder can be decomposed in principle into a lagged process, using a prior variable for the outcome, Y_{t-1}, and two separate unobserved variables, U and V, as in Figure 3.6.

For the DAG in Figure 3.6, there are two back-door paths from D to Y: $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$ and $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$. The lagged outcome variable Y_{t-1} lies along both of these back-door paths, but Y_{t-1} does not satisfy the back-door criterion. Notice first that Y_{t-1} blocks the first back-door path $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$ because, for this path, Y_{t-1} is the middle variable of a chain of mediation $V \rightarrow Y_{t-1} \rightarrow Y$. But, for the second path $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$, Y_{t-1} is a collider because it is the middle variable in an inverted fork of mutual causation $V \rightarrow Y_{t-1} \leftarrow U$. And, as a result, the back-door criterion states that, after conditioning on Y_{t-1}, at least one back-door path from D to Y will remain unblocked. For this example, it is the second path that includes the collider.

Having seen how condition 3 of the back-door criterion is applied, consider why it is so important. Pearl would state that, for this last example,

[9] Because the "or" in the back-door criterion is inclusive, one can condition on colliders and still satisfy the back-door criterion if the back-door paths along which the colliders lie are otherwise blocked because Z satisfies condition 1 or condition 2 with respect to another variable on the same back-door path.

[10] Note the stipulation in condition 3 that neither C nor the descendants of C can be in Z. We do not make much of these "descendants" in our presentation. But, see Hernan, Hernandez-Diaz, and Robins (2004) for a discussion of examples in epidemiology for which the distinction is important. For their examples, the collider is "getting sick enough to be admitted to a hospital for treatment" but the variable that is conditioned on is "in a hospital." Conditioning on "in a hospital" (by undertaking a study of hospital patients) induces associations between the determinants of sickness that can spoil standard analyses.

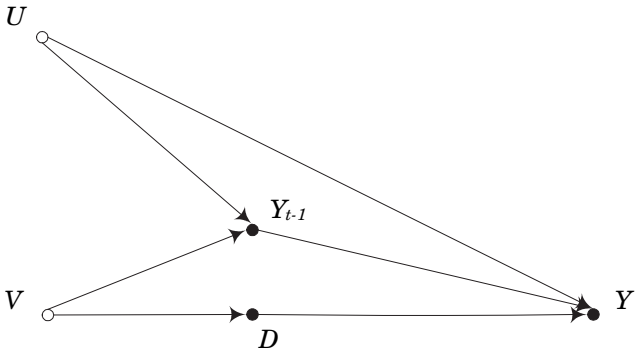


Figure 3.6: A causal diagram in which Y_{t-1} is a collider.

conditioning on Y_{t-1} eliminates part of the back-door association between D and Y because Y_{t-1} blocks the back-door path $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$. But, at the same time, conditioning on Y_{t-1} creates a new back-door association between D and Y because conditioning on Y_{t-1} unblocks the second back-door path $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$.

How can conditioning on a collider unblock a back-door path? To see the answer to this question, recall the simple characterization of colliders and the inverted fork of mutual causation in panel (c) of Figure 3.3. For that graph, the path $A \rightarrow C \leftarrow B$ contains a collider C. As we noted earlier, the indirect causal effects of A and B on each other are absorbed by C, and the path $A \rightarrow C \leftarrow B$ does not on its own generate an unconditional association between A and B. This basic result can be applied to back-door paths between D and Y that include colliders, as in the example presented in Figure 3.6. If a back-door path between a causal variable D and an outcome variable Y includes an intermediate variable that is a collider, that back-door path does not contribute to the unconditional association between D and Y. Because no back-door association between D and Y is generated, a back-door path that contains a collider does not confound the causal effect of D on Y.

At the same time, if a collider that lies along a back-door path is conditioned on, that conditioning will unblock the back-door path and thereby confound the causal effect. Recall the earlier discussion of conditioning in reference to panel (c) of Figure 3.3 and then as demonstrated in Figure 3.4. There, with the example of SAT and motivation effects on a hypothetical admissions decision to a college, we explained why conditioning on a collider variable induces an association between those variables that the collider is dependent on. That point applies here as well, when the causal effect of D on Y in Figure 3.6 is considered. Conditioning on a collider that lies along a back-door path unblocks the back-door path in the sense that it creates a net association between D and Y within at least one of the subgroups enumerated by the collider.

Consider the slightly more complex example that is presented in the DAG in Figure 3.7 (which is similar to Figure 1.1, except that the bidirected edges

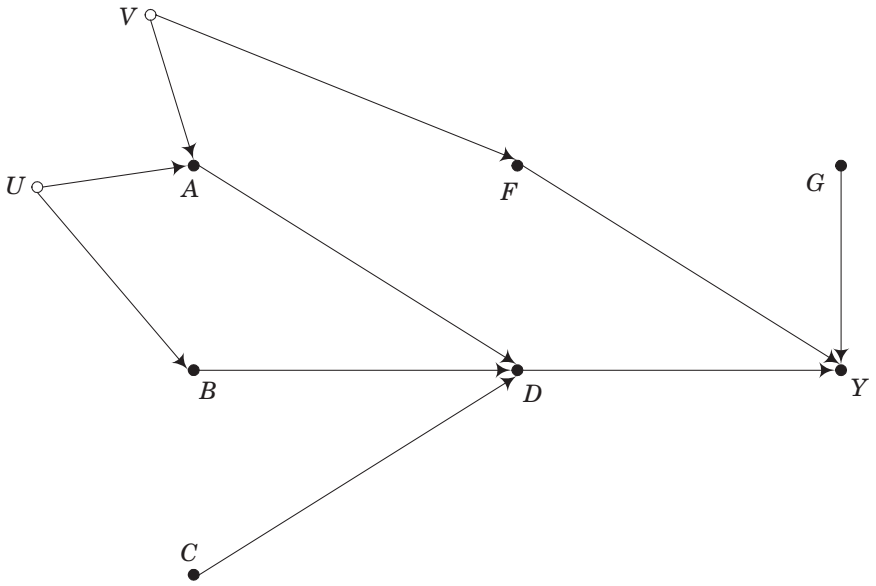


Figure 3.7: A causal diagram in which A is a collider.

that signified unspecified common causes have been replaced with two specific unobserved variables, U and V). Suppose, again, that we wish to estimate the causal effect of D on Y. For this DAG, there are two back-door paths between D and Y: (1) $D \leftarrow A \leftarrow V \rightarrow F \rightarrow Y$ and (2) $D \leftarrow B \leftarrow U \rightarrow A \leftarrow V \rightarrow F \rightarrow Y$. Notice that A is a collider variable in the second back-door path but not in the first back-door path. As a result, the first back-door path contributes to the association between D and Y, but the second back-door path does not contribute to the association between D and Y. We have to be careful that, whatever conditioning we enact to eliminate the confounding effect of the back-door path $D \leftarrow A \leftarrow V \rightarrow F \rightarrow Y$ does not unblock the back-door path $D \leftarrow B \leftarrow U \rightarrow A \leftarrow V \rightarrow F \rightarrow Y$ and thereby confound the causal effect in another way.

For this example, there are two entirely different and effective conditioning strategies available that will identify the causal effect (numbers 1 and 3 in the following list) and a third one that may appear to work but that will fail (number 2 in the following list):

1. F is the middle variable of a chain of mediation for both back-door paths, as in $V \rightarrow F \rightarrow Y$. As a result, F satisfies the back-door criterion, and conditioning on F identifies the causal effect of D on Y.

2. A is a middle variable of a chain of mediation for the first back-door path, as in $D \leftarrow A \leftarrow V$. But A is a collider variable for the second back-door path, as in $U \rightarrow A \leftarrow V$. As a result, A alone does not satisfy the

back-door criterion. Conditioning on A does not identify the causal effect of D on Y, even though A lies along both back-door paths. Conditioning on A would unblock the second back-door path and thereby create a new back-door association between D and Y.

3. A is a middle variable of a chain of mediation for the first back-door path, as in $D \leftarrow A \leftarrow V$. Likewise, B is a middle variable of a chain of mediation for the second back-door path, as in $D \leftarrow B \leftarrow U$. Thus, even though A blocks only the first back-door path (and, in fact, conditioning on it unblocks the second back-door path), conditioning on B blocks the second back-door path. As a result, A and B together (but not alone) satisfy the back-door criterion, and conditioning on them together identifies the causal effect of D on Y.

In sum, for this example the causal effect can be identified by conditioning in one of two minimally sufficient ways: either condition on F or condition on both A and B.[11]

The key point of this section is that conditioning on variables that lie along back-door paths can be an effective strategy to identify a causal effect. If all back-door paths between the causal variable and the outcome variable are blocked after the conditioning is enacted, then back-door paths do not contribute to the association between the causal variable and the outcome variable. And, as a result, the remaining association between the causal variable and outcome variable identifies the causal effect. Even so, it must be kept in mind that conditioning on a collider variable has the opposite effect. It unblocks an already blocked back-door path. And thus, as the last two examples show, when a conditioning strategy is evaluated, each back-door path must be assessed carefully because a variable can be a collider along one back-door path but not a collider along another.

Pearl's back-door criterion for evaluating conditioning strategies is a generalization (and therefore a unification) of various traditions for how to solve problems that are frequently attributed to omitted-variable bias. From our perspective, Pearl's framework is particularly helpful in two respects. It shows clearly that researchers do not need to condition on all omitted direct causes of an outcome variable in order to solve an omitted-variable bias problem. This claim is not new, of course, but Pearl's back-door criterion shows clearly why researchers need to condition on only a minimally sufficient set of variables that renders all back-door paths blocked. Moreover, Pearl's framework shows how to think clearly about the appropriateness of conditioning on endogenous variables. Writing down each back-door path and then determining whether or not each endogenous variable is a collider along any of these back-door paths is a much simpler way to begin to consider the full complications of a conditioning strategy than prior approaches.

[11] One can of course condition in three additional ways that also satisfy the back-door criterion: F and A, F and B, and F, A, and B. These conditioning sets include unnecessary and redundant conditioning.

In the next section, we consider models of causal exposure that have been used in the counterfactual tradition, starting first with the statistics literature and carrying on to the econometrics literature. We will show that the assumptions often introduced in these two traditions to justify conditioning estimation strategies – namely, ignorability and selection on the observables – can be thought of as more specific assertions of the general point that the average causal effect is identified when all back-door paths are blocked.

3.2 Models of Causal Exposure in the Counterfactual Tradition

With this general presentation of the conditioning strategy in mind, return to the familiar case of a binary cause D and an observed outcome variable Y. As discussed in Chapter 2, in the counterfactual tradition we consider Y to have been generated by a switching process between two potential outcome variables, as in $Y = DY^1 + (1 - D)Y^0$, where the causal variable D is the switch. To model variation in Y and relate it to the individual-level causal effects defined by the potential outcome variables Y^1 and Y^0, a model for the variation in D must be adopted. This is known, in the counterfactual framework, as modeling the treatment assignment mechanism or as modeling the treatment selection mechanism, based on which tradition of analysis is followed.

In this section, we first consider the notation and language developed by statisticians, and we then turn to the alternative notation and language developed by econometricians. Although both sets of ideas are equivalent, they each have some distinct conceptual advantages. In showing both, we hope to deepen the understanding of each.

3.2.1 Treatment Assignment Modeling in Statistics

The statistics literature on modeling the treatment assignment mechanism is an outgrowth of experimental methodology and the implementation of randomization research designs. Accordingly, we begin by considering a randomized experiment for which the phrase treatment assignment remains entirely appropriate.

As discussed earlier, if treatment assignment is completely random, then the treatment indicator variable D is completely independent of the potential outcomes Y^0 and Y^1 as well as any function of them, such as the distribution of δ [see the earlier discussion of Equation (2.4)]. In this case, the treatment assignment mechanism can be specified completely if $\Pr[D = 1]$ is set to a constant between 0 and 1. If a researcher desires treatment and control groups of approximately the same size, then $\Pr[D = 1]$ can be set to .5. Individual realized values of D for those in the study, denoted d_i generically, are then equal to 1 or 0. The values for d_i can be thought of as realized outcomes of a Bernoulli trial for the random variable D.

For more complex randomization schemes, more elaborate statements are required. If, for example, study subjects are stratified first by gender and then assigned with disproportionate probability to the treatment group if female, then the treatment assignment mechanism might instead be

$$\begin{aligned} \Pr[D = 1|\text{Gender} = \text{Female}] &= .7, \\ \Pr[D = 1|\text{Gender} = \text{Male}] &= .5. \end{aligned} \tag{3.1}$$

These conditional probabilities are often referred to as propensity scores, as they indicate the propensity that an individual with specific characteristics will be observed in the treatment group. Accordingly, for this example, Equations (3.1) are equivalent to stating that the propensity score for the treatment is .7 for females and .5 for males. In randomized experiments, the propensity scores are known to the researcher.

In contrast, a researcher with observational data only does not possess *a priori* knowledge of the propensity scores that apply to different types of individuals. However, she may know the characteristics of individuals that systematically determine their propensity scores. In this case, treatment selection patterns are represented by the general conditional probability distribution:

$$\Pr[D = 1|S], \tag{3.2}$$

where S now denotes all variables that systematically determine treatment assignment/selection. An observational researcher may know and have measures of all of the variables in S, even though he or she may not know the specific values of the propensity scores, which are defined as the values that $\Pr[D = 1|S]$ is equal to under different realized values s of the variables in S. Complete observation of S allows a researcher to assert that treatment selection is "ignorable" and then consistently estimate the average treatment effect, as we now explain.

The general idea here is that, within strata defined by S, the remaining variation in the treatment D is completely random and hence the process that generates this remaining variation is ignorable. The core of the concept of ignorability is the independence assumption that was introduced earlier in Equation (2.4):

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

where the symbol $\perp\!\!\!\perp$ denotes independence. As defined by Rubin (1978), ignorability of treatment assignment holds when the potential outcomes are independent of the treatment dummy indicator variable, as in this case all variation in D is completely random. But ignorability also holds in the weaker case,

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid S, \tag{3.3}$$

and when S is fully observed. The treatment assignment mechanism is ignorable when the potential outcomes (and any function of them, such as δ) are independent of the treatment variable within strata defined by all combinations

of values on the observed variables in S that determine treatment selection.[12]

If ignorability of treatment assignment is asserted for an observational study, then a researcher must (1) determine from related studies and supportable assumptions grounded in theory what the components of S are, (2) measure each of the component variables in S, and (3) collect enough data to be able to consistently estimate outcome differences on the observed variable Y within strata defined by S.[13] A researcher does not need to know the exact propensity scores (i.e., what $\Pr[D = 1|S = s]$ is equal to for all s), only that the systematic features of treatment selection can be exhaustively accounted for by the data in hand on the characteristics of individuals. The naive estimator can then be calculated within strata defined by values of the variables in S, and a weighted average of these stratified estimates can be formed as a consistent estimate of the average treatment effect.

Consider the Catholic school example. It is well known that students whose parents self-identify as Catholic are more likely to be enrolled in Catholic schools than students whose parents self-identify as non-Catholic. Suppose that parents' religious identity is the only characteristic of students that systematically determines whether they attend Catholic schools instead of public schools. In this case, a researcher can consistently estimate the average treatment effect by collecting data on test scores, students' school sector attendance, and parent's religious identification. A researcher would then estimate the effect of Catholic schooling separately by using the naive estimator within groups of students defined by parents' religious identification and then take a weighted average of these estimates based on the proportion of the population of interest whose parents self-identify as Catholic and as non-Catholic. This strategy is exactly the conditioning strategy introduced in the last section: Parents' religious identification blocks all back-door paths from Catholic school attendance to test scores.

Ignorability is thus directly related to conditioning on variables that satisfy the back-door criterion of Pearl. Suppose that we are confronted with the causal diagram in panel (a) of Figure 3.8, which includes the causal effect $D \rightarrow Y$ but also the bidirected edge $D \dashleftarrow\!\dashrightarrow Y$. The most common solution is to build an explicit causal model that represents the variables that generate the bidirected edge between D and Y in panel (a) of Figure 3.8. The simplest such model is presented in panel (b) of Figure 3.8, where $D \dashleftarrow\!\dashrightarrow Y$ has been replaced with

[12]Rosenbaum and Rubin (1983a) defined strong ignorability to develop the matching literature, which we will discuss later. To Rubin's ignorability assumption, Rosenbaum and Rubin (1983a) required for strong ignorability that each subject have a nonzero probability of being assigned to both the treatment and the control groups. Despite these clear definitions, the term ignorability is often defined in different ways in the literature. We suspect that this varied history of usage explains why Rosenbaum (2002) rarely uses the term in his monograph on observational data analysis, even though he is generally credited, along with Rubin, with developing the ignorability semantics in this literature. And it also explains why some of the most recent econometrics literature uses the words unconfoundedness and exogeneity for the same set of independence and conditional-independence assumptions (see Imbens 2004).

[13]This third step can be weakened if the data are merely sparse, as we discuss later when presenting propensity score techniques.

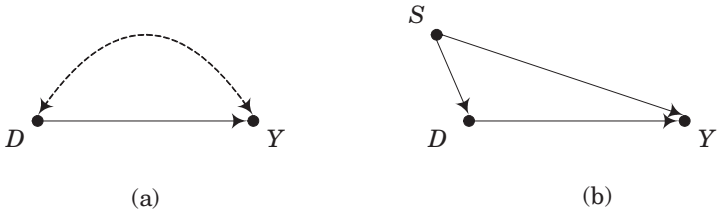


Figure 3.8: Causal diagrams in which treatment assignment is (a) nonignorable and (b) ignorable.

the back-door path $D \leftarrow S \rightarrow Y$. Thus, if S is observed, then conditioning on S will solve the causal inference problem. When identification by back-door conditioning is feasible, then treatment selection is ignorable.[14]

We will discuss these techniques in detail in Chapter 4, where we present matching estimators of causal effects. But the immediate complications of undertaking this strategy for the Catholic school example should be clear. How do we determine all of the factors that systematically determine whether a student enrolls in a Catholic school instead of a public school? And can we obtain measures of all of these factors? Attendance at a Catholic school is determined by more than just parents' religious self-identification, and some of these determinants are likely unmeasured. If this is the case, the treatment selection mechanism remains nonignorable, as treatment selection is then a function of unobserved characteristics of students.

3.2.2 Treatment Selection Modeling in Econometrics

The econometrics literature also has a long tradition of analyzing causal effects of these forms, and this literature may be more familiar to social scientists. Whereas concepts such as ignorability are used infrequently in the social sciences, the language of selection bias is commonly used throughout the social sciences. This usage is due, in large part, to the energy that economists have devoted to exploring the complications of self-selection bias.

[14]The potential outcome random variables are not represented in Figure 3.8. Pearl introduces the causal states in a different way, using the semantics of an intervention and introducing the $do(.)$ operator. For the causal diagram in panel (a) of Figure 3.8, the average values of Y that accompany the two values of D do not correspond to the two values of Y that one would obtain by calculating the average values of Y that would result from intervening to set D to its two possible values. For the causal diagram in panel (a) of Figure 3.8, $E[Y|D = 1] - E[Y|D = 0]$ does not equal $E[Y|do(D = 1] - E[Y|do(D = 0)]$. As should be clear, holding all semantic issues aside, this last statement is equivalent to saying that, for observations of variables D and Y for the model in panel (a) of Figure 3.8, the naive estimator does not equal the average difference in potential outcomes. Thus, the $do(.)$ operator is equivalent to the assertion of well-defined causal states, and that assertion is attached to the causal variable rather than to potential outcomes for each causal state. Thus, for Pearl, $E[Y^1] - E[Y^0]$ is equal to $E[Y|do(D = 1)] - E[Y|do(D = 0)]$.

The selection-bias literature in econometrics is vast, but the most relevant piece that we focus on here is James Heckman's specification of the random-coefficient model for the treatment effects of training programs (which he attributes, despite the difference in substance, to Roy 1951). The clearest specification of this model was presented in a series of papers that Heckman wrote with Richard Robb (see Heckman and Robb 1985, 1986, 1989), but Heckman worked out many of these ideas in the 1970s. Using the notation we have adopted in this book, take Equation (2.2),

$$Y = DY^1 + (1 - D)Y^0,$$

and then rearrange and relabel terms as follows:

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= Y^0 + \delta D \\ &= \mu^0 + \delta D + \upsilon^0, \end{aligned} \tag{3.4}$$

where $\mu^0 \equiv E[Y^0]$ and $\upsilon^0 \equiv Y^0 - E[Y^0]$. The standard outcome model from the econometrics of treatment evaluation simply reexpresses Equation (3.4) so that potential variability of δ across individuals in the treatment and control groups is relegated to the error term, as in

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}, \tag{3.5}$$

where $\mu^1 \equiv E[Y^1]$, $\upsilon^1 \equiv Y^1 - E[Y^1]$, and all else is as defined for Equation (3.4).[15] Note that, in evolving from Equation (2.2) to Equation (3.5), the

[15]The original notation is a bit different, but the ideas are the same. Without much usage of the language of potential outcomes, Heckman and Robb (1985; Section 1.4) offered the following setup for the random coefficient model of treatment effects to analyze posttreatment earnings differences for a fictitious manpower training example. For each individual i, the earnings of individual i if trained are

$$y_i^1 = \beta^1 + U_i^1,$$

and the earnings of individual i in the absence of training are

$$y_i^0 = \beta^0 + U_i^0,$$

(where we have suppressed subscripting on t for time from the original presentation and also shifted the treatment state descriptors from subscript to superscript position). With observed training status represented by a binary variable, d_i, Heckman and Robb then substitute the right-hand sides of these equations into the definition of the observed outcome in Equation (2.2) and rearrange terms to obtain

$$y_i = \beta^0 + (\beta^1 - \beta^0)d_i + U_i^0 + (U_i^1 - U_i^0)d_i,$$

which they then collapse into

$$y_i = \beta^0 + \bar{\alpha}d_i + \{U_i^0 + \varepsilon_i d_i\},$$

where $\bar{\alpha} \equiv \beta^1 - \beta^0$ and $\varepsilon_i \equiv U_i^1 - U_i^0$ (see Heckman and Robb 1985, Equation 1.13). As a result, $\bar{\alpha}$ is the average treatment effect, which we defined as $E[\delta]$ in Equation (2.3), and ε_i is the individual-level departure of δ_i from the average treatment effect $E[\delta]$. Although the notation in this last equation differs from the notation in Equation (3.5), the two equations are equivalent. Heckman and Vytlacil (2005) give a fully nonparametric version of this treatment

definition of the observed outcome variable Y has taken on the look and feel of a regression model.[16] The first μ^0 term is akin to an intercept, even though it is defined as $E[Y^0]$. The term $(\mu^1 - \mu^0)$ that precedes the first appearance of D is akin to a coefficient on the primary causal variable of interest D, even though $(\mu^1 - \mu^0)$ is defined as the true average causal effect $E[\delta]$. Finally, the term in braces, $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$, is akin to an error term, even though it represents both heterogeneity of the baseline no-treatment potential outcome and of the causal effect, δ, and even though it includes within it the observed variable D.[17]

Heckman and Robb use the specification of the treatment evaluation problem in Equation (3.5), and many others similar to it, to demonstrate all of the major problems created by selection bias in program evaluation contexts when simple regression estimators are used. Heckman and Robb show why a regression of Y on D does not in general identify the average treatment effect [in this case $(\mu^1 - \mu^0)$] when D is correlated with the population-level variant of the error term in braces in Equation (3.5), as would be the case when the size of the individual-level treatment effect [in this case $(\mu^1 - \mu^0) + \{\upsilon_i^0 + d_i(\upsilon_i^1 - \upsilon_i^0)\}$] differs among those who select the treatment and those who do not.

The standard regression strategy that prevailed in the literature at the time was to include additional variables in a regression model of the form of Equation (3.5), hoping to break the correlation between D and the error term.[18] Heckman and Robb show that this strategy is generally ineffective with the data available on manpower training programs because (1) some individuals are thought to enter the programs based on anticipation of the treatment effect itself and (2) none of the available data sources have measures of such anticipation. We will return to this case in detail in Chapter 5, where we discuss regression models.

To explain these complications, Heckman and Robb explore how effectively the dependence between D and the error term in Equation (3.5) can be broken. They proceed by proposing that treatment selection be modeled by specifying a latent continuous variable $\tilde{D}$:

$$\tilde{D} = Z\phi + U, \tag{3.6}$$

where Z represents all observed variables that determine treatment selection, ϕ is a coefficient (or a vector of coefficients if Z includes more than one vari-

selection framework, which we draw on later.

[16] Sometimes, Equation (3.5) is written as

$$Y = \mu^0 + [(\mu^1 - \mu^0) + (\upsilon^1 - \upsilon^0)]D + \upsilon^0$$

in order to preserve its random-coefficient interpretation. This alternative representation is nothing other than a fully articulated version of Equation (3.4).

[17] Statisticians sometimes object to the specification of "error terms" because, among other things, they are said to represent a hidden assumption of linearity. In this case, however, the specification of this error term is nothing other than an expression of the definition of the individual-level causal effect as the linear difference between y_i^1 and y_i^0.

[18] Barnow, Cain, and Goldberger (1980:52) noted that "the most common approach" is to "simply assume away the selection bias after a diligent attempt to include a large number of variables" in the regression equation.

able), and U represents both systematic unobserved determinants of treatment selection and completely random idiosyncratic determinants of treatment selection. The latent continuous variable $\tilde{D}$ in Equation (3.6) is then related to the treatment selection dummy, D, by

$$\begin{aligned} D &= 1 \text{ if } \tilde{D} \geq 0, \\ D &= 0 \text{ if } \tilde{D} < 0, \end{aligned}$$

where the threshold 0 is arbitrary because the term U has no inherent metric (because it is composed of unobserved and possibly unknown variables).

To see the connection between this econometric specification and the one from the statistics literature introduced in the last section, first recall that statisticians typically specify the treatment selection mechanism as the general conditional probability distribution $\Pr[D = 1|S]$, where S is a vector of all systematic observed determinants of treatment selection.[19] This is shown in the DAG in panel (b) of Figure 3.8. The corresponding causal diagram for the econometric selection equation is presented in two different graphs in Figure 3.9, as there are two scenarios corresponding to whether or not all elements of S have been observed as Z.

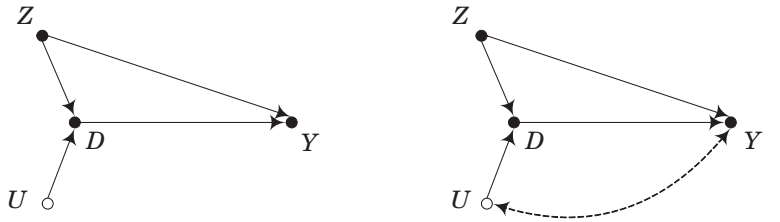
For the case in which Z in Equation (3.6) is equivalent to the set of variables in S in Equation (3.2), treatment selection is ignorable, as defined in Equation (3.3), because conditioning on Z is exactly equivalent to conditioning on S. In the econometric tradition, this situation would not, however, be referred to as a case for which treatment assignment/selection is ignorable. Rather, treatment selection would be characterized as "selection on the observables" because all systematic determinants of treatment selection are included in the observed treatment selection variables Z. This phrase is widely used by social scientists because it conveys the essential content of the ignorability assumption: All systematic determinants of treatment selection have been observed.

The scenario of selection on the observables is depicted in panel (a) of Figure 3.9. The variable S in panel (b) of Figure 3.8 is simply relabeled Z, and there are no back-door paths from D to Y other than the one that is blocked by Z. The remaining idiosyncratic random variation in D is attributed in the econometric tradition to a variable U, which is presented in panel (a) of Figure 3.9 as a cause of D that is conditionally independent of both Z and Y. This error term U represents nothing other than completely idiosyncratic determinants of treatment selection. It could therefore be suppressed in panel (a) of Figure 3.9, which would render this DAG the same as the one in panel (b) of Figure 3.8.[20]

Now consider the case in which the observed treatment selection variables in Z are only a subset of the variables in S. In this case, some components of

[19] When more specific, the basic model is usually a Bernoulli trial, in which $\Pr[D = 1|S = s]$ gives the specific probability of drawing a 1 and the complement of drawing a 0 for individuals with S equal to s.

[20] The latent variable specification in the econometric tradition can be made equivalent to almost all particular specifications of the statement $\Pr[D = 1|S]$ in the statistics tradition by the choice of an explicit probability distribution for U.



(a) Selection on the observables (b) Selection on the unobservables

Figure 3.9: Causal diagrams for the terminology from econometric modeling of treatment selection.

S enter into the treatment selection latent variable $\tilde{D}$ through the error term, U, of Equation (3.6). In this case, treatment selection is nonignorable. Or, in the words of econometricians, "selection is on the unobservables." The scenario of selection on the unobservables is depicted in panel (b) of Figure 3.9, where there is now a back-door path from D to Y: $D \leftarrow U \dashleftarrow\dashrightarrow Y$. Conditioning on Z for this causal diagram does not block all back-door paths.

In spite of differences in language and notation, there is little that differentiates the statistics and econometrics models of treatment selection, especially now that the outcome equations used by economists are often completely general nonparametric versions of Equation (3.5) (see Heckman and Vytlacil 2005, which we will discuss later). For now, the key point is that both the statistics and econometric specifications consider the treatment indicator variable, D, to be determined by a set of systematic treatment selection variables in S. When all of these variables are observed, the treatment selection mechanism is ignorable and selection is on the observables only. When some of the variables in S are unobserved, the treatment selection mechanism is nonignorable and selection is on the unobservables.

3.3 Conditioning to Balance versus Conditioning to Adjust

When presenting Pearl's back-door criterion for determining a sufficient set of conditioning variables, we noted that for some applications more than one set of conditioning variables is sufficient. In this section, we return to this point as a bridge to the following two chapters that present both matching and regression implementations of conditioning. Although we will show that both sets of techniques can be considered variants of each other, here we point to the different ways in which they are usually invoked in applied research. Matching is most often considered a technique to balance the determinants of the causal

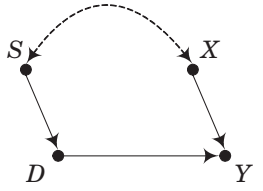


Figure 3.10: A causal diagram in which sufficient conditioning can be performed with respect to S or X.

variable, and regression is most often considered a technique to adjust for other causes of the outcome.

To frame this discussion, consider first the origins of the balancing approach in the randomized experiment tradition. Here, the most familiar approach is a randomized experiment that ensures that treatment status is unassociated with all observed and unobserved variables that determine the outcome (although only in expectation). When treatment status is unassociated with an observed set of variables W, the data are balanced with respect to W. More formally, the data are balanced if

$$\Pr[W|D = 1] = \Pr[W|D = 0], \tag{3.7}$$

which requires that the probability distribution of W be the same within the treatment and control groups.

Now consider the graph presented in Figure 3.10. A back-door path $D \leftarrow S \dashleftarrow\dashrightarrow X \rightarrow Y$ is present from D to Y, where S represents the complete set of variables that are direct causes of treatment assignment/selection, X represents the complete set of variables other than D that are direct causes of Y, and the bidirected edge between S and X signifies that they are mutually caused by some set of common unobserved causes.[21]

Because neither S nor X is a collider, all back-door paths in the graph can be blocked by conditioning on either S or X (and we write "paths" because there may be many paths signified by the bidirected edge between S and X). Conditioning on S is considered a balancing conditioning strategy whereas conditioning on X is considered an adjustment-for-other-causes conditioning strategy. If one observes and then conditions on S, the variables in S and D are no longer associated within the subgroups defined by the conditioning. The treatment and control groups are thereby balanced with respect to the distribution of S. Alternatively, if one conditions on X, the resulting subgroup differences

[21] For this example, we could have motivated the same set of conclusions with other types of causal graphs. The same basic conclusions would hold even if X and S include several variables within them in which some members of X cause D directly and some members of S cause Y directly. In other words, all that we need to make the distinction between balancing and adjustment for other direct causes is two sets of variables that are related to each other, with at least one variable in one set that causes D but not Y and at least one variable in the other set that causes Y but not D.

in Y across D within X can be attributed to D alone. In this case, the goal is not to balance X but rather to partial out its effects on Y in order to isolate the net effect of D on Y.

The distinction between balancing and adjustment for other causes is somewhat artificial. For the graph in Figure 3.10, balancing X identifies the causal effect. Thus it is technically valid to say that one can identify a causal effect by balancing a sufficient set of other causes of Y. Nonetheless, the graph in Figure 3.10 demonstrates why the distinction is important. The ultimate set of systematic causes that generates the relationship between S and X is unobserved, as it often is in many applied research situations. Because one cannot condition on these unobserved variables, one must condition on either S or X in order to identify the causal effect. These two alternatives may be quite different in their practical implementation.

Should one balance the determinants of a cause, or should one adjust for other causes of the outcome? The answer to this question is situation specific, and it depends on the quality of our knowledge and measurement of the determinants of D and Y. Perhaps the best answer is that one should do both.[22] Nonetheless, there is a specific advantage of balancing that may tip the scales in its favor if both strategies are feasible: It diminishes the inferential problems that can be induced by data-driven specification searches. We will discuss these issues in the course of presenting matching and regression conditioning strategies in the next two chapters.

3.4 Point Identification of Conditional Average Treatment Effects by Conditioning

At the beginning of this chapter, we indicated that we would implicitly focus our presentation of graphical causal models and identification issues on the estimation of the unconditional average treatment effect. This narrow focus is entirely consistent with the graphical tradition, in which parameters such as the average treatment effect for the treated in Equation (2.5) and the average treatment effect for the untreated in Equation (2.6) are given considerably less attention than in the counterfactual modeling tradition in both statistics and econometrics. Some comments on the connections may be helpful at this point to foreshadow some of the specific material on causal effect heterogeneity that we will present in the next two chapters.

Identification When the Unconditional Average Treatment Effect is Identified

If one can identify and consistently estimate the unconditional average treatment effect with conditioning techniques, then one can usually estimate some of the

[22] As we discuss in Subsection 5.3.4, many scholars have argued for conditioning on both S and X. Robins, for example, argues for this option as a double protection strategy that offers two chances to effectively break the back-door path between Y and D.

conditional average treatment effects that may be of interest as well. As we will show in the next two chapters, consistent estimates of conditional average treatment effects can usually be formed by specification of alternative weighted averages of the average treatment effects for subgroups defined by values of the conditioning variables. Thus, calculating average effects other than the unconditional average treatment effect may be no more complicated than simply adding one step to the more general conditioning strategy we have presented in this chapter.

Consider again the graph presented in Figure 3.10. The back-door path from D to Y is blocked by both S and X. As a result, a consistent estimate of the average treatment effect in Equation (2.3) can be obtained by conditioning on either S or X. But, in addition, consistent estimates of the average treatment effect for the treated in Equation (2.5) and the average treatment effect for the untreated in Equation (2.6) can be obtained by properly weighting conditional differences in the observed values of Y. In particular, if conditioning is performed with respect to S, first calculate the sample analogs to the differences $E[Y|D = 1, S = s] - E[Y|D = 0, S = s]$ for all values s of S. Then, weight these differences by the conditional distributions $\Pr[S|D = 1]$ and $\Pr[S|D = 0]$ to calculate the average treatment effect for the treated and the average treatment effect for the untreated, respectively. If, on the other hand, conditioning is performed with respect to X, then alternative quantities are calculated as sample analogs to $E[Y|D = 1, X = x] - E[Y|D = 0, X = x]$, $\Pr[X|D = 1]$, and $\Pr[X|D = 0]$. These estimated quantities will differ from those that are generated by conditioning on S, but they can still be used in an analogous way to form consistent estimates of the average treatment effect for the treated and the average treatment effect for the untreated. We will present examples of these sorts of stratification and weighting estimators in the next chapter.

But note that the ingredients utilized to estimate the average treatment effect for the treated and the average treatment effect for the untreated in these two conditioning routines are quite different. If S is observed, then conditional average treatment effects can be calculated for those who are subject to the cause for different reasons, based on the values of S that determine D. If X is observed, then conditional average treatment effects can be calculated holding other causes of Y at chosen values of X. Each of these sets of conditional average treatment effects has its own appeal, with the relative appeal of each depending on the application. In the counterfactual tradition, average treatment effects conditional on S would likely be of more interest than average treatment effects conditional on X. But for those who are accustomed to working within an all-cause regression tradition, then average treatments effects conditional on X might be more appealing.

Identification When the Unconditional Average Treatment Effect is Not Identified

If selection is on the unobservables, conditioning strategies will generally fail to identify unconditional average treatment effects. Nonetheless, weaker as-

sumptions may still allow for the identification and subsequent estimation by conditioning of various conditional average treatment effects. We will present these specific weaker assumptions in the course of explaining matching and regression techniques in the next two chapters, but for now we give a brief overview of the identification issues in relation to the graphical models presented in this chapter. (See also the prior discussion in Subsection 2.6.4 of similar issues with regard to the bias of the naive estimator.)

Suppose, for example, that the graph in panel (b) of Figure 3.9 now obtains, and hence that a back-door path from D to Y exists via unobserved determinants of the cause, U. In this case, conditioning on Z will not identify the unconditional average treatment effect. Nonetheless, conditioning on Z may still identify a conditional average treatment effect of interest, as narrower effects can be identified if weaker assumptions can be maintained even though unblocked back-door paths may still exist between D and Y.

Consider a case for which partial ignorability holds, such that $Y^0 \perp\!\!\!\perp D|S$ is true but $(Y^0, Y^1) \perp\!\!\!\perp D \mid S$ is not. Here, conditioning on S generates a consistent estimate of the average treatment effect for the treated even though S does not block the back-door path from D to Y. The opposite is, of course, also true. If partial ignorability holds in the other direction, such that $Y^1 \perp\!\!\!\perp D|S$ holds but $(Y^0, Y^1) \perp\!\!\!\perp D \mid S$ does not, then the average treatment effect for the untreated can be estimated consistently.[23]

Consider the first case, in which only $Y^0 \perp\!\!\!\perp D \mid S$ holds. Even after conditioning on S, a back-door path remains between D and Y because Y^1 still differs systematically between those in the treatment and control groups and Y is determined in part by Y^1 [see Equation (2.2)]. Nonetheless, if, after conditioning on S, the outcome under the no-treatment-state, Y^0, is independent of exposure to the treatment, then the average treatment effect for the treated can be estimated consistently. The average values of Y, conditional on S, can be used to consistently estimate the average what-if values for the treated if they were instead in the control state. This type of partial ignorability is akin to Assumption 2 in Equation (2.14), except that it is conditional on S. We will give a full explanation of the utility of such assumptions when discussing matching estimates of the treatment effect for the treated and the treatment effect for the untreated in the next chapter.

3.5 Conclusions

In the next two chapters, we present details and connections between the two main types of conditioning estimation strategies: matching and regression. We show how they generally succeed when selection is on the observables and fail when selection is on the unobservables. We lay out the specific assumptions that

[23] And, as we will show later, the required assumptions are even simpler, as the entire distributions of Y^0 and Y^1 need not be conditionally independent of D. As long as SUTVA holds, only mean independence must be maintained.

allow for the identification of unconditional average treatment effects, as well as the weaker assumptions that allow for the identification of narrower conditional average treatment effects. In later chapters, we then present more complex methods for identifying and estimating causal effects when simple conditioning methods are insufficient.

Chapter 4

Matching Estimators of Causal Effects

Written with David Harding[1]

The rise of the counterfactual model to prominence has increased the popularity of data analysis routines that are most clearly useful for estimating the effects of causes. The matching estimators that we will review and explain in this chapter are perhaps the best example of a classic technique that has reemerged in the past two decades as a promising procedure for estimating causal effects.[2] Matching represents an intuitive method for addressing causal questions, primarily because it pushes the analyst to confront the process of causal exposure as well as the limitations of available data. Accordingly, among social scientists who adopt a counterfactual perspective, matching methods are fast becoming an indispensable technique for prosecuting causal questions, even though they usually prove to be the beginning rather than the end of causal analysis on any particular topic.

We begin with a brief discussion of the past use of matching methods. Then, we present the fundamental concepts underlying matching, including stratification of the data, weighting to achieve balance, and propensity scores. Thereafter, we discuss how matching is usually undertaken in practice, including an overview of various matching algorithms. Finally, we discuss how the assumptions behind matching estimators often break down in practice, and we introduce some of the remedies that have been proposed to address the resulting problems.

[1] This chapter is based on Morgan and Harding (2006).

[2] Matching techniques can be motivated as estimators without invoking causality. Just as with regression modeling, which we discuss in detail in the next chapter, matching can be used to adjust the data in search of a meaningful descriptive fit to the data in hand. Given the nature of this book, we will focus on matching as an estimator of causal effects. We will, however, discuss the descriptive motivation for regression estimators in the next chapter.

In the course of presentation, we will offer four hypothetical examples that demonstrate some of the essential claims of the matching literature, progressing from idealized examples of stratification and weighting to the implementation of alternative matching algorithms on simulated data for which the treatment effects of interest are known by construction. As we offer these examples, we add real-world complexity in order to demonstrate how such complexity can overwhelm the power of the techniques.

4.1 Origins of and Motivations for Matching

Matching techniques have origins in experimental work from the first half of the twentieth century. Relatively sophisticated discussions of matching as a research design can be found in early methodological texts in the social sciences (e.g., Greenwood 1945) and also in attempts to adjudicate between competing explanatory accounts in applied demography (Freedman and Hawley 1949). This early work continued in sociology (e.g., Althauser and Rubin 1970, 1971; Yinger, Ikeda, and Laycock 1967) right up to the key foundational literature in statistics (Rubin 1973a, 1973b, 1976, 1977, 1979, 1980a) that provided the conceptual foundation for the new wave of matching techniques that we will present in this chapter.

In the early 1980s, matching techniques, as we conceive of them now, were advanced in a set of papers by Rosenbaum and Rubin (1983a, 1984, 1985a, 1985b) that offered solutions to a variety of practical problems that had limited matching techniques to very simple applications in the past. Variants of these new techniques found some use immediately in sociology (Berk and Newton 1985; Berk, Newton, and Berk 1986; Hoffer et al. 1985), continuing with work by Smith (1997). In the late 1990s, economists and political scientists joined in the development of matching techniques (e.g., Heckman et al. 1999; Heckman, Ichimura, Smith, and Todd 1998; Heckman, Ichimura, and Todd 1997, 1998 in economics and Ho, Imai, King, and Stuart 2005 and Diamond and Sekhon 2005 in political science). Given the growth of this literature, and the applications that are accumulating, we expect that matching will complement other types of modeling in the social sciences with greater frequency in the future.

In the methodological literature, matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among two larger pools of such cases or (2) as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted.

For the first motivation, the archetypical example is an observational biomedical study in which a researcher is called on to assess what can be learned about a particular treatment. The investigator is given access to two sets of data, one for individuals who have been treated and one for individuals who have not. Each dataset includes a measurement of current symptoms, Y, and a set of characteristics of individuals, as a vector of variables X, that are drawn from

demographic profiles and health histories. Typically, the treatment cases are not drawn from a population by means of any known sampling scheme. Instead, they emerge as a result of the distribution of initial symptoms, patterns of access to the health clinic, and then decisions to take the treatment. The control cases, however, may represent a subsample of health histories from some known dataset. Often, the treatment is scarce, and the control dataset is much larger than the treatment dataset.

In this scenario, matching is a method of strategic subsampling from among treated and control cases. The investigator selects a nontreated control case for each treated case based on the characteristics observed as x_i. All treated cases and matched control cases are retained, and all nonmatched control cases are discarded. Differences in the observed y_i are then calculated for treated and matched cases, with the average difference serving as the treatment effect estimate for the group of individuals given the treatment.[3]

The second motivation has no archetypical substantive example, as it is similar in form to any attempt to use regression to estimate causal effects with survey data. Suppose, for a general example, that an investigator is interested in the causal effect of an observed dummy variable, D, on an observed outcome, Y. For this example, it is known that a simple bivariate regression, $Y = \alpha + \delta D + \varepsilon$, will yield an estimated coefficient $\hat{\delta}$ that is a biased and inconsistent estimate of the causal effect of interest because the causal variable D is associated with variables included in the error term, ε. For a particular example, if D is the receipt of a college degree and Y is a measure of economic success, then the estimate of interest is the causal effect of obtaining a college degree on subsequent economic success. However, family background variables are present in ε that are correlated with D, and this relationship produces omitted-variable bias for a college-degree coefficient estimated from a bivariate ordinary least squares (OLS) regression of Y on D.

In comparison with the biomedical example just presented, this motivation differs in two ways: (1) In most applications of this type, the data represent a random sample from a well-defined population and (2) the common practice in the applied literature is to use regression to estimate effects. For the education example, a set of family background variables in X is assumed to predict both D and Y. The standard regression solution is to estimate an expanded regression equation: $Y = \alpha + \delta D + X\beta + \varepsilon^*$. With this strategy (which we will discuss in detail in the next chapter), the goal is to estimate simultaneously the causal effects of X and D on the outcome, Y.

In contrast, a matching estimator nonparametrically balances the variables in X across D solely in the service of obtaining the best possible estimate of the causal effect of D on Y. The most popular technique is to estimate the probability of D for each individual i as a function of X and then to select

[3] A virtue of matching, as developed in this tradition, is cost effectiveness for prospective studies. If the goal of a study is to measure the evolution of a causal effect over time by measuring symptoms at several points in the future, then discarding nontreated cases unlike any treated cases can cut expenses without substantially affecting the quality of causal inferences that a study can yield.

for further analysis only matched sets of treatment and control cases that contain individuals with equivalent values for these predicted probabilities. This procedure results in a subsampling of cases, comparable with the matching procedure described for the biomedical example, but for a single dimension that is a function of the variables in X. In essence, the matching procedure throws away information from the joint distribution of X and Y that is unrelated to variation in the treatment variable D until the remaining distribution of X is equivalent for both the treatment and control cases. When this equivalence is achieved, the data are said to be balanced with respect to X.[4] Under specific assumptions, the remaining differences in the observed outcome between the treatment and matched control cases can then be regarded as attributable solely to the effect of the treatment.[5]

For the remainder of this chapter, we will adopt this second scenario because research designs in which data are drawn from random-sample surveys are much more common in the social sciences.[6] Thus, we will assume that the data in hand were generated by a relatively large random-sample survey (in some cases an infinite sample to entirely remove sampling error from consideration), in which the proportion and pattern of individuals who are exposed to the cause are fixed in the population by whatever process generates causal exposure.

4.2 Matching as Conditioning via Stratification

In this section we introduce matching estimators in idealized research conditions, drawing connections with the broad perspective on conditioning introduced in Chapter 3. Thereafter, we proceed to a discussion of matching in more realistic scenarios, which is where we explain the developments of matching techniques that have been achieved in the past three decades.

4.2.1 Estimating Causal Effects by Stratification

Suppose that those who take the treatment and those who do not are very much unlike each other, and yet the ways in which they differ are captured exhaustively by a set of observed treatment assignment/selection variables S. For the language we will adopt in this chapter, knowledge and observation of S allow for a "perfect stratification" of the data. By "perfect," we mean precisely that individuals within groups defined by values on the variables in S are entirely indistinguishable from each other in all ways except for (1) observed treatment

[4] As we will discuss later, in many applications balance can be hard to achieve without some subsampling from among the treatment cases. In this case, the causal parameter that is identified is narrower even than the average treatment effect for the treated (and is usually a type of marginal treatment effect pinned to the common support of treatment and control cases).

[5] A third motivation, which is due to Ho, Imai, King, and Stuart (2005), has now emerged. Matching can be used as a data preprocessor that prepares a dataset for further causal modeling with a parametric model. We discuss this perspective along with others that seek to combine matching and regression approaches later, especially in Chapter 5.

[6] See our earlier discussion in Section 1.4 of this random-sample setup.

status and (2) differences in the potential outcomes that are independent of treatment status. Under such a perfect stratification of the data, even though we would not be able to assert Assumptions 1 and 2 in Equations (2.13) and (2.14), we would be able to assert conditional variants of those assumptions:

$$\text{Assumption 1-S:} \quad E[Y^1|D=1,S] = E[Y^1|D=0,S], \tag{4.1}$$

$$\text{Assumption 2-S:} \quad E[Y^0|D=1,S] = E[Y^0|D=0,S]. \tag{4.2}$$

These assumptions would suffice to enable consistent estimation of the average treatment effect, as the treatment can be considered randomly assigned within groups defined by values on the variables in S.

When in this situation, researchers often assert that the naive estimator in Equation (2.7) is subject to bias (either generic omitted-variable bias or individually generated selection bias). But, because a perfect stratification of the data can be formulated, treatment assignment is ignorable [see the earlier discussion of Equation (3.2)] or treatment selection is on the observable variables S only [see the earlier discussion of Equation (3.6)]. This is a bit imprecise, however, because Assumptions 1-S and 2-S are implied by ignorability and selection on the observables (assuming S is observed). For ignorability and selection on the observables to hold more generally, the full distributions of Y^1 and Y^0 (and any functions of them) must be independent of D conditional on S [see the discussion of Equation (3.3)]. Thus Assumptions 1-S and 2-S are weaker than assumptions of ignorability and selection on the observables, but they are sufficient to identify the three average causal effects of primary interest.

Recall the DAG in panel (b) of Figure 3.8, where S lies along the only back-door path from D to Y. As discussed there, conditioning on S allows for consistent estimation of the unconditional average treatment effect, as well as the average treatment effects for the treated and for the untreated. Although we gave a conceptual discussion in Chapter 3 of why conditioning works in this scenario, we will now explain more specifically with a demonstration. First note why everything works out so cleanly when a set of perfect stratifying variables is available. If Assumption 1-S is valid, then

$$\begin{aligned} E[\delta|D=0,S] &= E[Y^1 - Y^0|D=0,S] \\ &= E[Y^1|D=0,S] - E[Y^0|D=0,S] \\ &= E[Y^1|D=1,S] - E[Y^0|D=0,S]. \end{aligned} \tag{4.3}$$

If Assumption 2-S is valid, then

$$\begin{aligned} E[\delta|D=1,S] &= E[Y^1 - Y^0|D=1,S] \\ &= E[Y^1|D=1,S] - E[Y^0|D=1,S] \\ &= E[Y^1|D=1,S] - E[Y^0|D=0,S]. \end{aligned} \tag{4.4}$$

The last line of Equation (4.3) is identical to the last line of Equation (4.4), and neither line includes counterfactual conditional expectations. Accordingly, one

can consistently estimate the difference in the last line of Equation (4.3) and the last line of Equation (4.4) for each value of S. To then form consistent estimates of alternative average treatment effects, one simply averages the stratified estimates over the distribution of S, as we show in the following demonstration.

Matching Demonstration 1

Consider a completely hypothetical example in which Assumptions 1 and 2 in Equations (2.13) and (2.14) cannot be asserted because positive self-selection ensures that those who are observed in the treatment group are more likely to benefit from the treatment than those who are not. But assume that a three-category perfect stratifying variable S is available that allows one to assert Assumptions 1-S and 2-S in Equations (4.1) and (4.2). Moreover, suppose for simplicity of exposition that our sample is infinite so that sampling error is zero. In this case, we can assume that the sample moments in our data equal the population moments (i.e., $E_N[y_i|d_i = 1] = E[Y|D = 1]$ and so on).

If it is helpful, think of Y as a measure of an individual's economic success at age 40, D as an indicator of receipt of a college degree, and S as a mixed family-background and preparedness-for-college variable that completely accounts for the pattern of self-selection into college that is relevant for lifetime economic success. Note, however, that no one has ever discovered such a variable as S for this particular causal effect.

Suppose now that, for our infinite sample, the sample mean of the outcome for those observed in the treatment group is 10.2 whereas the sample mean of the outcome for those observed in the control group is 4.4. In other words, we have data that yield $E_N[y_i|d_i = 1] = 10.2$ and $E_N[y_i|d_i = 0] = 4.4$, and for which the naive estimator would yield a value of 5.8 (i.e., $10.2 - 4.4$).

Consider, now, an underlying set of potential outcome variables and treatment assignment patterns that could give rise to a naive estimate of 5.8. Table 4.1 presents the joint probability distribution of the treatment variable D and the stratifying variable S in its first panel as well as expectations, conditional on S, of the potential outcomes under the treatment and control states. The joint distribution in the first panel shows that individuals with S equal to 1 are more likely to be observed in the control group, individuals with S equal to 2 are equally likely to be observed in the control group and the treatment group, and individuals with S equal to 3 are more likely to be observed in the treatment group.

As shown in the second panel of Table 4.1, the average potential outcomes conditional on S and D imply that the average causal effect is 2 for those with S equal to 1 or S equal to 2 but 4 for those with S equal to 3 (see the last column). Moreover, as shown in the last row of the table, where the potential outcomes are averaged over the within-D distribution of S, $E[Y|D = 0] = 4.4$ and $E[Y|D = 1] = 10.2$, matching the initial setup of the example based on a naive estimate of 5.8 from an infinite sample.

Table 4.2 shows what can be calculated from the data, assuming that S offers a perfect stratification of the data. The first panel presents the sample

Table 4.1: The Joint Probability Distribution and Conditional Population Expectations for Matching Demonstration 1

	Joint probability distribution of S and D		
	$D=0$	$D=1$	
$S=1$	$\Pr[S=1, D=0]=.36$	$\Pr[S=1, D=1]=.08$	$\Pr[S=1]=.44$
$S=2$	$\Pr[S=2, D=0]=.12$	$\Pr[S=2, D=1]=.12$	$\Pr[S=2]=.24$
$S=3$	$\Pr[S=3, D=0]=.12$	$\Pr[S=3, D=1]=.2$	$\Pr[S=3]=.32$
	$\Pr[D=0]=.6$	$\Pr[D=1]=.4$	
	Potential outcomes		
	Under the control state	Under the treatment state	
$S=1$	$E[Y^0\|S=1]=2$	$E[Y^1\|S=1]=4$	$E[Y^1-Y^0\|S=1]=2$
$S=2$	$E[Y^0\|S=2]=6$	$E[Y^1\|S=2]=8$	$E[Y^1-Y^0\|S=2]=2$
$S=3$	$E[Y^0\|S=3]=10$	$E[Y^1\|S=3]=14$	$E[Y^1-Y^0\|S=3]=4$
	$E[Y^0\|D=0]$ $=\frac{.36}{.6}(2)+\frac{.12}{.6}(6)$ $+\frac{.12}{.6}(10)$ $=4.4$	$E[Y^1\|D=1]$ $=\frac{.08}{.4}(4)+\frac{.12}{.4}(8)$ $+\frac{.2}{.4}(14)$ $=10.2$	

expectations of the observed outcome variable conditional on D and S. The second panel of Table 4.2 presents corresponding sample estimates of the conditional probabilities of S given D.

The existence of a perfect stratification (and the supposed availability of data from an infinite sample) ensures that the estimated conditional expectations in the first panel of Table 4.2 equal the population-level conditional expectations of the second panel of Table 4.1. When stratifying by S, the average observed outcome for those in the control/treatment group with a particular value of S is equal to the average potential outcome under the control/treatment state for those with a particular value of S. Conversely, if S were not a perfect stratifying variable, then the sample means in the first panel of Table 4.2 would not equal the expectations of the potential outcomes in the second panel of Table 4.1. The sample means would be based on heterogeneous groups of individuals who differ systematically within the strata defined by S in ways that are correlated with individual-level treatment effects.

If S offers a perfect stratification of the data, then one can estimate from the numbers in the cells of the two panels of Table 4.2 both the average treatment effect among the treated as

$$(4-2)(.2)+(8-6)(.3)+(14-10)(.5)=3$$

Table 4.2: Estimated Conditional Expectations and Probabilities for Matching Demonstration 1

	Estimated mean observed outcome conditional on $s_i$and d_i	
	Control group	Treatment group
$s_i = 1$	$E_N[y_i \mid s_i = 1, d_i = 0] = 2$	$E_N[y_i \mid s_i = 1, d_i = 1] = 4$
$s_i = 2$	$E_N[y_i \mid s_i = 2, d_i = 0] = 6$	$E_N[y_i \mid s_i = 2, d_i = 1] = 8$
$s_i = 3$	$E_N[y_i \mid s_i = 3, d_i = 0] = 10$	$E_N[y_i \mid s_i = 3, d_i = 1] = 14$
	Estimated probability of S conditional on D	
$s_i = 1$	$\Pr_N[s_i = 1 \mid d_i = 0] = .6$	$\Pr_N[s_i = 1 \mid d_i = 1] = .2$
$s_i = 2$	$\Pr_N[s_i = 2 \mid d_i = 0] = .2$	$\Pr_N[s_i = 2 \mid d_i = 1] = .3$
$s_i = 3$	$\Pr_N[s_i = 3 \mid d_i = 0] = .2$	$\Pr_N[s_i = 3 \mid d_i = 1] = .5$

and the average treatment effect among the untreated as

$$(4-2)(.6) + (8-6)(.2) + (14-10)(.2) = 2.4.$$

Finally, if one calculates the appropriate marginal distributions of S and D (using sample analogs for the marginal distribution from the first panel of Table 4.1), one can perfectly estimate the unconditional average treatment effect either as

$$(4-2)(.44) + (8-6)(.24) + (14-10)(.32) = 2.64$$

or as

$$3(.6) + 2.4(.4) = 2.64.$$

Thus, for this hypothetical example, the naive estimator would be (asymptotically) upwardly biased for the average treatment effect among the treated, the average treatment effect among the untreated, and the unconditional average treatment effect. But, by appropriately weighting stratified estimates of the treatment effect, one can obtain unbiased and consistent estimates of these average treatment effects.

In general, if a stratifying variable S completely accounts for all systematic differences between those who take the treatment and those who do not, then conditional-on-S estimators yield consistent estimates of the average treatment effect conditional on a particular value s of S:

$$\{E_N[y_i \mid d_i = 1, s_i = s] - E_N[y_i \mid d_i = 0, s_i = s\} \xrightarrow{p} E[Y^1 - Y^0 \mid S = s] = E[\delta \mid S = s]. \tag{4.5}$$

Weighted sums of these stratified estimates can then be taken, such as for the unconditional average treatment effect:

$$\sum_s \{E_N[y_i \mid d_i = 1, s_i = s] - E_N[y_i \mid d_i = 0, s_i = s]\} \Pr{}_N[s_i = s] \xrightarrow{p} E[\delta]. \tag{4.6}$$

Substituting into this last expression the distributions of S conditional on the two possible values of D (i.e., $\Pr_N[s_i = s|d_i = 1]$ or $\Pr_N[s_i = s|d_i = 0]$), one can obtain consistent estimates of the average treatment effect among the treated and the average treatment effect among the untreated.

The key to using stratification to solve the causal inference problem for all three causal effects of primary interest is twofold: finding the stratifying variable and then obtaining the marginal probability distribution $\Pr[S]$ as well as the conditional probability distribution $\Pr[S|D]$. Once these steps are accomplished, obtaining consistent estimates of the within-strata treatment effects is straightforward. Then, consistent estimates of other average treatment effects can be formed by taking appropriate weighted averages of the stratified estimates.

This simple example shows all of the basic principles of matching estimators. Treatment and control subjects are matched together in the sense that they are grouped together into strata. Then, an average difference between the outcomes of treatment and control subjects is estimated, based on a weighting of the strata (and thus the individuals within them) by a common distribution. The imposition of the same set of stratum-level weights for those in both the treatment and control groups ensures that the data are balanced with respect to the distribution of S across treatment and control cases.

4.2.2 Overlap Conditions for Stratifying Variables

Suppose now that a perfect stratification of the data is available, but that there is a stratum in which no member of the population ever receives the treatment. Here, the average treatment effect is undefined. A hidden stipulation is built into Assumptions 1-S and 2-S if one wishes to be able to estimate the average treatment effect for the entire population. The "perfect" stratifying variables must not be so perfect that they sort deterministically individuals into either the treatment or the control. If so, the range of the stratifying variables will differ fundamentally for treatment and control cases, necessitating a redefinition of the causal effect of interest.[7]

Matching Demonstration 2

For the example depicted in Tables 4.3 and 4.4, S again offers a perfect stratification of the data. The setup of these two tables is exactly equivalent to that of the prior Tables 4.1 and 4.2 for Matching Demonstration 1. We again assume that the data are generated by a random sample of a well-defined population, and for simplicity of exposition that the sample is infinite. The major difference is evident in the joint distribution of S and D presented in the first panel of Table 4.3. As shown in the first cell of the second column, no individual with S equal to 1 would ever be observed in the treatment group of a dataset of

[7] In this section, we focus on the lack of overlap that may exist in a population (or superpopulation). For now, we ignore the lack of overlap that can emerge in observed data solely because of the finite size of a dataset. We turn to these issues in the next section, where we discuss solutions to sparseness.

Table 4.3: The Joint Probability Distribution and Conditional Population Expectations for Matching Demonstration 2

	Joint probability distribution of S and D		
	$D=0$	$D=1$	
$S=1$	$\Pr[S=1, D=0]=.4$	$\Pr[S=1, D=1]=0$	$\Pr[S=1]=.4$
$S=2$	$\Pr[S=2, D=0]=.1$	$\Pr[S=2, D=1]=.13$	$\Pr[S=2]=.23$
$S=3$	$\Pr[S=3, D=0]=.1$	$\Pr[S=3, D=1]=.27$	$\Pr[S=3]=.37$
	$\Pr[D=0]=.6$	$\Pr[D=1]=.4$	

	Potential outcomes		
	Under the control state	Under the treatment state	
$S=1$	$E[Y^0\|S=1]=2$		
$S=2$	$E[Y^0\|S=2]=6$	$E[Y^1\|S=2]=8$	$E[Y^1-Y^0\|S=2]=2$
$S=3$	$E[Y^0\|S=3]=10$	$E[Y^1\|S=3]=14$	$E[Y^1-Y^0\|S=3]=4$
	$E[Y^0\|D=0]$ $=\frac{.4}{.6}(2)+\frac{.1}{.6}(6)+\frac{.1}{.6}(10)$ $=4$	$E[Y^1\|D=1]$ $=\frac{.13}{.4}(8)+\frac{.27}{.4}(14)$ $=12.05$	

any size because the joint probability of S equal to 1 and D equal to 1 is zero. Corresponding to this structural zero in the joint distribution of S and D, the second panel of Table 4.3 shows that there is no corresponding conditional expectation of the potential outcome under the treatment state for those with S equal to 1. And, thus, as shown in the last column of the second panel of Table 4.3, no causal effect exists for individuals with S equal to 1.[8]

Adopting the college degree causal effect framing of the last hypothetical example in Matching Demonstration 1, this hypothetical example asserts that there is a subpopulation of individuals from such disadvantaged backgrounds that no individuals with $S=1$ have ever graduated from college. For this group of individuals, we assume in this example that there is simply no justification for using the wages of those from more advantaged social backgrounds to extrapolate to the what-if wages of the most disadvantaged individuals if they had somehow overcome the obstacles that prevent them from obtaining college degrees.

Table 4.4 shows what can be estimated for this example. If S offers a perfect stratification of the data, one could consistently estimate the treatment effect

[8]The naive estimate can be calculated for this example, and it would equal 8.05 for a very large sample because $[8(.325)+14(.675)]-[2(.667)+6(.167)+10(.167)]$ is equal to 8.05. See the last row of the table for the population analogs to the two pieces of the naive estimator.

Table 4.4: Estimated Conditional Expectations and Probabilities for Matching Demonstration 2

	Estimated mean observed outcome conditional on $s_i$and d_i	
	Control group	Treatment group
$s_i = 1$	$E_N[y_i \mid s_i = 1, d_i = 0] = 2$	
$s_i = 2$	$E_N[y_i \mid s_i = 2, d_i = 0] = 6$	$E_N[y_i \mid s_i = 2, d_i = 1] = 8$
$s_i = 3$	$E_N[y_i \mid s_i = 3, d_i = 0] = 10$	$E_N[y_i \mid s_i = 3, d_i = 1] = 14$
	Estimated probability of S conditional on D	
$s_i = 1$	$\Pr_N[s_i = 1 \mid d_i = 0] = .667$	$\Pr_N[s_i = 1 \mid d_i = 1] = 0$
$s_i = 2$	$\Pr_N[s_i = 2 \mid d_i = 0] = .167$	$\Pr_N[s_i = 2 \mid d_i = 1] = .325$
$s_i = 3$	$\Pr_N[s_i = 3 \mid d_i = 0] = .167$	$\Pr_N[s_i = 3 \mid d_i = 1] = .675$

for the treated as

$$(8 - 6)(.325) + (14 - 10)(.675) = 3.35.$$

However, there is no way to consistently estimate the treatment effect for the untreated, and hence no way to consistently estimate the unconditional average treatment effect.

Are examples such as this one ever found in practice? For an example that is more realistic than the causal effect of a college degree on economic success, consider the evaluation of a generic program in which there is an eligibility rule. The benefits of enrolling in the program for those who are ineligible cannot be estimated from the data, even though, if some of those individuals were enrolled in the program, they would likely be affected by the treatment in some way.[9]

Perhaps the most important point of this last example, however, is that the naive estimator is entirely misguided for this hypothetical application. The average treatment effect is undefined for the population of interest. More generally, not all causal questions have answers worth seeking even in best-case data availability scenarios, and sometimes this will be clear from the data and contextual knowledge of the application. However, at other times, the data may appear to suggest that no causal inference is possible for some group of individuals even though the problem is simply a small sample size. There is a clever solution to sparseness of data for these types of situations, which we discuss in the next section.

[9] Developing such estimates would require going well beyond the data, introducing assumptions that allow for extrapolation off of the common support of S.

4.3 Matching as Weighting

As shown in the last section, if all of the variables in S have been observed such that a perfect stratification of the data would be possible with an infinitely large random sample from the population, then a consistent estimator is available in theory for each of the average causal effects of interest defined in Equations (2.3), (2.5), and (2.6). However, in many (if not most) datasets of finite size, it may not be possible to use the simple estimation methods of the last section to generate consistent estimates. Treatment and control cases may be missing at random within some of the strata defined by S, such that some strata contain only treatment or only control cases. In this situation, some within-stratum causal effect estimates cannot be calculated. We now introduce a set of weighting estimators that rely on estimated propensity scores to solve the sort of data sparseness problems that afflict samples of finite size.

4.3.1 The Utility of Known Propensity Scores

An estimated propensity score is the estimated probability of taking the treatment as a function of variables that predict treatment assignment. Before the attraction of estimated propensity scores is explained, there is value in understanding why known propensity scores would be useful in an idealized context such as a perfect stratification of the data.

Within a perfect stratification, the true propensity score is nothing other than the within-stratum probability of receiving the treatment, or $\Pr[D = 1|S]$. For the hypothetical example in Matching Demonstration 1 (see Subsection 4.2.1), the propensity scores are:

$$\Pr[D = 1|S = 1] = \frac{.08}{.44} = .182,$$
$$\Pr[D = 1|S = 2] = \frac{.12}{.24} = .5,$$
$$\Pr[D = 1|S = 3] = \frac{.2}{.32} = .625.$$

Why is the propensity score useful? As shown earlier for Matching Demonstration 1, if a perfect stratification of the data is available, then the final ingredient for calculating average treatment effect estimates for the treated and for the untreated is the conditional distribution $\Pr[S|D]$. One can recover $\Pr[S|D]$ from the propensity scores by applying Bayes' rule using the marginal distributions of D and S. For example, for the first stratum,

$$\Pr[S = 1|D = 1] = \frac{\Pr[D = 1|S = 1]\Pr[S = 1]}{\Pr[D = 1]} = \frac{(.182)(.44)}{(.4)} = .2.$$

Thus, the true propensity scores encode all of the necessary information about the joint dependence of S and D that is needed to estimate and then combine conditional-on-S treatment effect estimates into estimates of the treatment effect

for the treated and the treatment effect for the untreated. Known propensity scores are thus useful for unpacking the inherent heterogeneity of causal effects and then averaging over such heterogeneity to calculate average treatment effects.

Of course, known propensity scores are almost never available to researchers working with observational rather than experimental data. Thus, the literature on matching more often recognizes the utility of propensity scores for addressing an entirely different concern: solving comparison problems created by the sparseness of data in any finite sample. These methods rely on estimated propensity scores, as we discuss next.

4.3.2 Weighting with Propensity Scores to Address Sparseness

Suppose again that a perfect stratification of the data exists and is known. In particular, Assumptions 1-S and 2-S in Equations (4.1) and (4.2) are valid, and the true propensity score is greater than 0 and less than 1 for every stratum defined by S. But, suppose now that (1) there are multiple variables in S and (2) some of these variables take on many values. In this scenario, there may be many strata in the available data in which no treatment or control cases are observed, even though the true propensity score is between 0 and 1 for every stratum in the population.

Can average treatment effects be consistently estimated in this scenario? Rosenbaum and Rubin (1983a) answer this question affirmatively. The essential points of their argument are the following (see the original article for a formal proof): First, the sparseness that results from the finiteness of a sample is random, conditional on the joint distribution of S and D. As a result, within each stratum for a perfect stratification of the data, the probability of having a zero cell in the treatment or the control state is solely a function of the propensity score. Because such sparseness is conditionally random, strata with identical propensity scores (i.e., different combinations of values for the variables in S but the same within-stratum probability of treatment) can be combined into a more coarse stratification. Over repeated samples from the same population, zero cells would emerge with equal frequency across all strata within these coarse propensity-score-defined strata.

Because sparseness emerges in this predictable fashion, stratifying on the propensity score itself (rather than more finely on all values of the variables in S) solves the sparseness problem because the propensity score can be treated as a single stratifying variable. In fact, as we show in the next hypothetical example, one can obtain consistent estimates of treatment effects by weighting the individual-level data by an appropriately chosen function of the estimated propensity score, without ever having to compute any stratum-specific causal effect estimates.

But how does one obtain the propensity scores for data from a random sample of the population of interest? Rosenbaum and Rubin (1983a) argue that, if one has observed the variables in S, then the propensity score can be

estimated using standard methods, such as logit modeling. That is, one can estimate the propensity score, assuming a logistic distribution,

$$\Pr[D = 1|S] = \frac{\exp(S\phi)}{1 + \exp(S\phi)}, \tag{4.7}$$

and invoke maximum likelihood to estimate a vector of coefficients $\hat{\phi}$. One can then stratify on the index of the estimated propensity score, $e(s_i) = s_i\hat{\phi}$, or appropriately weight the data as we show later, and all of the results established for known propensity scores then obtain.[10] Consider the following hypothetical example, in which weighting is performed only with respect to the estimated propensity score, resulting in unbiased and consistent estimates of average treatment effects even though sparseness problems are severe.

Matching Demonstration 3

Consider the following Monte Carlo simulation, which is an expanded version of the hypothetical example in Matching Demonstration 1 (see Subsection 4.2.1) in two respects. First, for this example, there are two stratifying variables, A and B, each of which has 100 separate values. As for Matching Demonstration 1, these two variables represent a perfect stratification of the data and, as such, represent all of the variables in the set of perfect stratifying variables, defined earlier as S. Second, to demonstrate the properties of alternative estimators, this example utilizes 50,000 samples of data, each of which is a random realization of the same set of definitions for the constructed variables and the stipulated joint distributions between them.

Generation of the 50,000 Datasets. For the simulation, we gave the variables A and B values of .01, .02, .03, and upward to 1. We then cross-classified the two variables to form a 100×100 grid and stipulated a propensity score, as displayed in Figure 4.1, that is a positive, nonlinear function of both A and B.[11] We then populated the resulting 20,000 constructed cells (100×100 for the $A \times B$ grid multiplied by the two values of D) using a Poisson random number generator with the relevant propensity score as the Poisson parameter for the 10,000 cells for the treatment group and one minus the propensity score as the Poisson parameter for the 10,000 cells for the control group. This sampling scheme generates (on average across simulated datasets) the equivalent of 10,000

[10] As Rosenbaum (1987) later clarified (see also Rubin and Thomas 1996), the estimated propensity scores do a better job of balancing the observed variables in S than the true propensity scores would in any actual application, because the estimated propensity scores correct for the chance imbalances in S that characterize any finite sample. This insight has led to a growing literature that seeks to balance variables in S by various computationally intensive but powerful nonparametric techniques. We discuss this literature later, and for now we present only parametric models, as they dominate the foundational literature on matching.

[11] The parameterization of the propensity score is a constrained tensor product spline regression for the index function of a logit. See Ruppert, Wand, and Carroll (2003) for examples of such parameterizations. Here, $S\phi$ in Equation (4.7) is equal to $-2 + 3(A) - 3(A - .1) + 2(A - .3) - 2(A - .5) + 4(A - .7) - 4(A - .9) + 1(B) - 1(B - .1) + 2(B - .7) - 2(B - .9) + 3(A - .5)(B - .5) - 3(A - .7)(B - .7)$.

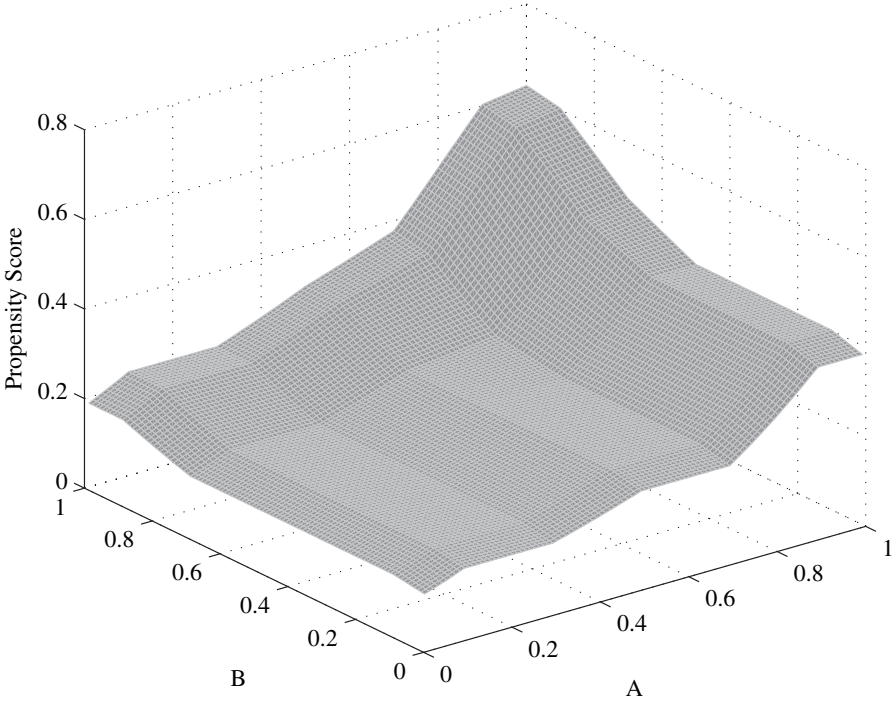


Figure 4.1: The propensity score specification for Matching Demonstration 3.

sample members, assigned to the treatment instead of the control as a function of the probabilities plotted in Figure 4.1.

Across the 50,000 simulated datasets, on average 7,728 of the 10,000 possible combinations of values for both A and B had no individuals assigned to the treatment, and 4,813 had no individuals assigned to the control. No matter the actual realized pattern of sparseness for each simulated dataset, all of the 50,000 datasets are afflicted, such that a perfect stratification on all values for the variables A and B would result in many strata within each dataset for which only treatment or control cases are present.

To define treatment effects for each dataset, two potential outcomes were defined as linear functions of individual values for A and B:

$$y_i^1 = 102 + 6a_i + 4b_i + \upsilon_i^1, \tag{4.8}$$
$$y_i^0 = 100 + 3a_i + 2b_i + \upsilon_i^0, \tag{4.9}$$

where both υ_i^1 and υ_i^0 are independent random draws from a normal distribution with expectation 0 and standard deviation of 5. Then, as in Equation (2.2), individuals from the treatment group were given an observed y_i equal to their simulated y_i^1, and individuals from the control group were given an observed y_i equal to their simulated y_i^0.

Table 4.5: Monte Carlo Means and Standard Deviations of Treatment Effects and Treatment Effect Estimates for Matching Demonstration 3

	Average treatment effect	Average treatment effect for the treated	Average treatment effect for the untreated
True treatment effects	4.525	4.892	4.395
	(.071)	(.139)	(.083)
Propensity-score-based weighting estimators:			
Misspecified propensity score estimates	4.456	4.913	4.293
	(.122)	(.119)	(.128)
Perfectly specified propensity score estimates	4.526	4.892	4.396
	(.120)	(.127)	(.125)
True propensity scores	4.527	4.892	4.396
	(.127)	(.127)	(.132)

With this setup, the simulation makes available 50,000 datasets for which the individual treatment effects can be calculated exactly, as true values of y_i^1 and y_i^0 are available for all simulated individuals. Because the true average treatment effect, treatment effect for the treated, and treatment effect for the untreated are thus known for each simulated dataset, these average effects can then serve as baselines against which alternative estimators that use data only on y_i , d_i, a_i, and b_i can be compared. The first row of Table 4.5 presents true Monte Carlo means and standard deviations of the three average treatments effects, calculated across the 50,000 simulated datasets. The mean of the average treatment effect across datasets is 4.525, whereas the means of the average treatment effects for the treated and for the untreated are 4.892 and 4.395, respectively. Similar to the hypothetical example in Matching Demonstration 1, this example represents a form of positive selection, in which those who are most likely to be in the treatment group are also those most likely to benefit from the treatment.[12] Accordingly, the treatment effect for the treated is larger than the treatment effect for the untreated.

Methods for Treatment Effect Estimation. The last three rows of Table 4.5 present results for three propensity-score-based weighting estimators. For the estimates in the second row, it is (wrongly) assumed that the propensity score can be estimated consistently with a logit model with linear terms for A and B

[12]It can also be represented by the DAG in Figure 3.8.

[i.e., assuming that, for Equation (4.7), a logit with $S\phi$ specified as $\alpha + \phi_A A + \phi_B B$ will yield consistent estimates of the propensity score surface plotted in Figure 4.1]. After the logit model was estimated for each of the 50,000 datasets with the wrong specification, the estimated propensity score for each individual was then calculated,

$$\hat{p}_i = \frac{\exp(\hat{\alpha} + \hat{\phi}_A a_i + \hat{\phi}_B b_i)}{1 + \exp(\hat{\alpha} + \hat{\phi}_A a_i + \hat{\phi}_B b_i)}, \tag{4.10}$$

along with the estimated odds of the propensity of being assigned to the treatment:

$$\hat{r}_i = \frac{\hat{p}_i}{1 - \hat{p}_i}, \tag{4.11}$$

where $\hat{p}_i$ is as constructed in Equation (4.10).

To estimate the treatment effect for the treated, we then implemented a weighting estimator by calculating the average outcome for the treated and subtracting from this average outcome a counterfactual average outcome using weighted data on those from the control group:

$$\hat{\delta}_{\text{TT,weight}} \equiv \left(\frac{1}{n^1} \sum_{i:d_i=1} y_i \right) - \left(\frac{\sum_{i:d_i=0} \hat{r}_i y_i}{\sum_{i:d_i=0} \hat{r}_i} \right), \tag{4.12}$$

where n^1 is the number of individuals in the treatment group and $\hat{r}_i$ is the estimated odds of being in the treatment group instead of in the control group, as constructed in Equations (4.10) and (4.11). The weighting operation in the second term gives more weight to control group individuals equivalent to those in the treatment group (see Rosenbaum 1987, 2002).[13] To estimate the treatment effect for the untreated, we then implemented a weighting estimator that is the mirror image of the one in Equation (4.12):

$$\hat{\delta}_{\text{TUT,weight}} \equiv \left(\frac{\sum_{i:d_i=1} y_i/\hat{r}_i}{\sum_{i:d_i=1} n^1/\hat{r}_i} \right) - \left(\frac{1}{n^0} \sum_{i:d_i=0} y_i \right), \tag{4.13}$$

where n^0 is the number of individuals in the control group. Finally, the corresponding estimator of the unconditional average treatment effect is

$$\hat{\delta}_{\text{ATE,weight}} \equiv \left(\frac{1}{n} \sum_i d_i \right) \left(\hat{\delta}_{\text{TT,weight}} \right) + \left[\left(1 - \frac{1}{n} \sum_i d_i \right) \right] \left(\hat{\delta}_{\text{TUT,weight}} \right), \tag{4.14}$$

[13] As we will describe later when discussing the connections between matching and regression, the weighting estimator in Equation (4.12) can be written as a weighted OLS regression estimator.

where $\hat{\delta}_{\text{TT,weight}}$ and $\hat{\delta}_{\text{TUT,weight}}$ are as defined in Equations (4.12) and (4.13), respectively. Accordingly, an average treatment effect estimate is simply a weighted average of the two conditional average treatment effect estimates.

The same basic weighting scheme is implemented for the third row of Table 4.5, but the estimated propensity score utilized to define the estimated odds of treatment, $\hat{r}_i$, is instead based on results from a flawlessly estimated propensity score equation (i.e., one that uses the exact same specification that was fed to the random-number generator that assigned individuals to the treatment; see prior note on page 100 for the specification). Finally, for the last row of Table 4.5, the same weighting scheme is implemented, but, in this case, the estimated odds of treatment, $\hat{r}_i$, are replaced with the true odds of treatment, r_i, as calculated with reference to the exact function that generated the propensity score for Figure 4.1.

Monte Carlo Results. The naive estimator would yield a value of 5.388 for this example, which is substantially larger than each of the three true average treatment effects presented in the first row of Table 4.5. The second row of the table presents three estimates from the weighting estimators in Equations (4.12)–(4.14), using weights based on the misspecified logit described earlier. These estimates are closer to the true values presented in the first row (and much closer than the naive estimate), but the misspecification of the propensity-score-estimating equation leads to some systematic bias in the estimates. The third row of the table presents another three weighting estimates, using a perfect specification of the propensity-score-estimating equation, and now the estimates are asymptotically shown to be unbiased and consistent for the average treatment effect, the treatment effect for the treated, and the treatment effect for the untreated. Finally, the last row presents weighting estimates that utilize the true propensity scores and are also asymptotically unbiased and consistent (but, as shown by Rosenbaum 1987, more variable than those based on the flawlessly estimated propensity score; see also Hahn 1998; Hirano, Imbens, and Ridder 2003; Rosenbaum 2002).

The last two rows demonstrate the most important claim of the literature: If one can obtain consistent estimates of the true propensity scores, one can solve entirely the problems created by sparseness of data.

This example shows the potential power of propensity-score-based modeling. If treatment assignment can be modeled perfectly, one can solve the sparseness problems that afflict finite datasets, at least in so far as offering estimates that are consistent. On the other hand, this simulation also develops an important qualification of this potential power. Without a perfect specification of the propensity-score-estimating equation, one cannot rest assured that unbiased and consistent estimates can be obtained. Because propensity scores achieve their success by "undoing" the treatment assignment patterns, analogous to weighting a stratified sample, systematically incorrect estimated propensity scores can generate systematically incorrect weighting schemes that yield biased and inconsistent estimates of treatment effects.[14]

[14]There is also the larger issue of whether the challenges of causal inference can be reduced to mere concerns about conditionally random sparseness, and this will depend entirely on

Given the description of matching estimators offered in Section 4.1 (i.e., algorithms for mechanically identifying matched sets of equivalent treatment and control cases), in what sense are the individual-level weighting estimators of the hypothetical example in Matching Demonstration 3 equivalent to matching estimators?

As emphasized earlier for the hypothetical examples in Matching Demonstrations 1 and 2, stratification estimators have a straightforward connection to matching. The strata that are formed represent matched sets, and a weighting procedure is then used to average stratified treatment effect estimates in order to obtain the average treatment effects of interest. The propensity score weighting estimators, however, have a less straightforward connection. Here, the data are, in effect, stratified coarsely by the estimation of the propensity score (i.e., because all individuals in the same strata, as defined by the stratifying variables in S, are given the same estimated propensity scores), and then the weighting is performed directly across individuals instead of across the strata. This type of individual-level weighting is made necessary because of sparseness (as some of the fine strata for which propensity scores are estimated necessarily contain only treatment or control cases, thereby preventing the direct calculation of stratified treatment effect estimates). Nonetheless, the same principle of balancing holds: Individuals are weighted within defined strata in order to ensure that the distribution of S is the same within the treatment and control cases that are then used to estimate the treatment effects.

In the opposite direction, it is important to recognize that the algorithmic matching estimators that we summarize in the next section can be considered weighting estimators. As we show next, these data analysis procedures warrant causal inference by achieving an as-if stratification of the data that results in a balanced distribution of covariates across matched treatment and control cases. Although it is sometimes easier to represent matching estimators as algorithmic data analysis procedures that mechanically match seemingly equivalent cases to each other, it is best to understand matching as a method to weight the data in order to warrant causal inference by balancing S across the treatment and control cases.

4.4 Matching as a Data Analysis Algorithm

Algorithmic matching estimators differ primarily in (1) the number of matched cases designated for each to-be-matched target case and (2) how multiple matched cases are weighted if more than one is utilized for each target case. In this section, we describe the four main types of matching estimators.

Heckman, Ichimura, and Todd (1997, 1998) and Smith and Todd (2005) outline a general framework for representing alternative matching estimators, and we follow their lead. With our notation, all matching estimators of the

whether one is justified in imposing assumptions on the potential outcomes and treatment assignment patterns, as outlined earlier.

treatment effect for the treated would be defined in this framework as

$$\hat{\delta}_{\text{TT,match}} = \frac{1}{n^1} \sum_i \left[(y_i | d_i = 1) - \sum_j \omega_{i,j} (y_j | d_j = 0) \right], \tag{4.15}$$

where n^1 is the number of treatment cases, i is the index over treatment cases, j is the index over control cases, and $\omega_{i,j}$ represents a set of scaled weights that measure the distance between each control case and the target treatment case. In Equation (4.15), the weights are entirely unspecified.

Alternative matching estimators of the treatment effect for the treated can be represented as different procedures for deriving the weights represented by $\omega_{i,j}$. As we will describe next, the weights can take on many values, indeed as many $n^1 \times n^0$ different values, because alternative weights can be used when constructing the counterfactual value for each target treatment case. The difference in the propensity score is the most common distance measure used to construct weights. Other measures of distance are available, including the estimated odds of the propensity score, the difference in the index of the estimated logit, and the Mahalanobis metric.[15]

Before describing the four main types of matching algorithms, we note two important points. First, for simplicity of presentation, in this section we will focus on matching estimators of the treatment effect for the treated. Each of the following matching algorithms could be used in reverse, instead focusing on matching treatment cases to control cases in order to construct an estimate of the treatment effect for the untreated. We mention this, in part, because it is sometimes implied in the applied literature that the matching techniques that we are about to summarize are useful for estimating only the treatment effect for the treated. This is false. If (1) all variables in S are known and observed, such that a perfect stratification of the data could be formed with a suitably large dataset because both Assumptions 1-S and 2-S in Equations (4.1) and (4.2) are valid and (2) the ranges of all of the variables in S are the same for both treatment and control cases, then simple variants of the matching estimators that we will present in this section can be formed that are consistent for the treatment effect among the treated, the treatment effect among the untreated, and the average treatment effect.

Moreover, to consistently estimate the treatment effect for the treated, one does not need to assume full ignorability of treatment assignment or that both Assumptions 1-S and 2-S in Equations (4.1) and (4.2) are valid. Instead, only Assumption 2-S (i.e., $E[Y^0|D = 1, S] = E[Y^0|D = 0, S]$) must hold.[16] In

[15]The Mahalanobis metric is $(S_i - S_j)'\Sigma^{-1}(S_i - S_j)$, where Σ is the covariance matrix of the variables in S (usually calculated for the treatment cases only). There is a long tradition in this literature of using Mahalanobis matching in combination with propensity score matching.

[16]To estimate the treatment effect for the treated, the ranges of the variables in S must be the same for the treatment and control cases. We do not mention this requirement in the text, as there is a literature (see Heckman, Ichimura, and Todd 1997, 1998), which we discuss later, that defines the treatment effect for the treated on the common support and argues that this is often the central goal of analysis. Thus, even if the support of S is not the same in the

other words, to estimate the average treatment effect among the treated, it is sufficient to assume that, conditional on S, the average level of the outcome under the control for those in the treatment is equal, on average, to the average level of the outcome under the control for those in the control group.[17] This assumption is still rather stringent, in that it asserts that those in the control group do not disproportionately gain from exposure to the control state more than would those in the treatment group if they were instead in the control group. But it is surely weaker than having to assert Assumptions 1-S and 2-S together.[18]

Second, as we show in a later section, the matching algorithms we summarize next are data analysis procedures that can be used more generally even when some of the variables in S are unobserved. The matching estimators may still be useful, as argued by Rosenbaum (2002), as a set of techniques that generates a provisional set of causal effect estimates that can then be subjected to further analysis.

4.4.1 Basic Variants of Matching Algorithms

Exact Matching

For the treatment effect for the treated, exact matching constructs the counterfactual for each treatment case using the control cases with identical values on the variables in S. In the notation of Equation (4.15), exact matching uses weights equal to $1/k$ for matched control cases, where k is the number of matches selected for each target treatment case. Weights of 0 are given to all unmatched control cases. If only one match is chosen randomly from among possible exact matches, then $\omega_{i,j}$ is set to 1 for the randomly selected match (from all available exact matches) and to 0 for all other control cases. Exact matching may be combined with any of the matching methods described later.

Nearest-Neighbor Matching

For the treatment effect for the treated, nearest-neighbor matching constructs the counterfactual for each treatment case using the control cases that are closest to the treatment case on a unidimensional measure constructed from the variables in S, usually an estimated propensity score but sometimes variants of propensity scores (see Althauser and Rubin 1970; Cochran and Rubin 1973; Rosenbaum and Rubin 1983a, 1985a, 1985b; Rubin 1973a, 1973b, 1976,

treatment and control groups, an average treatment effect among a subset of the treated can be estimated.

[17] There is an ignorability variant of this mean-independence assumption: D is independent of Y^0 conditional on S. One would always prefer a study design in which this more encompassing form of independence holds. Resulting causal estimates would then hold under transformations of the potential outcomes. This would be particularly helpful if the directly mapped Y [defined as $DY^1 + (1 - D)Y^0$] is not observed but some monotonic transformation of Y is observed (as could perhaps be generated by a feature of measurement).

[18] And this is again weaker than having to assert an assumption of ignorability of treatment assignment.

1980a,1980b). The traditional algorithm randomly orders the treatment cases and then selects for each treatment case the control case with the smallest distance. The algorithm can be run with or without replacement. With replacement, a control case is returned to the pool after a match and can be matched later to another treatment case. Without replacement, a control case is taken out of the pool once it is matched.[19]

If only one nearest neighbor is selected for each treatment case, then $\omega_{i,j}$ is set equal to 1 for the matched control case and 0 for all other control cases. One can also match multiple nearest neighbors to each target treatment case, in which case $\omega_{i,j}$ is set equal to $1/k_i$ for the matched nearest neighbors, where k_i is the number of matches selected for each target treatment case i. Matching more control cases to each treatment case results in lower expected variance of the treatment effect estimate but also raises the possibility of greater bias, because the probability of making more poor matches increases.

A danger with nearest-neighbor matching is that it may result in some very poor matches for treatment cases. A version of nearest-neighbor matching, known as caliper matching, is designed to remedy this drawback by restricting matches to some maximum distance. With this type of matching, some treatment cases may not receive matches, and thus the effect estimate will apply to only the subset of the treatment cases matched (even if ignorability holds and there is simply sparseness in the data).[20]

Interval Matching

Interval matching (also referred to as subclassification and stratification matching) sorts the treatment and control cases into segments of a unidimensional metric, usually the estimated propensity score, and then calculates the treatment effect within these intervals (see Cochran 1968; Rosenbaum and Rubin 1983a, 1984; Rubin 1977). For each interval, a variant of the matching estimator in Equation (4.15) is estimated separately, with $\omega_{i,j}$ chosen to give the same amount of weight to the treatment cases and control cases within each interval. The average treatment effect for the treated is then calculated as the mean of the interval-specific treatment effects, weighted by the number of treatment cases in each interval. This method is nearly indistinguishable from nearest-neighbor caliper matching with replacement when each of the intervals includes exactly one treatment case.

[19] One weakness of the traditional algorithm when used without replacement is that the estimate will vary depending on the initial ordering of the treatment cases. A second weakness is that without replacement the sum distance for all treatment cases will generally not be the minimum because control cases that might make better matches to later treatment cases may be used early in the algorithm. See our discussion of optimal matching later.

[20] A related form of matching, known as radius matching (see Dehejia and Wahba 2002), matches all control cases within a particular distance – the "radius" – from the treatment case and gives the selected control cases equal weight. If there are no control cases within the radius of a particular treatment case, then the nearest available control case is used as the match.

Kernel Matching

Developed by Heckman, Ichimura, Smith, and Todd (1998) and Heckman, Ichimura, and Todd (1997, 1998) kernel matching constructs the counterfactual for each treatment case using all control cases but weights each control case based on its distance from the treatment case. The weights represented by $\omega_{i,j}$ in Equation (4.15) are calculated with a kernel function, $G(.)$, that transforms the distance between the selected target treatment case and all control cases in the study. When the estimated propensity score is used to measure the distance, kernel-matching estimators define the weight as

$$\omega_{ij} = \frac{G[\frac{\hat{p}(s_j)-\hat{p}(s_i)}{a_n}]}{\sum_j G[\frac{\hat{p}(s_j)-\hat{p}(s_i)}{a_n}]}, \tag{4.16}$$

where a_n is a bandwidth parameter that scales the difference in the estimated propensity scores based on the sample size and $\hat{p}(.)$ is the estimated propensity score as a function of its argument.[21] The numerator of this expression yields a transformed distance between each control case and the target treatment case. The denominator is a scaling factor equal to the sum of all the transformed distances across control cases, which is needed so that the sum of $\omega_{i,j}$ is equal to 1 across all control cases when matched to each target treatment case.

Although kernel-matching estimators appear complex, they are a natural extension of interval and nearest-neighbor matching: All control cases are matched to each treatment case but weighted so that those closest to the treatment case are given the greatest weight. Smith and Todd (2005) offer an excellent intuitive discussion of kernel matching along with generalizations to local linear matching (Heckman, Ichimura, Smith, and Todd 1998; Heckman, Ichimura, and Todd 1997, 1998) and local quadratic matching (Ham, Li, and Reagan 2003).

4.4.2 Which of These Basic Matching Algorithms Works Best?

There is very little specific guidance in the literature on which of these matching algorithms works best, and the answer very likely depends on the substantive application. Smith and Todd (2005), Heckman, Ichimura, Smith, and Todd (1998), and Heckman, Ichimura, and Todd (1997, 1998) have experimental data against which matching estimators can be compared, and they argue for the advantages of kernel matching (and a particular form of robust kernel matching). To the extent that a general answer to this question can be offered, we would suggest that nearest-neighbor caliper matching with replacement, interval matching, and kernel matching are all closely related and should be preferred to nearest-neighbor matching without replacement. If the point of a matching estimator is to minimize bias by comparing target cases with similar matched

[21] Increasing the bandwidth increases bias but lowers variance. Smith and Todd (2005) find that estimates are fairly insensitive to the size of the bandwidth.

cases, then methods that make it impossible to generate poor matches should be preferred.[22] Matching on both the propensity score and the Mahalanobis metric has also been recommended for achieving balance on higher-order moments (see Diamond and Sekhon 2005; Rosenbaum and Rubin 1985b).[23] Because there is no clear guidance on which of these matching estimators is "best," we constructed a fourth hypothetical example to give a sense of how often alternative matching estimators yield appreciably similar estimates.

Matching Demonstration 4

For this example, we use simulated data for which we defined the potential outcomes and treatment assignment patterns so that we can explore the relative performance of alternative matching estimators. The former are estimated under alternative scenarios with two different specifications of the propensity-score-estimating equation. Unlike the hypothetical example in Matching Demonstration 3, we do not repeat the simulation for multiple samples but confine ourselves to results on a single sample, as would be typical of any real-world application.

Generation of the Dataset. The dataset that we constructed mimics the dataset from the National Education Longitudinal Study (NELS) analyzed by Morgan (2001). For that application, regression and matching estimators were used to estimate the effect of Catholic schooling on the achievement of high school students in the United States (for a summary of research on this question, see Chapter 1). For our simulation here, we generated a dataset of 10,000 individuals with values for 13 baseline variables that resemble closely the joint distribution of the similar variables in Morgan (2001). The variables for respondents include dummy variables for race, region, urbanicity, whether they have their own bedrooms, whether they live with two parents, an ordinal variable for number of siblings, and a continuous variable for socioeconomic status. Then we created an entirely hypothetical cognitive skill variable, assumed to reflect innate and acquired skills in unknown proportions.[24]

[22] Another criterion for choosing among alternative matching estimators is relative efficiency. Our reading of the literature suggests that little is known about the relative efficiency of these estimators (see especially Abadie and Imbens 2006; Hahn 1998; Imbens 2004), even though there are claims in the literature that kernel-based methods are the most efficient. The efficiency advantage of kernel-matching methods is only a clear guide to practice if kernel-based methods are known to be no more biased than alternatives. But the relative bias of kernel-based methods is application dependent and should interact further with the bandwidth of the kernel. Thus, it seems that we will know for sure which estimators are most efficient for which types of applications only when statisticians discover how to calculate the sampling variances of all alternative estimators. Thereafter, it should be possible to compute mean-squared-error comparisons across alternative estimators for sets of typical applications.

[23] One method for matching on both the Mahalanobis metric and the propensity score is to include the propensity score in the Mahalanobis metric. A second is to use interval matching and divide the data into blocks by use of one metric and then match on the second metric within blocks.

[24] To be precise, we generated a sample using a multinomial distribution from a race-by-region-by-urbanicity grid from the data in Morgan (2001). We then simulated socioeconomic status as random draws from normal distributions with means and standard deviations

We then defined potential outcomes for all 10,000 individuals, assuming that the observed outcome of interest is a standardized test taken at the end of high school. For the potential outcome under the control (i.e., a public school education), we generated what-if test scores from a normal distribution, with an expectation as a function of race, region, urbanicity, number of siblings, socioeconomic status, family structure, and cognitive skills. We then assumed that the what-if test scores under the treatment (i.e., a Catholic school education) would be equal to the outcome under the control plus a boosted outcome under the treatment that is function of race, region, and cognitive skills (under the assumption, based on the dominant position in the extant literature, that black and Hispanic respondents from the north, as well as all of those with high preexisting cognitive skills, are disproportionately likely to benefit from Catholic secondary schooling).

We then defined the probability of attending a Catholic school using a logit with 26 parameters, based on a specification from Morgan (2001) along with an assumed self-selection dynamic in which individuals are slightly more likely to select the treatment as a function of the relative size of their individual-level treatment effect.[25] This last component of the logit creates a nearly insurmountable challenge, because in any particular application one would not have such a variable with which to estimate a propensity score. That, however, is our point in including this term, as individuals are thought, in many real-world applications, to be selecting from among alternative treatments based on accurate expectations, unavailable as measures to researchers, of their likely gains from alternative treatment regimes. The probabilities defined by the logit were then passed to a binomial distribution, which resulted in 986 of the 10,000 simulated students attending Catholic schools. Finally, observed outcomes were assigned according to treatment status.

With the sample divided into the treatment group and the control group, we calculated from the prespecified potential outcomes the true baseline average treatment effects. The treatment effect for the treated is 6.96 in the simulated data, whereas the treatment effect for the untreated is 5.9. In combination, the average treatment effect is then 6.0.

Methods for Treatment Effect Estimation. In Table 4.6, we offer 12 separate types of matching estimates. These are based on routines written for Stata by three sets of authors: Abadie, Drukker, Herr, and Imbens (2004), Becker

estimated separately for each of the race-by-region-by-urbanicity cells. Then, we generated all other variables iteratively, building on top of these variables, using joint distributions (where possible) based on estimates from the NELS data. Because we relied on standard parametric distributions, the data are smoother than the original NELS data.

[25] The index of the assumed logit was $-4.6 - .69$(Asian)$+.23$(Hispanic) $-.76$(black) $-.46$ (native American) $+2.7$(urban) $+1.5$(northeast) $+ 1.3$(north central) $+ .35$(south) $- .02$(siblings) $-.018$(bedroom) $+ .31$(two parents) $+ .39$(socioeconomic status) $+.33$(cognitive skills) $-.032$(socioeconomic status squared) $-.23$(cognitive skills squared) $-.084$(socioeconomic status)(cognitive skills) $-.37$(two parents)(black) $+ 1.6$(northeast)(black) $-.38$(north central)(black) $+ .72$(south)(black) $+ .23$(two parents)(Hispanic) $-.74$(northeast)(Hispanic) -1.3(north central)(Hispanic) $-.13$(south)(Hispanic) $+ .25$(individual treatment effect $-$ average treatment effect).

and Ichino (2002), and Leuven and Sianesi (2003).[26] We estimate all matching estimators under two basic scenarios. First, we offer a set of estimates based on poorly estimated propensity scores, derived from an estimating equation from which we omitted nine interaction terms along with the cognitive skill variable. The last specification error is particularly important because the cognitive skill variable has a correlation of 0.401 with the outcome variable and 0.110 with the treatment variable in the simulated data. For the second scenario, we included the cognitive skill variable and the nine interaction terms. Both scenarios lack an adjustment for the self-selection dynamic, in which individuals select into the treatment partly as a function of their expected treatment effect.

Regarding the specific settings for the alternative matching estimators, which are listed in the row headings of Table 4.6, the interval matching algorithm began with five blocks and subdivided blocks until each block achieved balance on the estimated propensity score across treatment and control cases. Nearest-neighbor matching with replacement was implemented with and without a caliper of 0.001, in both one- and five-nearest-neighbor variants. Radius matching was implemented with a radius of 0.001. For the kernel-matching estimators, we used two types of kernels – Epanechnikov and Gaussian – and the default bandwidth of 0.06 for both pieces of software. For the local linear matching estimator, we used the Epanechnikov kernel with the default bandwidth of 0.08.

Results. We estimated treatment effects under the assumption that self-selection on the individual-level Catholic school effect is present, and yet cannot be adjusted for using a statistical model without a measure of individuals' expectations. Thus, we operate under the assumption that only the treatment effect for the treated has any chance of being estimated consistently, as in the study by Morgan (2001) on which this example is based. We therefore compare all estimates to the true simulated treatment effect for the treated, identified earlier as 6.96

Estimates based on the poorly estimated propensity scores are reported in the first column of Table 4.6, along with the implied bias as an estimate of the treatment effect for the treated in the second column (i.e., the matching estimate minus 6.96). As expected, all estimates have a substantial positive bias. Most of the positive bias results from the mistaken exclusion of the cognitive skill variable from the propensity-score-estimating equation.

Matching estimates made with the well-estimated propensity scores are reported in the third column of Table 4.6, along with the expected bias in the fourth column. On the whole, these estimates are considerably better. Having the correct specification reduces the bias in those estimates with the largest bias from column three, and on average all estimates oscillate around the true treatment effect for the treated of 6.96.

We have demonstrated three basic points with this example. First, looking across the rows of Table 4.6, one clearly sees that matching estimators and different software routines yield different treatment effect estimates (even ones

[26] We do not provide a review of software routines because such a review would be immediately out-of-date on publication. At present, three additional sets of routines seem to be in use in the applied literature (see Hansen 2004b; Ho et al. 2004; Sekhon 2005).

Table 4.6: Matching Estimates for the Simulated Effect of Catholic Schooling on Achievement, as Specified in Matching Demonstration 4

	Poorly specified propensity score-estimating equation		Well-specified propensity score-estimating equation	
Method	TT estimate	Bias	TT estimate	Bias
Interval with variable blocks (B&I)	7.93	0.97	6.73	−0.23
One nearest-neighbor with caliper = 0.001 (L&S)	8.16	1.20	6.69	−0.27
One nearest-neighbor without caliper (A)	7.90	0.94	6.62	−0.34
Five nearest-neighbors with caliper = 0.001 (L&S)	7.97	1.01	7.04	0.08
Five nearest-neighbors without caliper (A)	7.85	0.89	7.15	0.19
Radius with radius = 0.001 (L&S)	8.02	1.06	6.90	−0.06
Radius with radius = 0.001 (B&I)	8.13	1.17	7.29	0.33
Kernel with Epanechnikov kernel (L&S)	7.97	1.01	6.96	0.00
Kernel with Epanechnikov kernel (B&I)	7.89	0.93	6.86	−0.10
Kernel with Gaussian kernel (L&S)	8.09	1.13	7.18	0.22
Kernel with Gaussian kernel (B&I)	7.97	1.01	7.03	0.07
Local linear with Epanechnikov kernel (L&S)	7.91	0.95	6.84	−0.12

Notes: B&I denotes the software of Becker and Ichino (2002); L&S denotes the software of Leuven and Sianesi (2003); A denotes the software of Abadie et al. (2004).

that are thought to be mathematically equivalent). Thus, at least for the near future, it will be crucial for researchers to examine multiple estimates of the same treatment effect across estimators and software packages. The lack of similarity across seemingly equivalent estimators from alternative software routines is surprising, but we assume that this unexpected variation will dissipate with software updates.

Second, matching estimators cannot compensate for an unobserved covariate in S, which leads to comparisons of treatment and control cases that are not identical in all relevant aspects other than treatment status. The absence of the cognitive skill variable in this example invalidates both Assumption 1-S and

2-S. The matching routines still balance the variables included in the propensity-score-estimating equation, but the resulting matching estimates remain biased and inconsistent for both the average treatment effect and the average treatment effect for the treated.

Third, the sort of self-selection dynamic built into this example – in which individuals choose Catholic schooling as a function of their expected gains from Catholic schooling – makes estimation of both the average treatment effect among the untreated and the average treatment effect impossible (because Assumption 1-S cannot be maintained). Fortunately, if all variables in S other than anticipation of the individual-level causal effects are observed (i.e., including cognitive skill in this example), then the average treatment effect among the treated can be estimated consistently.[27]

Unfortunately, violation of the assumption of ignorable treatment assignment (and of both Assumptions 1-S and 2-S) is the scenario in which most analysts will find themselves, and this is the scenario to which we turn in the next section. Before discussing what can be done in these situations, we first close the discussion on which types of matching may work best.

4.4.3 Matching Algorithms That Seek Optimal Balance

For the hypothetical example in Matching Demonstration 4, we judged the quality of matching algorithms by examining the distance between the treatment effect estimates that we obtained and the true treatment effects that we stipulated in constructing our hypothetical data. Because we generated only one sample, these differences are not necessarily a very good guide to practice, even though our main goal of the example was to show that alternative matching estimators generally yield different results and none of these may be correct. That point aside, it is generally recognized that the best matching algorithms are those that optimize balance in the data being analyzed. Building on this consensus, a broader set of matching algorithms is currently in development, which grows out of the optimal matching proposals attributed to Rosenbaum (1989).

Matching is generally judged to be successful if, for both the treatment and matched control groups, the distribution of the matching variables is the same. When this result is achieved, the data are said to be balanced, as noted earlier. [See also our discussion of Equation (3.7).] Assessing balance, however, can be difficult for two reasons. First, evaluating the similarity of full distributions necessitates going beyond an examination of differences in means (see Abadie

[27] At the same time, this example shows that even our earlier definition of a "perfect stratification" is somewhat underspecified. According to the definition stated earlier, if self-selection on the causal effect occurs, a perfect stratification is available only if variables that accurately measure anticipation of the causal effect for each individual are also available and duly included in S. Thus, perhaps it would be preferable to refer to three types of perfect stratification: one for which Assumption 1-S is valid (which enables estimation of the average treatment effect for the untreated), one for which Assumption 2-S is valid (which enables estimation of the average treatment for the treated), and one for which both are valid (which enables estimation of the average treatment effect, as well as the average treatment effects for the treated and the untreated).

2002). Second, the use of any hypothesis test of similarity has two associated dangers. With small samples, the null hypothesis of no difference may be accepted when in fact the data are far from balanced (i.e., a generic type II error). Second, with very large datasets, almost any difference, however small, is likely to be statistically significant. As such, hypothesis tests are generally less useful for assessing balance than standardized differences and their generalizations.[28] Imai, King, and Stuart (2006) provide a full discussion of these issues.

If the covariates are not balanced, the estimation model for the propensity score can be changed, for example, by the addition of interaction terms, quadratic terms, or other higher-order terms. Or, matching can be performed on the Mahalanobis metric in addition to the propensity score, perhaps nesting one set of matching strategies within another. This respecification is not considered data mining because it does not involve examining the effect estimate. But it can be labor intensive, and there is no guarantee that one will find the best possible balance by simply reestimating the sorts of matching algorithms introduced earlier, or by combining them in novel ways.

For this reason, two more general forms of matching have been proposed, each of which is now fairly well developed. Rosenbaum (2002, Chapter 10) reports on recent results for full optimal matching algorithms that he has achieved with colleagues since Rosenbaum (1989). His algorithms seek to optimize balance and efficiency of estimation by searching through all possible matches that could be made, after stipulating the minimum and maximum number of matches for matched sets of treatment and control cases. Although full optimal matching algorithms vary (see also Hansen 2004a), they are based on the idea of minimizing the average distance between the estimated propensity scores among matched cases. If the estimated propensity scores are correct, then this minimization problem should balance S.

Diamond and Sekhon (2005) propose a general multivariate matching method that uses a genetic algorithm to search for the match that achieves the best possible balance. Although their algorithm can be used to carry out matching after the estimation of a propensity score, their technique is more general and can almost entirely remove the analyst from having to make any specification choices other than designating the matching variables. Diamond and Sekhon (2005) show that their matching algorithms provide superior balance in both Monte Carlo simulations and a test with genuine data.[29]

[28] The standardized difference for a matching variable X is $\frac{|E_N[x_i|d_i=1]-E_N[x_i|d_i=0]|}{\sqrt{\frac{1}{2}\mathrm{Var}_N[x_i|d_i=1]+\frac{1}{2}\mathrm{Var}_N[x_i|d_i=0]}}$. Because this index is a scaled absolute difference in the means of the X across treatment and control cases, it can be compared across alternative Xs. It is generally a better criterion for balance assessment than t statistics are. However, like t statistics, this index considers only differences in the mean of X across matched treatment and control cases. Indices of higher moments should be considered as well.

[29] For Diamond and Sekhon (2005), balance is assessed by using t-tests of differences in means and also bootstrapped Kolmogorov–Smirnov tests for the full distributions of the matching variables. It is unclear how sensitive their results are to the usage of balance tests that are insensitive to sample size. Their algorithm, however, appears general enough that such modifications can be easily incorporated.

Although there is good reason to expect that these types of matching algorithms can outperform the nearest-neighbor, interval, and kernel-matching algorithms by the criteria of balance, they are considerably more difficult to implement in practice. With software developments underway, these disadvantages will be eliminated.

4.5 Matching When Treatment Assignment is Nonignorable

What if neither Assumption 1-S nor Assumption 2-S is valid because we observe only a subset of the variables in S, which we will now denote by X? We can still match on X using the techniques just summarized, as we did for the first column of Table 4.6 in the hypothetical example for Matching Demonstration 4.

Consider, for example, the working paper of Sekhon (2004), in which a matching algorithm is used to balance various predictors of voting at the county level in an attempt to determine whether or not John Kerry lost votes in the 2004 presidential election campaign because optical scan voting machines were used instead of direct electronic voting machines in many counties (see Subsection 1.3.2 on voting technology effects in Florida for the 2000 election). Sekhon shows that it is unlikely that voting technology caused John Kerry to lose votes. In this analysis, ignorability is not asserted in strict form, as it is quite clear that unobserved features of the counties may well have been correlated with both the distribution of votes and voting technology decisions. Nonetheless, the analysis is convincing because the predictors of treatment assignment are quite rich, and it is hard to conceive of what has been left out.

When in this position, however, it is important to concentrate on estimating only one type of treatment effect (usually the treatment effect for the treated, although perhaps the unconditional average treatment effect). Because a crucial step must be added to the project – assessing the level of bias that may arise from possible nonignorability of treatment – focusing on a very specific treatment effect of primary interest helps to ground a discussion of an estimate's limitations. Then, after using one of the matching estimators of the last section, one should use the data to minimize bias in the estimates and, if possible, proceed thereafter to a sensitivity analysis (which we will discuss later in Chapter 6).

4.6 Remaining Practical Issues in Matching Analysis

In this section, we discuss the remaining practical issues that analysts who consider using matching estimators must confront. First, we discuss the issue of how to identify empirically the common support of the matching variables. Then

we discuss what is known about the sampling variance of alternative matching estimators, and we give a guide to usage of the standard errors provided by existing software. Finally, we consider multivalued treatments.

4.6.1 Assessing the Region of Common Support

In practice, there is often good reason to believe that some of the lack of observed overlap of S for the treatment and control cases may have emerged from systematic sources, often related to the choice behavior of individuals (see Heckman, Ichimura, Smith, and Todd 1998). In these situations, it is not a sparseness problem that must be corrected. Instead, a more fundamental mismatch between the observed treatment and control cases must be addressed, as in our earlier hypothetical example in Matching Demonstration 2. Treatment cases that have no possible counterpart among the controls are said to be "off the support" of S for the control cases, and likewise for control cases who have no possible counterparts among the treatment cases.[30]

When in this situation, applied researchers who use matching techniques to estimate the treatment effect for the treated often estimate a narrower treatment effect. Using one of the variants of the matching estimators outlined earlier, analysis is confined only to treatment cases whose propensity scores fall between the minimum and maximum propensity scores in the control group. Resulting estimates are then interpreted as estimates of a narrower treatment effect: the common-support treatment effect for the treated (see Heckman, Ichimura, and Todd 1997, 1998; see also Crump, Hotz, Imbens, and Mitnik 2006).

The goal of these sorts of techniques is to exclude at the outset those treatment cases that are beyond the observed minima and maxima of the probability distributions of the variables in S among the control cases (and vice versa). Although using the propensity score to find the region of overlap may not capture all dimensions of the common support (as there may be interior spaces in the joint distribution defined by the variables in S), subsequent matching is then expected to finish the job.

Sometimes matching on the region of common support helps to clarify and sharpen the contribution of a study. When estimating the average treatment effect for the treated, there may be little harm in throwing away control cases outside the region of common support if all treatment cases fall within the support of the control cases. And, even if imposing the common-support condition results in throwing away some of the treatment cases, this can be considered an important substantive finding, especially for interpreting the treatment effect estimate. In this case, the resulting estimate is the treatment effect for a subset of the treated only, and, in particular, a treatment effect estimate that is informative only about those in the treatment and control groups who are equivalent with respect to observed treatment selection variables. In some applications, this is precisely the estimate needed (e.g., when evaluating whether

[30]Support is often given slightly different definitions depending on the context, although most definitions are consistent with a statement such as this: the union of all intervals of a probability distribution that have true nonzero probability mass.

a program should be expanded in size in order to accommodate more treatment cases but without changing eligibility criteria).[31] We will discuss these marginal treatment effects later in Chapter 7.

4.6.2 The Expected Variance of Matching Estimates

After computing a matching estimate of some form, most researchers naturally desire a measure of its expected variability across samples of the same size from the same population, either to conduct hypothesis tests or to offer an informed posterior distribution for the causal effect that can guide subsequent research. We did not, however, report standard errors for the treatment effect estimates reported in Table 4.6 for the hypothetical example in Matching Demonstration 4.

Most of the available software routines provide such estimates. For example, for the software of Abadie and his colleagues, the one- and five-nearest-neighbor matching estimates of 7.90 and 7.85 in the first column of Table 4.6 have estimated standard errors of .671 and .527, respectively. Nonetheless, each of the software routines we used relies on a different methodology for calculating such estimates, and given their lack of agreement we caution against too strong of a reliance on the standard error estimates produced by any one software routine, at least at present. Much remains to be worked out before commonly accepted standards for calculating standard errors are available. For now, our advice is to report a range of standard errors produced by alternative software for corresponding matching estimates.[32]

We recommend caution for the following reasons. In some simple cases, there is widespread agreement on how to properly estimate standard errors for matching estimators. For example, if a perfect stratification of the data can be found, the data can be analyzed as if they are a stratified random sample with the treatment randomly assigned within each stratum. In this case, the variance

[31] Coming to terms with these common-support issues has become somewhat of a specialized art form within the empirical matching literature, and some guidance is available. Heckman, Ichimura, and Todd (1998; see also Smith and Todd 2005) recommend trimming the region of common support to eliminate cases in regions of the common support with extremely low density (and not just with respect to the propensity score but for the full distribution of S). This involves selecting a minimum density (labeled the "trimming level") that is greater than zero. Heckman and his colleagues have found that estimates are rather sensitive to the level of trimming in small samples, with greater bias when the trimming level is lower. However, increasing the trimming level excludes more treatment cases and results in higher variance. More recently, Crump et al. (2006) have developed alternative optimal weighting estimators that are more general but designed to achieve the same goals.

[32] Two of the three matching software routines that we utilized allow one to calculate bootstrapped standard errors in Stata. This is presumably because these easy-to-implement methods were once thought to provide a general framework for estimating the standard errors of alternative matching estimators and hence were a fair way to compare the relative efficiency of alternative matching estimators (see Tu and Zhou 2002). Unfortunately, Abadie and Imbens (2004) show that conventional bootstrapping is fragile and will not work in general for matching estimators. Whether generalized forms of bootstrapping may still be used effectively remains to be determined.

estimates from stratified sampling apply. But rarely is a perfect stratification available in practice without substantial sparseness in the data at hand. Once stratification is performed with reference to an estimated propensity score, the independence that is assumed within strata for standard error estimates from stratified sampling methodology is no longer present. And, if one adopts a Bayesian perspective, the model uncertainty of the propensity-score-estimating equation must be represented in the posterior.[33]

Even so, there is now also widespread agreement that convergence results from nonparametric statistics can be used to justify standard error estimates for large samples. A variety of scholars have begun to work out alternative methods for calculating such asymptotic standard errors for matching estimators, after first rewriting matching estimators as forms of nonparametric regression (see Abadie and Imbens 2006; Hahn 1998; Heckman, Ichimura, and Todd 1998; Hirano et al. 2003; Imbens 2004). For these large-sample approaches, however, it is generally assumed that matching is performed directly with regard to the variables in S, and the standard errors are appropriate only for large samples in which sparseness is vanishing. Accordingly, the whole idea of using propensity scores to solve rampant sparseness problems is almost entirely dispensed with, and estimated propensity scores then serve merely to clean up whatever chance variability in the distribution of S across treatment and control cases remains in a finite sample.

Abadie and Imbens (2006) show that one can use brute force computational methods to estimate sample variances at points of the joint distribution of S. When combined with nonparametric estimates of propensity scores, one can obtain consistent estimates of all of the pieces of their proposed formulas for asymptotic standard errors. And, yet, none of this work shows that the available variance estimators remain good guides for the expected sampling variance of matching estimators under different amounts of misspecification of the propensity-score-estimating equation, or when matching is attempted only with regard to the estimated propensity score rather than completely on the variables in S. Given that this literature is still developing, it seems prudent to report alternative standard errors from alternative software routines and to avoid drawing conclusions that depend on accepting any one particular method for calculating standard errors.

[33]There is also a related set of randomization inference techniques, built up from consideration of all of the possible permutations of treatment assignment patterns that could theoretically emerge from alternative enactments of the same treatment assignment routine (see Rosenbaum 2002). These permutation ideas generate formulas for evaluating specific null hypotheses, which, from our perspective, are largely uncontroversial. They are especially reasonable when the analyst has deep knowledge of a relatively simple treatment assignment regime and has reason to believe that treatment effects are constant in the population. Although Rosenbaum provides large-sample approximations for these permutation-based tests, the connections to the recent econometrics literature that draws on nonparametric convergence results have not yet been established.

4.6.3 Matching Estimators for Many-Valued Causes

Given the prevalence of studies of many-valued causes, it is somewhat odd to place this section under the more general heading of practical issues. But this is appropriate because most of the complications of estimating many-valued treatment effects are essentially practical, even though very challenging in some cases.[34]

Recall the setup for many-valued causes from Chapter 2, Appendix B, where we have a set of J treatment states, a set of J causal exposure dummy variables, $\{Dj\}_{j=1}^{J}$, and a corresponding set of J potential outcome random variables, $\{Y^{Dj}\}_{j=1}^{J}$. The treatment received by each individual is Dj^*, and the outcome variable for individual i, y_i, is then equal to $y_i^{Dj^*}$. For $j \neq j^*$, the potential outcomes of individual i exist as $J-1$ counterfactual outcomes y_i^{Dj}.

There are two basic approaches to matching with many-valued treatments (see Rosenbaum 2002, Section 10.2.4). The most straightforward and general approach is to form a series of two-way comparisons between the multiple treatments, estimating a separate propensity score for each contrast between each pair of treatments.[35] After the estimated propensity scores are obtained, treatment effect estimates are calculated pairwise between treatments. Care must be taken, however, to match appropriately on the correct estimated propensity scores. The observed outcomes for individuals with equivalent values on alternative propensity scores cannot be meaningfully compared (see Imbens 2000, Section 5).

For example, for three treatments with J equal to 1, 2, and 3, one would first estimate three separate propensity scores, corresponding to three contrasts for the three corresponding dummy variables: $D1$ versus $D2$, $D1$ versus $D3$, and $D2$ versus $D3$. One would obtain three estimated propensity scores: $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d2_i = 1, s_i]$, $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d3_i = 1, s_i]$, and $\Pr_N[d2_i = 1 | d2_i = 1 \text{ or } d3_i = 1, s_i]$. One would then match separately for each of the three contrasts leaving, for example, those with $d3_i = 1$ unused and unmatched when matching on the propensity score for the comparison of treatment 1 versus treatment 2. At no point would one match together individuals with equivalent values for alternative estimated propensity scores. For example, there is no meaningful causal comparison between two individuals, in which for the first individual $d2_i = 1$ and $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d2_i = 1, s_i] = .6$ and for the second individual $d3_i = 1$ and $\Pr_N[d1_i = 1 | d1_i = 1 \text{ or } d3_i = 1, s_i] = .6$.

When the number of treatments is of modest size, such as only four or five alternatives, there is much to recommend in this general approach. However, if

[34] Although we could present these methods with reference to methods of stratification as well, we consider the most general case in which propensity score methods are used to address sparseness issues as well.

[35] Some simplification of the propensity score estimation is possible. Rather than estimate propensity scores separately for each pairwise comparison, one can use multinomial probit and logit models to estimate the set of propensity scores (see Lechner 2002a, 2000b; see also Hirano and Imbens 2004; Imai and van Dyk 2004; Imbens 2000). One must still, however, extract the right contrasts from such a model in order to obtain an exhaustive set of estimated propensity scores.

the number of treatments is considerably larger, then this fully general approach may be infeasible. One might then choose to simply consider only a subset of causal contrasts for analysis, thereby reducing the aim of the causal analysis.

If the number of treatments can be ordered, then a second approach developed by Joffe and Rosenbaum (1999) and implemented in Lu, Zanutto, Hornik, and Rosenbaum (2001) is possible. These models generally go by the name of dose-response models because they are used to estimate the effects of many different dose sizes of the same treatment, often in comparison with a base dosage of 0 that signifies no treatment.

Rather than estimate separate propensity scores for each pairwise comparison, an ordinal probability model is estimated and the propensity score is defined as a single dimension of the predictors of the model (i.e., ignoring the discrete shifts in the odds of increasing from one dosage level to the next that are parameterized by the estimated cut-point parameters for each dosage level). Thereafter, one then performs a slightly different form of matching in which optimal matched sets are formed by two criteria, which Lu et al. (2001:1249) refer to as "close on covariates; far apart on doses." The idea here is to form optimal contrasts between selected sets of comparable individuals to generate estimates of counterfactually defined responses. The goal is to be able to offer a predicted response to any shift in a dosage level from any k' to k'', where both k' and k'' are between the smallest and largest dosage values observed.

Again, however, these methods assume that the treatment values can be ordered, and further that the propensity scores can be smoothed across dose sizes after partialing out piecewise shifts. Even so, these assumptions are no more severe than what is typically invoked implicitly in regression modeling approaches to causality, as we discuss later. Thus, ordered probability models can be used to consistently estimate treatment effects for many-valued causes of a variety of types (see also Hirano and Imbens 2004 and Imai and van Dyk 2004 for further details).

4.7 Conclusions

We conclude this chapter by discussing the strengths and weaknesses of matching as a method for causal inference from observational data. Some of the advantages of matching methods are not inherent or unique to matching itself but rather are the result of the analytical framework in which most matching analyses are conducted. Matching focuses attention on the heterogeneity of the causal effect. It forces the analyst to examine the alternative distributions of covariates across those exposed to different levels of the causal variable. The process of examining the region of common support helps the analyst to recognize which cases in the study are incomparable, such as which control cases one should ignore when estimating the treatment effect for the treated and which treatment cases may have no meaningful counterparts among the controls.

Although these are the advantages of matching, it is important that we not oversell the potential power of the techniques. First, even though the extension

of matching techniques to multivalued treatments has begun, readily available matching estimators can be applied only to treatments or causal exposures that are binary. Second, as we just discussed, our inability to estimate the variance of most matching estimators with commonly accepted methods is a genuine weakness (although it is reasonable to expect that this weakness can be overcome in the near future). Third, as the hypothetical example in Matching Demonstration 4 showed, different matching estimators can lead to somewhat different estimates of causal effects, and as yet there is little guidance on which types of matching estimators work best for different types of applications.

Finally, we close by drawing attention to a common misunderstanding about matching estimators. In much of the applied literature on matching, the propensity score is presented as a single predictive dimension that can be used to balance the distribution of important covariates across treatment and control cases, thereby warranting causal inference. As we showed in the hypothetical example in Matching Demonstration 4, perfect balance on important covariates does not necessarily warrant causal claims. If one does not know of variables that, in an infinite sample, would yield a perfect stratification, then simply predicting treatment status from the observed variables with a logit model and then matching on the estimated propensity score does not solve the causal inference problem. The estimated propensity scores will balance those variables across the treatment and control cases. But the study will remain open to the sort of "hidden bias" explored by Rosenbaum (2002) but that is often labeled selection on the unobservables in the social sciences. Matching is thus a statistical method for analyzing available data, which may have some advantages in some situations.

Chapter 5

Regression Estimators of Causal Effects

Regression models are perhaps the most common form of data analysis used to evaluate alternative explanations for outcomes of interest to quantitatively oriented social scientists. In the past 40 years, a remarkable variety of regression models have been developed by statisticians. Accordingly, most major data analysis software packages allow for regression estimation of the relationships between interval and categorical variables, in cross sections and longitudinal panels, and in nested and multilevel patterns. In this chapter, however, we restrict our attention to OLS regression, focusing mostly on the regression of an interval-scaled variable on a binary causal variable. As we will show, the issues are complicated enough for these models. And it is our knowledge of how least squares models work that allows us to explain the complexity.

In this chapter, we present least squares regression from three different perspectives: (1) regression as a descriptive modeling tool, (2) regression as a parametric adjustment technique for estimating causal effects, and (3) regression as a matching estimator of causal effects. We give more attention to the third of these three perspectives on regression than is customary in methodological texts because this perspective allows one to understand the others from a counterfactual perspective. At the end of the chapter, we will draw some of the connections between least squares regression and more general models, and we will discuss the estimation of causal effects for many-valued causes.

5.1 Regression as a Descriptive Tool

Least squares regression can be justified without reference to causality, as it can be considered nothing more than a method for obtaining a best-fitting descriptive model under entailed linearity constraints. Goldberger (1991), for example, motivates least squares regression as a technique to estimate a best-fitting

linear approximation to a conditional expectation function that may be nonlinear in the population.

Consider this descriptive motivation of regression a bit more formally. If X is a collection of variables that are thought to be associated with Y in some way, then the conditional expectation function of Y, viewed as a function in X, is denoted $E[Y|X]$. Each particular value of the conditional expectation for a specific realization x of X is then denoted $E[Y|X = x]$.

Least squares regression yields a predicted surface $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is a vector of estimated coefficients from the regression of the realized values y_i on x_i. The predicted surface, $X\hat{\beta}$, does not necessarily run through the specific points of the conditional expectation function, even for an infinite sample, because (1) the conditional expectation function may be a nonlinear function of one or more of the variables in X and (2) a regression model can be fit without parameterizing all nonlinearities in X. An estimated regression surface simply represents a best-fitting linear approximation of $E[Y|X]$ under whatever linearity constraints are entailed by the chosen parameterization.[1]

The following demonstration of this usage of regression is simple. Most readers know this material well and can skip ahead to the next section. But, even so, it may be worthwhile to read the demonstration quickly because we will build directly on it when shifting to the consideration of regression as a causal effect estimator.

Regression Demonstration 1

Recall the stratification example presented as Matching Demonstration 1 (see page 92 in Chapter 4). Suppose that the same data are being analyzed, as generated by the distributions presented in Tables 4.1 and 4.2; features of these distributions are reproduced in Table 5.1 in more compact form. As before, assume that well-defined causal states continue to exist and that S serves as a perfect stratification of the data.[2] Accordingly, the conditional expectations in the last three panels of Table 5.1 are equal as shown.

But, for this demonstration of regression as a descriptive tool, assume that a cautious researcher does not wish to rush ahead and attempt to estimate the specific underlying causal effect of D on Y, either averaged across all individuals or averaged across particular subsets of the population. Instead, the researcher is cautious and is willing to assert only that the variables S, D, and Y constitute some portion of a larger system of causal relationships. In particular, the researcher is unwilling to assert anything about the existence or nonexistence of other variables that may also lie on the causal chain from S, through D, to Y. This is tantamount to doubting the claim that S offers a perfect stratification of the data, even though that claim is true by construction for this example.

[1] One can fit a large variety of nonlinear surfaces with regression by artful parameterizations of the variables in X, but these surfaces are always generated by a linear combination of a coefficient vector and values on some well-defined coding of the variables in X.

[2] For this section, we will also stipulate that the conditional variances of the potential outcomes are constant across both of the potential outcomes and across levels of S.

Table 5.1: The Joint Probability Distribution and Conditional Population Expectations for Regression Demonstration 1

	Joint probability distribution of S and D	
	Control group: $D = 0$	Treatment group: $D = 1$
$S = 1$	$\Pr[S = 1, D = 0] = .36$	$\Pr[S = 1, D = 1] = .08$
$S = 2$	$\Pr[S = 2, D = 0] = .12$	$\Pr[S = 2, D = 1] = .12$
$S = 3$	$\Pr[S = 3, D = 0] = .12$	$\Pr[S = 3, D = 1] = .2$
	Potential outcomes under the control state	
$S = 1$	$E[Y^0 \mid S = 1, D = 0] = 2$	$E[Y^0 \mid S = 1, D = 1] = 2$
$S = 2$	$E[Y^0 \mid S = 2, D = 0] = 6$	$E[Y^0 \mid S = 2, D = 1] = 6$
$S = 3$	$E[Y^0 \mid S = 3, D = 0] = 10$	$E[Y^0 \mid S = 3, D = 1] = 10$
	Potential outcomes under the treatment state	
$S = 1$	$E[Y^1 \mid S = 1, D = 0] = 4$	$E[Y^1 \mid S = 1, D = 1] = 4$
$S = 2$	$E[Y^1 \mid S = 2, D = 0] = 8$	$E[Y^1 \mid S = 2, D = 1] = 8$
$S = 3$	$E[Y^1 \mid S = 3, D = 0] = 14$	$E[Y^1 \mid S = 3, D = 1] = 14$
	Observed outcomes	
$S = 1$	$E[Y \mid S = 1, D = 0] = 2$	$E[Y \mid S = 1, D = 1] = 4$
$S = 2$	$E[Y \mid S = 2, D = 0] = 6$	$E[Y \mid S = 2, D = 1] = 8$
$S = 3$	$E[Y \mid S = 3, D = 0] = 10$	$E[Y \mid S = 3, D = 1] = 14$

In this situation, suppose that the researcher simply wishes to estimate the best linear approximation to the conditional expectation $E[Y|D, S]$ and does not wish to then give a causal interpretation to any of the coefficients that define the linear approximation. The six true values of $E[Y|D, S]$ are given in the last panel of Table 5.1. Notice that the linearity of $E[Y|D, S]$ in D and S is present only when $S \leq 2$. The value of 14 for $E[Y|D = 1, S = 3]$ makes $E[Y|D, S]$ nonlinear in D and S over their full distributions.

Now consider the predicted surfaces that would result from the estimation of two alternative least squares regression equations with data from a sample of infinite size (to render sampling error zero). A regression of Y on D and S that treats D as a dummy variable and S as an interval-scaled variable would yield a predictive surface of

$$\hat{Y} = -2.71 + 2.69(D) + 4.45(S). \tag{5.1}$$

This model constrains the partial association between Y and S to be linear. It represents a sensible predicted regression surface because it is a best-fitting,

linear-in-the-parameters model of the association between Y and the two variables D and S, where "best" is defined as minimizing the average squared differences between the fitted values and the true values of the conditional expectation function.

For this example, one can offer a better descriptive fit at little interpretive cost by using a more flexible parameterization of S. An alternative regression that treats S as a discrete variable represented in the estimation routine by dummy variables $S2$ and $S3$ (for S equal to 2 and S equal to 3, respectively) would yield a predictive surface of

$$\hat{Y} = 1.86 + 2.75(D) + 3.76(S2) + 8.92(S3). \tag{5.2}$$

Like the predicted surface for the model in Equation (5.1), this model is also a best linear approximation to the six values of the true conditional expectation $E[Y|D, S]$. The specific estimated values are

$$\begin{aligned}
D = 0, S = 1 &: \ \hat{Y} = 1.86, \\
D = 0, S = 2 &: \ \hat{Y} = 5.62, \\
D = 0, S = 3 &: \ \hat{Y} = 10.78, \\
D = 1, S = 1 &: \ \hat{Y} = 4.61, \\
D = 1, S = 2 &: \ \hat{Y} = 8.37, \\
D = 1, S = 3 &: \ \hat{Y} = 13.53.
\end{aligned}$$

In contrast to the model in Equation (5.1), for this model the variable S is given a fully flexible coding. As a result, parameters are fit that uniquely represent all values of S.[3] The predicted change in Y for a shift in S from 1 to 2 is 3.76

[3]The difference between a model in which a variable is given a fully flexible coding and one in which it is given a more constrained coding is clearer for a simpler conditional expectation function. For $E[Y|S]$, consider the values in the cells of Table 5.1. The three values of $E[Y|S]$ can be obtained from the first and fourth panels of Table 5.1 as follows:

$$\begin{aligned}
E[Y|S=1] &= \frac{.36}{(.36+.08)}(2) + \frac{.08}{(.36+.08)}(4) = 2.36, \\
E[Y|S=2] &= \frac{.12}{(.12+.12)}(6) + \frac{.12}{(.12+.12)}(8) = 7, \\
E[Y|S=3] &= \frac{.12}{(.12+.2)}(10) + \frac{.2}{(.12+.2)}(14) = 12.5.
\end{aligned}$$

Notice that these three values of $E[Y|S]$ do not fall on a straight line; the middle value of 7 is closer to 2.36 than it is to 12.5.

For $E[Y|S]$, a least squares regression of Y on S, treating S as an interval-scaled variable, would yield a predictive surface of

$$\hat{Y} = -2.78 + 5.05(S).$$

The three values of this estimated regression surface lie on a straight line −2.27, 7.32, and 12.37 – and they do not match the corresponding true values of 2.36, 7, and 12.5. A regression of Y on S, treating S as a discrete variable with dummy variables $S2$ and $S3$, would yield an alternative predictive surface of

$$\hat{Y} = 2.36 + 4.64(S2) + 10.14(S3).$$

(i.e., $5.62 - 1.86 = 3.76$ and $8.37 - 4.61 = 3.76$) whereas the predicted change in Y for a shift in S from 2 to 3 is 5.16 (i.e., $10.78 - 5.62 = 5.16$ and $13.53 - 8.37 = 5.16$).

Even so, the model in Equation (5.2) constrains the parameter for D to be the same without regard to the value of S. And, because the level of Y depends on the interaction of S and D, specifying more than one parameter for the three values of S does not bring the predicted regression surface into alignment with the six values of $E[Y|D, S]$ presented in the last panel of Table 5.1. Thus, even when S is given a fully flexible coding (and even for an infinitely large sample), the fitted values do not equal the true values of $E[Y|D, S]$.[4] As we discuss later, a model that is saturated fully in both S and D – that is, one that adds two additional parameters for the interactions between D and both $S2$ and $S3$ – would yield predicted values that would exactly match the six true values of $E[Y|D, S]$ in a dataset of sufficient size.

Recall the more general statement of the descriptive motivation of regression analysis presented earlier, in which the predicted surface $\hat{Y} = X\hat{\beta}$ is estimated for the sole purpose of obtaining a best-fitting linear approximation to the true conditional expectation function $E[Y|X]$. When the purposes of regression are so narrowly restricted, the outcome variable of interest, Y, is not generally thought to be a function of potential outcomes associated with well-defined causal states. Consequently, it would be inappropriate to give a causal interpretation to any of the estimated coefficients in $\hat{\beta}$.

This perspective implies that if one were to add more variables to the predictors, embedding X in a more encompassing set of variables W, then a new set of least squares estimates $\hat{\gamma}$ could be obtained by regressing Y on W. The estimated surface $W\hat{\gamma}$ then represents a best-fitting, linear-in-the-parameters, descriptive fit to a more encompassing conditional expectation function, $E[Y|W]$. Whether one then prefers $W\hat{\gamma}$ to $X\hat{\beta}$ as a description of the variation in Y depends on whether one finds it more useful to approximate $E[Y|W]$ than $E[Y|X]$. The former regression approximation is often referred to as the long regression, with the latter representing the short regression. These labels are aptly chosen, when regression is considered nothing more than a descriptive tool, as there is no inherent reason to prefer a short to a long regression if neither is meant to

This second model uses a fully flexible coding of S, and each value of the conditional expectation function is a unique function of the parameters in the model (that is, $2.36 = 2.36$, $4.64 + 2.36 = 7$, and $10.14 + 2.36 = 12.5$). Thus, in this case, the regression model would, in a suitably large sample, estimate the three values of $E[Y|S]$ exactly.

[4] Why would one ever prefer a constrained regression model of this sort? Consider a conditional expectation function, $E[Y|X]$, where Y is earnings and X is years of education (with 21 values from 0 to 20). A fully flexible coding of X would fit 20 dummy variables for the 21 values of X. This would allow the predicted surface to change only modestly between some years (such as between 7 and 8 and between 12 and 13) and more dramatically between other years (such as between 11 and 12 and between 15 and 16). However, one might wish to treat X as an interval-scaled variable, smoothing these increases from year to year by constraining them to a best-fitting line parameterized only by an intercept and a constant slope. This constrained model would not fit the conditional expectation function as closely as the model with 20 dummy variables, but it might be preferred in some situations because it is easier to present and easier to estimate for a relatively small sample.

be interpreted as anything other than a best-fitting linear approximation to its respective true conditional expectation function.

In many applied regression textbooks, the descriptive motivation of regression receives no direct explication. And, in fact, many textbooks state that the only correct specification of a regression model is one that includes all explanatory variables. Goldberger (1991) admonishes such textbook writers, countering their claims with:

> An alternative position is less stringent and is free of causal language. Nothing in the CR [classical regression] model itself requires an exhaustive list of explanatory variables, nor any assumption about the direction of causality. (Goldberger 1991:173)

Goldberger is surely correct, but his perspective nonetheless begs an important question on the ultimate utility of descriptively motivated regression. Clearly, if one wishes to know only predicted values of the outcome Y for those not originally studied but whose variables in X are known, then being able to form the surface $X\hat{\beta}$ is a good first step (and perhaps a good last step). And, if one wishes to build a more elaborate regression model, allowing for an additional variable in W or explicitly accounting for multilevel variability by modeling the nested structure of the data, then regression results will be useful if the aim is merely to generate descriptive reductions of the data. But, if one wishes to know the value of Y that would result for any individual in the population if a variable in X were shifted from a value k to a value k', then regression results may be uninformative.

Many researchers (perhaps a clear majority) who use regression models in their research are very much interested in causal effects. Knowing the interests of their readers, many textbook presentations of regression sidestep these issues artfully by, for example, discussing how biased regression coefficients result from the omission of important explanatory variables but without introducing explicit, formal notions of causality into their presentations. Draper and Smith (1998:236), for example, write of the bias that enters into estimated regression coefficients when only a subset of the variables in the "true response relationship" are included in the fitted model. Similarly, Greene (2000:334) writes of the same form of bias that results from estimating coefficients for a subset of the variables from the "correctly specified regression model."[5] And, in his presentation of regression models for social scientists, Stolzenberg (2004:188) equivocates:

> Philosophical arguments about the nature of causation notwithstanding (see Holland, 1986), in most social science uses of regression, the *effect* of an independent variable on a dependent variable is the *rate* at which differences in the independent variable are associated with (or cause) differences or changes in the dependent variable. [Italics in the original.]

[5]There are, of course, other textbooks that do present a more complete perspective, such as Berk (2004), Freedman (2005), and Gelman and Hill (2007).

We also assume that the readers of our book are interested in causal effect estimators. And thus, although we recognize the classical regression tradition, perhaps best defended by Goldberger (1991) as interpretable merely as a descriptive data reduction tool, we will consider regression primarily as a causal effect estimator in the following sections of this chapter. And we further note that, in spite of our reference to Goldberger (1991), in other writing Goldberger has made it absolutely clear that he too was very much interested in the proper usage of regression models to offer warranted causal claims. This is perhaps most clear in work in which he criticized what he regarded as unwarranted causal claims generated by others using regression techniques, such as in his robust critique of Coleman's Catholic schools research that we summarized in Subsection 1.3.2 (see Goldberger and Cain 1982). We will return to a discussion of the notion of a correct specification of a regression model at the end of the chapter, where we discuss the connections between theoretical models and regressions as all-cause perfect specifications. Until then, however, we return to the same basic scenario considered in our presentation of matching in Chapter 4: the estimation of a single causal effect that may be confounded by other variables.

5.2 Regression Adjustment as a Strategy to Estimate Causal Effects

In this section, we consider the estimation of causal effects in which least squares regression is used to adjust for variables thought to be correlated with both the causal and the outcome variables. We first consider the textbook treatment of the concept of omitted-variable bias, with which most readers are probably well acquainted. Thereafter, we consider the same set of ideas after specifying the potential outcome variables that the counterfactual tradition assumes lie beneath the observed data.

5.2.1 Regression Models and Omitted-Variable Bias

Suppose that one is interested in estimating the causal effect of a binary variable D on an observed outcome Y. This goal can be motivated as an attempt to obtain an unbiased and consistent estimate of a coefficient δ in a generic bivariate regression equation:

$$Y = \alpha + \delta D + \varepsilon, \tag{5.3}$$

where α is an intercept and ε is a summary random variable that represents all other causes of Y (some of which may be related to the causal variable of interest, D). When Equation (5.3) is used to represent the causal effect of D on Y without any reference to individual-varying potential outcome variables, the parameter δ is implicitly cast as an invariant, structural causal effect that applies to all members of the population of interest.[6]

[6] Although this is generally the case, there are of course introductions to regression that explicitly define δ as the mean effect of D on Y across units in the population of interest or,

The OLS estimator of this bivariate regression coefficient is then:

$$\hat{\delta}_{\text{OLS, bivariate}} \equiv \frac{\text{Cov}_N(y_i, d_i)}{\text{Var}_N(d_i)}, \tag{5.4}$$

where $\text{Cov}_N(.)$ and $\text{Var}_N(.)$ are unbiased, sample-based estimates from a sample of size N of the population-level covariance and variance of the variables that are their arguments.[7] Because D is a binary variable, $\hat{\delta}_{\text{OLS, bivariate}}$ is exactly equivalent to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$, presented earlier in Equation (2.7) (i.e., sample mean of y_i for those in the treatment group minus the sample mean of y_i for those in the control group). Our analysis thus follows quite closely the prior discussion of the naive estimator in Subsection 2.6.3. The difference is that here we will develop the same basic claims with reference to the relationship between D and ε rather than the general implications of heterogeneity of the causal effect.

Consider first a case in which D is randomly assigned, as when individuals are randomly assigned to the treatment and control groups. In this case, D would be uncorrelated with ε in Equation (5.3), even though there may be a chance correlation between D and ε in any finite set of study subjects.[8] The literature on regression, when presented as a causal effect estimator, maintains that, in this case, (1) the estimator $\hat{\delta}_{\text{OLS, bivariate}}$ is unbiased and consistent for δ in Equation (5.3) and (2) δ can be interpreted as the causal effect of D on Y.

To understand this claim, it is best to consider a counterexample in which D is correlated with ε in the population because D is correlated with other causes of Y that are implicitly embedded in ε. For a familiar example, consider again the effect of education on earnings. Individuals are not randomly assigned to the treatment "completed a bachelor's degree." It is generally thought that those who complete college would be more likely to have had high levels of earnings

as was noted in the last section, without regard to causality at all.

[7]Notice that we are again focusing on the essential features of the methods, and thus we maintain our perfect measurement assumption (which allows us to avoid talking about measurement error in D or in Y, the latter of which would be embedded in ε). We also ignore degree-of-freedom adjustments because we assume that the available sample is again large. To be more precise, of course, we would want to indicate that the sample variance of D does not equal the population-level variance of D in the absence of such a degree-of-freedom adjustment, and so on. We merely label $\text{Var}_N(.)$ as signifying such an unbiased estimate of the population-level-variance of that which is its argument. Thus, $\text{Var}_N(.)$ implicitly includes the proper degree-of-freedom adjustment, which would be $N/(N-1)$ and which would then be multiplied by the average of squared deviations from the sample mean.

[8]We will frequently refer to D and ε as being uncorrelated for this type of assumption, as this is the semantics that most social scientists seem to use and understand when discussing these issues. Most textbook presentations of regression discuss very specific exogeneity assumptions for D that imply a correlation of 0 between D and ε. Usually, in the social sciences, the assumption is defined either by mean independence of D and ε or as an assumed covariance of 0 between D and ε. Both of these imply a correlation between D and ε of 0. In statistics, one often finds a stronger assumption: D and ε must be completely independent of each other. The argument in favor of this stronger assumption, which is convincing to statisticians, is that an inference is strongest when it holds under any transformation of Y (and thus any transformation of ε). When full independence of D and ε holds, mean independence of D and ε, a covariance of 0 between D and ε, and a 0 correlation between D and ε are all implied.

(a) A graph in which the causal effect of D on Y is unidentified

(b) An equivalent regression representation of the same graph

Figure 5.1: Graphs for a regression equation of the causal effect of D on Y.

in the absence of a college education. If this is true, D and the population-level error term ε are correlated because those who have a 1 on D are more likely to have high values rather than low values for ε. For this example, the least squares regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) would not yield a consistent and unbiased estimate of δ that can be regarded as an unbiased and consistent estimate of the causal effect of D on Y. Instead, $\hat{\delta}_{\text{OLS,bivariate}}$ must be interpreted as an upwardly biased and inconsistent estimate of the causal effect of D on Y. In the substance of the college-degree example, $\hat{\delta}_{\text{OLS, bivariate}}$ would be a poor estimate of the causal effect of a college degree on earnings, as it would suggest that the effect of obtaining a college degree is larger than it really is.[9]

Figure 5.1 presents two causal graphs. In panel (a), D and Y are connected by two types of paths, the direct causal effect $D \rightarrow Y$ and an unspecified number of back-door paths signified by $D \dashleftarrow\!\dashrightarrow Y$. (Recall that bidirected edges $\dashleftarrow\!\dashrightarrow$ represent an unspecified number of common causes of the two variables that they connect.) For the graph in panel (a), the causal effect of D on Y is unidentified because no observable variables block the back-door paths represented by $D \dashleftarrow\!\dashrightarrow Y$.

The graph in panel (b) is the regression analog to the causal graph panel (a). It contains three edges: $D \rightarrow Y$, $\varepsilon \rightarrow Y$, and $D \dashleftarrow\!\dashrightarrow \varepsilon$, where the node for ε is represented by a hollow circle $\circ$ rather than a solid circle $\bullet$ in order to indicate that ε is an unobserved variable. The back-door paths from D to Y now run through the error term ε, and the dependence represented by the bidirected edge

[9] Consider for one last time the alternative and permissible descriptive interpretation: The least squares regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) could be interpreted as an unbiased and consistent estimate of δ, in which the regression surface generated by the estimation of δ in Equation (5.3) can be interpreted as only a descriptively motivated, best linear prediction of the conditional expectation function, $E[Y|D]$ (i.e., where $\hat{\alpha}$ is an unbiased and consistent estimate of $E[Y|D = 0]$ and $\hat{\alpha} + \hat{\delta}$ *is* an unbiased and consistent estimate of $E[Y|D = 1]$). And, in the substance of the college-degree example, it could be regarded as an efficient estimate of the mean difference between the earnings of those who have obtained a college degree and those who have not. For this second type of interpretation, see the last section of this chapter.

contaminates the bivariate least squares regression coefficient for the regression of Y on D. Bivariate regression results, when interpreted as warranted causal effect estimates, assume that there are no unblocked back-door paths from the causal variable to the outcome variable.

For many applications in the social sciences, a correlation between D and ε is conceptualized as a problem of omitted variables. For the example in this section, a bivariate OLS estimate of the effect of a college degree on labor market earnings would be said to be biased because intelligence is unobserved but is correlated with both education and earnings. Its omission from Equation (5.3) leads the estimate of the effect of a college degree on earnings from that equation to be larger than it would have been if a variable for intelligence were instead included in the equation.

This perspective, however, has led to much confusion, especially in cases in which a correlation between D and ε emerges because subjects choose different levels of D based on their expectations about the variability of Y, and hence their own expectations of the causal effect itself. For example, those who attend college may be more likely to benefit from college than those who do not, even independent of the unobserved ability factor. Although this latent form of anticipation can be labeled an omitted variable, it is generally not. Instead, the language of research shifts toward notions such as self-selection bias, and this is less comfortable territory for the typical applied researcher.

To clarify the connections between omitted-variable bias and self-selection bias within a more general presentation, we draw on the counterfactual model in the next section. We break the error term in Equation (5.3) into component pieces defined by underlying potential outcome variables and allow for the more general forms of causal effect heterogeneity that are implicitly ruled out by constant-coefficient models.

5.2.2 Potential Outcomes and Omitted-Variable Bias

Consider the same set of ideas but now use the potential outcome framework to define the observable variables. We build directly on the variant of the counterfactual model presented in Subsection 3.2.2. From that presentation, recall Equation (3.5), which we reintroduce here as

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}, \tag{5.5}$$

where $\mu^0 \equiv E[Y^0]$, $\mu^1 \equiv E[Y^1]$, $\upsilon^0 \equiv Y^0 - E[Y^0]$, and $\upsilon^1 \equiv Y^1 - E[Y^1]$. We could rewrite this equation to bring it into closer alignment with Equation (5.3) by stipulating that $\alpha = \mu^0$, $\delta = (\mu^1 - \mu^0)$, and $\varepsilon = \upsilon^0 + D(\upsilon^1 - \upsilon^0)$. But note that this is not what is typically meant by the terms α, δ, and ε in Equation (5.3). The parameters α and δ in Equation (5.3) are not considered to be equal to $E[Y^0]$ or $E[\delta]$ for two reasons: (1) models are usually asserted in the regression tradition (e.g., in Draper and Smith 1998) without any reference to underlying causal states tied to potential outcomes and (2) the parameters α and δ are usually implicitly held to be constant structural effects that do not

vary over individuals in the population. Similarly, the error term, ε, in Equation (5.3) is not separated into two pieces as a function of the definition of potential outcomes and their relationship to D. For these reasons, Equation (5.5) is quite different from the traditional bivariate regression in Equation (5.3), in the sense that it is more finely articulated but also irretrievably tied to a particular formalization of a causal effect.

Suppose that we are interested in estimating the average treatment effect, denoted $(\mu^1 - \mu^0)$ here. D could be correlated with the population-level variant of the error term $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ in Equation (5.5) in two ways. First, suppose that there is a net baseline difference in the hypothetical no-treatment state that is correlated with membership in the treatment group, but the size of the individual-level treatment effect does not differ on average between those in the treatment group and those in the control group. In this case, υ^0 would be correlated with D, generating a correlation between $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ and D, even though the $D(\upsilon^1 - \upsilon^0)$ term in $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ would be equal to zero on average because $\upsilon^1 - \upsilon^0$ does not vary with D. Second, suppose there is a net treatment effect difference that is correlated with membership in the treatment group, but there is no net baseline difference in the absence of treatment. Now, $D(\upsilon^1 - \upsilon^0)$ would be correlated with D, even though υ^0 is not, because the average difference in $\upsilon^1 - \upsilon^0$ varies across those in the treatment group and those in the control group. In either case, an OLS regression of the realized values of Y on D would yield a biased and inconsistent estimate of $(\mu^1 - \mu^0)$.

It may be helpful to see precisely how these sorts of bias come about with reference to the potential outcomes of individuals. Table 5.2 presents three simple two-person examples in which the least squares bivariate regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) is biased. Each panel presents the potential outcome values for two individuals and then the implied observed data and error term in the braces from Equation (5.5). Assume for convenience that there are only two types of individuals in the population, both of which are homogeneous with respect to the outcomes under study and both of which comprise one half of the population. For the three examples in Table 5.2, we have sampled one of each of these two types of individuals for study.

For the example in the first panel, the true average treatment effect is 15, because for the individual in the treatment group δ_i is 10 whereas for the individual in the control group δ_i is 20. The values of υ_i^1 and υ_i^0 are deviations of the values of y_i^1 and y_i^0 from $E[Y^1]$ and $E[Y^0]$, respectively. Because these expectations are equal to 20 and 5, the values of υ_i^1 are both equal to 0 because each individual's value of y_i^1 is equal to 20. In contrast, the values of υ_i^0 are equal to 5 and -5 for the individuals in the treatment and control groups, respectively, because their two values of y_i^0 are 10 and 0.

As noted earlier, the bivariate regression estimate of the coefficient on D is equal to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$. Accordingly, a regression of the values for y_i on d_i would yield a value of 0 for the intercept and a value of 20 for the coefficient on D. This estimated value of 20 is an upwardly biased estimate for the true average causal effect because the values of d_i are positively correlated with the values of the error term $\upsilon_i^0 + d_i(\upsilon_i^1 - \upsilon_i^0)$. In this

Table 5.2: Examples of the Two Basic Forms of Bias for Least Squares Regression

			Differential baseline bias only				
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	0	5	20	1	0
In control group	20	0	0	−5	0	0	−5
			Differential treatment effect bias only				
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	2.5	0	20	1	2.5
In control group	15	10	−2.5	0	10	0	0
			Both types of bias				
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	25	5	5	−2.5	25	1	5
In control group	15	10	−5	2.5	10	0	2.5

case, the individual with a value of 1 for d_i has a value of 0 for the error term whereas the individual with a value of 0 for d_i has a value of −5 for the error term.

For the example in the second panel, the relevant difference between the individual in the treatment group and the individual in the control group is in the value of y_i^1 rather than y_i^0. In this variant, both individuals would have had the same outcome if they were both in the control state, but the individual in the treatment group would benefit relatively more from being in the treatment state. Consequently, the values of d_i are correlated with the values of the error term in the last column because the true treatment effect is larger for the individual in the treatment group than for the individual in the control group. A bivariate regression would yield an estimate of 10 for the average causal effect, even though the true average causal effect is only 7.5 in this case.

Finally, in the third panel of the table, both forms of baseline and net treatment effect bias are present, and in opposite directions. In combination, however, they still generate a positive correlation between the values of d_i and the error term in the last column. This pattern results in a bivariate regression estimate of 15, which is upwardly biased for the true average causal effect of 12.5.

For symmetry, and some additional insight, now consider two additional two-person examples in which regression gives an unbiased estimate of the average causal effect. For the first panel of Table 5.3, the potential outcomes are independent of D, and as a result a bivariate regression of the values y_i on d_i would

Table 5.3: Two-Person Examples in Which Least Squares Regression Estimates are Unbiased

	Independence of (Y^1, Y^0) from D						
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	0	0	20	1	0
In control group	20	10	0	0	10	0	0
	Offsetting dependence of $Y^1$and Y^0 on D						
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	5	−5	20	1	5
In control group	10	20	−5	5	20	0	5

yield an unbiased estimate of 10 for the true average causal effect. But the example in the second panel is quite different. Here, the values of v_i^1 and v_i^0 are each correlated with the values of d_i, but they cancel each other out when they jointly constitute the error term in the final column. Thus, a bivariate regression yields an unbiased estimate of 0 for the true average causal effect of 0. And, yet, with knowledge of the values for y_i^1 and y_i^0, it is clear that these results mask important heterogeneity of the causal effect. Even though the average causal effect is indeed 0, the individual-level causal effects are equal to 10 and −10 for the individuals in the treatment group and control group, respectively. Thus, regression gives the right answer, but it hides the underlying heterogeneity that one would almost certainly wish to know.

Having considered these examples, we are now in a position to answer, from within the counterfactual framework, the question that so often confounds students when first introduced to regression as a causal effect estimator: What is the error term of a regression equation? Compare the third and fourth columns with the final column in Tables 5.2 and 5.3. The regression error term, $v^0 + D(v^1 - v^0)$, is equal to v^0 for those in the control group and v^1 for those in the treatment group. This can be seen without reference to the examples in the tables. Simply rearrange $v^0 + D(v^1 - v^0)$ as $Dv^1 + (1 - D)v^0$ and then rewrite Equation (5.5) as

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{Dv^1 + (1 - D)v^0\}. \tag{5.6}$$

It should be clear that the error term now appears very much like the observability of Y definition presented earlier as $DY^1 + (1-D)Y^0$ in Equation (2.2). Just as Y switches between Y^1 and Y^0 as a function of D, the error term switches between v^1 and v^0 as a function of D. Given that v^1 and v^0 can be interpreted as Y^1 and Y^0 centered around their respective population-level expectations $E[Y^1]$ and $E[Y^0]$, this should not be surprising.

Even so, few presentations of regression characterize the error term of a bivariate regression in this way. Some notable exceptions do exist. The connection is made to the counterfactual tradition by specifying Equation (5.3) as

$$Y = \alpha + \delta D + \varepsilon_{(D)}, \tag{5.7}$$

where the error term $\varepsilon_{(D)}$ is considered to be an entirely different random variable for each value of D (see Pratt and Schlaifer 1988). Consequently, the error term ε in Equation (5.3) switches between $\varepsilon_{(1)}$ and $\varepsilon_{(0)}$ in Equation (5.7) depending on whether D is equal to 1 or 0.[10]

Before moving on to adjustment techniques, it seems proper to ask one final question. If both υ^1 and υ^0 are uncorrelated with D, will the bivariate least squares regression coefficient for D be an unbiased and consistent estimate of the average causal effect? Yes, but two qualifications should be noted, both of which were revealed in the second example in Table 5.3. First, bivariate regression can yield an unbiased and consistent estimate in other cases, as when the nonzero correlations that υ^1 and υ^0 have with D "cancel out" in the construction of the combined error term $D\upsilon^1 + (1 - D)\upsilon^0$. Second, an unbiased and consistent regression estimate of the average causal effect may still mask important heterogeneity of causal effects. The first of these qualifications would rarely apply to real-world applications, but the second qualification, we suspect, obtains widely and is less frequently recognized than it should be.

5.2.3 Regression as Adjustment for Otherwise Omitted Variables

How well can regression adjust for an omitted variable if that variable is observed and included in an expanded regression equation? The basic strategy behind regression analysis as an adjustment technique to estimate a causal effect is to add a sufficient set of "control variables" to the bivariate regression in Equation (5.3) in order to break a correlation between the treatment variable D and the error term ε, as in

$$Y = \alpha + \delta D + X\beta + \varepsilon^*, \tag{5.8}$$

where X represents one or more variables, β is a coefficient (or a conformable vector of coefficients if X represents more than one variable), ε^* is a residualized version of the original error term ε from Equation (5.3), and all else is as defined for Equation (5.3).

For the multiple regression analog to the least squares bivariate regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4), the observed data values d_i and x_i are embedded in an all-encompassing $\mathbf{Q}$ matrix, which is $N \times K$, where N is the number of respondents and K is the number of variables in X plus 2 (one

[10] This is the same approach taken by Freedman (see Berk 2004, Freedman 2005), and he refers to Equation (5.7) as a response schedule. See also the discussion of Sobel (1995). For a continuous variable, Garen (1984) notes that there would be an infinite number of error terms (see discussion of Garen's Equation 10).

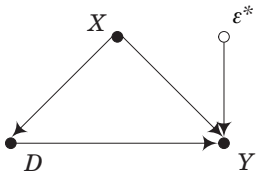


Figure 5.2: A causal graph for a regression equation in which the causal effect of D on Y is identified by conditioning on X.

for the constant and one for the treatment variable D). The OLS estimator for the parameters in Equation (5.8) is then written in matrix notation as

$$\hat{\delta}_{\text{OLS, multiple}} \equiv (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}, \tag{5.9}$$

where $\mathbf{y}$ is an $N \times 1$ vector for the observed outcomes y_i. As all regression textbooks show, there is nothing magical about these least squares computations, even though the matrix representation may appear unfamiliar to some readers. OLS regression is equivalent to the following three-step regression procedure with reference to Equation (5.8) [and without reference to the perhaps overly compact Equation (5.9)]: (1) Regress y_i on x_i and calculate $y_i^* = y_i - \hat{y}_i$; (2) regress d_i on x_i and calculate $d_i^* = d_i - \hat{d}_i$; (3) regress y_i^* on d_i^*. The regression coefficient on d_i^* yielded by step (3) is the OLS estimate of δ, which is typically declared unbiased and consistent for δ in Equation (5.8) if the true correlation between D and ε^* is assumed to be equal to zero. Thus, in this simple example, OLS regression is equivalent to estimating the relationship between residualized versions of Y and D from which their common dependence on other variables in X has been "subtracted out."

Even though the variables in X might be labeled control variables in a regression analysis of a causal effect, this label expresses the intent rather than the outcome of their utilization. The goal of such a regression adjustment strategy is to find variables in X that can be used to redraw the causal graph in panel (b) of Figure 5.1 as the DAG in Figure 5.2. If this can be done, then one can condition on X in order to consistently estimate the causal effect of D on Y because X blocks the only back-door path between D and Y.

If D is uncorrelated with ε^* (i.e., the error term net of adjustment for X), then least squares regression yields an estimate that is ostensibly freed of the bias generated by the correlation of the treatment D with the error term ε in Equation (5.3). However, even in this case some complications remain when one invokes the potential outcome model.

First, if one assumes that δ is truly constant across individuals (i.e., that $y_i^1 - y_i^0$ is equal to the same constant for all individuals i), then the OLS estimate is unbiased and consistent for δ and for $(\mu^1 - \mu^0)$. If, however, $y_i^1 - y_i^0$ is not constant, then the OLS estimate represents a conditional-variance-weighted estimate of the underlying causal effects of individuals, δ_i, in which the weights are a function of the conditional variance of D (see Angrist 1998, as well as our

explanation of this result in the next section). Under these conditions, the OLS estimate is unbiased and consistent for this particular weighted average, which is usually not a causal parameter of interest.

Second, note that the residualized error term, ε^*, in Equation (5.8) is not equivalent to either ε from Equation (5.3) or to the multipart error term $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ from Equation (5.5). Rather, it is defined by whatever adjustment occurs within Equation (5.8), as represented by the term $X\beta$. Consequently, the residualized error term ε^* cannot be interpreted independently of decisions about how to specify the vector of adjustment variables in X, and this can make it difficult to define when a net covariance between D and ε^* can be assumed to be zero.

We explain these two complications and their important implications in the following sections of this chapter, where we consider a variety of examples that demonstrate the connections between matching and regression estimators of causal effects. Before developing these explanations, however, we conclude this section with two final small-N examples that demonstrate how the regression adjustment strategy does and does not work.

Table 5.4 presents two six-person examples. For both examples, a regression of Y on D yields a biased estimate of the true average treatment effect. And, in fact, both examples yield the same biased estimate because the observed values y_i and d_i are the same for both examples. Moreover, an adjustment variable X is also available for both examples, and its observed values x_i have the same associations with the observed values y_i and d_i for both examples. But the underlying potential outcomes differ substantially between the two examples. These differences render regression adjustment by X effective for only the first example.

For the example in the first panel, a regression of Y on D would yield an estimate of the coefficient for D of 11.67, which is an upwardly biased estimate of the true average causal effect of 10. The bias arises because the correlation between the error term in the last column and the realized values for d_i is not zero but is instead .33.

For the example in the second panel, a regression of Y on D would yield an estimate of the coefficient for D of 11.67 because the values for y_i and d_i are exactly the same as for the example in the first panel. Moreover, this estimate is also upwardly biased because the error term in the last column is positively correlated with the realized values of d_i. However, here the patterns are more complex. The underlying potential outcomes are different, and individual-level heterogeneity of the causal effect is now present. One member of the control group has an individual-level treatment effect of only 8, and as a result the true average treatment effect is only 9.67. Consequently, the same bivariate regression coefficient of 11.67 has a larger upward bias in this second example, and the correlation between the values of d_i and the error term in the last column is now .39 rather than .33.[11]

[11] Moreover, the correlation between the values of d_i and both υ_i^1 and υ_i^0 differs, with the former generating a correlation coefficient of .51 and the latter generating a correlation coefficient of .33.

Table 5.4: Two Six-Person Examples in Which Regression Adjustment is Differentially Effective

					Regression adjustment with X generates an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In treatment group	15	5	−2.5	−2.5	15	1	0	−2.5
In control group	20	10	2.5	2.5	10	0	1	2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5

					Regression adjustment with X does not generate an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In treatment group	15	5	−2.17	−2.5	15	1	0	−2.17
In control group	18	10	.83	2.5	10	0	1	2.5
In control group	15	5	−2.17	−2.5	5	0	0	−2.5
In control group	15	5	−2.17	−2.5	5	0	0	−2.5

This underlying difference in potential outcomes also has consequences for the capacity of regression adjustment to effectively generate unbiased estimates of the average treatment effect. This is easiest to see by rearranging the rows in Table 5.4 for each of the two examples based on the values of X for each individual, as in Table 5.5. For the first example, the values of d_i are uncorrelated with the error term within subsets of individuals defined by the two values of X. In contrast, for the second example, the values of d_i remain positively correlated with the error term within subsets of individuals defined by the two values of X. Thus, conditioning on X breaks the correlation between D and the error term in the first example but not in the second example. Because the observed data are the same for both examples, this difference is entirely a function of the underlying potential outcomes that generate the data.

This example demonstrates an important conceptual point. Recall that the basic strategy behind regression analysis as an adjustment technique is to estimate

$$Y = \alpha + \delta D + X\beta + \varepsilon^*,$$

Table 5.5: A Rearrangement to Show How Regression Adjustment is Differentially Effective

					Regression adjustment with X generates an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
					For those with $X=1$			
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In control group	20	10	2.5	2.5	10	0	1	2.5
					For those with $X=0$			
In treatment group	15	5	−2.5	−2.5	15	1	0	−2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5
					Regression adjustment with X does not generate an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
					For those with $X=1$			
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In control group	18	10	.83	2.5	10	0	1	2.5
					For those with $X=0$			
In treatment group	15	5	−2.17	−2.5	15	1	0	−2.17
In control group	15	5	−2.17	−2.5	5	0	0	−2.5
In control group	15	5	−2.17	−2.5	5	0	0	−2.5

where X represents one or more control variables, β is a coefficient (or a conformable vector of coefficients if X represents more than one variable), and ε^* is a residualized version of the original error term ε from Equation (5.3) [see our earlier presentation of Equation (5.8)]. The literature on regression often states that an estimated coefficient $\hat{\delta}$ from this regression equation is unbiased and consistent for the average causal effect if ε^* is uncorrelated with D. But, because the specific definition of ε^* is conditional on the specification of X, many researchers find this requirement of a zero correlation difficult to interpret and hence difficult to evaluate.

The crux of the idea, however, can be understood without reference to the error term ε^* but rather with reference to the simpler (and, as we have argued earlier) more clearly defined error term $v^0 + D(v^1 - v^0)$ from Equation (5.5) [or, equivalently, $Dv^1 + (1-D)v^0$ from Equation (5.6)]. Regression adjustment

by X in Equation (5.8) will yield an unbiased and consistent estimate of the average causal effect of D when

1. D is uncorrelated with $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ for each subset of respondents identified by distinct values on the variables in X,
2. the causal effect of D does not vary with X, and
3. a fully flexible parameterization of X is used.[12]

Consider the relationship between this set of conditions and what was described earlier in Subsection 3.2.1 as an assumption that treatment assignment is ignorable. Switching notation from S to X in Equation (3.3) results in

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid X, \tag{5.10}$$

where, again, the symbol $\perp\!\!\!\perp$ denotes independence. Now, rewrite the assumption, deviating Y^0 and Y^1 from their population-level expectations:

$$(\upsilon^0, \upsilon^1) \perp\!\!\!\perp D \mid X. \tag{5.11}$$

This switch from (Y^0, Y^1) to (υ^0, υ^1) does not change the assumption, at least insofar as is relevant here (because we have defined the individual-level causal effect as a linear difference, because the expectation operator is linear, and because $E[Y^0]$ and $E[Y^1]$ do not depend on who is in the treatment state and who is in the control state). Consequently, ignorability of treatment assignment can be defined only with respect to individual-level departures from the true average potential outcomes across all members of the population under the assumptions already introduced.

Given that an assumption of ignorable treatment assignment can be written as Equation (5.11), the connections between this assumption and the set of conditions that justify a regression estimator as unbiased and consistent for the effect of D on Y should now be clear. If treatment assignment is ignorable as defined in Equation (5.11), then a regression equation that conditions fully on all values of X by including a fully flexible coding of X as a set of dummy variables will yield unbiased and consistent regression estimates of the average causal effect of D on Y. Even so, ignorability is not equivalent to the set of conditions just laid out. Instead, υ^0 and υ^1 [as well as functions of them, such as $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$] must only be mean independent of D conditional on X, not fully independent of D conditional on X.

Stepping back from this correspondence, we should note that this is not the only set of conditions that would establish least squares estimation unbiased and consistent for the average causal effect, but it is the most common

[12] Here again, we use the word uncorrelated to characterize the necessary association between D and $\upsilon^0 + D\,(\upsilon^1 - \upsilon^0)$. More formally, it would be best to state that D and $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ must be mean independent, so that a 0 covariance of D and $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ is implied.

set of conditions that would apply in most research situations.[13] Our point in laying it out is not to provide a rigid guideline applicable to all types of regression models but instead to show why the earlier statement that "ε^* must be uncorrelated with D" is insufficiently articulated from a counterfactual perspective.

A larger point of this section, however, is that much of the received wisdom on regression modeling breaks down in the presence of individual-level heterogeneity of a causal effect, as would be present in general when causal effects are defined with reference to underlying potential outcomes tied to well-defined causal states. In the next section, we begin to explain these complications more systematically, starting from the assumption, as in prior chapters, that causal effects are inherently heterogeneous and likely to vary systematically between those in the treatment and control groups. We then present the connections among regression, matching, and stratification, building directly on our presentation of matching as conditioning by stratification in Chapter 4.

5.3 The Connections Between Regression and Matching

In this section, we return to the demonstrations utilized to motivate matching estimators in Chapter 4. Our goal is to establish when matching and regression yield different results, even though a researcher is attempting to adjust for the same set of variables. We then show how regression estimators can be reformulated to yield the same results as matching estimators – as a full parameterization of a perfect stratification of the data and then as weighted least squares estimators in which the weights are a function of the propensity score. In these cases, we show that regression is an effective estimator of causal effects defined by potential outcomes.

5.3.1 Regression as Conditional-Variance-Weighted Matching

We first show why least squares regression can yield misleading causal effect estimates in the presence of individual-level heterogeneity of causal effects, even though the only variable that needs to be adjusted for is given a fully flexible coding (i.e., when the adjustment variable is parameterized with a dummy variable for each of its values, save one for the reference category).[14] When

[13] For example, the second condition can be dropped if the heterogeneity of the causal effect is modeled as a function of X (i.e., the parameterization is fully saturated in both D and X). In this case, however, regression then becomes a way of enacting a stratification of the data, as for the matching techniques presented in the last chapter.

[14] When we write of a fully flexible coding of a variable, we are referring to a dummy variable coding of that variable only. As we will discuss later, a saturated model entails a fully flexible coding of each variable *as well as all interactions between them.* For the models discussed

a single parameter is calculated for the causal effect of D on Y, least squares estimators implicitly invoke conditional-variance weighting of individual-level causal effects. This weighting scheme generates a conditional-variance-weighted estimate of the average causal effect, which is not an average causal effect that is often of any inherent interest to a researcher.[15] Angrist (1998) provides a more formal explanation of the following results, which is then placed in the context of a larger class of models in Angrist and Krueger (1999).

Regression Demonstration 2

Reconsider Regression Demonstration 1, beginning on page 124. But now step back from the cautious mindset of the fictitious descriptively oriented researcher. Suppose that a causality-oriented researcher had performed the same exercise and obtained, in particular, the results for the regression model reported in Equation (5.2):

$$\hat{Y} = 1.86 + 2.75(D) + 3.76(S2) + 8.92(S3). \tag{5.12}$$

We know from Matching Demonstration 1 (beginning on page 92), on which Regression Demonstration 1 is based, that for this hypothetical example the average treatment effect among the treated is 3, the average treatment effect among the untreated is 2.4, and the unconditional average treatment effect is 2.64. If the causality-oriented researcher were to declare that the coefficient on D of 2.75 in Equation (5.12) is a good estimate of the causal effect of D on Y, then the researcher would be incautious but not appreciably incorrect. The value of 2.75 is indeed close to the true average treatment effect of 2.64, and we know from the setup of Regression Demonstration 1 that the variable S continues to serve as a perfect stratifying variable. Thus, if the researcher were to state that the regression model in Equation (5.2) statistically controls for the common effect of S on both D and Y, as in Equation (5.8), where S is specified as the sole element of X but as two dummy variables $S2$ and $S3$, then the researcher is not horribly off the mark. The researcher has offered an adjustment for S, and gotten close to the true average treatment effect.

Unfortunately, the closeness of the estimate to the true average treatment effect is not a general feature of this type of a regression estimator. Under this particular specification of the regression equation, the OLS estimator yields

here, a saturated model would include interactions between the causal variable D and each dummy variable for all but one of the values of S. For a model with only a fully flexible coding of S, these interactions are left out.

[15] It could be of interest to a researcher who seeks a minimum-variance estimate and who has reason to believe that the bias of the regression estimate is modest. We discuss this point later, but we hope to show that most applied researchers have good reason to want unbiased and consistent estimates rather than minimum mean-squared-error estimates of their causal parameters of interest.

precisely the value of 2.75 in an infinite sample as the sum of sample analogs to three terms:

$$\frac{\text{Var}[D|S=1]\Pr[S=1]}{\sum_S \text{Var}[D|S=s]\Pr[S=s]}\{E[Y|D=1,S=1]-E[Y|D=0,S=1]\} \quad (5.13)$$
$$+\frac{\text{Var}[D|S=2]\Pr[S=2]}{\sum_S \text{Var}[D|S=s]\Pr[S=s]}\{E[Y|D=1,S=2]-E[Y|D=0,S=2]\}$$
$$+\frac{\text{Var}[D|S=3]\Pr[S=3]}{\sum_S \text{Var}[D|S=s]\Pr[S=s]}\{E[Y|D=1,S=3]-E[Y|D=0,S=3]\}.$$

These three terms are not as complicated as they may appear. First, note that the differences in the braces on the right-hand side of each term are simply the stratum-specific differences in the outcomes, which in this case are

$$E[Y|D=1,S=1]-E[Y|D=0,S=1] = 4-2, \quad (5.14)$$
$$E[Y|D=1,S=2]-E[Y|D=0,S=2] = 8-6, \quad (5.15)$$
$$E[Y|D=1,S=3]-E[Y|D=0,S=3] = 14-10. \quad (5.16)$$

The left-hand portion of each term is then just a weight, exactly analogous to the stratum-specific weights that were used for Matching Demonstration 1 to average the stratum-specific causal effect estimates in various ways to obtain unbiased and consistent estimates of the average treatment effect, the average treatment effect for the treated, and the average treatment effect for the untreated. But, rather than use the marginal distribution of S, $\Pr[S]$, or the two conditional distributions of S, $\Pr[S|D=1]$ and $\Pr[S|D=0]$, a different set of weights is implicitly invoked by the least squares operation. In this case, the weights are composed of three pieces: (1) the variance of the treatment variable within each stratum, $\text{Var}[D|S=s]$, (2) the marginal probability of S for each stratum, $\Pr[S=s]$, and (3) a summation of the product of these two terms across S so that the three weights sum to 1.

Accordingly, the only new piece of this estimator that was not introduced and examined for Matching Demonstration 1 is the conditional variance of the treatment, $\text{Var}[D|S=s]$. Recall that the treatment variable is distributed within each stratum solely as a function of the stratum-specific propensity score, $\Pr[D|S=s]$. Thus, the treatment variable is a Bernoulli distributed random variable within each stratum. As can be found in any handbook of statistics, the variance of a Bernoulli distributed random variable is $p(1-p)$, where p is the Bernoulli probability of success (in this case D equal to 1) instead of failure (in this case D equal to 0). Accordingly, the expected variance of the within-stratum treatment variable D is simply $(\Pr[D|S=s])\,(1-\Pr[D|S=s])$.

For this example, the conditional variances $\text{Var}[D|S = s]$ contribute to the numerator of each weight as follows:

$$\text{Var}[D|S = 1]\Pr[S = 1] = \left[\left(\frac{.08}{.08 + .36}\right)\left(1 - \frac{.08}{.08 + .36}\right)\right](.08 + .36)\,, \quad (5.17)$$

$$\text{Var}[D|S = 2]\Pr[S = 2] = \left[\left(\frac{.12}{.12 + .12}\right)\left(1 - \frac{.12}{.12 + .12}\right)\right](.12 + .12)\,, \quad (5.18)$$

$$\text{Var}[D|S = 3]\Pr[S = 3] = \left[\left(\frac{.2}{.2 + .12}\right)\left(1 - \frac{.2}{.2 + .12}\right)\right](.2 + .12)\,. \quad (5.19)$$

The terms in brackets on the right-hand sides of Equations (5.17)–(5.19) are $\text{Var}[D|S = 1]$, $\text{Var}[D|S = 2]$, and $\text{Var}[D|S = 3]$. The terms in parentheses on the right-hand sides of Equations (5.17)–(5.19) are the marginal probability of S for each stratum, $\Pr[S = 1]$, $\Pr[S = 2]$, and $\Pr[S = 3]$. For example, for the stratum with $S = 1$, $\text{Var}[D|S = 1] = \left(\frac{.08}{.08+.36}\right)\left(1 - \frac{.08}{.08+.36}\right)$ and $\Pr[S = 1] = (.08 + .36)$. Finally, the denominator of each of the three stratum-specific weights in Equation (5.13) for this example is the sum of Equations (5.17)–(5.19). The denominator is constant across all three weights and simply scales the weights so that they sum to 1.

With an understanding of the implicit stratum-specific weights of least squares regression, the regression estimator can be seen clearly as an estimator for the average treatment effect but with supplemental conditional-variance weighting. Weighting is performed with respect to the marginal distribution of individuals across strata, but weighting is also performed with respect to the conditional variance of the treatment variable across strata as well. Thus, net of the weight given to stratum-specific effects solely as a function of $\Pr[S]$, the conditional-variance terms give more weight to stratum-specific causal effects in strata with propensity scores close to .5 and less weight to stratum-specific causal effects in strata with propensity scores close to either 0 or 1.

Why would the OLS estimator implicitly invoke conditional-variance weighting as a supplement to weighting simply by the marginal distribution of S? OLS is a minimum-variance-based estimator of the parameter of interest. As a result, it gives more weight to stratum-specific effects with the lowest expected variance, and the expected variance of each stratum-specific effect is an inverse function of the stratum-specific variance of the treatment variable D. Thus, if the two pieces of the weighting scheme are not aligned (i.e., the propensity score is close to 0 or 1 for strata that have high total probability mass but close to .5 for strata with low probability mass), then a regression estimator of this form, even under a fully flexible coding of S, can yield estimates that are far from the true average treatment effect even in an infinite sample.

To see the effects that supplemental weighting by the conditional variance of the treatment can have on a regression estimate, consider the alternative joint distributions for S and D presented in Table 5.6. For this example, suppose that the values of $E[Y^0|S, D]$, $E[Y^1|S, D]$, and $E[Y|S, D]$ in the final three panels of Table 5.1 again obtain, such that S continues to offer a perfect stratification

Table 5.6: The Joint Probability Distribution for Two Variants of the Stratifying and Treatment Variables in Prior Regression Demonstration 1

	Joint probability distribution of S and D	
	Control group: $D = 0$	Treatment group: $D = 1$
	Variant I	
$S = 1$	$\Pr[S = 1, D = 0] = .40$	$\Pr[S = 1, D = 1] = .04$
$S = 2$	$\Pr[S = 2, D = 0] = .20$	$\Pr[S = 2, D = 1] = .04$
$S = 3$	$\Pr[S = 3, D = 0] = .16$	$\Pr[S = 3, D = 1] = .16$
	Variant II	
$S = 1$	$\Pr[S = 1, D = 0] = .40$	$\Pr[S = 1, D = 1] = .04$
$S = 2$	$\Pr[S = 2, D = 0] = .12$	$\Pr[S = 2, D = 1] = .12$
$S = 3$	$\Pr[S = 3, D = 0] = .03$	$\Pr[S = 3, D = 1] = .29$

of the data. Note that, for the two alternative joint distributions of S and D in Table 5.6, the marginal distribution of S remains the same as in Regression Example 1: $\Pr[S = 1] = .44$, $\Pr[S = 2] = .24$, and $\Pr[S = 3] = .32$. As a result, the unconditional average treatment effect is the same for both variants of the joint distribution of S and D depicted in Table 5.6, and it matches the unconditional average treatment effect for the original example represented fully in Table 5.1. In particular, the same distribution of stratum-specific causal effects results in an unconditional average treatment effect of 2.64.

The difference represented by each variant of the joint distributions in Table 5.6 is in the propensity score for each stratum of S, which generates an alternative marginal distribution for D and thus alternative true average treatment effects for the treated and for the untreated (and, as we will soon see, alternative regression estimates from the same specification).

For Variant I in Table 5.6, those with S equal to 1 or 2 are much less likely to be in the treatment group, and those with S equal to 3 are now only equally likely to be in the treatment group and the control group. As a result, the marginal distribution of D is now different, with $\Pr[D = 0] = .76$ and $\Pr[D = 1] = .24$. The average treatment effect for the treated is now 3.33 whereas the average treatment effect among the untreated is 2.42. Both of these are larger than was the case for the example represented by Table 5.1 because (1) a greater proportion of those in the control group have $S = 3$ (i.e., $\frac{.16}{.76} > \frac{.12}{.6}$), (2) a greater proportion of those in the treatment group have $S = 3$ (i.e., $\frac{.16}{.24} > \frac{.2}{.4}$), and (3) those with $S = 3$ gain the most from the treatment.

For Variant II, those with S equal to 1 are still very unlikely to be in the treatment group, but those with S equal to 2 are again equally likely to be in the treatment group. But those with S equal to 3 are now very likely to be in

the treatment group. As a result, the marginal distribution of D is now different again, with $\Pr[D = 0] = .55$ and $\Pr[D = 1] = .45$, and the average treatment effect for the treated is now 3.29 whereas the average treatment effect among the untreated is 2.11. Both of these are smaller than for Variant I because a smaller proportion of both the treatment group and the control group have $S = 3$.

For these two variants of the joint distribution of S and D, we have examples in which the unconditional average treatment effect is the same as it was for the example in Table 5.1, but the underlying average treatment effects for the treated and for the untreated differ considerably. Does the reestimation of Equation (5.12) for these variants of the example still generate an estimate for the coefficient on D that is (a) relatively close to the true unconditional average treatment effect and (b) closer to the unconditional average treatment effect than either the average treatment effect for the treated or the average treatment effect for the untreated?

For Variant I, the regression model yields

$$\hat{Y} = 1.90 + 3.07(D) + 3.92(S2) + 8.56(S3) \tag{5.20}$$

for an infinite sample. In this case, the coefficient of 3.07 on D is not particularly close to the unconditional average treatment effect of 2.64, and in fact it is closer to the average treatment effect for the treated of 3.33 (although still not particularly close). For Variant II, the regression model yields

$$\hat{Y} = 1.96 + 2.44(D) + 3.82(S2) + 9.45(S3). \tag{5.21}$$

In this case, the coefficient of 2.44 on D is closer to the unconditional average treatment effect of 2.64, but not as close as was the case for the example in Regression Demonstration 1. It is now relatively closer to the average treatment effect for the untreated, which is 2.11 (although, again, still not particularly close).

For Variant I, the regression estimator is weighted more toward the stratum with $S = 3$, for which the propensity score is .5. For this stratum, the causal effect is 4. For Variant II, in contrast, the regression estimator is weighted more toward the stratum with $S = 2$, for which the propensity score is .5. And, for this stratum, the causal effect is 2.[16]

What is the implication of these alternative setups of the same basic demonstration? Given that the unconditional average treatment effect is the same for all three joint distributions of S and D, it would be unwise for the incautious researcher to believe that this sort of a regression specification will provide a reliably close estimate to the unconditional average treatment effect, the average treatment effect for the treated, or the unconditional average treatment effect when there is reason to believe that these three average causal effects differ because of individual-level heterogeneity. The regression estimate will be weighted

[16]Recall that, because the marginal distribution of S is the same for all three joint distributions of S and D by construction of the example, the $\Pr[S = s]$ pieces of the weights remain the same for all three alternatives. Thus, the differences between the regression estimates are produced entirely by differences in the $\text{Var}[D|S = s]$ pieces of the weights.

toward stratum-specific effects for which the propensity score is closest to .5, net of all else.

In general, regression models do not offer consistent estimates of the average treatment effect when causal effect heterogeneity is present, even when a fully flexible coding is given to the only necessary adjustment variable(s). Regression estimators with fully flexible codings of the adjustment variables do provide consistent estimates of the average treatment effect if either (a) the true propensity score does not differ by strata or (b) the average stratum-specific causal effect does not vary by strata.[17] Condition (a) would almost never be true (because, if it were, one would not even think to adjust for S because it is already independent of D). And condition (b) is probably not true in most applications, because rarely are investigators willing to assert that all consequential heterogeneity of a causal effect has been explicitly modeled.

Instead, for this type of a regression specification, in which all elements of a set of perfect stratifying variables S are given fully flexible codings (i.e., a dummy variable coding for all but one of the possible combinations of the values for the variables in S), the OLS estimator $\hat{\delta}_{\text{OLS, multiple}}$ in Equation (5.9) is equal to

$$\frac{1}{c}\sum_{s}\text{Var}_N[d_i|s_i=s]\ \Pr{}_N[s_i=s]\,\{E_N[y_i|d_i=1,s_i=s]-E_N[y_i|d_i=0,s_i=s]\} \tag{5.22}$$

in a sample of size N. Here, c is a scaling constant equal to the sum (over all combinations of values s of S) of the terms $\text{Var}_N[d_i|s_i=s]\Pr_N[s_i=s]$.

There are two additional points to emphasize. First, the weighting scheme for stratified estimates in Equation (5.22) applies only when the fully flexible parameterization of S is specified. Under a constrained specification of S [e.g., in which some elements of S are constrained to have linear effects, as in Equation (5.1)] the weighting scheme is more complex. The weights remain a function of the marginal distribution of S and the stratum-specific conditional variance of D, but the specific form of each of these components becomes conditional on the specification of the regression model (see Angrist and Krueger 1999, Section 2.3.1). The basic intuition here is that a linear constraint on a variable in S in a regression model represents an implicit linearity assumption about true underlying propensity score that may not be linear in S.[18]

[17] As a by-product of either condition, the average treatment effect must be equal to the average treatment effect for the treated and the average treatment effect for the untreated. Thus, the regression estimator would be consistent for both of these as well.

[18] For a binary causal exposure variable D, a many-valued variable S that is treated as an interval-scaled variable, and a regression equation

$$\hat{Y}=\hat{\alpha}+\hat{\delta}(D)+\hat{\beta}(S)\ ,$$

the OLS estimator $\hat{\delta}$ is equal to

$$\frac{1}{l}\sum_{s}\widetilde{\text{Var}}_N[\hat{d}_i|s_i=s]\ \widetilde{\Pr}_N[s_i=s]\,\{E_N[y_i|d_i=1,s_i=s]-E_N[y_i|d_i=0,s_i=s]\}$$

Second, regression can make it all too easy to overlook the same sort of fundamental mismatch problems that were examined for Matching Demonstration 2 in Subsection 4.2.2. Regression will implicitly drop strata for which the propensity score is either 0 or 1 in the course of forming its weighted average by Equation (5.22). As a result, a researcher who interprets a regression result as a decent estimate of the average treatment effect, but with supplemental conditional-variance weighting, may be entirely wrong. No meaningful average causal effect may exist in the population. The second point is best explained by the following illustration.

Regression Demonstration 3

Reconsider the hypothetical example presented as Matching Demonstration 2 from Chapter 4, beginning on page 95. The assumed relationships that generate the data are very similar to those for Regression Demonstrations 1 and 2, but, as shown in Table 5.7, no individual for whom S is equal to 1 in the population is ever exposed to the treatment because $\Pr[S = 1, D = 1] = 0$ whereas $\Pr[S = 1, D = 0] = .4$. As a result, not even an infinite sample from the population would ever include an individual in the treatment group with $s_i = 1$.[19] Because

in a sample of size N, where l is a scaling constant equal to the sum over all s of S the terms $\widetilde{\text{Var}}_N[\hat{d}_i|s_i = s]\ \widetilde{\Pr}_N[s_i = s]$.

The distinction between $\widetilde{\text{Var}}_N[\hat{d}_i|s_i = s]\ \widetilde{\Pr}_N[s_i = s]$ and $\text{Var}_N[d_i|s_i = s]\ \Pr_N[s_i = s]$ in the main text results from a constraint on the propensity score that is implicit in the regression equation. In specifying S as an interval-scaled variable, least squares implicitly assumes that the true propensity score $\Pr[D|S]$ is linear in S. As a result, the first portion of the stratum-specific weight is

$$\widetilde{\text{Var}}_N[\hat{d}_i|s_i = s] \equiv \hat{p}_s(1 - \hat{p}_s),$$

where $\hat{p}_s$ is equal to the predicted stratum-specific propensity score from a linear regression of d_i on s_i: $\hat{p}_s = \hat{\xi} + \hat{\phi}_s s$.

Perhaps somewhat less clear, the term $\widetilde{\Pr}_N[s_i = s]$ is also a function of the constraint on S. $\widetilde{\Pr}_N[s_i = s]$ is not simply the marginal distribution of S in the sample, as $\Pr_N[s_i = s]$ is. Rather, one must use Bayes' rule to determine the implied marginal distribution of S, given the assumed linearity of the propensity score across levels of S. Rearranging

$$\Pr[d_i = 1|s_i] = \frac{\Pr[s_i|d_i = 1]\Pr[d_i = 1]}{\Pr[s_i]}$$

as

$$\Pr[s_i] = \frac{\Pr[s_i|d_i = 1]\Pr[d_i = 1]}{\Pr[d_i = 1|s_i]},$$

and then substituting $\hat{p}_s$ for $\Pr[d_i = 1|s_i]$, we then find that

$$\widetilde{\Pr}_N[s_i = s] = \frac{\Pr_N[s_i = s|d_i = 1]\Pr_N[d_i = 1]}{\hat{p}_s}.$$

The terms $\Pr_N[s_i = s|d_i = 1]$ and $\Pr_N[d_i = 1]$ are, however, unaffected by the linearity constraint on the propensity score. They are simply the true conditional probability of S equal to s given D equal to d as well as the marginal probability of D equal to d for a sample of size N.

Note that, if the true propensity score is linear in S, then the weighting scheme here is equivalent to the one in the main text.

[19] Again, recall that we assume no measurement error in general in this book. In the presence of measurement error, some individuals might be misclassified and therefore might show up in the data with $s_i = 1$ and $d_i = 1$.

Table 5.7: The Joint Probability Distribution and Conditional Population Expectations for Regression Demonstration 3

	Joint probability distribution of S and D	
	Control group: $D = 0$	Treatment group: $D = 1$
$S = 1$	$\Pr[S = 1, D = 0] = .4$	$\Pr[S = 1, D = 1] = 0$
$S = 2$	$\Pr[S = 2, D = 0] = .1$	$\Pr[S = 2, D = 1] = .13$
$S = 3$	$\Pr[S = 3, D = 0] = .1$	$\Pr[S = 3, D = 1] = .27$
	Potential outcomes under the control state	
$S = 1$	$E[Y^0 \mid S = 1, D = 0] = 2$	
$S = 2$	$E[Y^0 \mid S = 2, D = 0] = 6$	$E[Y^0 \mid S = 2, D = 1] = 6$
$S = 3$	$E[Y^0 \mid S = 3, D = 0] = 10$	$E[Y^0 \mid S = 3, D = 1] = 10$
	Potential outcomes under the treatment state	
$S = 1$	$E[Y^1 \mid S = 1, D = 0] = 4$	
$S = 2$	$E[Y^1 \mid S = 2, D = 0] = 8$	$E[Y^1 \mid S = 2, D = 1] = 8$
$S = 3$	$E[Y^1 \mid S = 3, D = 0] = 14$	$E[Y^1 \mid S = 3, D = 1] = 14$
	Observed outcomes	
$S = 1$	$E[Y \mid S = 1, D = 0] = 2$	
$S = 2$	$E[Y \mid S = 2, D = 0] = 6$	$E[Y \mid S = 2, D = 1] = 8$
$S = 3$	$E[Y \mid S = 3, D = 0] = 10$	$E[Y \mid S = 3, D = 1] = 14$

of this structural zero in the joint distribution of S and D , the three conditional expectations, $E[Y^0 \mid S = 1, D = 0]$, $E[Y^1 \mid S = 1, D = 0]$, and $E[Y \mid S = 1, D = 0]$, are properly regarded as undefined and hence are omitted from the last three panels of Table 5.7.

As shown earlier in Subsection 4.2.2, the naive estimator can still be calculated for this example and will be equal to 8.05 in an infinite sample. Moreover, the average treatment effect for the treated can be estimated consistently as 3.35 by considering only the values for those with S equal to 2 and 3. But there is no way to consistently estimate the treatment effect for the untreated, and hence no way to consistently estimate the unconditional average treatment effect.

Consider now the estimated values that would be obtained with data arising from this joint distribution for a regression model specified equivalently as in Equations (5.12), (5.20), and (5.21):

$$\hat{Y} = 2.00 + 3.13(D) + 3.36(S2) + 8.64(S3). \tag{5.23}$$

In this case, the OLS estimator is still equivalent to Equation (5.22), which in an infinite sample would then be equal to Equation (5.13). But, with reference to Equation (5.13), note that the weight for the first term,

$$\frac{\text{Var}[D|S=1]\Pr[S=1]}{\sum_S \text{Var}[D|S=s]\Pr[S=s]},$$

is equal to zero because $\text{Var}[D|S=1]$ is equal to 0 in the population by construction. Accordingly, the numerator of the stratum-specific weight is zero, and it enters into the summation of the denominator of the other two stratum-specific weights as zero. As a result, the regression estimator yields a coefficient on D that is 3.13, which is biased downward as an estimate of the average treatment effect for the treated and has no relationship with the undefined average treatment effect. If interpreted as an estimate of the average treatment effect for the treated, but with supplemental conditional-variance weighting, then the coefficient of 3.13 is interpretable. But it cannot be interpreted as a meaningful estimate of the average treatment effect in the population once one commits to the potential outcome framework because the average treatment effect does not exist.

The importance of this demonstration is only partly revealed in this way of presenting the results. Imagine that a researcher simply observes $\{y_i, d_i, s_i\}_{i=1}^N$ and then estimates the model in Equation (5.23) without first considering the joint distribution of S and D as presented in Table 5.7. It would be entirely unclear to such a researcher that there are no individuals in the sample (or in the population) whose values for both D and S are 1. Such a researcher might therefore be led to believe that the coefficient estimate for D is a meaningful estimate of the causal effect of D for all members of the population.

All too often, regression modeling, at least as practiced in the social sciences, makes it too easy for an analyst to overlook fundamental mismatches between treatment and control cases. And, thus, one can obtain average treatment effect estimates with regression techniques even when no meaningful average treatment effect exists. Even though this is the case, we do not want to push this argument too far. Therefore, in the next section we make the (perhaps obvious) point that regression can be used as a technique to execute a perfect stratification of the data under the same assumptions that justified matching as a stratification estimator in Chapter 4.

5.3.2 Regression as an Implementation of a Perfect Stratification

Matching and regression can both be used to carry out a perfect stratification of the data. Consider how the matching estimates presented in Matching Demonstration 1 (see page 92) could have been generated by standard regression routines. For that hypothetical example, an analyst could specify S as two dummy

variables and D as one dummy variable. If all two-way interactions between S and D are then included in a regression model predicting the observed outcome Y, then the analyst has enacted the same perfect stratification of the data by fitting a model that is saturated in both S and D to all of the cells of the first panel of Table 4.2 (or see instead the reproduction in Table 5.1 on page 125):

$$\hat{Y} = 2 + 2(D) + 4(S2) + 8(S3) + 0(D \times S2) \ + 2(D \times S3). \tag{5.24}$$

The values of each of the six cells of the panel are unique functions of the six estimated coefficients from the regression model. Accordingly, by use of the marginal distribution of S and the joint distribution of S given D, coefficient contrasts can be averaged across the relevant distributions of S in order to obtain consistent estimates of the average treatment effect, the treatment effect among the treated, and the treatment effect among the untreated.

Nevertheless, for many applications, such a saturated model may not be possible, and in some cases this impossibility may be misinterpreted. For Regression Demonstration 3 (see page 149), if one were to fit the seemingly saturated model with the same six parameters as in Equation (5.24), the coefficient on D would be dropped by standard software routines. One might then attribute this to the size of the dataset and then instead use a more constrained parameterization [i.e., either enter S as a simple linear term interacted with D or instead specify the model in Equation (5.23)]. These models must then be properly interpreted, and in no case could they be interpreted as yielding unbiased and consistent estimates of the average treatment effect. In a sense, this problem is simply a matter of model misspecification. But, at a deeper level, it may be that regression as a method tends to encourage the analyst to oversimplify these important model specification issues.[20]

What if a zero cell in the joint distribution of S and D occurred by chance in any single dataset? In other words, what if there is no fundamental overlap problem in the distribution of S across D, but instead the only problem is a finite dataset? In this case, regression can be reformulated as a weighting estimator, as we describe in the next subsection, in order to solve the sparseness problem.

5.3.3 Matching as Weighted Regression

In this subsection, we explore further the connections between matching and regression estimators, demonstrating how weighted regression can be used to estimate causal effects. Imbens (2004) reviews alternative ways to calculate the same sorts of weighted averages that we present here, and he fully accounts for the connections between inverse-probability-weighting procedures and nonparametric regression.

As was shown for the hypothetical example in Matching Demonstration 3 (see page 100), matching can be considered a method to weight the data in order

[20] Rubin (1977) provides simple and elegant examples of all such complications, highlighting the importance of assumptions about the relationships between covariates and outcomes (see also Holland and Rubin 1983 and Rosenbaum 1984a, 1984b).

to balance predictors of treatment selection and thereby calculate contrasts that can be given causal interpretations. In this section, we show that the three propensity-score-weighting estimators in Equations (4.12) – (4.14) can be specified as three weighted OLS regression estimators. In fact, if one defines a weighting variable appropriately, then any standard software package that estimates weighted regression can be used.

To see how to do this, note first that the naive estimator in Equation (2.7) can be written as an OLS estimator, $(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}$, where (1) $\mathbf{Q}$ is an $n \times 2$ matrix that contains a vector of 1s in its first column and a vector of the values of d_i for each individual in its second column and (2) $\mathbf{y}$ is an $n \times 1$ column vector containing values of y_i for each individual. To estimate each of the propensity-score-weighting estimators in Equations (4.12)–(4.14), simply estimate a weighted OLS estimator:

$$\hat{\delta}_{\text{OLS, weighted}} \equiv (\mathbf{Q}'\mathbf{P}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{P}\mathbf{y}, \tag{5.25}$$

where $\mathbf{P}$ is an appropriately chosen weight matrix, depending on the average treatment effect of interest.

For the treatment effect for the treated, specify $\mathbf{P}$ as an $n \times n$ diagonal matrix with 1 in the $i \times i$th place for members of the treatment group and $\hat{p}_i/(1-\hat{p}_i)$ in the $i \times i$th place for members of the control group (where, as defined earlier for the hypothetical example in Matching Demonstration 3, $\hat{p}_i$ is the estimated propensity score; see Subsection 4.3.2). For the treatment effect for the untreated, specify $\mathbf{P}$ as an $n \times n$ diagonal matrix with $(1-\hat{p}_i)/\hat{p}_i$ in the $i \times i$th place for members of the treatment group and 1s in the $i \times i$th place for members of the control group. Finally, for the unconditional average treatment effect, specify $\mathbf{P}$ as an $n \times n$ diagonal matrix with $1/\hat{p}_i$ in the $i \times i$th place for members of the treatment group and $1/(1-\hat{p}_i)$ in the $i \times i$th place for members of the control group. Consider the following demonstration, which builds directly on the prior presentation of matching in Section 4.3.

Regression Demonstration 4

Consider first how the matching estimates in the hypothetical example in Matching Demonstration 3 (beginning on page 100) could have been generated by a standard regression routine. As shown there for Equations (4.8) and (4.9), the potential outcomes were specified as functions of individual values for A and B:

$$y_i^1 = 102 + 6a_i + 4b_i + \upsilon_i^1, \tag{5.26}$$
$$y_i^0 = 100 + 3a_i + 2b_i + \upsilon_i^0, \tag{5.27}$$

where A and B are distributed as independent uniform random variables with a minimum of .1 and a maximum of 1, and where υ_i^1 and υ_i^0 are independent random draws from a normal distribution with expectation 0 and a standard deviation of 5. For each individual, y_i is then equal to $y_i^1(d_i) + (1-d_i)y_i^0$, where

the value of d_i is determined by a Bernoulli distribution with the probability of 1 rather than 0 as the nonlinear function in A and B that is presented in Figure 4.1 in Subsection 4.3.2. The first panel of Table 5.8 reproduces the true average treatment effects for this example from the prior Table 4.5; the unconditional average treatment effect is 4.53 whereas the average treatment effects for the treated and the untreated are 4.89 and 4.40, respectively.

The second panel of Table 5.8 introduces a second variant on this basic setup. For this variant, Equations (5.26) and (5.27) are replaced with

$$y_i^1 = 102 + 3a_i + 2b_i + 6(a_i \times b_i) + \upsilon_i^1, \tag{5.28}$$

$$y_i^0 = 100 + 2a_i + 1b_i - 2(a_i \times b_i) + \upsilon_i^0, \tag{5.29}$$

but everything else remains that same. These alternative potential outcome definitions result in a slightly more dramatic pattern for the average treatment effects. As shown in the second panel of Table 5.8, the unconditional average treatment effect is 5.05 whereas the average treatment effects for the treated and the untreated are now 5.77 and 4.79, respectively. Although this difference is notable, the nonlinearity of the individual-level treatment effects is of most consequence here. The opposite-signed parameters specified for the cross-product interaction of A and B ensures that those with high levels of A and B together have much larger individual-level treatment effects than others. For Variant I in the first panel of Table 5.8, the differential sizes of the treatment effects were separable into simple linear pieces that could be independently attributed to A and B.

For each variant of this example within its corresponding panel, three coefficients on D (for an infinite sample so that sampling error is zero) are presented for three OLS regression specifications: (1) Y regressed on D, (2) Y regressed on D, A, and B, and (3) Y regressed on D, A, A-squared, B, and B-squared. All three of these OLS estimates are placed in the column for the average treatment effect because such estimates are commonly interpreted as average treatment effect estimates. But, of course, this is somewhat misleading, because regression estimates such as these are often presented without any reference to the average treatment effect (i.e., often as an estimate of an implicit constant structural effect of D on Y). None of these OLS estimates is particularly close to its respective true average treatment effect. And, although the quadratic specifications of A and B help to some degree, the estimates are still far from their targets.

For the last row of each panel, we then implemented the weighted regression models specified earlier in Equation (5.25). For these estimates, we take the estimated propensity scores (i.e., the $\hat{p}_i$ used to construct each of the three $\mathbf{P}$ matrices) from the row labeled "Perfectly specified propensity score estimates" in Table 4.5. Each of the estimates lands exactly on the targeted parameter, as will always be the case in a sufficiently large dataset if the propensity score is estimated flawlessly.

As this demonstration shows, the relationships between matching and regression are now well established in the literature. In this section, we have offered a demonstration to show some of these connections. But, more generally,

Table 5.8: OLS and Weighted OLS Estimates of Treatment Effects for Regression Demonstration 4

	Average treatment effects		
	$E[\delta]$	$E[\delta\|D=1]$	$E[\delta\|D=0]$
Variant I: Y^1 and Y^0 linear in A and B			
True treatment effects	4.53	4.89	4.40
OLS regression estimates:			
Y regressed on D	5.39		
Y regressed on D and linear A and B	4.75		
Y regressed on D and quadratic A and B	4.74		
Weighted OLS regression of Y on D	4.53	4.89	4.40
Variant II: Y^1 and Y^0 nonlinear in A and B			
True treatment effects	5.05	5.77	4.79
OLS regression estimates:			
Y regressed on D	5.88		
Y regressed on D and linear A and B	5.47		
Y regressed on D and quadratic A and B	5.44		
Weighted OLS regression of Y on D	5.05	5.77	4.79

it is now known that most matching estimators can be rewritten as forms of nonparametric regression (see Abadie and Imbens 2006; Hahn 1998; Heckman, Ichimura, and Todd 1998; Hirano et al. 2003; Imbens 2004; Lunceford and Davidian 2004). And, as shown earlier in this chapter, OLS regression, under certain specifications, can be seen as a form of matching with supplemental conditional-variance weighting (which may or may not be useful, depending on the application). We have further shown with this demonstration one particular advantage of a matching estimator, whether carried out as specified earlier in Chapter 4, or as a weighted regression estimator as shown in this subsection. Matching may significantly outperform regression models when the true functional form of a regression is nonlinear but a simple linear specification is used mistakenly. However, as shown earlier, the superior performance requires that the propensity score be estimated effectively, and perhaps flawlessly, in the case of weighting estimators.

5.3.4 Regression as Supplemental Adjustment When Matching

For Equation (5.25), we defined the **Q** matrix as containing a column of 1*s* and a column of individual-level values d_i. In fact, there is a literature that argues that all variables that predict treatment selection should be included as additional columns in **Q** as well. The idea is to offer what James Robins refers to as a "doubly robust" or "doubly protected" estimator (see Bang and Robins 2005; Robins and Rotnitzky 2001). Robins and Rotnitzky reflect on the fallibility of both standard regression methods and propensity-score-based weighting estimators:

> There has been considerable debate as to which approach to confounder control is to be preferred, as the first is biased if the outcome regression model is misspecified while the second approach is biased if the treatment regression, i.e., propensity, model is misspecified. This controversy could be resolved if an estimator were available that was guaranteed to be consistent for θ whenever at least one of the two models was correct under an asymptotic sequence in which the outcome and treatment regression models remain fixed as the sample size n increases to infinity. We refer to such combined methods as doubly-robust or doubly-protected as they can protect against misspecification of either the outcome or treatment model, although not against simultaneous misspecification of both. (Robins and Rotnitzky 2001:922)

The basic motivation of this practice is to give the analyst two chances to "get it right," in hopes that misspecifications of the propensity-score-estimating equation and the final regression equation will neutralize each other. And, although Robins is credited with developing the recent asymptotic justification for a variety of specific procedures (see Robins and Ritov 1997; Robins, Rotnitzky, and Zhao 1994; Scharfstein, Rotnitzky, and Robins 1999; van der Laan and Robins 2003), the idea of using matching and regression together is quite general and has a long legacy in applied work (see Cochran and Rubin 1973; Gelman and King 1990; Heckman, Ichimura, and Todd 1998; Hirano and Imbens 2001; Rubin and Thomas 1996, 2000; Smith and Todd 2005). Consider the following demonstration that shows some of these possibilities.

Regression Demonstration 5

Recall Matching Demonstration 4, beginning on page 110, and consider now how regression can be used to supplement a matching algorithm.[21] Recall that for Matching Demonstration 4, we presented matching estimates of the Catholic school effect on achievement for simulated data. As we discuss there, the treatment effect for the treated is 6.96 in the simulated data, whereas the treatment

[21] The results for this demonstration are drawn from Morgan and Harding (2006), and we thereby thank David Harding for his contribution to this section. More details on these results are presented in the article as well.

Table 5.9: Combined Matching and Regresison Estimates for the Simulated Effect of Catholic Schooling on Achievement, as Specified in Matching Demonstration 4

	Poorly specified propensity-score-estimating equation		Well-specified propensity-score-estimating equation	
	TT Estimate	Bias	TT Estimate	Bias
OLS Regression:				
Not restricted to region of common support	7.79	0.83	6.81	-0.15
Restricted to region of common support	7.88	0.92	6.80	-0.16
Matching with regression adjustment:				
Interval with variable blocks (B&I)	7.95	0.99	6.70	-0.26
One Nearest Neighbor with caliper = 0.001 (L&S)	8.05	1.09	7.15	0.19
One Nearest Neighbor without caliper (Abadie)	7.78	0.82	6.88	-0.08
Five Nearest Neighbors with caliper = 0.001 (L&S)	7.92	0.96	7.17	0.21
Five Nearest Neighbors without caliper (Abadie)	7.82	0.86	7.20	0.24

Notes: B&I denotes the software of Becker and Ichino; L&S denotes the software of Leuven and Sianesi; Abadie denotes the software of Abadie et al.

effect for the untreated is 5.9. In combination, the average treatment effect is then 6.0.

In this demonstration, we present in the first two lines of Table 5.9 least squares regression estimates of the treatment effect under two specifications, including the same variables for the propensity-score-estimating equation directly in the regression equation (and in the two different specifications used for the earlier Table 4.6). We present regression estimates in two variants: (1) without regard to the distributions of the variables and (2) based on a subsample restricted to the region of common support, as defined by the propensity score estimated from the covariates utilized for the respective scenario. Comparing these estimates with the values from Table 4.6, linear regression does about as well as the matching algorithms as an estimator of the treatment effect for the treated. In some cases, these estimates outperform some of the matching estimates. In fairness to the matching estimates, however, it should be pointed out that the data analyzed for this example are well suited to regression because the assumed functional form of each potential outcome variable is linear and hence relatively simple. Although we believe that this is reasonable for the simulated application, there are surely scenarios in which matching can be shown to clearly outperform regression because of nonlinearities that are not parameterized by the relevant regression model.

In the second panel of Table 5.9, we provide five examples of matching combined with regression adjustment. Interval matching with regression adjustment calculates the treatment effect within blocks after adjusting for the same covariates included in the propensity-score-estimating equation for the particular

scenario, averaging over blocks to produce an overall treatment effect estimate. With nearest-neighbor matching, one accomplishes regression adjustment by regressing the outcome on the treatment and covariates using the matched sample, with appropriate weights for duplicated observations in the matched control group and for multiple neighbor matching. When the results of Tables 4.6 and 5.9 are compared, supplemental regression adjustment reduces the bias for only the nearest-neighbor matching with one match, whereas it offers no improvement for the other three matching estimators.

As this demonstration shows, supplemental regression adjustment may provide a slight improvement over an analogous matching estimator implemented without regression adjustment. But this is not true in all cases, especially when the matching estimator chosen already uses many cases to match to each target case.

Another advantage of combining matching and regression has emerged recently. Rather than consider regression as a supplement to a possibly faulty matching routine, one can consider matching a remedy to artifactual regression results that have been produced by incautious data mining. Ho et al. (2005) suggest that the general procedure one should carry out in any multivariate analysis is to first balance one's data with a matching routine and then to estimate a regression model on the balanced data. From this perspective, matching is a data preprocessor, which can be used to prepare the data for subsequent analysis with something such as a regression routine.[22]

5.4 Extensions and Other Perspectives

In this chapter, we have focused almost exclusively on the estimation of the effect of a binary cause on an interval-scaled outcome, and we have considered only least squares adjustments. Before carrying on to discuss least squares estimation of the effects of many-valued causes, we of course must concede what the reader is surely aware of: We have considered only a tiny portion of what falls under the general topic of regression modeling. We have not considered categorical outcome variables, time series analysis, nested data structures, variance-component models, and so on. One can gain a full perspective of the variants of regression modeling in just sociology and economics by consulting Agresti (2002), Arminger, Clogg, and Sobel (1995), Berk (2004), Hamilton (1994), Hayashi (2000), Hendry (1995), Long (1997), Powers and Xie (2000), Raudenbush and Bryk (2002), Ruud (2000), and Wooldridge (2002).

In the next subsection, we consider only one modest extension: least squares regression models for many-valued causes. This presentation then leads naturally to a discussion that follows in the next subsection of what might be labelled the "all-cause correct specification" tradition of regression analysis. Informed by the demonstrations offered in this chapter, we discuss the attractiveness of the promise of this alternative perspective but also the implausibility of the

[22] As we note in the conclusion to this chapter, our perspective is that matching and regression should be used together, and to the extent that the distinction between them fades away. Brand and Halaby (2006) is an example of this approach.

perspective as a general guide for either causal analysis or regression practice in the social sciences.

5.4.1 Regression Estimators for Many-Valued Causes

We suspect that the vast majority of published regression estimates of causal effects in the social sciences are for causes with more than two values. Accordingly, as in Subsection 4.6.3 on matching estimators for many-valued causes, we must discuss the additional complexities of analogous regression estimators. We will again, however, restrict attention to an interval-scaled outcome.

First, again recall the basic setup for many-valued causes from Chapter 2, Appendix B, in which we have a set of J treatment states, a corresponding set of J causal exposure dummy variables, $\{Dj\}_{j=1}^{J}$, and a corresponding set of J potential outcome random variables, $\{Y^{Dj}\}_{j=1}^{J}$. The treatment received by each individual is Dj^*.

How would one estimate the causal effect of such a J-valued cause with regression methods? The first answer should be clear from our presentation in the last section: Because regression can be seen as a form of matching, one can use the same basic strategies outlined for matching estimators of many-valued causes in Subsection 4.6.3. One could form a series of two-way comparisons between the values of the cause, estimate a separate propensity score for each contrast, and then use a weighted regression model to estimate each pairwise causal effect.

If the number of causal states is relatively large, then this general strategy is infeasible. Some smoothing across pairwise comparisons would be necessary, either by collapsing some of the J causal states or by imposing an ordering on the distribution of the causal effect across the J causal states. The most common parametric restriction would be to assume that the causal effect is linear in j for each individual i. For example, for a set of causal states (such as years of schooling) enumerated by values from 0 to 1, 2, 3, to J, the linearity assumption is the assumption that $y_i^{Dj} = y_i^{D0} + \beta_i(j)$ for all $j > 0$, which requires that the difference $y_i^{Dj} - y_i^{Dj-1}$ for each individual i be equal to a constant β_i for all $j > 0$. In this case, the individual-level causal effect is then a slope β_i [rather than the simple difference in potential outcomes, δ_i, specified earlier in Equation (2.1)]. This setup is analogous to the dose-response models for matching estimators discussed in Subsection 4.6.3, but it explicitly leaves open the possibility that the dose-response relationship varies across individuals even though it remains linear.

Angrist and Krueger (1999) show in a very clear example how both a linearity assumption on the individual-specific, dose-response relationship and a fully flexible coding of adjustment variables results in an OLS weighting scheme for the average value of β_i in a sample that is even more complex than what we discussed earlier for simple binary causes (see Regression Demonstration 2 in Subsection 5.3.1). A form of conditional-variance weighting is present again, but now the weighting is in multiple dimensions because least squares must calculate average derivatives across the linearly ordered causal variable (see Angrist and

Krueger 1999, Equation 34). Because one cannot intuitively grasp how these weights balance out across all the dimensions of the implicit weighting scheme (at least we cannot do so), Angrist and Krueger help by offering a familiar example: an OLS estimate of the average causal effect of an additional year of schooling on labor market earnings, assuming linearity in years of schooling and using a fully flexible coding of adjustment variables for age, race, and residence location. They show that, for this example, OLS implicitly gives more weight to the causal effect of shifting from 13 to 14 years of schooling and from 14 to 15 years of schooling than for much more common differences such as the shift from 11 to 12 years of schooling (primarily because the net conditional unexplained variance of schooling is greatest for the contrasts between 13 and 14 years and between 14 and 15 years). They also show that, for this example, the piecewise increases in average earnings happen to be largest for the years of schooling that OLS systematically weights downward. The result is a least squares estimate under the linearity constraint of .094, which is smaller than the weighted average estimate of .144 that one can calculate by dropping the linearity constraint and then averaging year-specific estimates over the marginal distribution of years of schooling.

For other examples, the weighting schemes may not generate sufficiently different estimates, as the overall weighting is a complex function of the relationship between the unaccounted for variance of the causal variable within strata of the adjustment variables and the level of nonlinearity of the conditional expectation function. But the general point is clear and should be sobering: Linearity constraints across causal states may lead OLS models to generate nonintuitive (and sometimes misleading) averages of otherwise easily interpretable stratum-specific causal effects.

5.4.2 Data Mining and the Challenge of Regression Specification

In this subsection, we discuss the considerable appeal of what can be called the all-cause, complete-specification tradition of regression analysis. We argue that this orientation is impractical for most of the social sciences, for which theory is too weak and the disciplines too contentious to furnish perfect specifications that can be agreed on. At the same time, we agree that inductive approaches to discovering flawless regression models that represent all causes are mostly a form of self-deception, even though some computer programs now exist that can prevent the worst forms of abuse.

Consider first a scenario in which one has a theoretical model that one believes is true. It suggests all of the inputs that determine the outcome of interest, as a set of observable variables, and it is in the form of a specific function that relates all inputs to the outcome. In this case, one can claim to have the correct specification for a regression of the outcome on some function of the variables suggested by the theoretical model. The only remaining challenges are then measurement, sampling, and observation.

The weakness of this approach is that critics can claim that the model is not true and hence that the entailed regression specification is wrong. Fighting off any such critics with empirical results can then be difficult, given that the regression specification used to generate the empirical results has been called into question.

In general, if members of a community of competing researchers assert their own true models and then offer up purportedly flawless regression models, the result may be a war of attrition in which no scientific progress is possible. It is therefore natural to ask: Can the *data* generate an all-cause, complete-specification regression model that all competing researchers can jointly adopt?

The first step in answering this question is to determine what an all-cause, complete specification would be, which is sometimes simply labeled a "correct specification."[23] In his 1978 book, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Edward Leamer lays out the following components of what he labels The Axiom of Correct Specification:

> (a) The set of explanatory variables that are thought to determine (linearly) the dependent variable must be (1) unique, (2) complete, (3) small in number, and (4) observable. (b) Other determinants of the dependent variable must have a probability distribution with at most a few unknown parameters. (3) All unknown parameters must be constant. (Leamer 1978:4)

But Leamer then immediately undermines the axiom as it applies to observational data analysis in the social sciences:

> If this axiom were, in fact, accepted, we would find one equation estimated for every phenomenon, and we would have books that compiled these estimates published with the same scientific fanfare that accompanies estimates of the speed of light or the gravitational constant. Quite the contrary, we are literally deluged with regression equations, all offering to "explain" the same event, and instead of a book of findings we have volumes of competing estimates. (Leamer 1978:4)

One can quibble with Leamer's axiom [e.g., that component (3) is not essential and so on], but the literature seems to provide abundant support for his conclusion. Few examples of flawless regression models suggested by true theoretical models can be found in the social science literature. One might hope for such success in the future, but the past 40 years of research do not give much reason for optimism.

Leamer instead argues that most regression models are produced by what he labels a data-instigated specification search, which he characterizes as a Sherlock Holmes form of inference wherein one refrains from developing a model

[23] The literature has never clearly settled on a definition that has achieved consensus, but Leamer's is as good as any. Part of the confusion arises from the recognition that a descriptively motivated regression model is always correct, no matter what its specification happens to be.

or any firm hypotheses before first considering extensively all the facts of a case. Leamer argues that this approach to variable selection and specification is fraught with potential danger and invalidates traditional notions of inference.

Consider the example of the Catholic school effect on learning, and in particular the research of James Coleman and his colleagues. In seeking to estimate the effect of Catholic schooling on achievement, Coleman did not draw a complete specification for his regression models from a specific theoretical model of human learning. This decision was not because no such models existed, nor because Coleman had no appreciation for the need for such models. He was, in contrast, well aware of classic behaviorist models of learning (see Bush and Mosteller 1955, 1959) that specified complex alternative mechanisms for sequences of responses to learning trials. Although he appreciated these models, he recognized (see Coleman 1964:38) that they could not be deployed effectively in the complex environments of secondary schooling in the United States, the context of which he had already studied extensively (see Coleman 1961).

As a result, Coleman did not specify a learning model that justified the regression models that he and his colleagues presented (see Sørensen 1998; Sørensen and Morgan 2000).[24] Their basic specification strategy was instead to attempt to adjust for a sufficient subset of other causes of learning so that, net of these effects, it could be claimed that Catholic and public schools students were sufficiently equivalent. The specific variables that Coleman and his colleagues chose to include in their models were based in part on Coleman's deep knowledge of what predicts learning in high school (and one could argue that Coleman was the most knowledgeable social scientist on the topic in the world at the time). But he and his colleagues also adopted an empirical approach, as indicated parenthetically at the end of the following account of their selection of adjustment variables:

> In order to minimize the effects of differences in initial selection masquerading as effects of differences in the sectors themselves, achievement subtests were regressed, by sector and grade, on a larger number of background variables that measure both objective and subjective differences in the home. Some of these subjective differences may not be prior to the student's achievement, but may in part be consequences of it, so that there may be an overcompensation for

[24] When the 1982 data on seniors became available to supplement the 1980 data on sophomores, Coleman and his colleagues did move toward a stronger foundation for their specifications, providing an underlying model for the lagged achievement gain regression model that was an outgrowth of Coleman's early work on Markov chains and his proposals for longitudinal data analysis (Coleman 1964, 1981). In Hoffer et al. (1985:89–91), he and his colleagues showed that (subject to restrictions on individual heterogeneity) the lagged test score model is a linearized reduced-form model of two underlying rates (learning and forgetting) for the movement between two states (know and don't know) for each item on the cognitive test. Although the model is plausible, it is clearly constrained so that it can be estimated with simple regression techniques (see Coleman 1981:8–9 for an explanation of his *modus operandi* in such situations), and this is of course not the sort of constraint that one must adopt if one is truly interested in laying out the correct theoretical model of learning.

> background differences. It was felt desirable to do this so as to compensate for possible unmeasured differences in family background; but of course the results may be to artificially depress the resulting levels of background-controlled achievement in Catholic and other private schools. (A few additional background variables were initially included; those that showed no effects beyond the ones listed in the following paragraph were eliminated from the analysis.) (Coleman et al. 1982:147)

Coleman and his colleagues then reported that the final list of variables included 10 they considered "clearly prior" to school sector – including family income, parents' education, number of siblings, and number of rooms in the home – as well as 7 other variables that they considered "not clearly prior" to school sector – including more than 50 books in the home, owning a pocket calculator, and having a mother who thinks the student should go to college after high school.

As so often occurs in causal controversies of public importance, critics found this resulting list inadequate. From the perspective of their critics, Coleman and his colleagues had not provided a clear enough accounting of why some students were observed in Catholic schools whereas others were observed in public schools and why levels of learning should be considered a linear function of background and the specific school characteristics selected. After arguing that more would be known when follow-up data were collected and test score gains from sophomore to senior year could be analyzed, Alexander and Pallas (1983) argued that Coleman and his colleagues should have searched harder for additional adjustment variables:

> Failing this [estimating models with pretest and posttest data], another possibility would be to scout about for additional controls that might serve as proxies for student input differences that remain after socioeconomic adjustments. One candidate is the student's curriculum placement in high school. (Alexander and Pallas 1983:171)

Alexander and Pallas then laid out a rationale for this proxy approach, and they offered models that showed that the differences between public and private schools are smaller after conditioning on type of curriculum.

As this example shows, it is often simply unclear how one should go about selecting a sufficient set of conditioning variables to include in a regression equation when adopting the "adjustment for all other causes" approach to causal inference. Coleman and colleagues clearly included some variables that they believed that perhaps they should not, and they presumably tossed out some variables that they thought they should perhaps include but that proved to be insufficiently powerful predictors of test scores. Even so, Alexander and Pallas criticized Coleman and his colleagues for too little scouting.[25]

[25] Contrary to the forecasts of Coleman and his critics, after the 1982 data were released, the specification debate did not end. It simply moved on to new concerns, primarily how

Leamer, as mentioned earlier, would characterize such scouting as a Sherlock-Holmes-style, data-driven specification search. Leamer argues that this search strategy turns classical inference on its head:

> ... if theories are constructed after having studied the data, it is difficult to establish by how much, if at all, the data favor the data-instigated hypothesis. For example, suppose I think that a certain coefficient ought to be positive, and my reaction to the anomalous result of a negative estimate is to find another variable to include in the equation so that the estimate is positive. Have I found evidence that the coefficient is positive? (Leamer 1983:40)

Taken to its extreme, the Sherlock Holmes regression approach may discover relationships between candidate independent variables and the outcome variable that are due to sampling variability and nothing else. David Freedman showed this possibility in a simple simulation exercise, in which he sought to demonstrate that " ... in a world with a large number of unrelated variables and no clear a priori specifications, uncritical use of standard [regression] methods will lead to models that appear to have a lot of explanatory power" (Freedman 1983:152). To show the plausibility of this conclusion, Freedman constructed an artificial dataset with 100 individuals, one outcome variable Y, and 50 other variables X_1 through X_{50}. The 100 values for each of these 51 variables were then independent random draws from the standard normal distribution. Thus, the data represent complete noise with only chance dependencies between the variables that mimic what any real-world sampling procedure would produce. The data were then subjected to regression analysis, with Y regressed on X_1 through X_{50}. For these 50 variables, 1 variable yielded a coefficient with a p value of less than .05 and another 14 had p values of less than .25. Freedman then ran a second regression of Y on the 15 variables that had p values of less than .25, and in this second pass, 14 of them again turned up with p values of less than .25. Most troubling, 6 of them now had p values of less than .05, and the model as a whole had an R^2 of .36. From pure noise and simulated sampling variability, Freedman produced a regression model that looks similar to any number of those published in social science articles. It had six coefficients that passed conventional standards of statistical significance, and it explained a bit more than one third of the variance of the outcome variable.[26]

The danger of data-driven specification searches is important to recognize, but not all procedures are similarly in danger, especially given developments since Leamer first presented his critique in the 1970s and 1980s. There is a new literature on data mining and statistical learning that has devised techniques to avoid the problems highlighted by Freedman's simulation (see Hastie,

to adjust for sophomore test scores (with or without a family background adjustment, with or without curriculum differences, with only a subset of sophomore test scores, and with or without adjustment for attenuation that is due to measurement error).

[26] Raftery (1995) repeated Freedman's simulation experiment and obtained even more dramatic results.

Tibshirani, and Friedman 2001). For a very clear overview of these methods, see Berk (2006, 2007).[27]

Even so, data-instigated specifications of regression equations remain a problem in practice, because few applied social scientists use the fair and disciplined algorithms in the statistical learning literature. The Catholic school example is surely a case in which scouting led to the inclusion of variables that may not have been selected by a statistical learning algorithm. But, nonetheless, none of the scholars in the debate dared to reason backwards from their regression models in order to declare that they had inductively constructed a true model of learning. And, in general, it is hard to find examples of complete inductive model building in the published literature; scholars are usually driven by some theoretical predilections, and the results of mistaken induction are often fragile enough to be uncovered in the peer review process.[28] Milder forms of misspecification are surely pervasive.

5.5 Conclusions

Regression models, in their many forms, remain one of the most popular techniques for the evaluation of alternative explanations in the social sciences. In this chapter, we have restricted most of our attention to OLS regression of an interval-scaled variable on a binary causal variable. And, although we have considered how regression modeling can be used as a descriptive data reduction tool, we have focused mostly on regression as a parametric adjustment technique for estimating causal effects, while also presenting the deep connections between regression and matching as complementary forms of a more general conditioning estimation strategy.

We conclude this chapter by discussing the strengths and weaknesses of regression as a method for causal inference from observational data. The main strengths of regression analysis are clearly its computational simplicity, its myriad forms, its familiarity to a wide range of social scientists, and the ease with which one can induce computer software to generate standard errors. These are all distinct advantages over the matching techniques that we summarized in Chapter 4.

[27] And, as we noted earlier in Chapter 4, there are cases in which a data-driven specification search is both permissive and potentially quite useful. Consider again the causal graph in Figure 3.10 and suppose that one has a large number of variables that may be associated with both D and Y in one's dataset and that one presumes may be members of either S or X. Accordingly, one has the choice of conditioning on two different types of variables that lie along the back-door path from D to Y: the variables in S that predict D or the variables in X that predict Y. Engaging in a data-driven specification search for variables that predict Y will fall prey to inferential difficulties about the causal effect of D on Y for exactly the reasons just discussed. But a data-driven specification search for variables that predict D will not fall prey to the same troubles, because in this search one does not use any direct information about the outcome Y.

[28] That being said, predictions about the behavior of financial markets can come close. See Krueger and Kennedy (1990) for discussion and interpretation of the apparent effect of Super Bowl victories on the U.S. stock market.

But, as we have shown in this chapter, regression models have some serious weaknesses. Their ease of estimation tends to suppress attention to features of the data that matching techniques force researchers to consider, such as the potential heterogeneity of the causal effect and the alternative distributions of covariates across those exposed to different levels of the cause. Moreover, the traditional exogeneity assumption of regression (e.g., in the case of least squares regression that the independent variables must be uncorrelated with the regression error term) often befuddles applied researchers who can otherwise easily grasp the stratification and conditioning perspective that undergirds matching. As a result, regression practitioners can too easily accept their hope that the specification of plausible control variables generates an as-if randomized experiment.

Focusing more narrowly on least squares models, we have shown through several demonstrations that they generate causal effect estimates that are both nonintuitive and inappropriate when consequential heterogeneity has not been fully parameterized. In this sense, the apparent simplicity of least squares regression belies the complexity of how the data are reduced to a minimum mean-squared-error linear prediction. For more complex regression models, the ways in which such heterogeneity is implicitly averaged are presently unknown. But no one seems to suspect that the complications of unparameterized heterogeneity are less consequential for fancier maximum-likelihood-based regression models in the general linear modeling tradition.

Our overall conclusion is thus virtually the same as for matching: Regression is a statistical method for analyzing available data, and for the estimation of causal effects it may have some advantages in some situations. The joint implication of these conclusions for causal analysis is that matching and regression are probably best used together, or at least used in ways such that the distinction between them fades away.

Part 3: Estimating Causal Effects When Simple Conditioning Is Ineffective

Chapter 6

Identification in the Absence of a Complete Model of Causal Exposure

In this chapter, we introduce strategies to estimate causal effects when simple conditioning methods will not suffice. After reviewing the related concepts of a nonignorable treatment assignment and selection on the unobservables, we then consider sensitivity analysis and partial identification approaches. Thereafter, we introduce three strategies to identify and then estimate causal effects: (1) conditioning on a prior value of the observed outcome variable, (2) using an instrumental variable (IV) for the causal variable, and (3) estimating an isolated and exhaustive mechanism (or set of mechanisms) that relates the causal variable to the outcome variable. Under very specific assumptions, these strategies will identify a specific average causal effect of interest even though selection is on the unobservables and treatment assignment is nonignorable. These strategies are then explained more completely in the following three chapters, where the specific details of estimation are laid out.

6.1 Nonignorability and Selection on the Unobservables Revisited

As demonstrated earlier in Subsections 3.2.1 and 3.2.2, the concept of ignorable treatment assignment is closely related to the concept of selection on the observables. In many cases, they can both be represented by the same causal graph. Recall Figure 3.8, in which in panel (a) there are two types of paths between D and Y: the causal effect of D on Y represented by $D \rightarrow Y$ and an unspecified set of back-door paths represented collectively by the bidirected edge $D \dashleftarrow\!\!\dashrightarrow Y$. If this diagram can be elaborated, as in panel (b), by replacing $D \dashleftarrow\!\!\dashrightarrow Y$ with a fully articulated back-door path $D \leftarrow S \rightarrow Y$, then

the graph becomes a full causal model. And observation of S ensures that the matching and regression conditioning strategies of the prior two chapters can be used to generate consistent estimates of the causal effect of D on Y. In this scenario, selection is on the observables – the variables in S – and the remaining treatment assignment mechanism is composed only of random variation that is completely ignorable.

If, in contrast, as shown in panel (b) of Figure 3.9 rather than Figure 3.8, only a subset Z of the variables in S is observed, then selection is on the unobservables because some components of S are now embedded in U. Conditioning on Z in this causal diagram leaves unblocked back-door paths represented by $D \leftarrow U \dashleftarrow\dashrightarrow Y$ untouched. And, as a result, any observed association between D and Y within strata defined by Z cannot be separated into the genuine causal effect of D on Y and the back-door association between D and Y that is generated by the unobserved determinants of selection, U.

As noted in Chapter 3, and then as shown in demonstrations in Chapters 4 and 5, the concepts of ignorability and selection on the observables are a bit more subtle when potential outcomes are introduced. Weaker forms of conditional independence can be asserted about the joint distribution of Y^1, Y^0, D, and S than can be conveyed as in Figures 3.8 and 3.9, in which the observable variable Y is depicted instead of the potential outcome variables Y^1 and Y^0. For example, the average treatment effect for the treated can be estimated consistently by asserting only that Y^0 is independent of D conditional on S, even though full ignorability does not hold [see discussion of Equation (3.3) in Subsection 3.2.1].[1]

6.2 Sensitivity Analysis for Provisional Causal Effect Estimates

What if a researcher suspects, based on substantive knowledge from other studies or a particular theoretical perspective, that treatment selection is nonignorable because selection is, in part, on unobserved variables? In this case, the researcher can still offer a causal effect estimate under the assumption that treatment selection is ignorable. She must then judge how wrong the results may be and offer interpretations that are appropriately cautious.

Although Hubert Blalock is often accused of inspiring naive causal analysis in sociology (and we have engaged, as well, in our own share of Blalock criticism in Chapter 1), he did warn against overconfident causal claims based on conditioning with regression methods. At the end of his influential 1964 book *Causal Inferences in Nonexperimental Research*, he wrote:

[1] Furthermore, as discussed in Sections 4.2.1 and 4.4 in the presentation of matching techniques, one need not assume full conditional independence Y^0 in order to consistently estimate the treatment effect for the treated, as an equality of the two conditional expectations, $E[Y^0|D = 1, S] = E[Y^0|D = 0, S]$, will suffice [see discussion of Assumption 2-S in Equation (4.2)].

> It seems safe to conclude that the problem of making causal inferences on the basis of nonexperimental data is by no stretch of the imagination a simple one. A number of simplifying assumptions must be made, perhaps so many that the goal would seem almost impossible to achieve. The temptation may very well be to give up the entire venture. But, as we have suggested at a number of points, there may not be any satisfactory alternatives. Most likely, social scientists will continue to make causal inferences, either with or without an explicit awareness of these problems. (Blalock 1964:184)

His recommendation was to report a range of plausible results, and he lamented the pressures to settle in on only one favored set of results:

> ... it would be extremely helpful if social scientists would develop the habit of contrasting the results of several *different* procedures. ... If conclusions or results differed, one might gain valuable insights as to why specific differences have occurred. ... Present practices and publication policies are undoubtedly unfavorable to such a strategy. ... there are undoubtedly pressures on the individual investigator to report only his "best" results in instances where different techniques have not given the same results. (Blalock 1964:184-5)

In the decades since scholars such as Blalock appealed for the multimodel reporting of results, a number of scholars have attempted to systematize this approach. In the counterfactual tradition, the approach known as sensitivity analysis has been the most influential.[2] Based largely on the work of statistician Paul Rosenbaum and his colleagues (see Rosenbaum 1991, 1992, 2002; Gastwirth, Krieger, and Rosenbaum 1998, 2000), the guiding principle of the approach is simple: When reporting and interpreting an estimate of a causal effect, researchers should analyze and report how sensitive the estimates and interpretations are to the maintained assumptions of the analysis.

For Rosenbaum, sensitivity analysis of this form is quite distinct from other types of robustness checks on one's results because of its concern with estimates of causal effects.[3] Rosenbaum (1999:275) writes "Assumptions are of three kinds: (i) the scientific model or hypothesis, which is the focus of scientific

[2]There is a related literature on sensitivity analysis that is tied to simulation methods and deterministic modeling more generally. In this tradition, for example, the parameters of a deterministic simulation of an outcome are varied systematically. If variation in nuisance parameters of no inherent interest does not change the basic results of the model, then the model is robust. And, as described in Saltelli, Tarantola, Campolongo, and Ratto (2004), such methods can be used to assess the uncertainty of model predictions based on the uncertainty of model inputs. Such results can direct future effort optimally, so as to reduce the uncertainty of those inputs that generate the most uncertainty of prediction.

[3]For example, Rosenbaum (1999:271) writes that "A sensitivity analysis asks to what extent plausible changes in assumptions change conclusions. In contrast, a stability analysis asks how ostensibly innocuous changes in analytical decisions change conclusions. A sensitivity analysis typically examines a continuous family of departures from a critical assumption, in which the magnitude of the departure and the magnitude of the change in conclusions are the focus of attention. In contrast, a stability analysis typically examines a discrete decision, and the

interest, (ii) incidental assumptions, needed for statistical inference but of little or no scientific interest and (iii) pivotal assumptions which rival the scientific hypothesis and are the focus of scientific controversy."

Sensitivity analysis is usually focused narrowly on the specific and pivotal assumption of ignorability of treatment assignment. Here, the question of interest is, How sensitive are estimates of a causal effect to the potential effects of unobservable treatment selection patterns?

Rosenbaum (2002) devotes a large portion of his excellent book on observational data analysis to strategies for performing sensitivity analysis, especially for matching designs.[4] Examples in the literature demonstrate the utility of the approach. Harding (2003), for example, assesses the strength of the relationship that a composite of unobserved variables would have to have with a treatment and an outcome variable in order to challenge his causal inferences on the effect of neighborhood context on the odds of completing high school.

If, after performing a sensitivity analysis, one can show that all plausible violations of ignorability would not change the qualitative conclusions, then one should report the level of sensitivity and confidently stick to the conclusions. If, however, plausible violations of ignorability would change the qualitative conclusions, then one should step back from strong conclusions and consider other methods of analysis. In the next section, we consider the most obvious first step – an analysis of bounds.

6.3 Partial Identification with Minimal Assumptions

Before considering alternative types of assumptions that allow for the point identification of causal effects of interest even though selection is on the unobservables, such as IV estimators and causal effect estimation by Pearl's front-door criterion, we first consider what can be learned about the causal effects while imposing the weakest assumptions conceivable.

In a series of articles and books that have built on less formalized work of the past, Charles Manski (1994, 1995, 1997, 2003) has investigated plausible values for treatment effect parameters, such as the average treatment effect, that are consistent with the data when weak assumptions alone are maintained. In this section, we introduce Manski's work to show that, in most research situations, the observed data themselves provide some information on the size of the treatment effect without any auxiliary assumptions.

---

hope and expectation is that the conclusions are largely unaltered by changing this decision. Stability analyses are necessary in most complex analyses; however, the extent to which they are needed varies markedly from study to study."

[4] For expositions of the basic perspective written for social scientists, see DiPrete and Gangl (2004) as well as Frank (2000).

6.3.1 No-Assumptions Bounds

To see why the collection of data can bound the range of permissible values for the average treatment effect, consider a hypothetical example in which we know that both Y^1 and Y^0 are bounded by 0 and 1, and thus that Y is also bounded by 0 and 1. Examples could be dichotomous potential outcomes for graduating high school or not in response to two causal states, such as Catholic schooling or public schooling discussed extensively already. Or the potential outcomes could be whether one voted for Al Gore, depending on whether or not one received a butterfly ballot (see Subsection 1.3.2).[5]

In this case, we know from the definitions of Y^1 and Y^0, even without collecting data, that the average treatment effect $E[\delta]$ cannot be greater than 1 or less than -1. This is a straightforward implication of the obvious point that no individual treatment effect can be greater than 1 or less than -1. Accordingly, the maximum average treatment effect of 1 would occur if $E[Y^1|D=1] = E[Y^1|D=0] = 1$ and $E[Y^0|D=1] = E[Y^0|D=0] = 0$ whereas the minimum average treatment effect of -1 would occur instead if $E[Y^1|D=1] = E[Y^1|D=0] = 0$ and $E[Y^0|D=1] = E[Y^0|D=0] = 1$. Thus, we know that $E[\delta]$ must be contained in an interval of length 2, which can be stated as a known bound on $E[\delta]$, using the notation

$$-1 \leq E[\delta] \leq 1 \tag{6.1}$$

or that $E[\delta]$ lies in an interval with closed bounds, as in

$$E[\delta] \in [-1, 1]. \tag{6.2}$$

Manski has shown that we can improve on these bounded intervals considerably without making additional assumptions by collecting data on Y and D. By following a systematic sampling strategy from a well-defined population, we can assert the specific convergence results stated earlier in Equations (2.9)–(2.11). These would ensure that, for an infinite sample, $E_N[d_i] = E[D]$, $E_N[y_i|d_i=1] = E[Y^1|D=1]$, and $E_N[y_i|d_i=0] = E[Y^0|D=0]$. Knowledge of these three quantities allows one to narrow the bound of width 2 in Equations (6.1) and (6.2) to one with a width of only 1.

Consider the hypothetical example depicted in Table 6.1, where, as shown in the first panel, we stipulate that $E[Y^1|D=1] = .7$ and $E[Y^0|D=0] = .3$. The naive estimator, $E_N[y_i|d_i=1] - E_N[y_i|d_i=0]$, would yield values that converge to .4 as the sample size increases. Note that the naive estimator does not use the information about the distribution of the sample across the observed values of D. And, we leave question marks to stand in for the unknown values of the counterfactual conditional expectations $E[Y^1|D=0]$ and $E[Y^0|D=1]$.

[5] This latter example would be especially appropriate to analyze from a bounds perspective, as the goal of such a causal analysis would be to determine whether the causal effect is plausibly large enough to have flipped the election results. With that goal clearly in focus, concerns over putative statistical significance are not terribly relevant.

Table 6.1: A Hypothetical Example of the Calculation of Bounds for the Average Treatment Effect

	$E[Y^1 \mid .]$	$E[Y^0 \mid .]$
Naive estimator suggests $E[\delta] = .4$		
Treatment group	$E[Y^1 \mid D = 1] = .7$	$E[Y^0 \mid D = 1] = ?$
Control group	$E[Y^1 \mid D = 0] = ?$	$E[Y^0 \mid D = 0] = .3$
Largest possible $E[\delta] = .7$		
Treatment group	$E[Y^1 \mid D = 1] = .7$	$E[Y^0 \mid D = 1] = 0$
Control group	$E[Y^1 \mid D = 0] = 1$	$E[Y^0 \mid D = 0] = .3$
Smallest possible $E[\delta] = -.3$		
Treatment group	$E[Y^1 \mid D = 1] = .7$	$E[Y^0 \mid D = 1] = 1$
Control group	$E[Y^1 \mid D = 0] = 0$	$E[Y^0 \mid D = 0] = .3$

Suppose that $E[D] = .5$ and that our sample is infinite such that sampling error is zero. And, for simplicity of notation and consistency with the decomposition in Chapter 2, allow π to again stand in for $E[D]$. For the second and third panels of Table 6.1, the minimum and maximum values of 0 and 1 for the counterfactual conditional expectations $E[Y^1|D = 0]$ are $E[Y^0|D = 1]$ are substituted for the question marks in the first panel.

If half of the sample is in the treatment group, and if $E[Y^1|D = 1] = .7$ and $E[Y^0|D = 0] = .3$, then the largest possible treatment effect is .7 whereas the smallest possible treatment effect is $-.3$. This result is a straightforward calculation of what-if values for the decomposition of the average treatment effect presented in Equation (2.8), which was used earlier to discuss the bias of the naive estimator:

$$\begin{aligned} E[\delta] &= \{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &\quad - \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\}. \end{aligned} \tag{6.3}$$

Plugging in the values from the second panel of Table 6.1 into the decomposition in Equation (6.3) yields the largest possible treatment effect as

$$\begin{aligned} E[\delta] &= \{(.5)(.7) + (1 - .5)(1)\} - \{(.5)(0) + (1 - .5)(.3)\} \\ &= .85 - .15 \\ &= .7, \end{aligned}$$

whereas plugging in the values from the third panel yields the smallest possible treatment effect as

$$\begin{aligned} E[\delta] &= \{(.5)(.7) + (1 - .5)(0)\} - \{(.5)(1) + (1 - .5)(.3)\} \\ &= .35 - .65 \\ &= -.3. \end{aligned}$$

Thus, the constraints implied by the observed data alone guarantee, for this example, that $E[\delta] \in [-.3, .7]$, which is an interval of length 1 and is half the length of the maximum interval calculated before estimates of π, $E[Y^1|D = 1]$, and $E[Y^0|D = 0]$ were obtained from the data.

Manski labels this interval the "no-assumptions bound" because it requires knowledge only of the bounds on Y^1 and Y^0, as well as collection of data on Y and D from a systematic random sample of a well-defined population. Consider now a more general development of these same ideas.

The treatment effect can be bounded by finite values only when the potential outcomes Y^1 and Y^0 are bounded by finite values. In other words, because $E[Y^1|D = 0]$ and $E[Y^0|D = 1]$ are both unobserved, they can take on any values from $-\infty$ to ∞ in the absence of known restrictions on the ranges of Y^1 and Y^0. Thus, in the absence of any known restrictions on $E[Y^1|D = 0]$ and $E[Y^0|D = 1]$, $E[\delta]$ is contained in the completely uninformative interval between $-\infty$ and ∞.

Manski (1994, 1995, 2003) shows that with potential outcome variables bounded by 1 and 0, the no-assumptions bound will always be of length 1. This result is obtained by a more general manipulation of Equation (6.3), for which the lower bound is derived as

$$\{\pi E[Y^1|D = 1] + (1 - \pi)(0)\} - \{\pi(1) + (1 - \pi)E[Y^0|D = 0]\}, \tag{6.4}$$

and the upper bound is derived as

$$\{\pi E[Y^1|D = 1] + (1 - \pi)(1)\} - \{\pi(0) + (1 - \pi)E[Y^0|D = 0]\}. \tag{6.5}$$

Simplifying and then combining Equations (6.4) and (6.5) yields the no-assumptions bound for potential outcomes bounded by 1 and 0:

$$\begin{aligned} &\pi E[Y^1|D = 1] - (1 - \pi)E[Y^0|D = 0] - \pi \\ &\quad \leq E[\delta] \\ &\quad \leq \pi E[Y^1|D = 1] - (1 - \pi)E[Y^0|D = 0] + (1 - \pi). \end{aligned} \tag{6.6}$$

The length of the bound is 1 because the upper and lower bounds differ only by two complementary probabilities π and $1 - \pi$ that sum to 1. The location of the bound in the $[-1, 1]$ interval is set by the common term $\pi E[Y^1|D = 1] - (1 - \pi)E[Y^0|D = 0]$, to which $-\pi$ and $1 - \pi$ are then added to form the bound.

More generally, if Y^1 and Y^0 are bounded by any finite values a and b, where $b > a$, such as theoretical minimum and maximum scores on a standardized test, then Equation (6.1) can be written more generally as

$$-b + a \leq E[\delta] \leq b - a, \tag{6.7}$$

which then generates the more general no-assumptions bound for the observed data:

$$\begin{aligned} &\pi E[Y^1|D=1] - (1-\pi)E[Y^0|D=0] + a(1-\pi) \ - b\pi \\ &\leq E[\delta] \\ &\leq \pi E[Y^1|D=1] - (1-\pi)E[Y^0|D=0] + b(1-\pi) - a\pi. \end{aligned} \tag{6.8}$$

The same basic results hold as before. The no-assumptions bound always includes zero, and it is only half as wide as the bound in Equation (6.7) implied only by the bounds on the potential outcomes. In this case, the no-assumptions bound is of length $b - a$ rather than of length $2(b - a)$.[6] As with the case for potential outcomes bounded by 1 and 0, the particular location of the interval of length $b - a$ in $[-b + a, b - a]$ is again determined by the same term, $\pi E[Y^1|D = 1] - (1 - \pi)E[Y^0|D = 0]$.

6.3.2 Bounds Under Additional Weak Assumptions

For Manski, calculation of the no-assumptions bound is only the starting point of an analysis of bounds. The primary goal is to analyze how additional assumptions can narrow the no-assumptions bound. Manski's basic perspective is summarized nicely in following passage:

> Empirical researchers should be concerned with both the logic and the credibility of their inferences. Credibility is a subjective matter, yet I take there to be wide agreement on a principle I shall call:
>
> *The Law of Decreasing Credibility*: The credibility of inference decreases with the strength of the assumptions maintained.
>
> This principle implies that empirical researchers face a dilemma as they decide what assumptions to maintain: Stronger assumptions yield inferences that may be more powerful but less credible. (Manski 2003:1)

Manski has shown that weak and often plausible assumptions can substantially narrow the no-assumptions bound. Consider the following simple assumptions about the direction of causality and the direction of self-selection (and see his empirical applications for more detail, such as Manski and Nagin 1998; Manski, Sandefur, McLanahan, and Powers 1992).

[6] To see this, note that $[b - a] + [-1(-b + a)]$ simplifies to $2(b - a)$ and $[b(1 - \pi) - a\pi] + [-1(a(1 - \pi) \ - b\pi)]$ simplifies to $b - a$.

Monotone Treatment Response

In many situations, it may be reasonable to assume that the individual-level treatment effect cannot be negative, such that $\delta \geq 0$ for every individual i. Manski labels this assumption the monotone treatment response (MTR) assumption.

Under MTR (in the direction where $\delta \geq 0$), the lower bound for the average treatment effect must be 0. MTR implies that members of the control group have counterfactual values of y_i^1 that are at least as high as their observed values of y_i. The reasoning here is simple: If $y_i = y_i^0$ for members of the control group, then the MTR assumption that $y_i^1 \geq y_i^0$ for all individuals i implies that $y_i^1 \geq y_i$ for members of the control group. The opposite is likewise true for members of the treatment group; their counterfactual values for y_i^0 are no higher than their observed values of y_i. Under MTR for the hypothetical example presented in Table 6.1, one can therefore replace the extreme values of 1 and 0 in the no-assumptions lower bound:

$$\{(.5)(.7) + (1 - .5)(0)\} \; - \{(.5)1 + (1 - .5)(.3)\} = -.3$$

with less extreme values of .7 and .3. As a result, one obtains a new lower bound:

$$\{(.5)(.7) + (1 - .5)(.3)\} \; - \{(.5)(.7) + (1 - .5)(.3)\} = 0,$$

implying that the bound for the average treatment effect, assuming MTR, is $[0, .7]$ rather than $[-.3, .7]$.

Monotone Treatment Selection

It may also be possible to assume that those who receive the treatment have higher average outcomes under potential exposure to both the treatment and control, which Manski labels the monotone treatment selection (MTS) assumption. In most cases, one would assume jointly that $E[Y^1|D = 1] \geq E[Y^1|D = 0]$ and $E[Y^0|D = 1] \geq E[Y^0|D = 0]$, which is traditionally thought of as positive self-selection. (But one could flip the direction of the assumption, just as we also noted for MTR.) Manski and Pepper (2000) present this type of MTS assumption for the example of the effect of education on earnings, for which it is equivalent to assuming that individuals with higher education would on average receive higher wages than individuals with lower education, under counterfactual conditions in which they had the same levels of education.

For the hypothetical example presented in Table 6.1, MTS implies that the naive estimator would be an upper bound for the average treatment effect. In short, MTS in this direction stipulates that members of the treatment/control group could not have done any worse/better in the control/treatment state than those observed in the control/treatment group. Maintaining MTS allows one to replace the extreme values of 1 and 0 in the no-assumptions upper bound,

$$\{(.5)(.7) + (1 - .5)(1)\} \; - \{(.5)0 + (1 - .5)(.3)\} = .7,$$

with the less extreme values of .7 and .3, yielding for this example,

$$\{(.5)(.7) + (1 - .5)(.7)\} \; - \{(.5)(.3) + (1 - .5)(.3)\} = .4.$$

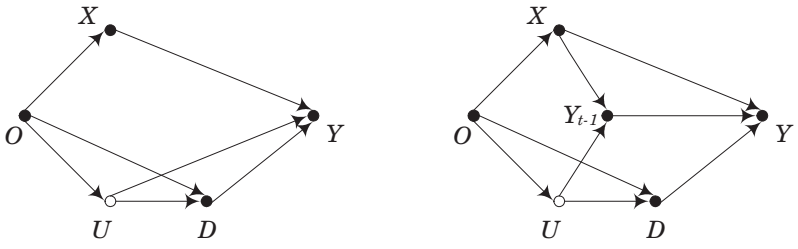
MTS thereby implies that the bound for the average treatment effect is $[-.3, .4]$ rather than $[-.3, .7]$.

Combinations of Weak Assumptions

The power of Manski's approach comes from the ability to apply different assumptions at the same time in order to narrow the no-assumptions bound to a bound that is considerably more informative. For the hypothetical example presented in Table 6.1, one can narrow the no-assumptions bound from $[-.3, .7]$ to $[0, .4]$ by invoking the MTR and MTS assumptions together. The resulting bound still includes 0, but in some applications such narrowing may well be helpful in ruling out some unreasonable and extreme causal claims.

In his 2003 book, *Partial Identification of Probability Distributions*, Manski formalizes the bounds approach presented here into a complete theory of partial identification. For this formalization, he relabels the bounded intervals as identification regions and states: "I begin with the identification region obtained using the empirical evidence alone [the no-assumptions bound represented by Equation (6.8)] and study how distributional assumptions may shrink this region" (Manski 2003:5).[7] We will return to Manski's partial identification approach later, when we discuss IVs as here he has developed a set of weaker assumptions that are of considerable interest. Next, we present three point-identification strategies that do not require that selection be on the observables only.

[7]Manski continues this quotation: "A mathematically complementary approach is to begin with some point-identifying assumption and examine how identification decays as this assumption is weakened in specific ways" (Manski 2003:5). This alternative is sensitivity analysis, as described earlier with reference to the work of Paul Rosenbaum (see Rosenbaum 2002). Manski regards Rosenbaum's sensitivity analysis approach as too narrow, writing, "Another form of prior information that may yield nonintersecting bounds is proposed by Rosenbaum in his ... discussion of sensitivity analysis. Here Rosenbaum, who views the assumption of ignorable treatment assignment as critical to the interpretation of observational studies, considers weakening this assumption in a particular manner. ... Where Rosenbaum and I differ is that I do not view the assumption of ignorable treatment selection to have a special status in observational studies of treatment effects. As an economist, I usually am inclined to think that treatments are purposefully selected and that comparison of outcomes plays an important role in the selection process. Perhaps the departures from ignorable treatment selection that Rosenbaum entertains in his sensitivity analysis can be interpreted behaviorally in terms of some model of purposeful treatment selection, but for now I do not see how" (Manski 1999:281). Sensitivity analysis can be and is being generalized, to the extent that the boundary between sensitivity analysis and partial identification will gradually disappear. Consider, for example, savvy applied work such as Berk and de Leeuw (1999).



(a) The identification puzzle (b) Coleman's solution

Figure 6.1: Coleman's strategy for the identification of the causal effect of Catholic schooling on learning.

6.4 Additional Strategies for the Point Identification of Causal Effects

Although Manski shows that there are many potential strategies to point identify causal effects, we focus in this section on the three most prominent strategies for cases in which treatment selection is on unobserved variables. The following three strategies have already been foreshadowed in the two prior sections in Chapters 1 and 3 on Pearl's approach to the identification of causal effects. Here, with more context, we present them with reference to the regression equations offered by James Coleman and his colleagues for the estimation of the causal effect of Catholic schooling on high school achievement.

We first note that point identification can (sometimes) be achieved by conditioning on a lagged value of the outcome variable. Then we introduce IV strategies and conclude with a presentation of estimation via an isolated and exhaustive mechanism.[8] Each of these strategies is then laid out with substantial detail in Chapters 7–9.

6.4.1 Conditioning on a "Pretest"

The basic identification challenge that Coleman and his colleagues confronted is depicted in panel (a) of Figure 6.1. Although somewhat simplified, the causal diagram represents the basic types of causal variables that Coleman and his colleagues contemplated in their original analyses and then in debate with their critics. For the DAG in panel (a), Y is an observed score on a standardized achievement test, and D is an observed dichotomous causal variable equal to 1 for those who attend Catholic schools. The primary goal of the research was to estimate the causal effect of D on Y, conceptualized as the average causal effect of Catholic schooling on achievement.

[8] We also offer an appendix to this chapter, in which a latent variable model of self-selection is used to point identify causal effects.

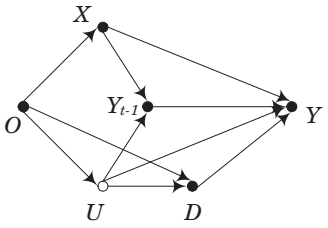
The remaining variables in the DAG represent the basic types of variables that Coleman and his colleagues decided should be specified in their regression equations. The variables in X represent determinants of achievement test scores that have no direct causal effects on sector selection. The variables in O represent ultimate background factors that determine both X and D. Coleman and his colleagues identified many variables that they believed were contained in O and X, although they were unsure of whether to consider them as members of O or X. Finally, U represents the problematic variables that Coleman and his colleagues recognized were probably unobserved. These variables – intrinsic motivation to learn, anticipation of the causal effect itself, and subtle features of the home environment – were thought to determine both school sector choice and achievement. Because these variables were assumed to be unobserved, they are represented collectively by the variable U with a node that is a hollow circle $\circ$.

In the initial research, the analysis strategy of Coleman and his colleagues was to condition on observed variables in X and/or O that blocked two separate back-door paths from D to Y: $D \leftarrow U \leftarrow O \rightarrow X \rightarrow Y$ and $D \leftarrow O \rightarrow X \rightarrow Y$. Unfortunately, these variables in X and O did not block the third back-door path from D to Y, which is $D \leftarrow U \rightarrow Y$. As a result, the causal effect of D on Y remained formally unidentified in their analysis.

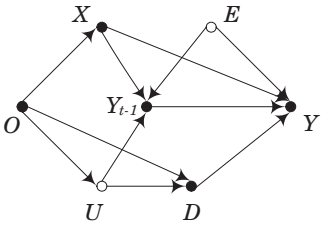
Their solution to this predicament is presented in panel (b) of Figure 6.1, which became possible to estimate when the second round of survey data became available two years later. Rather than treat the tenth-grade test scores as Y, they relabeled these scores as Y_{t-1} and considered the twelfth-grade test scores as a new Y. They then argued that the tenth-grade test scores could serve as an effective pretest variable that could be used to screen off the effects of the variables in U on Y.

In particular, their strategy is equivalent to asserting that conditioning on X, O, and Y_{t-1} blocks all five back-door paths between D and Y. Two of these back-door paths are blocked by Y_{t-1}: $D \leftarrow U \rightarrow Y_{t-1} \rightarrow Y$ and $D \leftarrow U \leftarrow O \rightarrow X \rightarrow Y_{t-1} \rightarrow Y$. The remaining three paths are blocked by X (in some cases with supplemental but unnecessary conditioning on O). The first two of these are easily recognizable as $D \leftarrow U \leftarrow O \rightarrow X \rightarrow Y$ and $D \leftarrow O \rightarrow X \rightarrow Y$. But the final back-door path may appear hidden: $D \leftarrow U \rightarrow Y_{t-1} \leftarrow X \rightarrow Y$. Note that this last path would be blocked in the absence of conditioning, as Y_{t-1} is a collider variable for this path (see our earlier discussion of collider variables in Subsection 3.1.3). Conditioning on Y_{t-1}, however, unblocks this path, and thus the path must be blocked by conditioning on X.

As shown in Figure 6.2, there are two basic criticisms of this approach. The first was seized on by their critics: The variables in U were not plausibly screened off by Y_{t-1}, as shown in panel (a) of Figure 6.2. If highly motivated students were more likely to pay tuition to enroll in Catholic schools, then enhanced motivation would contribute directly to learning in both the tenth grade and the twelfth grade. Similarly, it is unreasonable to assume that the motivation that is correlated with willingness to pay tuition would exert only a one-time boost early in the Catholic school careers of students. Thus, the critics argued,



(a) Criticism 1: The lagged Y is an imperfect screen



(b) Criticism 2: The lagged Y is a collider, and conditioning on it does not block a back-door path through a variable such as E

Figure 6.2: Criticism of Coleman's estimates of the effect of Catholic schooling on learning.

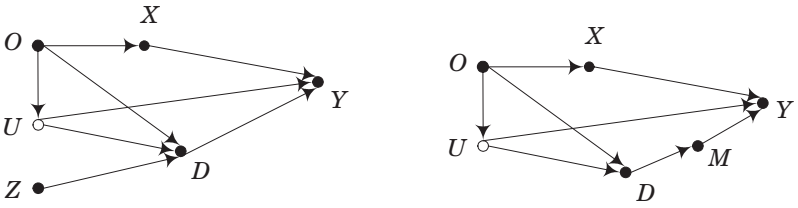
in effect, that a sixth back-door path $D \leftarrow U \rightarrow Y$ exists that cannot be blocked by conditioning on Y_{t-1}.

The second criticism is more subtle. It could be argued that there is a common unobserved cause of both Y_{t-1} and Y, which is represented in panel (b) of Figure 6.2 by E. Consider E to be an unobserved and completely random characteristic of students that measures their taste for tests. Independent of all else, it may be that some students enjoy and/or perform well when taking tests. If such a taste exists, it creates a new back-door path from D to Y: $D \leftarrow U \rightarrow Y_{t-1} \leftarrow E \rightarrow Y$. For this path, Y_{t-1} is a collider. In the absence of conditioning on Y_{t-1}, it would be blocked and would not create a net association between D and Y. Instead, by conditioning on Y_{t-1} in an attempt to block other back-door paths in the graph, Coleman and his colleagues unblocked this path, thereby creating a new net association between D and Y. This new association could be eliminated by additional conditioning on either U or E. But such supplemental conditioning is impossible because both U and E are unobserved.

Many examples in the social science literature have the same basic features. The plausibility of the criticism in both panels of Figure 6.2 should serve as a caution, which is a point we take up in considerable detail in Chapter 9. There, we discuss a variety of issues, including a more complete discussion of the sorts of dynamic processes that generate dependencies such as the one represented by E in panel (b) of Figure 6.2.

6.4.2 Instrumental Variables and Naturally Occurring Variation

An alternative strategy to identify the causal effect of Catholic schooling on achievement is presented in panel (a) of Figure 6.3. For this DAG, a new variable Z has been specified from among those variables that implicitly determined D, independent of O and U, in Figures 6.1 and 6.2. Z is commonly labeled an IV for D. For now, Z can be thought of as a shock to D that is independent of



(a) Alternative solution 1: An instrumental variable

(b) Alternative solution 2: A complete mechanism

Figure 6.3: Alternative identification strategies for the causal effect of Catholic schooling on learning.

O, U, and X. If such an IV exists, one can obtain an estimate of the causal effect of D on Y by first calculating the association between Z and Y and then dividing by the association between Z and D. As we will discuss in Chapter 7, for this strategy to work, Z must have an association with D and with Y, but Z cannot have a causal effect on Y except through its causal effect on D.

One example that we will use when discussing this sort of a technique in Chapter 7 is a variable Z for whether a student in a given geographic area was randomly assigned a tuition voucher that could be redeemed at a Catholic school. In this case, because Z is randomly assigned, it is unlikely to have a direct causal effect on Y. Although not impossible, it would seem that receiving a voucher in the mail would not affect achievement-relevant behavior. For the IV to identify the causal effect, the only causal effect of Z on Y must come about because (1) the voucher induces some students who would otherwise attend public schools to instead attend Catholic schools, and (2) Catholic schools raise the achievement of some of the students who have attended Catholic schools because of the voucher. The critical assumption here – that there is no other pathway, causal or otherwise, between Z and Y, except the path through D – is a strong and untestable assumption, as we discuss extensively with several examples in Chapter 7. Although this approach has considerable appeal, for many applications it can be hard to find a suitable IV. And, even when they are available, they usually do not identify the average causal effect if individual-level heterogeneity of the causal effect is present as well.

6.4.3 Mechanisms and the Front-Door Criterion

As we discussed in Chapter 3, and then demonstrated in Chapters 4 and 5, Pearl's back-door criterion for a sufficient set of conditioning variables offers a powerful strategy for prosecuting causal analysis. Pearl has a second closely related conditioning strategy: the front-door criterion. This type of strategy has been used informally in the social sciences for decades in attempts to explain how causal effects come about. But only recently has it been used as a formal

identification strategy in the social sciences (see, for example, Winship and Harding, forthcoming).

The front-door criterion is perhaps most easily understood with an example used by Pearl. The question of interest is whether smoking causes cancer. As decades of court cases proved, one cannot convince a jury that smoking causes cancer by measuring the association between smoking and cancer because there are many plausible back-door paths between both, such as common genetic propensities to smoke and succumb to cancer. Pearl shows that another analysis strategy was much more effective in making the case that smoking does indeed cause cancer: (1) estimating the effect of smoking on the amount of tar in the lungs, (2) estimating the effect of the amount of tar in the lungs on rates on lung cancer, and (3) multiplying together these two effects to form an estimate of the effect of smoking on cancer.

Consider how this approach can be applied to the Catholic school example, as presented in panel (b) of Figure 6.3. As before, no set of observed variables satisfies the back-door criterion. The problem, again, is that U is unobserved, and thus the back-door path $D \leftarrow U \rightarrow Y$ remains unblocked after conditioning on every subset of observed variables in the graph. Suppose, however, that there is a variable M that intercepts the full causal effect of D on Y. The variable M can be represented as the full mechanism (or causal pathway) that relates D to Y. Observation of M identifies the causal effect of D on Y by a double application of the back-door criterion to estimation of both the causal effect of D on M and the causal effect of M on Y. For the the causal effect $D \rightarrow M$, note that there are no back-door paths between D and M, and therefore the causal effect of D on M can be estimated by their simple association without any conditioning whatsoever. For the causal effect $M \rightarrow Y$, there are three back-door paths between M and Y: $M \leftarrow D \leftarrow U \rightarrow Y$, $M \leftarrow D \leftarrow O \rightarrow X \rightarrow Y$, and $M \leftarrow D \leftarrow U \leftarrow O \rightarrow X \rightarrow Y$. The variable D blocks all three of these back-door paths. Thus, because the causal effects $D \rightarrow M$ and $M \rightarrow Y$ can be estimated consistently, the full causal effect of D on Y can be estimated consistently by combining these causal effect estimates.[9]

In general, one can consistently estimate the effect of a causal variable on an outcome variable by estimating the effect as it propagates through an isolated and exhaustive mechanism. The mechanism need not be (and typically would not be) a single variable. The key requirements of this estimation strategy are that the mechanism (or set of variables that comprise the mechanism) be isolated and exhaustive.[10] Like other point-identification approaches, these are strong assumptions, which we discuss fully in Chapter 8.

[9] How these effects are combined generally involves some form of multiplication, depending on the scales of M and Y and whether the effects are linear.

[10] Before leaving this example, consider the exhaustiveness assumption. For panel (b) of Figure 6.3, if one assumes mistakenly that M represents the full mechanism that relates D to Y, when in fact it is only one isolated piece of a more complete mechanism, then one will likely underestimate the causal effect of D on Y using M. Because one mistakenly ignores some of the true causal pathways that relate D to Y, one attributes net associations generated by the alternative mechanisms to the unblocked back-door paths between D and Y, as in $D \leftarrow U \leftarrow O \rightarrow X \rightarrow Y$.

6.5 Conclusions

In this chapter, we have made the transition from "easy" to "hard" instances of causal effect estimation. No longer does simple conditioning on the determinants of the cause or all other direct causes of the outcome allow for identification and consistent estimation of the causal effect of interest. Instead, we have considered examples in which important variables that might have been used to mount an effective conditioning strategy are unobserved. Thus, selection of the treatment of interest is on unobserved characteristics of individuals that have unknown but suspected relationships to the outcome.

For these more trying situations, we have first considered minimal weak assumptions that can be used to bound plausible values for the size of the causal effect. We then showed how these bounds can be narrowed by the systematic introduction of additional assumptions. Then, we introduced three further sets of assumptions that point identify causal effects when they are valid: (1) conditioning on a lagged outcome variable, as in classic pretest/posttest designs, (2) using an IV, as one finds in the recent surge of interest in natural experiments, and (3) specifying an isolated and exhaustive mechanism.

We discuss these strategies in the remaining chapters, but in a slightly different order than suggested by this chapter. We first turn in Chapter 7 to the IV strategy, which we present in detail to explicate its strengths and weaknesses. Thereafter, we return to identification via mechanisms in Chapter 8, where we will discuss Pearl's perspective in relation to recent proposals for similar strategies from the social sciences. Finally, we consider methods that use longitudinal data in Chapter 9, where we will revisit the pretest estimation strategy considered here and more general models.

Appendix: Latent Variable Selection-Bias Models

Heckman's early work in the late 1970s on selection bias, particularly his lambda method, has received some attention in sociology. Heckman's closely related work on dummy endogenous variables (Heckman 1978), pursued at the same time, has received less attention (but see Winship and Mare 1984).

Although completed before most of Rubin and Rosenbaum's work on propensity scores, Heckman's work on the dummy endogenous variable problem can be understood in relation to the propensity score approach.[11] Recall the treatment selection tradition, and Equation (3.6) in particular:

$$\tilde{D} = Z\phi + U,$$

[11] The general selection model considered by Heckman (1979) can also be estimated by maximum likelihood or nonlinear least squares, although this involves stronger distributional assumptions than the lambda method discussed here. See Winship and Mare (1992) and Fu, Winship, and Mare (2004) for further discussion.

where $\tilde{D}$ is a latent continuous variable, Z includes variables that determine treatment selection, ϕ is a coefficient (or a vector of coefficients if Z includes more than one variable), and U represents both systematic unobserved determinants of treatment selection and completely random idiosyncratic determinants of treatment selection. As before, the latent continuous variable $\tilde{D}$ in Equation (3.6) is then related to the treatment selection dummy, D, by

$$D = 1 \text{ if } \tilde{D} \geq 0,$$
$$D = 0 \text{ if } \tilde{D} < 0,$$

where the threshold 0 is arbitrary because U has no inherent metric (as it is composed of unobservable and some unknown variables).

Rather than focus on the conditional probability of treatment selection (i.e., directly on $\Pr[D = 1|Z]$), Heckman focuses on the conditional expectation of the latent variable $\tilde{D}$. Using the linearity of Equation (3.6), he is interested in

$$E[\tilde{D}|Z\phi, D] = Z\phi + E[U|Z\phi, D]. \tag{6.9}$$

Note that the expected value of $\tilde{D}$ is a function of both $Z\phi$ and D. This allows Heckman to take account of selection that may be a function of both the observables Z and the unobservables U.

Heckman's insight was the recognition that, although one could not observe U directly, one could calculate its expected value from Equation (6.9). The expected value of U could then be used as a control variable (or a control function in Z and D) in order to consistently estimate the average causal effect of D with a standard multiple regression routine.

To calculate the expected value of U in Equation (6.9), one needs to make an assumption about the distribution of U. Typically, U is assumed to be normally distributed (although, as we discuss later, this assumption is the Achilles' heel of the method). If $f(.)$ is the normal density function and $F(.)$ is the corresponding cumulative distribution function, then

$$E[U|Z\phi, D = 1] = \frac{f(Z\phi)}{[1 - F(Z\phi)]}, \tag{6.10}$$

$$E[U|Z\phi, D = 0] = \frac{-f(Z\phi)}{F(Z\phi)}. \tag{6.11}$$

Equation (6.10) simply gives the formula for Heckman's λ in a standard sample selection problem. In the treatment context, a λ for those in the treatment group ($D = 1$) is calculated with Equation (6.10) and a second λ for those in the control group ($D = 0$) is calculated with Equation (6.11). These two λ's would then be entered into a regression equation as control variables analogous to those in X in Equation (5.8). Thus, the procedure here is identical to Heckman's lambda method for correcting for selection bias, except that two distinct λ's are utilized.

As Heckman and many others have come to recognize, estimates from this method can be very sensitive to assumptions about the distribution of U. This

recognition has led to a second wave of selection-bias estimators, which rely on semiparametric estimation of the treatment selection equation (see Honoré and Powell 1994; Pagan and Ullah 1999). Moreover, it has led to recognition of a point that Heckman made clearly in the original development of the model, that sample selection models are most effectively estimated when the variables in Z include instruments. As we discuss next in Chapter 7, Heckman and Vytlacil (2005) offer a framework that represents these sorts of latent variable selection models, semiparametric extensions, IV estimators, and others, within a common counterfactual framework.

Chapter 7

Instrumental Variable Estimators of Causal Effects

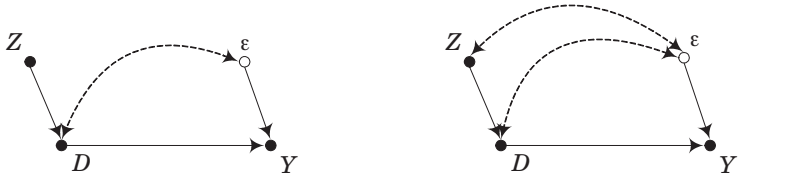
If a perfect stratification of the data cannot be achieved with available data, and thus neither matching nor regression nor any other type of basic conditioning technique can be used to effectively estimate a causal effect of D on Y, one solution is to find an exogenous source of variation that affects Y only by way of the causal variable D. The causal effect is then estimated by measuring how Y varies with the portion of the total variation in D that is attributable to the exogenous variation. The variable that indexes the portion of the total variation in D that is used to estimate the causal effect is an instrumental variable.

In this chapter, we orient the reader to IV estimation of causal effects by presenting examples of binary instruments, some of which are natural experiments because the IVs are "gifts of nature" (Rosenzweig and Wolpin 2000:872). We then return to the origins of IV techniques, and we contrast this estimation strategy with the perspective on regression that was presented in Chapter 5. We then develop the same ideas using the potential outcome framework, showing how the counterfactual perspective has led to a new literature on how to interpret IV estimates. This new literature suggests that IV techniques are more effective for estimating narrowly defined causal effects than for estimating the average causal effects that they are often mistakenly thought to inform. We conclude with a brief discussion of marginal treatment effects identified by local IVs, as well as the use of monotone IVs in an analysis of bounds.

7.1 Causal Effect Estimation with a Binary IV

Recall the causal regression setup in Equation (5.3):

$$Y = \alpha + \delta D + \varepsilon, \tag{7.1}$$



(a) Z is a valid instrumental variable for D

(b) Z is not a valid instrumental variable for D

Figure 7.1: Causal diagrams for Z as a potential IV.

where Y is the outcome variable, D is a binary causal exposure variable, α is an intercept, δ is the causal effect of D on Y, and ε is a summary random variable that represents all other causes of Y. As noted before, when Equation (7.1) is used to represent the causal effect of D on Y, the parameter δ is usually considered an invariant, structural causal effect that applies to all members of the population of interest.

Suppose that the probability that D is equal to 1 rather than 0 is a function of a binary variable Z that takes on values of 0 and 1. Figure 7.1 presents two possible ways in which the variable Z could be related to both D and Y. Note first that, for both causal diagrams, the presence of the set of back-door paths denoted by $D \dashleftarrow\!\dashrightarrow \varepsilon \rightarrow Y$ prevents a least squares regression of Y on D (or any other conditioning strategy) from generating a consistent estimate of the effect of D on Y. However, for the graph in panel (a), Z has an association with Y only through D whereas for the graph in panel (b) Z has an association with Y through D and also through common causes that determine both Z and ε.[1]

Consider how Z can be used to estimate the causal effect of D on Y in the first but not the second causal diagram. We know that for the set of causal relationships in both panels of Figure 7.1, the probability that D is equal to 1 rather than 0 is a function of the value of Z. But D still varies when conditioning on the value of Z because there are common causes that determine both D and ε as well as implicit independent causes that give D its probability distribution. If we continue to maintain the assumption that the effect of D on Y is a constant structural effect δ, then it is not necessary to relate all of the variation in D to all of the variation in Y in order to obtain a consistent estimate of the causal effect. The covariation in D and Y that is generated by the common causes of D and ε can be ignored if a way of isolating the variation in D and Y that is causal can be found.

Can the variation in D and Y that is generated by variation in Z be used to consistently estimate the causal effect of D on Y in this way? The answer to

[1] We have drawn the two bidirected edges separately for the diagram in panel (b) for simplicity. The same reasoning holds in more complex situations in which some of the common causes of Z and ε are also common causes of D and/or Y (and so on).

this question depends crucially on whether or not Z has an association with Y independent of its association with Y through D. For the graph in panel (a), this strategy will succeed because Z is associated with Y only through D. For the graph in panel (b), this strategy will fail because Z and Y share common causes, as represented by the bidirected edge $Z \dashleftarrow\dashrightarrow \varepsilon$. As a result, the variation that Z appears to generate in both D and Y may instead be generated by one or more unobserved common causes of Z and Y.

To see this result a bit more formally, take the population-level expectation of Equation (7.1), $E[Y] = E[\alpha + \delta D + \varepsilon] = \alpha + \delta E[D] + E[\varepsilon]$, and rewrite it as a difference equation in Z:

$$\begin{aligned} &E[Y|Z=1] - E[Y|Z=0] \\ &\quad = \delta(E[D|Z=1] - E[D|Z=0]) + (E[\varepsilon|Z=1] - E[\varepsilon|Z=0]) \ . \end{aligned} \tag{7.2}$$

Equation (7.2) is now focused narrowly on the variation in Y, D, and ε that exists across levels of Z.[2] Now, take Equation (7.2) and divide both sides by $E[D|Z=1] - E[D|Z=0]$, yielding

$$\begin{aligned} &\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} \\ &\quad = \frac{\delta(E[D|Z=1] - E[D|Z=0]) + (E[\varepsilon|Z=1] - E[\varepsilon|Z=0])}{E[D|Z=1] - E[D|Z=0]} \ . \end{aligned} \tag{7.3}$$

If the data are generated by the set of causal relationships depicted in panel (a) of Figure 7.1, then Z has no association with ε, and $E[\varepsilon|Z=1] - E[\varepsilon|Z=0]$ in Equation (7.3) is equal to 0. Consequently, the right-hand side of Equation (7.3) simplifies to δ:

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = \delta \ . \tag{7.4}$$

Under these conditions, the ratio of the population-level association between Y and Z and between D and Z is equal to the causal effect of D on Y. This result suggests that, if Z is in fact associated with D but not associated with ε (or with Y, except through D), then the following sample-based estimator will equal δ in an infinite sample:

$$\hat{\delta}_{\text{IV,WALD}} \equiv \frac{E_N[y_i|z_i=1] - E_N[y_i|z_i=0]}{E_N[d_i|z_i=1] - E_N[d_i|z_i=0]} \ . \tag{7.5}$$

[2] Equation (7.2) is generated in the following way. First, write $E[Y] = \alpha + \delta E[D] + E[\varepsilon]$ conditional on the two values of Z, yielding $E[Y|Z=1] = \alpha + \delta E[D|Z=1] + E[\varepsilon|Z=1]$ and $E[Y|Z=0] = \alpha + \delta E[D|Z=0] + E[\varepsilon|Z=0]$. Note that, because α and δ are considered constant structural effects for this motivation of IV estimation, they do not vary with Z. Now, subtract $E[Y|Z=0] = \alpha + \delta E[D|Z=0] + E[\varepsilon|Z=0]$ from $E[Y|Z=1] = \alpha + \delta E[D|Z=1] + E[\varepsilon|Z=1]$. The parameter α is eliminated by the subtraction, and δ can be factored out of its two terms, resulting in Equation (7.2).

As suggested by its subscript, this is the IV estimator, which is known as the Wald estimator when the instrument is binary. Although the Wald estimator is consistent for δ in this scenario, the assumption that δ is an invariant structural effect is crucial for this result.[3] From a potential outcomes perspective, in which we generally assume that causal effects vary meaningfully across individuals, this assumption is very limiting and quite likely unreasonable. Explaining when and how this assumption can be relaxed is one of the main goals of this chapter.[4]

For completeness, return to consideration of panel (b) of Figure 7.1, in which Z has a nonzero association with ε. In this case, $E[\varepsilon|Z = 1] - E[\varepsilon|Z = 0]$ in Equations (7.2) and (7.3) cannot be equal to 0, and thus Equation (7.3) does not reduce further to Equation (7.4). Rather, it reduces only to

$$\frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} = \delta + \frac{E[\varepsilon|Z = 1] - E[\varepsilon|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} \ . \tag{7.6}$$

In this case, the ratio of the population-level association between Y and Z and between D and Z does not equal the causal effect of D on Y but rather the causal effect of D on Y plus the the last term on the right-hand side of Equation (7.6). The Wald estimator in Equation (7.5) is not consistent or asymptotically unbiased for δ in this case. Instead, it converges to the right-hand side of Equation (7.6), which is equal to δ plus a bias term that is a function of the net association between Z and ε.

More generally, an IV estimator is a ratio that is a joint projection of Y and D onto a third dimension Z. In this sense, an IV estimator isolates a specific portion of the covariation in D and Y. For that selected covariation to be the basis of a valid causal inference for the effect of D on Y, it cannot be attributable to any extraneous common causes that determine both Z and Y. And, to justify the causal effect estimate generated by a subset of the covariation in D and Y as a consistent estimate of the population-level causal effect of D on Y, it is typically assumed in this tradition that δ is a constant for all members of the population.[5] Consider the following hypothetical demonstration, which is an extension of the school voucher example introduced earlier in Chapter 1.

[3] As for the origin of the Wald estimator, it is customarily traced to Wald (1940) by authors such as Angrist and Krueger (1999). As we discuss later, the Wald estimator is not generally unbiased in a finite sample and instead is only consistent.

[4] The reasons may well be clear to the reader already. In moving from Equation (7.1) to Equation (7.2), covariation in Y and D within levels of Z was purged from the equation. But if D has a causal effect on Y that varies across all individuals, it may be that whatever contrast is identified through Z may not be relevant to all members of the population.

[5] There is one trivial way around this assumption. If the naive estimator of D on Y is consistent for the average causal effect of D on Y, then D is its own IV. The subset of the covariation in D and Y that is projected onto Z is then the full set of covariation in D and Y because Z is equal to D. In this case, no extrapolation is needed and the constant treatment effect assumption can be avoided. As we will discuss later, there are slightly less trivial ways to avoid the assumption as well. One alternative is to assert that δ is a constant among the treated and then stipulate instead that the IV identifies only the treatment effect for the treated. Although plausible, there are better ways to handle heterogeneity, as we will discuss as this chapter unfolds.

IV Demonstration 1

Suppose that a state education department wishes to determine whether private high schools outperform public high schools in a given metropolitan area, as measured by the achievement of ninth graders on a standardized test. For context, suppose that a school voucher program is operating in the city and that the state is considering whether to introduce the program in other areas in order to shift students out of public schools and into private schools.

To answer this policy question, the state department of education uses a census of the population in the metropolitan area to select a random sample of 10,000 ninth graders. They then give a standardized test to each sampled ninth grader at the end of the year, and they collect data as $\{y_i, d_i\}_{i=1}^{10,000}$, where Y is the score on the standardized test and D is equal to 1 for students who attend private high schools and equal to 0 for students who attend public high schools.

After the data are collected, suppose that the values of y_i are regressed on the values of d_i and that a predicted regression surface is obtained:

$$\hat{Y} = 50.0 + 9.67\,(D)\,. \tag{7.7}$$

The state officials recognize that private school students typically have more highly educated parents and therefore are more likely to have higher test scores no matter what curriculum and school culture they have been exposed to. Accordingly, they surmise that 9.67 is likely a poor causal effect estimate, or at least not one that they would want to defend in public.

The state officials therefore decide to merge the data with administrative records on the school voucher program in the area. For this program, all eighth graders in the city (both in public and private schools) are entered into a random lottery for \$3,000 school vouchers that are redeemable at a private high school. By mandate, 10 percent of all eligible students win the voucher lottery.

After merging the data, the state officials cross-tabulate eighth-grade lottery winners (where $z_i = 1$ for those who won the lottery and $z_i = 0$ for those who did not) by school sector attendance in the ninth grade, d_i. As shown in Table 7.1, 1000 of the 10,000 sampled students were voucher lottery winners.[6] Of these 1000 students, 200 were observed in private schools in the ninth grade. In comparison, of the 9000 sampled students who were not voucher lottery winners, only 1000 were observed in private schools.

The researchers conclude that the dummy variable Z for winning the voucher lottery is a valid IV for D because (1) the randomization of the lottery renders Z independent of ε in the population-level causal regression equation $Y = \alpha + \delta D + \varepsilon$ and (2) Z is associated with D, as shown for the sample data in Table 7.1. They therefore estimate $\hat{\delta}_{\text{IV,WALD}}$ in Equation (7.5) as

$$\frac{E_N[y_i|z_i = 1] - E_N[y_i|z_i = 0]}{E_N[d_i|z_i = 1] - E_N[d_i|z_i = 0]} = \frac{51.6 - 51.111}{.2 - .111} = 5.5, \tag{7.8}$$

[6] For simplicity, we have assumed that the sample percentage of lottery winners is the same as the population percentage. Of course, some random variation would be expected in any sample.

Table 7.1: The Distribution of Voucher Winners by School Sector for IV Demonstration 1

		Public school $d_i = 0$	Private school $d_i = 1$
Voucher loser	$z_i = 0$	8000	1000
Voucher winner	$z_i = 1$	800	200

and conclude that the true causal effect of private schooling on ninth grade achievement is 5.5 rather than 9.67. Operationally, the Wald estimator takes the average difference in test scores among those students who have won a voucher and those who have not won a voucher and divides that difference by a corresponding difference in the proportion of high school students who attend private schools among those who have won a voucher and the proportion of high school students who attend private schools among those who have not won a voucher. The numerator of Equation (7.8) is equal to $51.6 - 51.111$ by an assumed construction of the outcome Y, which we will present later when we repeat this demonstration in more detail (as IV Demonstration 2, beginning on page 204). The denominator, however, can be calculated directly from Table 7.1. In particular, $E_N[d_i = 1|z_i = 1] = 200/1000 = .2$ whereas $E_N[d_i = 1|z_i = 0] = 1000/9000 = .111$.

For this demonstration, Z is a valid instrument by traditional standards. It is randomly assigned to students in the population, and it predicts enrollment in private high schools. The resulting estimator yields a point estimate that can be given a causal interpretation by traditional standards. However, as we will show later when we reintroduce this demonstration as IV Demonstration 2, the particular causal effect that this IV identifies is quite a bit different when individual-level causal effect heterogeneity is present. It does not identify the average causal effect for all students. Instead, it identifies only the average causal effect for the subset of all students who would attend a private school if given a voucher but who would not attend a private school in the absence of a voucher. This means, for example, that the IV estimate is uninformative about the average causal effect among those who would enroll in private high schools in the absence of a voucher. This group of students represents the vast majority of private school students (in this example 1000/1200 or 83 percent).

The potential outcome literature has provided the insight that allows one to craft such a precise causal interpretation, and symmetrically to understand what an IV estimate does not inform. Before presenting that newer material, we return, as in Chapter 5 on regression, to a more complete accounting of the traditional IV literature.

7.2 Traditional IV Estimators

As detailed by Goldberger (1972), Bowden and Turkington (1984), and Heckman (2000), IV estimators were developed first in the 1920s by biologists and economists analyzing equilibrium price determination in market exchange (see E. Working 1927, H. Working 1925, and Wright 1921, 1925). After subsequent development in the 1930s and 1940s (e.g., Haavelmo 1943, Reiersøl 1941, Schultz 1938, and Wright 1934), IV estimators were brought into widespread use in economics by researchers associated with the Cowles commission (see Hood and Koopmans 1953, Koopmans and Reiersøl 1950).[7] The structural equation tradition in sociology shares similar origins to that of the IV literature (see Duncan 1975). The most familiar deployment of IV estimation in the extant sociological research is as the order condition for identification of a system of structural equations (see Bollen 1989, 1995, 1996a, 1996b, 2001; Fox 1984).

Consider the same basic ideas presented earlier for the Wald estimator in Equation (7.5). Again, recall the causal regression setup in Equation (7.1):

$$Y = \alpha + \delta D + \varepsilon. \tag{7.9}$$

The OLS estimator of the regression coefficient on D is

$$\hat{\delta}_{\text{OLS, bivariate}} \equiv \frac{\text{Cov}_N(y_i, d_i)}{\text{Var}_N(d_i)}, \tag{7.10}$$

where $\text{Cov}_N(.)$ and $\text{Var}_N(.)$ denote unbiased, sample-based estimates from a sample of size N of the population-level covariance and variance.

Now, again suppose that a correlation between D and ε renders the least squares estimator biased and inconsistent for δ in Equation (7.9). If least squares cannot be used to effectively estimate δ, an alternative IV estimator can be attempted, with an IV Z, as in

$$\hat{\delta}_{\text{IV}} \equiv \frac{\text{Cov}_N(y_i, z_i)}{\text{Cov}_N(d_i, z_i)}, \tag{7.11}$$

where Z can now take on more than two values. If the instrument Z is correlated with D but uncorrelated with ε, then the IV estimator in Equation (7.11) is consistent for δ in Equation (7.9).[8]

[7] The canonical example for this early development was the estimation of price determination in markets. For a market in equilibrium, only one price and one quantity of goods sold is observable at any point in time. To make prospective predictions about the potential effects of exogenous supply-and-demand shocks on prices and quantities for a new market equilibrium, the shape of latent supply-and-demand curves must be determined. To estimate points on such curves, separate variables are needed that uniquely index separate supply-and-demand shocks. Usually based on data from a set of exchangeable markets or alternative well-chosen equilibria for the same market from past time periods, these IVs are then used to identify different points of intersection between a shifted linear supply/demand curve and a fixed linear demand/supply curve.

[8] Notice that substituting d_i for z_i in Equation (7.11) results in the least squares regression estimator in Equation (7.10). Thus, the least squares regression estimator implicitly treats D as an instrument for itself.

One way to see why IV estimators yield consistent estimates is to again consider the population-level relationships between Y, D, and Z, as in Equations (7.1)–(7.4). Manipulating Equation (7.1) as before, one can write the covariance between the outcome Y and the instrument Z as

$$\text{Cov}(Y, Z) = \delta \, \text{Cov}(D, Z) + \text{Cov}(\varepsilon, Z), \tag{7.12}$$

again assuming that δ is a constant structural effect. Dividing by $\text{Cov}(D, Z)$ then yields

$$\frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)} = \frac{\delta \, \text{Cov}(D, Z) + \text{Cov}(\varepsilon, Z)}{\text{Cov}(D, Z)}, \tag{7.13}$$

which is directly analogous to Equation (7.3). When $\text{Cov}(\varepsilon, Z)$ is equal to 0 in the population, then the right-hand side of Equation (7.13) simplifies to δ. This suggests that

$$\frac{\text{Cov}_N(y_i, z_i)}{\text{Cov}_N(d_i, z_i)} \xrightarrow{p} \delta \tag{7.14}$$

if $\text{Cov}(\varepsilon, Z) = 0$ in the population and if $\text{Cov}(D, Z) \neq 0$. This would be the case for the causal diagram in panel (a) of Figure 7.1. But here the claim is more general and holds for cases in which Z is many valued [and, in fact, for cases in which D is many valued as well, assuming that the linear specification in Equation (7.9) is appropriate].

In the traditional econometrics literature, IVs are considered all-purpose remedies for least squares estimators that yield biased and inconsistent estimates, whether generated by obvious omitted variables or subtle patterns of self-selection. Consider the following examples of IV estimation in economics.

Three Examples of IV Estimation from Economics

For the effect of education on labor market earnings, as introduced earlier in Subsection 1.3.1, IVs have been used to estimate the causal effect of education. As reviewed by Card (1999) and by Angrist and Krueger (2001), the causal effect of years of schooling on subsequent earnings has been estimated with a variety of IVs, including proximity to college, regional and temporal variation in school construction, tuition at local colleges, temporal variation in the minimum school-leaving age, and quarter of birth. The argument here is that each of these variables predicts educational attainment but has no direct effect on earnings. For the quarter-of-birth instrument, Angrist and Krueger (1991:981-2) reason:

> If the fraction of students who desire to leave school before they reach the legal dropout age is constant across birthdays, a student's birthday should be expected to influence his or her ultimate educational attainment. This relationship would be expected because, in the absence of rolling admissions to school, students born in different months of the year start school at different ages. This fact, in

> conjunction with compulsory schooling laws, which require students to attend school until they reach a specified birthday, produces a correlation between date of birth and years of schooling. ... Students who are born early in the calendar year are typically older when they enter school than children born late in the year. ... Hence, if a fixed fraction of students is constrained by the compulsory attendance law, those born in the beginning of the year will have less schooling, on average, than those born near the end of the year.

As discussed in Subsection 1.3.2, much of the early debate on the effectiveness of private schooling relative to its alternatives was carried out with observational survey data on high school students from national samples of public and Catholic high schools. In attempts to resolve the endogeneity of school sector selection, Evans and Schwab (1995), Hoxby (1996), and Neal (1997) introduced plausible IVs for Catholic school attendance. Hoxby and Neal argued that the share of the local population that is Catholic is a valid IV for Catholic school attendance, maintaining that exogenous differences in population composition influence the likelihood of attending a Catholic school (by lowering the costs of opening such schools, which then lowers the tuition that schools need to charge and to which parents respond when making school sector selection decisions). Variation in the number of Catholics in each county was attributed to lagged effects from past immigration patterns and was therefore assumed to have no direct effect on learning.[9]

For another example, consider the literature on the effects of military service on subsequent labor market outcomes, which is a topic that both economists and sociologists have studied for several decades. Some have argued that military service can serve as an effective quasi-job-training program for young men likely to otherwise experience low earnings because of weak attachment to more traditional forms of education (Browning, Lopreato, and Poston 1973) or as a more general productivity signal that generates a veteran premium (De Tray 1982). In one of the earliest studies, which focused primarily on veterans who served around the time of the Korean War, Cutright (1974) concluded that:

> ...two years in the military environment, coupled with the extensive benefits awarded to veterans, do *not* result in a clear cut, large net positive effect of service. ... Therefore, one may question the likely utility of social programs that offer minority men and whites likely to have low earnings a career contingency in a bridging environment similar to that provided by military service. (Cutright 1974:326)

[9] Evans and Schwab (1995) use a student's religious identification as an IV. This is rather unconvincing, and it is easy to show with the same data that Catholic students in public schools outperform similar non-Catholic students. Thus, to many (see Altonji, Elder, and Taber 2005a, 2005b) the assumed exclusion restriction appears unreasonable. Neal (1997) also used this IV but only selectively in his analysis.

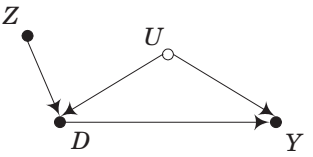


Figure 7.2: A DAG with an unblocked back-door path and a valid IV.

Following the Vietnam War, attention then focused on whether any military service premium had declined (see Rosen and Taubman 1982, Schwartz 1986; see also Berger and Hirsch 1983). Angrist (1990) used the randomization created by the draft lottery to estimate the Vietnam-era veteran effect. Here, the draft lottery turns date of birth into an IV, in much the same way that compulsory school entry and school-leaving laws turn date of birth into an IV for years of education. Angrist determined that veteran status had a negative effect on earnings, which he attributed to a loss of labor market experience.[10]

The IV Identifying Assumption Cannot be Tested with the Data

The basic assumption underlying the three studies just summarized – that Z has no causal effect on Y except indirectly through D – is a strong and untestable assumption. Some researchers believe mistakenly that this assumption is empirically testable in the following way. They believe that the assumption that Z has no direct effect on Y implies there is no association between Z and Y conditional on D. Thus, if an association between Z and Y is detected after conditioning on D, then the assumption must be incorrect.

Pearl's graphical approach shows clearly why individuals might believe that the identifying assumption can be tested in this way and also why it cannot. Consider the simplest possible causal DAG, presented in Figure 7.2, in which (1) an unblocked back-door path between D and Y exists because of the common cause U but (2) Z is a valid IV that identifies the causal effect of D on Y.[11] Here,

[10]We do not mean to imply that Angrist's work on this issue ended in 1990. Not only was his work very important for understanding what IV estimators accomplish in the presence of causal effect heterogeneity (see next section), he continued with subsequent substantive work on the military service effect. Angrist and Krueger (1994) estimated the World War II veteran effect, and, in an attempt to reconcile past estimates, concluded: "Empirical results using the 1960, 1970, and 1980 censuses support a conclusion that World War II veterans earn no more than comparable nonveterans, and may well earn less. These findings suggest that the answer to the question 'Why do World War II veterans earn more than nonveterans?' appears to be that World War II veterans would have earned more than nonveterans even had they not served in the military. Military service, in fact, may have reduced World War II veterans' earnings from what they otherwise would have been" (Angrist and Krueger 1994:92). Then, Angrist (1998) assessed the effects of voluntary service in the military in the 1980s, and he found mixed evidence.

[11]This DAG is consistent with the same basic diagram presented in panel (a) of Figure 7.1. We have replaced the bidirected edge from D to ε by (1) representing the association between D and ε as a back-door path between D and Y generated by a latent

the instrument Z is valid because it causes D and because it is unconditionally unassociated with U.

Why is the suggested test faulty? Again, the rationale for the test is that conditioning on D will block the indirect relationship between Z and Y through D. Accordingly, if the only association between Z and Y is indirect through D, then it is thought that there should be no association between Z and Y after conditioning on D. If such a net association is detected, then it may seem reasonable to conclude that the IV identifying assumption must be false.

Although this rationale feels convincing, it is incorrect. It is certainly true that if the IV assumption is invalid, then Z and Y will be associated after conditioning on D. But the converse is not true. Z and Y will always be associated after conditioning on D when the IV assumption is valid. The explanation follows from the fact that D in Figure 7.2 is a collider that is mutually caused by both Z and U. As shown earlier in Chapter 3 (see Subsection 3.1.3), conditioning on a collider variable creates dependence between the variables that cause it. Accordingly, conditioning on D in this DAG creates dependence between Z and U, even though the IV identifying assumption is valid. And, as a result, Z and Y will always be associated within at least one stratum of D even if the IV is valid. The faulty test yields an association between Z and Y when conditioning on D regardless of whether the IV identifying assumption is valid.[12]

7.3 Recognized Pitfalls of Traditional IV Estimation

The traditional IV literature suggests that, as long as there is an instrument that predicts the causal variable of interest but does not affect the outcome variable (except by way of the causal variable), then an IV estimator can be used to effectively estimate the causal effect. Even within this IV tradition, however, there are some recognized pitfalls of an IV estimation strategy. First, the assumption that an IV does not have a net direct effect on the outcome variable is often hard to defend. Second, even when an IV does not have a net direct effect on the outcome variable, IV estimators are biased in finite samples. Moreover, this bias can be substantial when an instrument only weakly predicts the causal variable. We discuss each of these weaknesses here.

Even "natural experiments" that generate compelling IVs are not immune from criticism. Consider date of birth as an instrument for veteran status in

common cause U and then (2) eliminating the remainder of the distribution of ε from the graph because it is unconditionally unassociated with Z and D. The latter is possible because the net variation in ε that is unrelated to U is irrelevant to the estimation of the causal effect of D on Y.

[12] For completeness, consider what the faulty test reveals when the IV assumption is invalid. Suppose that Figure 7.2 is augmented by an unobserved cause E and then two edges $E \rightarrow Z$ and $E \rightarrow U$. In this case, Z and Y would be associated within levels of D for two reasons: (1) conditioning on the collider D generates a net association between Z and U (and hence Z and Y) and (2) the common cause E of Z and U generates an unconditional association between Z and Y.

estimating the effect of military service in Vietnam on subsequent earnings (Angrist 1990). The case for excluding date of birth from the earnings equation that is the primary interest of the study is the randomization of the draft lottery, in which numbers were assigned by date of birth. Even though a type of randomization generates the IV, this does not necessarily imply that date of birth has no net direct effect on earnings. After the randomization occurs and the lottery outcomes are known, employers may behave differently with respect to individuals with different lottery numbers, investing more heavily in individuals who are less likely to be drafted. As a result, lottery number may be a direct, though probably weak, determinant of future earnings (see Heckman 1997, Moffitt 1996). IVs that are not generated by randomization are even more susceptible to causal narratives that challenge the assumption that the purported IV does not have a net direct effect on the outcome of interest.

Even in the absence of this complication, there are well-recognized statistical pitfalls. By using only a portion of the covariation in the causal variable and the outcome variable, IV estimators use only a portion of the information in the data. This represents a direct loss in statistical power, and as a result IV estimators tend to exhibit substantially more expected sampling variance than other estimators. By the criterion of mean-squared error, a consistent and asymptotically unbiased IV estimator is often outperformed by a biased and inconsistent regression estimator.

The problem can be especially acute in some cases. It has been shown that instruments that only weakly predict the causal variable of interest should be avoided entirely, even if they generate point estimates with acceptably small estimated standard errors (see Bound, Jaeger, and Baker 1995). In brief, the argument here is fourfold: (1) In finite samples, IV point estimates can always be computed because sample covariances are never exactly equal to zero; (2) as a result, an IV point estimate can be computed even for an instrument that is invalid because it does not predict the endogenous variable in the population [i.e., even if $\text{Cov}(D, Z) = 0$ in the population, rendering Equation (7.13) undefined because its denominator is equal to 0]; (3) at the same time, the formulas for calculating the standard errors of IV estimates fail in such situations, giving artificially small standard errors (when in fact the true standard error for the undefined parameter is infinity); and (4) the bias imparted by a small violation of the assumption that the IV affects the outcome variable only by way of the causal variable can explode if the instrument is weak.[13] To see this last result, consider Equation (7.6), which depicts the expected bias in the Wald estimator for a binary IV as the term

$$\frac{E[\varepsilon|Z = 1] - E[\varepsilon|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} \; . \tag{7.15}$$

When the identifying assumption is violated, the numerator of Equation (7.15) is nonzero because Z is associated with Y through ε. The bias is then an

[13] Complications (1), (2), (3), and (4) are all closely related. Situation (4) can be considered a less extreme version of the three-part predicament depicted in (1)–(3).

inverse function of the strength of the instrument; the weaker the instrument, the smaller the denominator and the larger the bias. If the denominator is close to zero, even a tiny violation of the identifying assumption can generate a large amount of bias. And, unfortunately, this relationship is independent of the sample size. Thus, even though a weak instrument may suggest a reasonable (or perhaps even intriguing) point estimate, and one with an acceptably small estimated standard error, the IV estimate may contain no genuine information whatsoever about the true causal effect of interest (see Hahn and Hausman 2003; Staiger and Stock 1997; Wooldridge 2002).[14]

Beyond these widely recognized pitfalls of standard IV estimation in economics, a third criticism is emerging of current practice. As noted by Angrist and Krueger (2001), the ways in which IVs are used has changed in the past 30 years. Because it has been hard to achieve consensus that particular no-net-direct-effect assumptions are credible, IVs that arise from naturally occurring variation have become more popular. Genuine "gifts of nature," such as variation in the weather and natural boundaries, have become the most prized sources of IVs (see Rosenzweig and Wolpin 2000 for a list of such instruments).

Not all economists see this shift in IV estimation techniques toward naturally occurring IVs as necessarily a step in the right direction. Rosenzweig and Wolpin (2000) offer one of the most cogent critiques (but see also Heckman 2000, 2005). They make two main points. First, the variation on which naturally occurring IVs capitalize is often poorly explained and/or does not reflect the variation that maintained theories posit should be important. As a result, IV estimates from natural experiments have a black-box character that lessens their appeal as informative estimates for the development of theory or policy guidance.

Second, the random variation created by a naturally occurring experiment does not necessarily ensure that an IV has no net direct effect on the outcome. Other causes of the outcome can respond to the natural event that generates the IV as well. Rosenzweig and Wolpin contend that natural experiments tend to lead analysts to ignore these issues because natural experiments appear falsely infallible.

We will explain these points further in this chapter, after introducing IV estimation in the presence of causal effect heterogeneity. (We will then revisit this critique in the next chapter on mechanisms as well.) The new IV literature, to be discussed next, addresses complications of the constant coefficient assumption implied by the stipulated constant value of δ in Equations (7.1) and (7.9). The issues are similar to those presented for regression estimators in Chapter 5, in that heterogeneity invalidates traditional causal inference from IV estimates. But IV estimators do identify specific narrow slices of average causal effects that

[14] There are no clear guidelines on how large an association between an IV and a treatment variable must be before analysis can proceed safely. Most everyone agrees that an IV is too weak if it does not yield a test statistic that rejects a null hypotheses of no association between Z and D. However, a main point of this literature is that the converse is not true. If a dataset is large enough, the small association between Z and D generated by a weak IV can still yield a test statistic that rejects a null hypotheses of no association. Even so, the problems in the main text are not vitiated, especially the explosion of the bias generated by a small violation of the identifying assumption.

may be of distinct interest, and as a result they represent a narrowly targeted estimation strategy with considerable appeal in some situations.

7.4 Instrumental Variable Estimators of Average Causal Effects

Following the adoption of a counterfactual perspective, a group of econometricians and statisticians has clarified considerably what IVs identify when individual-level causal effects are heterogeneous. In this section, we will emphasize the work that has developed the connections between traditional IV estimators and potential-outcome-defined treatment effects (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996; Imbens and Angrist 1994; Imbens and Rubin 1997). The key innovation here is the definition of a new treatment effect parameter: the local average treatment effect (LATE). We will also discuss other important work that has further clarified these issues, and some of this literature is more general than the LATE literature that we introduce first (see Heckman 1996, 1997, 2000; Heckman, Tobias, and Vytlacil 2003; Heckman, Urzua, and Vytlacil 2006; Heckman and Vytlacil 1999, 2000, 2005; Manski 2003; Vytlacil 2002).[15]

7.4.1 IV Estimation as LATE Estimation

Consider the following motivation of the Wald estimator in Equation (7.5), and recall the definition of Y presented in Equation (3.4):

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= Y^0 + \delta D \\ &= \mu^0 + \delta D + \upsilon^0, \end{aligned} \tag{7.16}$$

where $\mu^0 \equiv E[Y^0]$ and $\upsilon^0 \equiv Y^0 - E[Y^0]$. Note that δ is now defined as $Y^1 - Y^0$, unlike its structural representation in Equations (7.1) and (7.9) where δ was implicitly assumed to be constant across all individuals.

To understand when an IV estimator can be interpreted as an average causal effect estimator, Imbens and Angrist (1994) developed a counterfactual framework to classify individuals into those who respond positively to an instrument, those who remain unaffected by an instrument, and those who rebel against an instrument. Their innovation was to define potential treatment assignment variables, $D^{Z=z}$, for each state z of the instrument Z. When D and Z are binary variables, there are four possible groups of individuals in the population.[16] These can be summarized by a four-category latent variable $\tilde{C}$ for compliance

[15] Given the rapidity of these developments in the IV literature, some disagreement on the origins of these ideas pervades the literature. Heckman and Robb (1985, 1986) did provide extensive analysis of what IV estimators identify in the presence of heterogeneity. Subsequent work by others, as we discuss in this section, has clarified and extended these ideas, even while Heckman and his colleagues continued to refine their ideas.

[16] These four groups are considered principal strata in the framework of Frangakis and Rubin (2002).

status:

$$\begin{aligned}
\text{Compliers } (\tilde{C}=c): &\quad D^{Z=0}=0 \text{ and } D^{Z=1}=1,\\
\text{Defiers } (\tilde{C}=d): &\quad D^{Z=0}=1 \text{ and } D^{Z=1}=0,\\
\text{Always takers } (\tilde{C}=a): &\quad D^{Z=0}=1 \text{ and } D^{Z=1}=1,\\
\text{Never takers } (\tilde{C}=n): &\quad D^{Z=0}=0 \text{ and } D^{Z=1}=0.
\end{aligned}$$

Consider the private schooling example presented earlier in IV Demonstration 1. Students who would enroll in private schools only if offered the voucher are compliers ($\tilde{C}=c$). Students who would enroll in private schools only if not offered the voucher are defiers ($\tilde{C}=d$). Students who would always enroll in private schools, regardless of whether they are offered the voucher, are always takers ($\tilde{C}=a$). And, finally, students who would never enroll in private schools are never takers ($\tilde{C}=n$).

Analogous to the definition of the observed outcome, Y, the observed treatment indicator variable D can then be defined as

$$\begin{aligned}
D &= D^{Z=0}+(D^{Z=1}-D^{Z=0})Z \\
&= D^{Z=0}+\kappa Z,
\end{aligned} \tag{7.17}$$

where $\kappa \equiv D^{Z=1}-D^{Z=0}$.[17] The parameter κ in Equation (7.17) is the individual-level causal effect of the instrument on D, and it varies across individuals if $D^{Z=1}-D^{Z=0}$ varies across individuals (i.e., if observed treatment status varies as the instrument is switched from "off" to "on" for each individual). If the instrument represents encouragement to take the treatment, such as the randomly assigned school voucher in IV Demonstration 1, then κ can be interpreted as the individual-level compliance inducement effect of the instrument. Accordingly, $\kappa=1$ for compliers and $\kappa=-1$ for defiers. For always takers and never takers, $\kappa=0$ because none of them respond to the instrument.

Given these definitions of potential outcome variables and potential treatment variables, a valid instrument Z for the causal effect of D on Y must satisfy three assumptions in order to identify a LATE:

$$\text{independence assumption: } (Y^1,Y^0,D^{Z=1},D^{Z=0}) \perp\!\!\!\perp Z, \tag{7.18}$$

$$\text{nonzero effect of instrument assumption: } \kappa \neq 0 \text{ for all } i, \tag{7.19}$$

$$\text{monotonicity assumption: either } \kappa \geq 0 \text{ for all } i \text{ or } \kappa \leq 0 \text{ for all } i. \tag{7.20}$$

[17] Note also that D has now been counterfactually defined with reference to D^Z. Accordingly, the definition of Y in Equation (7.16) is conditional on the definition of D in Equation (7.17). Furthermore, although there is little benefit in doing so, it bears noting that this definition of D could be structured analogously to the definition of Y in Equation (7.16) so that the last line would become $D=\zeta+\kappa Z+\iota$, where $\zeta \equiv E[D^{Z=0}]$ and $\iota \equiv D^{Z=0}-E[D^{Z=0}]$. The parameter ζ would then be the expected probability of being in the treatment if all individuals were assigned to the state "instrument switched off," and ι would be the individual-level departure from this expected value, taking on values $1-E[D^{Z=0}]$ and $-E[D^{Z=0}]$ to balance the right-hand side of Equation (7.17) so that D is equal to either 1 or 0.

The independence assumption in Equation (7.18) is analogous to the assumption that $\text{Cov}(Z, \varepsilon) = 0$ in the traditional IV literature [see the earlier discussion of Equation (7.13)].[18] It stipulates that the instrument must be independent of the potential outcomes and potential treatments. Knowing the value of the instrument for individual i must not yield any information about the potential outcome of individual i under either treatment state. Moreover, knowing the realized value of the instrument for individual i must not yield any information about the probability of being in the treatment under alternative hypothetical values of the instrument. This latter point may well appear confusing, but it is exactly analogous to the independence assumption of potential outcomes from observed treatment status, discussed earlier for Equation (2.4). A valid instrument predicts observed treatment status (D), but it does not predict potential treatment status ($D^{Z=z}$).

The assumptions in Equations (7.19) and (7.20) are assumptions about individual responses to shifts in the instrument. The assumption of a nonzero effect of Z on D is a stipulation that the instrument must predict treatment assignment for at least some individuals. There must be at least some compliers or some defiers in the population of interest. The monotonicity assumption then further specifies that the effect of Z on D must be either weakly positive or weakly negative for all individuals i. Thus, there may be either defiers or compliers in the population but not both.[19]

If these three assumptions obtain, then an instrument Z identifies the LATE: the average causal effect of the treatment for the subset of the population whose treatment selection is induced by the instrument.[20] If $\kappa \geq 0$ for all i, then the Wald estimator from Equation (7.5) converges to a particular LATE:

$$\hat{\delta}_{\text{IV,WALD}} \xrightarrow{p} E[\delta | \tilde{C} = c], \tag{7.21}$$

which is equal to $E[Y^1 - Y^0 | D^{Z=1} = 1, D^{Z=0} = 0]$ and is therefore the average causal effect among compliers. In contrast, if $\kappa \leq 0$ for all i, then the Wald

[18] SUTVA must continue to hold, and now it must apply to potential treatments as well. In addition, as we noted earlier, our presentation here follows Imbens and Angrist (1994), Angrist and Imbens (1995), and Angrist et al. (1996). As we note later, IVs can be defined in slightly different (in some cases more general) ways. But for now, we restrict attention to the LATE literature, in which assumptions such as complete independence of the instrument are utilized.

[19] Manski defines this assumption as a MTS assumption in order to distinguish it from his MTR assumption (see Manski 1997; Manski and Pepper 2000). Vytlacil (2002) establishes its connections to the index structure model laid out in Heckman and Vytlacil (1999), as we discuss later. Heckman (2000), Heckman and Vytlacil (2005), and Heckman, Urzua, and Vytlacil (2006) provide a full accounting of the relationships between alternative IV estimators.

[20] The LATE is often referred to as the complier average causal effect to signify that the descriptor local is really defined by the restricted applicability of the LATE estimate to the average causal effect of compliers. The following example does not use the compliance language explicitly, except insofar as those who respond to the voucher are labeled compliers. The LATE literature that we cite provides the explicit connections between LATE and the large literature on noncompliance in randomized experiments (see in particular Imbens and Rubin 1997 and citations therein).

estimator from Equation (7.5) converges to the opposite LATE:

$$\hat{\delta}_{\text{IV,WALD}} \xrightarrow{p} E[\delta|\tilde{C} = d], \tag{7.22}$$

which is equal to $E[Y^1 - Y^0|D^{Z=1} = 0, D^{Z=0} = 1]$ and is therefore the average causal effect among defiers. In either case, the treatment effects of always takers and never takers are not informed in any way by the IV estimate.

In the next section, we will explain the claims in Equations (7.21) and (7.22) in considerable detail through a more elaborate version of IV Demonstration 1. But the basic intuition is straightforward. A valid IV is nothing more than an exogenous dimension across which the treatment and outcome variables are analyzed jointly. For a binary instrument, this dimension is a simple contrast, which is the ratio presented earlier [see Equations (7.5) and (7.8)]:

$$\frac{E_N[y_i|z_i = 1] - E_N[y_i|z_i = 0]}{E_N[d_i|z_i = 1] - E_N[d_i|z_i = 0]}.$$

The numerator is the naive estimate of the effect of Z on Y, and the denominator is the naive estimate of the effect of Z on D.

To give a causal interpretation to this ratio of differences across the third dimension indexed by Z, a model of individual treatment response must be specified and then used to interpret the causal effect estimate. The model of individual response adopted for a LATE analysis with a binary IV is the four-fold typology of compliance, captured by the latent variable $\tilde{C}$ defined earlier for always takers, never takers, compliers, and defiers. Within this model, the always takers and never takers do not respond to the instrument (i.e., they did not take part in the "experiment" created by the IV, and thus their treatment assignment is not determined by Z). This means that they are distributed in the same proportion within alternative values of the instrument Z. And, as a result, differences in the average value of Y, when examined across Z, are not a function of the outcomes of always takers and never takers.[21]

In contrast, defiers and compliers contribute all of the variation that generates the IV estimate because only their behavior is responsive to the instrument. For this reason, any differences in the average value of Y, when examined across Z, must result from treatment effects for those who move into and out of the causal states represented by D. If compliers are present but defiers are not, then the causal estimate offered by the ratio is interpretable as the average causal effect for compliers. If defiers are present but compliers are not, then the causal estimate offered by the ratio is interpretable as the average causal effect for defiers. If both compliers and defiers are present, then the estimate generated by the ratio does not have a well-defined causal interpretation. In the following demonstration, we will consider the most common case in the LATE literature

[21] In a sense, the outcomes of always takers and never takers represent a type of background noise that is ignored by the IV estimator. More precisely, always takers and never takers have a distribution of outcomes, but the distribution of these outcomes is balanced across the values of the instrument.

for which the monotonicity condition holds in the direction such that compliers exist in the population but defiers do not.

IV Demonstration 2

Recall IV Demonstration 1 in Section 7.1. In an attempt to determine whether private high schools outperform public high schools, a state education department assembles a dataset on a random sample of 10,000 ninth graders, $\{y_i, d_i, z_i\}_{i=1}^{10,000}$, where Y is a standardized test, and D is equal to 1 for students who attend private high schools and 0 for students who attend public high schools. Likewise, Z is equal to 1 for those who win a lottery for a \$3000 school voucher and 0 for those who do not.

As noted earlier for IV Demonstration 1, a bivariate regression of the values of y_i on d_i yielded a treatment effect estimate of 9.67 [see discussion of Equation (7.7)]. An alternative IV estimate with Z as an instrument for D yielded an estimate of 5.5 [see Equation (7.8)]. The hypothetical state officials relied on the IV estimate rather than on the regression estimate because they recognized that private school students have more advantaged social backgrounds. And they concluded that the randomization of the voucher lottery, in combination with the relationship between D and Z shown in Table 7.1, established the voucher lottery outcome as a valid IV.

We stated earlier that the estimate of 5.5 is properly interpreted as a particular LATE: the average causal effect for the subset of all students who would attend a private school if given a voucher but would not attend a private school in the absence of a voucher. To explain the reasoning behind this conclusion, we will first explain how the observed values for D and Y are generated as a consequence of variation in Z. Then we will introduce potential outcomes and use the treatment response model in order to explain why the IV estimate is interpretable as the average causal effect for compliers.

Recall that we reported the frequency distribution of D and Z in Table 7.1. The same information is presented again in Table 7.2, but now also as probability values (where, for example, the term $\Pr_N[.,.]$ in the upper left cell is equal to $\Pr_N[d_i = 0, z_i = 0]$ by plugging the row and column headings into the joint probability statement). We also now report the expectations of the outcome variable Y, conditional on D and Z.

First, consider how the least squares and IV estimates are calculated. The coefficient of 9.67 on D in Equation (7.7) is equal to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$. One can calculate these two conditional expectations from the elements of Table 7.2 by forming weighted averages within columns:

$$\begin{aligned} E_N[y_i|d_i = 1] &= \frac{.1}{.1 + .02}60 + \frac{.02}{.1 + .02}58 = 59.667, \\ E_N[y_i|d_i = 0] &= \frac{.8}{.8 + .08}50 + \frac{.08}{.8 + .08}50 = 50.0. \end{aligned}$$

Table 7.2: The Joint Probability Distribution and Conditional Expectations of the Test Score for Voucher Winner by School Sector for IV Demonstrations 1 and 2

		Public school $d_i = 0$	Private school $d_i = 1$
Voucher loser	$z_i = 0$	$N = 8000$ $\Pr_N[.,.] = .8$ $E_N[y_i \vert .,.] = 50$	$N = 1000$ $\Pr_N[.,.] = .1$ $E_N[y_i \vert .,.] = 60$
Voucher winner	$z_i = 1$	$N = 800$ $\Pr_N[.,.] = .08$ $E_N[y_i \vert .,.] = 50$	$N = 200$ $\Pr_N[.,.] = .02$ $E_N[y_i \vert .,.] = 58$

As shown earlier in Equation (7.8), the IV estimate of 5.5 is the ratio of two specific contrasts:

$$\frac{E_N[y_i|z_i = 1] - E_N[y_i|z_i = 0]}{E_N[d_i|z_i = 1] - E_N[d_i|z_i = 0]} = \frac{51.6 - 51.111}{.2 - .111} = 5.5. \tag{7.23}$$

Both of these contrasts are calculated within the rows of Table 7.2 rather than the columns. The contrast in the numerator is the naive estimate of the effect of Z on Y. It is calculated as the difference between

$$E_N[y_i|z_i = 1] = \frac{.08}{.08 + .02}50 + \frac{.02}{.08 + .02}58 = 51.6 \text{ and}$$
$$E_N[y_i|z_i = 0] = \frac{.8}{.8 + .1}50 + \frac{.1}{.8 + .1}60 = 51.111.$$

The contrast in the denominator is the naive estimate of the effect of Z on D. It is calculated as the difference between

$$E_N[d_i = 1|z_i = 1] = \frac{.02}{.08 + .02} = .2 \text{ and}$$
$$E_N[d_i = 1|z_i = 0] = \frac{.1}{.8 + .1} = .111.$$

Thus, calculating the IV estimate in this case is quite simple and does not require any consideration of the underlying potential outcomes or potential treatments.

But, to interpret the IV estimate when causal effect heterogeneity is present, the potential outcome and potential treatment framework are needed. Consider the three identifying assumptions in Equations (7.18)–(7.20). Because the voucher lottery is completely random, the voucher instrument Z is independent of the potential outcome and potential treatment variables. Also, as just shown,

Z predicts D, thereby sustaining the nonzero effect assumption. Thus, the first two assumptions are satisfied as explained earlier for IV Demonstration 1.

Only the monotonicity assumption in Equation (7.20) requires a new justification. Fortunately, there is no evidence in the school choice literature that students and their parents rebel against vouchers, changing their behavior to avoid private schooling only when offered a voucher.[22] Accordingly, it seems reasonable to assume that there are no defiers in this hypothetical population and hence that the monotonicity assumption obtains.

The joint implications of independence and monotonicity for the four groups of individuals in the cells of Table 7.2 should be clear. Monotonicity allows us to ignore defiers completely while independence ensures that the same distribution of never takers, always takers, and compliers is present among those who win the voucher lottery and those who do not. As a result, the proportion of always takers can be estimated consistently from the first row of Table 7.2 and the proportion of never takers can be estimated consistently from the second row of Table 7.2 as

$$\frac{\Pr_N[d_i = 1, z_i = 0]}{\Pr_N[z_i = 0]} \xrightarrow{p} \Pr[\tilde{C} = a], \tag{7.24}$$

$$\frac{\Pr_N[d_i = 0, z_i = 1]}{\Pr_N[z_i = 1]} \xrightarrow{p} \Pr[\tilde{C} = n] \ . \tag{7.25}$$

For this example, the proportion of always takers is $1000/9000 = .111$, and the proportion of never takers is $800/1000 = .8$. Because no defiers exist in the population with regard to this instrument, these two estimated proportions can be subtracted from 1 in order to obtain the proportion of compliers:

$$1 - \frac{\Pr_N[d_i = 1, z_i = 0]}{\Pr_N[z_i = 0]} - \frac{\Pr_N[d_i = 0, z_i = 1]}{\Pr_N[z_i = 1]} \xrightarrow{p} \Pr[\tilde{C} = c] \ . \tag{7.26}$$

For this example, the proportion of compliers in the population is $1 - .111 - .8 = .089$. Applying this distribution of always takers, never takers, and compliers (and a bit of rounding) to the frequencies from Table 7.2 yields the joint frequency distribution presented in Table 7.3.

For Table 7.3, notice the symmetry across rows that is generated by the independence of the instrument: $1000/7200 \approx 111/800$ (subject to rounding) and $800/9000 \approx 89/1000$ (again, subject to rounding). Of course, there is one major difference between the two rows: The compliers are in private schools among voucher winners but in public schools among voucher losers.

Before continuing, two important points should be noted. First, it is important to recognize that the calculations that give rise to the distribution of never takers, always takers, and compliers in Table 7.3 are not determined solely by the data. In a deeper sense, they are entailed by maintenance of the monotonicity

[22] Some parents might reason that private schooling is no longer as attractive if it is to be flooded with an army of voucher-funded children. Thus, defiers might emerge if the vouchers were widely distributed, and the monotonicity condition would then fail. In this case, however, the stable unit treatment value assumption that undergirds the entire counterfactual framework would likewise fail, necessitating deeper analysis in any case.

Table 7.3: The Distributions of Never Takers, Compliers, and Always Takers for IV Demonstration 2

		Public school $d_i = 0$	Private school $d_i = 1$
Voucher loser	$z_i = 0$	7200 Never takers 800 Compliers	1000 Always takers
Voucher winner	$z_i = 1$	800 Never takers	111 Always takers 89 Compliers

assumption. In the absence of that assumption, an unspecified number of defiers would be in the table as well, making the calculation of these proportions impossible.

Second, not all students in the dataset can be individually identified as always takers, never takers, or compliers. Consider the private school students for this example. Of these 1200 students, 1000 students are known to be always takers, as they have observed values of $d_i = 1$ and $z_i = 0$. The remaining 200 private school students are observationally equivalent, with observed values of $d_i = 1$ and $z_i = 1$. We know, based on the maintenance of the monotonicity and independence assumptions, that these 200 students include 111 always takers and 89 compliers. But it is impossible to determine which of these 200 students are among the 111 always takers and which are among the 89 compliers. The same pattern prevails for public school students. Here, we can definitively identify 800 students as never takers, but the 8000 public school students who are voucher losers cannot be definitively partitioned into the specific 7200 never takers and the 800 compliers.

Now, consider why the IV estimator yields 5.5 for this example, and then why 5.5 is interpretable as the average effect of private schooling for those who are induced to enroll in private schools because they have won vouchers. As noted already, because Z is independent of Y^1 and Y^0, the same proportion of always takers and never takers is present among both voucher winners and voucher losers. The difference in the expectation of Y across the two rows of Table 7.2 must arise from (1) the existence of compliers only in public schools in the first row and only in private schools in the second row and (2) the existence of a nonzero average causal effect for compliers.

To see this claim more formally, recall Equation (7.21), in which this particular LATE is defined. By the linearity of expectations and the definition of an individual-level causal effect as a linear difference between y_i^1 and y_i^0, the average causal effect among compliers, $E[\delta|\tilde{C} = c]$, is equal to the expectation $E[Y^1|\tilde{C} = c]$ minus the expectation $E[Y^0|\tilde{C} = c]$. To obtain a consistent estimate of the average causal effect for compliers, it is sufficient to obtain consistent estimates of $E[Y^1|\tilde{C} = c]$ and $E[Y^0|\tilde{C} = c]$ separately and then to subtract the latter from the former.

Fortunately, this strategy is feasible because the contribution of these two conditional expectations to the observed data can be written out in two equations and then solved. In particular, $E[Y^1|\tilde{C} = c]$ and $E[Y^0|\tilde{C} = c]$ contribute to the expectations of the observed outcome Y, conditional on D and Z, in the following two equations:

$$E[Y|D = 1, Z = 1] = \frac{\Pr[\tilde{C} = c]}{\Pr[\tilde{C} = c] + \Pr[\tilde{C} = a]} E[Y^1|\tilde{C} = c] + \frac{\Pr[\tilde{C} = a]}{\Pr[\tilde{C} = c] + \Pr[\tilde{C} = a]} E[Y^1|\tilde{C} = a], \tag{7.27}$$

$$E[Y|D = 0, Z = 0] = \frac{\Pr[\tilde{C} = c]}{\Pr[\tilde{C} = c] + \Pr[\tilde{C} = n]} E[Y^0|\tilde{C} = c] + \frac{\Pr[\tilde{C} = n]}{\Pr[\tilde{C} = c] + \Pr[\tilde{C} = n]} E[Y^0|\tilde{C} = n]. \tag{7.28}$$

These two equations are population-level decompositions of the conditional expectations for the observed data that correspond to the two cells of the diagonal of Table 7.2. These are the only two cells in which compliers are present, and thus the only two cells in which the observed data are affected by the outcomes of compliers.

How can we plug values into Equations (7.27) and (7.28) in order to solve for $E[Y^1|\tilde{C} = c]$ and $E[Y^0|\tilde{C} = c]$ and thereby obtain all of the ingredients of a consistent estimate of $E[\delta|\tilde{C} = c]$? We have already shown from applying the convergence assertions in Equations (7.24)–(7.26) that the terms $\Pr[\tilde{C} = c]$, $\Pr[\tilde{C} = a]$, and $\Pr[\tilde{C} = n]$ can be consistently estimated. And, in fact, these are given earlier for the example data as .089, .8, and .111. Thus, to solve these equations for $E[Y^1|\tilde{C} = c]$ and $E[Y^0|\tilde{C} = c]$, the only remaining pieces that need to be estimated are $E[Y^1|\tilde{C} = a]$ and $E[Y^0|\tilde{C} = n]$, which are the average outcome under the treatment for the always takers and the average outcome under the control for the never takers. Fortunately, the independence and monotonicity assumptions guarantee that voucher losers in private schools represent a random sample of always takers. Thus, $E[Y^1|\tilde{C} = a]$ is estimated consistently by $E_N[y_i|d_i = 1, z_i = 0]$, which is 60 for this example (see the upper right-hand cell in Table 7.2). Similarly, because voucher winners in public schools represent a random sample of never takers, $E[Y^0|\tilde{C} = n]$ is estimated consistently by $E_N[y_i|d_i = 0, z_i = 1]$, which is equal to 50 for this example (see the lower left-hand cell in Table 7.2). Plugging all of these values into Equations (7.27) and (7.28) then yields

$$58 = \frac{.089}{.089 + .8} E[Y^1|\tilde{C} = c] + \frac{.8}{.089 + .8} 60, \tag{7.29}$$

$$50 = \frac{.089}{.089 + .111} E[Y^0|\tilde{C} = c] + \frac{.111}{.089 + .111} 50. \tag{7.30}$$

Solving Equation (7.29) for $E[Y^1|\tilde{C} = c]$ results in 55.5 whereas solving Equation (7.30) for $E[Y^0|\tilde{C} = c]$ results in 50. The difference between these values is 5.5, which is the average causal effect for the subset of all students who would attend a private school if given a voucher but would not attend a private school in the absence of a voucher.[23] The value of 5.5 yields no information whatsoever about the effect of private schooling for the always takers and the never takers.[24]

Of course, the Wald estimate is also 5.5, as shown in Equations (7.8) and (7.23). And thus, in one sense, the Wald estimator can be thought of as a quick alternative method for calculating all of the steps just presented to solve exactly for $E[\delta|\tilde{C} = c]$. Even so, this correspondence does not explain when and how the Wald estimator can be interpreted as the average causal effect for compliers. For this example, the correspondence arises precisely because we have assumed that there are no defiers in the population, based on the substance of the application and the treatment response model that we adopted. As a result, the Wald estimate can be interpreted as a consistent estimate of the average effect of private schooling for those who comply with the instrument because it is equal to that value under the assumed model of treatment response we are willing to adopt.[25]

LATE estimators have been criticized because the identified effect is defined by the instrument under consideration. As a result, different instruments define different average treatment effects for the same group of treated individuals. And, when this is possible, the meanings of the labels for the latent compliance variable $\tilde{C}$ depend on the instrument, such that some individuals can be never takers for one instrument and compliers for another.

Although from one perspective this is a weakness, from another it is the most attractive feature of LATE estimation. For IV Demonstration 2, the IV estimate

[23] For completeness, consider how the naive estimate and the LATE would differ if all remained the same except the stipulated value of 50 for $E_N[d_i = 0, z_i = 0]$ in Table 7.2. If $E_N[d_i = 0, z_i = 0]$ were instead 50.25, then the naive estimate would be 9.44 and the LATE estimate would be 3.00. And, if $E_N[d_i = 0, z_i = 0]$ were instead 50.5, then the naive estimate would be 9.21 and the LATE estimate would be .5. Thus, for the example in the main text, compliers on average do no worse in public schools than never takers. But, for these two variants of the example, compliers on average do slightly better in public schools than never takers. As a result, the calculations in Equation (7.29) remain the same, but the values of Equation (7.30) change such that $E[Y^0|\tilde{C} = c]$ is equal to 52.5 and 55.0, respectively. The LATE estimate is therefore smaller in both cases because the performance of compliers in public schools is higher (whereas the performance of compliers in private schools remains the same).

[24] We can estimate $E[Y^1|\tilde{C} = a]$ and $E[Y^0|\tilde{C} = n]$ consistently with $E_N[y_i|d_i = 1, z_i = 0]$ and $E_N[y_i|d_i = 0, z_i = 1]$. But we have no way to effectively estimate their counterfactual analogs: the mean outcome in public schools for always takers and the mean outcome in privates schools for never takers.

[25] In other words, the Wald estimate of 5.5 is also a quick method for calculating an entirely different causal effect under a different set of assumptions. If monotonicity cannot be defended, then the IV estimate can be given a traditional structural interpretation, under the assumption that the causal effect is constant for all individuals. In this sense, because the Wald estimate has more than one possible causal interpretation, merely understanding how it is calculated does not furnish an explanation for how it can be interpreted.

does not provide any information about the average effect for individuals who would attend private schooling anyway (i.e., the always takers) or those who would still not attend the private schools if given a voucher (i.e., the never takers). Instead, the IV estimate is an estimate of a narrowly defined average effect only among those induced to take the treatment by the voucher policy intervention. But, for IV Demonstration 2, this is precisely what should be of interest to the state officials. If the policy question is "What is the effect of vouchers on school performance?" then they presumably care most about the average effect for compliers.

The limited power of the LATE interpretation of an IV estimate is thus, in some contexts, beneficial because of its targeted clarity. Moreover, when supplemented by a range of additional IV estimates (i.e., different voucher sizes, and so on), complementary LATE estimates may collectively represent an extremely useful set of parameters that describe variation in the causal effect of interest for different groups of individuals exposed to the cause for alternative (but related) reasons. Before summarizing the marginal treatment effect literature that more completely specifies the interrelationships among all types of average causal effect estimators, we first lay out the implications of the LATE perspective for traditional IV estimation.

7.4.2 Implications of the LATE Perspective for Traditional IV Estimation

The LATE literature specifies a set of assumptions under which it is permissible to give IV estimates an average causal effect interpretation using the counterfactual model of causality. In this sense, the new framework is mostly a set of guidelines for how to interpret IV estimates. As such, the LATE perspective has direct implications for traditional IV estimation, as introduced earlier in this chapter.

Monotonicity and Assumptions of Homogeneous Response

An important implication of the new LATE framework is that many conventional IV estimates lack a justifiable average causal effect interpretation if the IV does not satisfy a monotonicity condition. In the presence of causal effect heterogeneity, a conventional IV estimator yields a parameter estimate that has no clear interpretation, as it is an unknown and unidentifiable mixture of the treatment effects of compliers and defiers.

For IV Demonstration 2, we showed that the estimate of 5.5 is applicable to students whose families would change their child's enrollment choice from a public school to a private school for a $3000 voucher. Can an assumption be introduced that allows the estimate of 5.5 to be be interpreted as informative about other students who do (or who would) attend private schools?

Two variants of the same homogeneity assumption allow for such extrapolated inference: the assumption that the causal effect is a structural effect that is (a) constant across all members of the population or (b) constant across all

members of the population who typically take the treatment.[26] In its stronger form (a), the assumption simply asserts that the causal effect estimate is equally valid for all members of the population, regardless of whether or not the group of students whose enrollment status would change in response to the voucher is representative of the population of students as a whole. In its weaker form (b), the assumption pushes the assumed constancy of the effect only half as far, stipulating that the IV estimate is valid as an estimate of the treatment effect for the treated only. For IV Demonstration 2, the weaker variant of the homogeneity assumption is equivalent to asserting that the IV estimate provides information only about the achievement gains obtained by private school students.

Although weaker, this second homogeneity assumption is still quite strong, in that all individuals in private schools are considered homogeneous. In examples such as IV Demonstration 2, it is clear that there are two distinct groups within the treated: always takers and compliers. And there is little reason to expect that both groups respond in exactly the same way to private schooling. Thus, for examples such as this one, Manski (1995:44) argues that this homogeneity assumption "strains credibility" because there is almost certainly patterned heterogeneity in the effect among treated individuals.

One traditional way to bolster a homogeneity assumption is to condition on variables in a vector X that can account for all such heterogeneity and then assert a conditional homogeneity of response assumption. To do so, the Wald estimator must be abandoned in favor of a two-stage least squares (2SLS) estimator. As shown in any econometrics textbook (e.g., Greene 2000; Wooldridge 2002), the endogenous regressors D and X are embedded in a more all-encompassing $\mathbf{X}$ matrix, which is $n \times k$, where n is the number of respondents and k is the number of variables in X plus 2 (one for the constant and one for the treatment variable D). Then, a matrix $\mathbf{Z}$ is constructed that is equivalent to X, except that the column in $\mathbf{X}$ that includes the treatment variable D is replaced with its instrument Z. The 2SLS estimator is then

$$\hat{\delta}_{\text{IV,2SLS}} \equiv (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}, \tag{7.31}$$

where $\mathbf{y}$ is an $n \times 1$ vector containing the outcomes y_i. The strategy is to condition out all of the systematic variability in the observed response and then to simultaneously use the instrument Z to identify a pure net structural effect.

Is this strategy really a feasible solution? Probably not. If good measures of all of the necessary variables in X are available, a simple OLS estimator probably would have been feasible in the first place. Rarely would all possible necessary variables be available, save a single variable that has a net additive constant effect on the outcome. For IV Demonstration 2, a variable that measures "family responsiveness to incentives" would be needed, and it would have to flawlessly predict which families respond to $3000 vouchers and which families do not.

[26] This assumption is known as a constant-coefficient assumption, a homogeneous response assumption, or a shifted outcome assumption (see Manski 1995, 2003).

Other Difficulties of Interpretation

IV estimates are hard, if not impossible, to interpret as LATE estimates when the instrument measures something other than an incentive to which individuals can consciously respond by complying or defying. Instruments based on exogenous field variation (as championed in Angrist and Krueger 2001 but criticized in Rosenzweig and Wolpin 2000) can be particularly hard to interpret, because the shifts in costs and benefits that the natural variation is supposed to induce generally remain unspecified, thereby weakening a main link in the narrative that explains why some individuals take the treatment in response to the instrument.

Moreover, if two IVs are available, then the traditional econometric literature suggests that they should both be used to overidentify the model and obtain a more precise treatment effect estimate by the 2SLS estimator in Equation (7.31). Overidentified models, in which more than one instrument is used to identify the same treatment effect, generate a mixture-of-LATEs problem.

Consider the Catholic school example discussed earlier. Suppose that two IVs are used: the share of the county that identifies as Catholic and a student's religious identification (as in Neal 1997). Even if these potential IVs have no net direct effects on test scores (and further that the weak instrument problem discussed earlier is not applicable), can a theoretically meaningful LATE interpretation be given to the effect that the two instruments in this example jointly identify? For the religious identification instrument, the suggested LATE interpretation is the average effect of Catholic schooling among those who are more likely to attend Catholic schools because they are Catholic. However, when overidentified with the IV represented by the share of the local population that is Catholic, this suggested LATE interpretation is inextricably mixed with another LATE: the average effect of Catholic schooling among those who are more likely to attend Catholic schools because of the small difference in tuition that a high proportion of Catholics in the local population tends to generate. And thus the estimated causal effect would be averaged across two very different groups of hypothetical individuals, both of which likely deserve separate attention.[27]

In sum, if causal effect heterogeneity is present, then a constant-coefficient interpretation of an IV estimate is implausible. However, if the instrument satisfies a monotonicity condition and can be conceptualized as a relatively proximate inducement to take the treatment, then IV estimates can be given LATE interpretations. These fine-grained interpretations can be very illuminating about particular groups of individuals, even though they may provide no information whatsoever about other groups of individuals in the population (including other individuals who typically choose to take the treatment). Thus,

[27]There is also the possibility that, contra the justification of Hoxby (1996), the individuals who are induced to attend Catholic schooling because a greater share of the population is Catholic are not at all the same as those who are supposedly at the margin of a tuition-based cost calculation. This is one basic criticism that Rosenzweig and Wolpin (2000) level against all such naturally occurring IVs – there is often no evidence that the IV is inducing a group of individuals to take the treatment, or if it is, that it is a group of individuals whose outcomes are not particularly representative of the population of interest.

the limited nature of IV estimators when interpreted as LATE estimators shows both the potential strengths and weaknesses of IV estimation in general.

7.5 Two Additional Perspectives on the Identification of Causal Effects with IVs

In this section, we discuss two additional perspectives on the identifying power of IVs. The first shows how a perfect instrument can help to identify subtle patterns of causal effect heterogeneity. The second perspective shows that a putative IV that does not satisfy a traditional assumption of no-direct-causation nonetheless still retains some identifying power if the violation of that assumption is monotonic.

7.5.1 Local IVs and Marginal Treatment Effects

Heckman and Vytlacil (1999, 2005), building upon Heckman (1997), have shown that LATEs and many other average treatment effects can be seen as weighted averages of more fundamental marginal treatment effects.[28] Although the generality of their perspective is captivating, and the methodological content of the perspective unifies many strands of the literature in causal effect estimation, we summarize it only briefly here because the demands on data are quite substantial. The all-powerful IV that is needed to estimate a full schedule of marginal treatment effects will rarely be available to researchers.

The marginal treatment effect (MTE) perspective can be easily grasped with only a slight modification to the set up of IV Demonstration 2. Instead of 10 percent of students receiving a voucher that is exactly equal to $3000, suppose instead that these 10 percent of students receive a voucher that is a random draw from a uniform distribution with a minimum of $1 and a maximum equal to the tuition charged by the most expensive private school in the area.

For Heckman and Vytlacil, the size of each student's voucher is a valid instrument Z, maintaining the same assumptions as we did for IV Demonstration 2 (i.e., Z is randomly assigned, Z has a nonzero effect on D, and the effect of Z on D is monotonic). The monotonicity assumption is a little more complex than before, but it stipulates that the true probability of taking the treatment is higher for all individuals with values of Z equal to z'' rather than z' if $z'' > z'$. This fits cleanly into the notation introduced earlier by simply allowing Z to be many valued.

Heckman and Vytlacil then define two related concepts: a local instrumental variable (LIV) and an MTE. An LIV is the limiting case of a component binary IV drawn from Z in which z'' approaches z' for any two values of Z such that $z'' > z'$. Each LIV then defines a marginal treatment effect, which is the limiting form of a LATE, in which the IV is an LIV.

[28] See also Heckman, Urzua, and Vytlacil (2006).

Consider the more elaborate version of IV Demonstration 2 just introduced here. One could form LIVs from Z by stratifying the data by the values of Z and then considering adjacent strata. Given a large-enough sample for a large-enough voucher program, LIVs could be constructed for each dollar increase in the voucher. Each LIV could then be used to estimate a LATE, and these LIV-identified LATEs could then be considered MTEs.

Heckman and Vytlacil (2005) show that most average causal effect estimates can be represented as weighted averages of MTEs, identified by LIVs. But the weighting schemes differ based on the parameter of interest, some of which, as was the case in our regression chapter, may have no inherent interest. Heckman and Vytlacil therefore propose a more general strategy. They argue that researchers should define the policy-relevant treatment effect (PRTE) based on an assessment of how a contemplated policy would affect treatment selection. Then, MTEs should be estimated with LIVs and weighted appropriately to obtain the PRTE that is of primary interest.

There is much to recommend in this approach, and in fact it should not be considered an approach relevant only to policy research. The approach is quite easily extended to targeted theory-relevant causal effects, for which one wishes to weight marginal causal effects according to a foundational theoretical model. But, in spite of this appeal, the entire approach may well come to represent a gold standard for what ought to be done rather than what actually can be done in practice. If it is generally recognized that IVs satisfying the LATE assumptions are hard to find, then those that satisfy LIV assumptions must be harder still. In this regard, the next approach has an advantage.

7.5.2 Monotone IVs and Analyses of Bounds

Whereas the recent work of Heckman and his colleagues demonstrates how powerful the conclusions of a study can be if a perfect and finely articulated IV is available, Manski's recent work probes the opposite territory, following on the consideration in his early work of the capacity of traditional IVs to narrow the no-assumptions bound (see Manski 1994, 1995). Manski and Pepper (2000) investigate what can be learned about the average causal effect when both standard and weaker IV assumptions are maintained. Their results are then presented as a component of the general partial identification methodology laid out in Manski (2003; see especially Chapter 9).

Manski and Pepper (2000) first define the traditional IV assumption in terms of mean independence, after conditioning on a nuisance covariate X for generality. In our notation, the standard IV assumption of no direct effect of Z on Y (by either Y^1 or Y^0) is written out as

$$E[Y^1|X, Z = z^{'}] = E[Y^1|X, Z = z''], \tag{7.32}$$

$$E[Y^0|X, Z = z'] = E[Y^0|X, Z = z''], \tag{7.33}$$

for any two values z' and z'' of Z (and separately for strata defined by X).

In other words, Equations (7.32) and (7.33) require that the expectations of the potential outcomes be equal within strata defined by Z (conditional on X). Accordingly, the bounds analysis discussed in Section 6.3 then applies within each stratum of Z, and the bound on the average causal effect can then be defined as the intersection of the bounds across the strata defined by Z. This result implies that the no-assumptions bound can only be narrowed by an IV if the no-assumptions bounds differ across strata defined by Z.

Manski and Pepper (2000) then consider a weaker assumption for an IV analysis, known as a monotone IV (MIV) assumption. It states that for all values of the instrument Z, in which $z'' \geq z'$, the variable Z is an MIV if

$$E[Y^1|X, Z = z''] \geq E[Y^1|X, Z = z'], \tag{7.34}$$

$$E[Y^0|X, Z = z''] \geq E[Y^0|X, Z = z']. \tag{7.35}$$

In Equations (7.34) and (7.35), the expected values of both potential outcomes are weakly increasing in Z. Note that this usage of the concept of monotonicity is quite different than for the LATE analysis described earlier and for the LIV analysis of the last subsection. Here, monotonicity refers to the relationship between the instrument and the potential outcomes, not the relationship between the treatment and the instrument. The latter type of assumption is referred to as monotone treatment selection (MTS) by Manski and his colleagues (see, again, our earlier discussion in Section 6.3).

To what extent does an MIV narrow the bound for the average causal effect [or, in the words of Manski (2003), shrink the identification region for it]? The short answer: No more than a traditional IV (and usually considerable less), but still enough to have some identifying power.

It is easier to explain how an MIV bounds the expectation of each potential outcome than it is to demonstrate directly how an MIV bounds an average treatment effect that is a function of these expectations. Consider just the determination of the upper bound for $E[Y^1]$.

Under the standard IV assumption of mean independence, the upper bound for $E[Y^1]$ is equal to the smallest upper bound across the different subpopulations defined by the instrument Z. More precisely, the upper bound is the smallest value that $E[Y^1|Z = z]$ takes on across all values z of Z.

In contrast, under the weaker MIV assumption, the upper bound for $E[Y^1]$ is a weighted average of subpopulation upper bounds, for which each subpopulation upper bound is defined across the values of Z. The implicit calculation of this upper bound on $E[Y^1]$ can be described in a series of as-if algorithmic steps. First, a value of Z is selected as z'. The upper bound for $E[Y^1]$, with respect to z', is set as the smallest value that $E[Y^1|Z = z]$ takes on across all values z'' of Z where $z'' \geq z'$.[29] After this smallest upper bound is found with z' fixed, the

[29] Because this smallest upper bound relative to z' is selected from a smaller subset of possible upper bounds (only those greater than z' rather than all values of Z), the resulting weighted upper bound across z' is by definition no smaller (and usually larger) than would be the upper bound suggested by a more stringent mean-independence IV assumption. As a result, MIVs never shrink the identification region more than traditional IVs (and usually considerable less).

next value of z' is selected, and a smallest upper bound is found with respect to this next z'. After all values of Z have been selected as z', the upper bound for $E[Y^1]$ is set as the weighted average of the subpopulation upper bounds with respect to each value of Z, where the weights are the marginal distribution of Z attached pointwise to the smallest upper bounds set with respect to all z' of Z.

The determination of the lower bound on $E[Y^1]$ under MIV is the opposite of this procedure, in the sense that the greatest lower bound is first sought across subpopulations of Z, restricting the range of Z over which one searches in each step to be larger than the selected anchoring point z' of Z. To then find the bounds implied by an MIV for the average causal effect, the bounds on $E[Y^0]$ must be calculated with the same basic procedure. These bounds can then be substituted into the same basic framework introduced in Section 6.3, with the marginal distribution of D used to characterize the known distribution across treatment states.

Of course, as we noted in Chapter 6, the goal of a partial identification analysis is to invoke combinations of weak assumptions.[30] In that larger context, Manski and Pepper (2000) note that MTS is equivalent to an MIV assumption where the treatment variable is the MIV. The MIV approach to IV estimation is at least as convincing of a method for practice as the more general partial identification approach summarized in Chapter 6. Again, it is generally difficult to eliminate values of 0 for treatment effects in this tradition, but the methodology can be used, as Manski shows, to convincingly reject extreme causal effect assertions and build regions of credible inference to move the literature forward in any given area. Moreover, it is presumably much easier to find MIVs in any substantive area than IVs that allow for the identification of LATEs and or full schedules of MTEs.

7.6 Conclusions

The impressive development of the IV literature in econometrics and statistics in the past decade suggests a variety of recommendations for practice that differ from those found in the older IV literature. First, weak instruments yield estimates that are especially susceptible to finite sample bias. Consequently, natural experiments should be avoided if the implied IVs only weakly predict the causal variable of interest. No matter how seductive their claims to satisfy identification assumptions may be, resulting point estimates and standard errors may be very misleading.

[30] In a prior version of their manuscript, Manski and Pepper use the MTR, MTS, and MIV assumptions together to determine the bounds on the effect of education on the logged wages of respondents for the National Longitudinal Survey of Youth. When they invoked MTR and MTS assumptions, they found that the bound for the effect of a twelfth year of schooling was $[0, .199]$, that the bound for the effect of a fifteenth year of schooling was $[0, .255]$, and that the bound for the effect of a sixteenth year of schooling was $[0, .256]$. When they then used the Armed Forces Qualify Test as an MIV while still maintaining the MTR and MTS assumptions, they obtained narrower bounds respectively of $[0, .126]$, $[0, .162]$, and $[0, .167]$.

Second, if causal effect heterogeneity is present, a potential IV should be used only if it satisfies a monotonicity condition. IV estimates should then be interpreted as LATE estimates defined by the instrument.

Third, if causal effect heterogeneity is present, IVs should not be combined in a 2SLS model (except in the rare cases in which measures of all of the variables that account for the causal effect heterogeneity are available). Instead, IV estimates should be offered for those IVs that satisfy monotonicity conditions. These alternative estimates should then be interpreted as LATE estimates and reconciled with each other based on a narrative about why the causal effect varies for different units of analysis who are exposed to the cause (and/or different levels of the cause) for different reasons.

Finally, IVs should be used to examine general patterns of causal effect heterogeneity. Using IVs to estimate only the average treatment effect for the treated is too narrow of a purpose, as there is likely a good deal of variation in the treatment effect that is amenable to analysis with complementary IVs. The possibilities for this type of analysis are most clearly developed in the new literature on the identification of marginal treatment effects using local IVs.

Properly handled, there is much to recommend in the IV estimation strategy for causal analysis. But, of course, IVs may not be available. We now turn to other techniques that may allow for the identification and estimation of a causal effect when a complete model of causal exposure cannot be formulated because selection is determined by relevant unobserved variables.

Chapter 8

Mechanisms and Causal Explanation

Social scientists have recognized for decades that adequate explanations for how causes bring about their effects must, at some level, specify in empirically verifiable ways the causal pathways between causes and their outcomes. This requirement of depth of causal explanation applies to the counterfactual tradition as well. Accordingly, it is widely recognized that a consistent estimate of a counterfactually defined causal effect of D on Y may not qualify as a sufficiently deep causal account of how D effects Y, based on the standards that prevail in a particular field of study.

In this chapter, we first discuss the dangers of insufficiently deep explanations of causal effects, reconsidering the weak explanatory power of some of the natural experiments discussed already in Chapter 7. We then consider the older literature on intervening variables in the social sciences as a way to introduce the mechanism-based estimation strategy proposed by Pearl (2000). In some respects, Pearl's approach is completely new, as it shows in a novel and sophisticated way how causal mechanisms can be used to identify causal effects even when unblocked back-door paths between a causal variable and an outcome variable are present. In other respects, however, Pearl's approach is refreshingly familiar, as it helps to clarify the appropriate usage of intervening and mediating variables when attempting to deepen the explanation of a causal claim.

Independent of Pearl's important work, a diverse group of social scientists has appealed recently for the importance of mechanisms to all explanation in social science research. Although some of these appeals are not inconsistent with the basic counterfactual approach (e.g., Reskin 2003; Sørensen 1998), some of the more extended appeals (e.g., Goldthorpe 2000) claim to be at odds with some of the basic premises of the counterfactual model. We will argue instead that there is no incompatibility between causal mechanisms and counterfactual thinking. Finally, we draw on Machamer, Darden, and Craver (2000) and introduce their concept of a mechanism sketch as well as the process of bottoming out

in mechanistic explanation. This terminology helps to frame our final discussion of how mechanisms can be used to sustain and deepen causal explanation, which draws together Pearl's front-door criterion with standards for sufficient causal depth.

8.1 The Dangers of Insufficiently Deep Explanations

Before considering how mechanisms can be used to identify causal effects, we first discuss the importance of explanatory depth in counterfactual causal analysis. To do so, we return to the important critical work of Rosenzweig and Wolpin (2000) on the limited appeal of many natural experiments.

As discussed in Chapter 7, the natural experiment literature in economics uses naturally occurring forms of randomness as IVs in order to identify and then estimate causal effects of long-standing interest. Initially, this literature was heralded as the arrival of a new age of econometric analysis, in which it now appeared possible to consistently estimate some of the causal effects of greatest interest to economists, such as the effect of human capital investments in education on earnings. Looking back on these bold claims, Rosenzweig and Wolpin (2000:829–30) conclude, "The impression left by this literature is that if one accepts that the instruments are perfectly random and plausibly affect the variable whose effect is of interest, then the instrumental-variables estimates are conclusive." They then argue that these estimates are far from conclusive and, in fact, are far more shallow than was initially recognized.

To somewhat overstate the case, the initial overconfidence of the natural experiment movement was based on the mistaken belief that the randomness of a natural experiment allows one to offer valid causal inference in the absence of any explicit theory. One might say that some econometricians had been seduced by a position implicit in some writing in statistics: We do not need explicit theories in order to perform data analysis. Rosenzweig and Wolpin (2000) counter this position by arguing that the theory that underlies any model specification is critical to the interpretation of an estimate of a causal effect, and almost all examples of estimation by natural experiments have model specifications that make implicit theoretical claims.

Here, we provide an informal presentation of two of the examples analyzed by Rosenzweig and Wolpin – the effect of education on earnings and the effect of military service on earnings – each of which was already introduced and discussed briefly in Chapter 7. We will use causal diagrams here in order to demonstrate the issues involved.

Angrist and Krueger (1991, 1992) address the ability-bias issue in the estimation of the causal effect of schooling on subsequent labor market earnings. They assert that the quarter in which one is born is random but nonetheless predicts one's level of education because of compulsory school entry and dropout laws (see the quotation in Section 7.2 that gives the rationale). Angrist and

Krueger's estimates of the increase in log earnings for each year of education fall between .072 and .102, values that are consistent with those found by others using different methods (see Card 1999).

As discussed already in Chapter 7, the first limitation of their results is that their IV estimates apply to only a narrow segment of the population: those individuals whose schooling would have changed if their birthdate was in a different quarter. Again, this is a LATE estimate that applies to those individuals whose years of schooling are responsive to school entry and dropout laws. No one maintains that this narrow subpopulation is simply a random sample of all individuals in the population of interest, and most would argue that Angrist and Krueger estimated the returns to schooling only for disadvantaged youth prone to dropping out of high school for other reasons. There are, however, complications of how to interpret their estimates even for this subpopulation.

Consider the causal diagram presented in Figure 8.1, which is based loosely on the critique offered by Rosenzweig and Wolpin (2000) and is a summary of the debate that unfolded in the years following the publication of Angrist and Krueger (1991). The causal diagram is not meant to represent a full causal account of how education determines wages, but rather only the casual diagram that is applicable to compliers. As discussed in Chapter 7, for this example, compliers are those members of the population whose education is (or would be) responsive to a switch in their quarter of birth. Thus, this diagram takes it for granted that this particular LATE is the only parameter that is informed by this analysis.[1]

For the causal diagram in Figure 8.1, schooling has both direct and indirect effects on wages. Most important, as is maintained in the human capital literature, schooling is thought to have a negative indirect effect on wages through work experience; going to school longer reduces the amount of work experience one acquires by any particular age. Accordingly, there is a causal pathway from schooling to wages via work experience. The quarter-of-birth IV does not provide a separate estimate of this distinct pathway, because a change in schooling in response to one's birthdate also changes work experience. At best, the quarter-of-birth IV estimates only the total effect of schooling on wages, not its direct effect.[2] The IV yields a LATE estimate that likely mixes together two distinct and countervailing causal pathways: a positive direct effect of schooling on wages and a negative indirect effect via work experience. Given the long-standing interest of economists in the interaction between investments in formal schooling and the provision of on-the-job training, this total effect estimate is regarded by many as an insufficiently deep causal account of the

[1] This position is equivalent to assuming that a more encompassing DAG exists that is applicable to the entire population and that the diagram in Figure 8.1 applies only to the joint probability distribution for compliers.

[2] Angrist and Krueger could deal with this problem by conditioning on a measure of work experience, but in their data no such variable is available. The only alternative with their data, which is common in the literature, is to attempt to untangle the effects by conditioning on age (which is typically done within an age-heterogeneous sample). However, this form of conditioning does not completely explain away the association between schooling and work experience, as is widely recognized (see Card 1999).

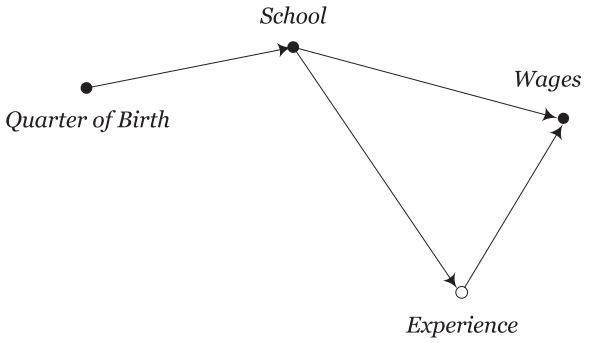


Figure 8.1: A causal diagram for compliers with quarter of birth as an IV for schooling.

effect of education on earnings (even if one is convinced, as we are, that learning about compliers in this case is still illuminating).

For a second example, consider the effect of military service on lifetime earnings, as analyzed by Angrist (1990) and introduced in Section 7.2. The question of interest here is whether military service provides important training that increases later earnings in civilian life or rather whether military service is simply a period of lost civilian work experience. Military service, however, is not random. On the one hand, individuals must pass a mental ability test and a health examination in order to enlist. On the other hand, individuals with attractive schooling opportunities and civilian labor market opportunities are less likely to enlist. To deal with the problem of nonrandom selection into the military, Angrist (1990) considers individuals who were potentially eligible to be drafted into the military by a lottery during the later stages of the Vietnam War. To ensure that the draft was fair, the U.S. government decided to draft individuals based on a lottery that selected birthdates (see our earlier discussion in Section 7.2).

Consider now the causal diagram in Figure 8.2, again based on the summary of critiques of the study compiled by Rosenzweig and Wolpin (2000). Here, there are three issues to consider. First, as we noted for the last example, the draft lottery identifies a LATE, and thus it is not applicable to the causal effect for always takers (those who voluntarily enlist) and never takers (those who are draft dodgers, those who failed the mental ability test, and those who failed the physical examination). Accordingly, as for Figure 8.1, the causal diagram in Figure 8.2 also applies to compliers only.

Second, note that there is a potential path from the draft lottery to civilian experience/training. If this path exists, then the draft lottery is not a valid IV for military service. As we noted earlier in Chapter 7, Heckman (1997) argues that employers would be likely to invest less in individuals with unfavorable lottery numbers. For example, it is plausible that employers may have given less on-the-job training to those most likely to be drafted and/or may have assigned

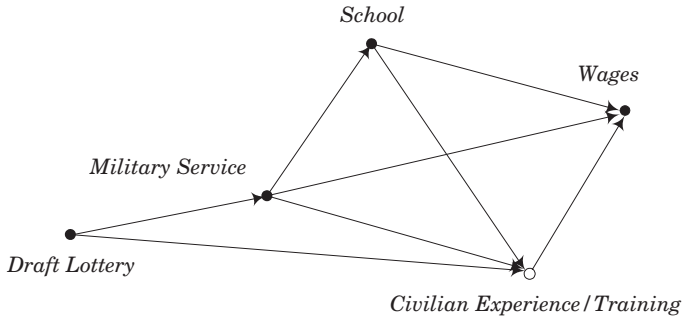


Figure 8.2: A causal diagram for compliers with the Vietnam draft lottery as an IV for military service.

such individuals to short-run tasks that did not require the accumulation of skill to master. If this effect exists, then the draft lottery would be an invalid instrument for military service.

Third, and most important for our consideration here, there are four separate causal pathways between military service and wages. In addition to the direct causal effect of military service on wages, there are two causal pathways solely mediated by civilian experience and schooling respectively. Here, it is generally thought that military service reduces civilian labor force experience and schooling, both of which then reduce wages. But there is then a countervailing effect of military service that snakes through the causal diagram in a fourth causal pathway: Military service reduces schooling, which then increases work experience, and which then increases wages. Because all four of these pathways are activated by the shock induced by the draft lottery, Angrist's only resort is to assert that his estimates are for the total effect of military service on wages. But, given the inherent interest in untangling how the military service effect interacts with both schooling and the accumulation of civilian experience, a total effect estimate is insufficiently deep to end all future research, even though this natural experiment was informative and an important contribution to the literature.

Rosenzweig and Wolpin (2000) consider many other examples, all of which emphasize the same basic points as these two examples. IV analyses of this type provide estimates of very specific parameters that are often not of fundamental interest because they apply to nonrandom segments of the population of interest. Moreover, IV analyses typically provide total causal effect estimates, often in substantive areas in which scholars have an inherent interest in the separable causal pathways that generate the outcome in response to the cause. Understanding the separable causal pathways that make up these total effects requires an explicit specification of additional intervening and mediating variables.

Now consider these two issues more generally, moving away from IV estimates toward more general estimates of average causal effects. Consider an

optimistic scenario for which one obtains what all agree is a consistent estimate of the average causal effect of D on Y (as warranted, for example, by a consensus that one has conditioned on all variables that block all back-door paths from D to Y). Even in this scenario, in which the causal claim is valid by the standards of the counterfactual model, there are two related ways in which such an estimate can be regarded as insufficiently deep.

First, the average causal effect may not be a parameter of any fundamental interest. This point has been made most forcefully by Heckman (2000, 2005). If treatment effects are heterogeneous, which often is the case as we have noted in prior chapters, then the average causal effect will be population specific. To oversimplify the argument, it may be that what is of fundamental interest are conditional average treatment effects of some form (perhaps even just the average treatment effect for the treated and the average treatment effect for the untreated). The overall average treatment effect is then simply a weighted average of any such underlying effects, in which the weights are a function of whatever pattern of causal exposure may prevail in the particular population from which one's sample is drawn. Heckman and many others note that the average causal effect of D on Y may be of limited use for predicting the outcomes of policy interventions, either for new populations or in different contexts. But the point is in fact much more general and is applicable outside of policy research. Average causal effects are tied to particular populations, and quite often social scientists desire explanations for outcomes that can be modified in straightforward ways when populations shift or contexts change in ways that are separable from the fundamental conditional average causal effects.

The second issue, which is our primary focus in the remainder of this chapter, is that a consistent estimate of the average causal effect of D on Y does not necessarily entail any particular mechanism that explains how D brings about Y. If a theory suggests why and how D brings about Y, then merely providing evidence of the amount that Y can be expected to shift in response to an intervention on D does not then provide any support for the underlying theory. If an assessment of the support for the underlying theory is desired, then a more narrowly focused analysis of the putative causal pathways that relate D to Y must be undertaken. We present Pearl's approach to this type of analysis in the next section, which we see as consistent with decades of research in the social sciences, although appreciably more clear as a guide for future research. We first return to the classic perspective on the importance of intervening variables in causal explanation.

8.2 Explanation and Identification of Causal Effects by Mechanisms

As we noted earlier, social scientists have generally considered the explication of mechanisms through the introduction of mediating and intervening variables to be essential to sound explanatory practice in causal analysis. Duncan and his

colleagues wrote in their 1972 book, *Socioeconomic Background and Achievement*:

> ... much of the scientific quest is concerned with the search for intervening variables that will serve to interpret or explain gross associations presumed to reflect a causal relationship. (Duncan, Featherman, and Duncan 1972:12)

The words "interpret" and "explain" are implicit references to the language of social research that is usually associated with Paul Lazarsfeld. More than two decades earlier, Kendall and Lazarsfeld (1950) had distinguished between alternative types of elaboration that can be carried out when investigating an association between a causal variable X and an outcome variable Y.[3] Each type of elaboration involves the introduction of a test factor (or test variable) T, after which the association between the causal variable X and the outcome variable Y is calculated for each value of T.

Kendall and Lazarsfeld considered two types of M-elaboration, which arise when the partial association between X and Y within strata defined by T is smaller than the original total association between X and Y.[4] The two possibilities are determined by whether T precedes X in time according to the maintained theory:

1. If T follows X in time, then an M-type elaboration can be represented by the diagram $X \rightarrow T \rightarrow Y$ (Kendall and Lazarsfeld 1950:157). This type of elaboration is referred to as an *interpretation* of the association between X and Y, using a test factor that is an intervening variable.

2. If T precedes X in time, then an M-type elaboration can be represented by the diagram $X \leftarrow T \rightarrow Y$ (Kendall and Lazarsfeld 1950:157). This type of elaboration is referred to as an *explanation* of the association between X and Y, using a test factor that is an antecedent variable.

Although Kendall and Lazarsfeld did not argue that causality can be established by interpretation, and although they use the word explanation in a rather limited sense (and with a different usage than Duncan and colleagues, as just quoted), Kendall and Lazarsfeld were clearly interested in mechanistic accounts of causal effects. For example, they wrote:

[3]The citations here are for a discussion of Samuel Stouffer's research in *The American Soldier* (which we also discuss in Chapter 1). Patricia Kendall was the lead author for this piece, and yet she is almost never cited in the derivative literature as a contributor to this language. Because it is often written that Lazarsfeld presented and discussed this basic typology in many places, it is possible that it is fair to credit him disproportionately for these ideas. Our reading of the literature, however, does not yield a clear interpretation. But it does suggest to us that the Kendall has received less credit than she deserved.

[4]They also lay out a third form of elaboration, referred to as P-type elaboration (or specification). Here, the goal is to focus "on the relative size of the partial relationship [between X and Y within strata of the test factor T] in order to specify the circumstances under which the original relation is more or less pronounced" (Kendall and Lazarsfeld 1950:157). This type

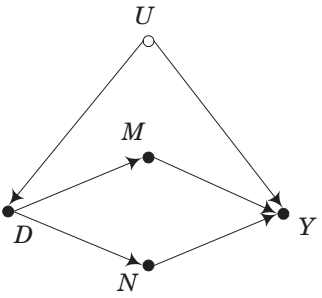


Figure 8.3: A causal diagram in which M and N represent an isolated and exhaustive mechanism for the causal effect of D on Y.

> When we interpret a result we try to determine the process through which the assumed cause is related to what we take to be its effect. How did the result come about? What are the "links" between the two variables? Answers to these questions are provided in the interpretation of the result. (Kendall and Lazarsfeld 1950:148)

A long tradition of empirical social science exists that follows these basic ideas, in which a central goal of investigation is the search for mediating, intervening, and processual variables that can account for an association that is thought to be causal. Even so, as we noted in Chapter 5, applications of the all-cause regression specification approach gradually obscured this goal, leading over the 1980s and 1990s to more frequent attempts to offer single-equation models in which all putative causes of an outcome are evaluated next to each other.

Fortunately, Pearl (2000) has developed an inside–out perspective on the identifying power of mechanisms that can help the social sciences recapture the inclination to pursue mechanistic explanation of putative causal effects. Pearl's approach goes well beyond Lazarsfeldian elaboration. Instead, he shows how one can consistently estimate the effect of a causal variable on an outcome variable by estimating the effect as it propagates through an isolated and exhaustive mechanism.

Consider the graph in Figure 8.3. For this DAG, the variable U is unobserved, and a back-door path $D \leftarrow U \rightarrow Y$ exists between D and Y. As a result, identification by the back-door criterion, as presented in Chapter 3, is infeasible. However, the variables M and N intercept the full causal effect of D on Y, and thus these two variables represent the full mechanism (or set of causal pathways) that relates D to Y. (Again, recall that there is no assumption of linearity in a DAG, and thus this model is consistent with any nonlinear function in D, M, N, and Y.)

---

of elaboration is appropriate when the test factor is related to either X or Y but not both. We ignore this form of elaboration here, as we focus on variables that are related to both the causal and outcome variables.

For this graph, observation of M and N identifies the causal effect of D on Y by a double application of the back-door criterion, first for the estimation of the causal effects $D \rightarrow M$ and $D \rightarrow N$ and then for the estimation of the causal effects $M \rightarrow Y$ and $N \rightarrow Y$. For the first two causal effects, note that there are no back-door paths between D and both M and N. Therefore one can consistently estimate $D \rightarrow M$ and $D \rightarrow N$ with their simple unconditional associations.

For $M \rightarrow Y$ and $N \rightarrow Y$, there are two back-door paths to consider for each effect, but in all cases these back-door paths can be blocked by conditioning on D. For example, for the causal effect of M on Y, the two back-door paths are $M \leftarrow D \leftarrow U \rightarrow Y$ and $M \leftarrow D \rightarrow N \rightarrow Y$, both of which are blocked by D.

Because one can obtain consistent estimates of the casual effects of D on both M and N and, in turn, of both M and N on Y, one can calculate the full causal effect of D on Y by combining these causal effect estimates. The presence of the back-door path $D \leftarrow U \rightarrow Y$ is inconsequential.

The front-door criterion is very simple, as it is nothing more than a straightforward two-step application of the back-door criterion introduced in Chapter 3. The front-door criterion, however, does not give direct guidance on how deep an identifying mechanism must be in order to qualify as a sufficiently deep causal explanation. Before addressing the latter issue, we clarify in the remainder of this section what the requirements of isolation and exhaustiveness entail.

The Assumption That the Mechanism is Isolated and Exhaustive

The crucial assumption of this approach, as we have noted already, is that the mechanism (or set of variables that comprise the mechanism) is isolated and exhaustive. To see the importance of isolation first, consider the alternative graph presented in Figure 8.4. Here, the variable U is again unobserved, and thus the back-door path $D \leftarrow U \rightarrow Y$ is unblocked. Also, the variable M intercepts the full causal effect of D on Y. But the mechanism represented by M is not isolated because U has a causal effect on M. This contamination from U generates a new back-door path from M to Y as $M \leftarrow U \rightarrow Y$. As a result, the causal effect of M on Y is not identified because D does not block all back-door paths from M to Y. And, even though a consistent estimate of the effect of D on M can be obtained, a consistent estimate of the effect of D on Y is not available.

This example also shows what is crucial in the criterion of isolation. The mechanistic variables must be isolated from otherwise unblocked back-door paths so that the back-door criterion can be applied twice in order to recover the full causal effect from the data. For Figure 8.4, if U were instead an observed variable, then the back-door path $M \leftarrow U \rightarrow Y$ could be blocked by conditioning on U. This would allow for the estimation of the effect of M on Y by standard conditioning techniques. And, as a result, the causal effect can be estimated consistently by first estimating the causal effect $D \rightarrow M$ with their unconditional association and then estimating the causal effect $M \rightarrow Y$ by conditioning on U.

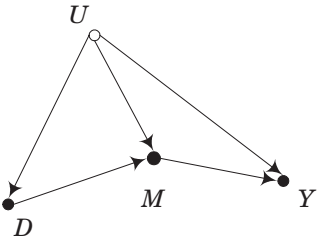


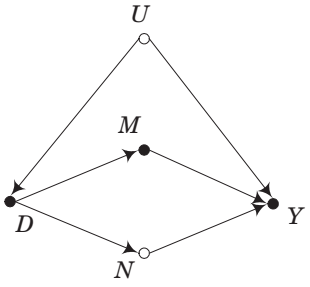
Figure 8.4: A causal diagram in which M is not an isolated mechanism for the causal effect of D on Y.

More generally, isolation comes in strong and weak forms, both of which are sufficient. For the strong form considered until now, no back-door paths exist between the mechanistic variables and the outcome variable that are not blocked by D. For the weak form, back-door paths do exist between the mechanistic variables and the outcome variable that are not blocked by D, but each of these back-door paths can then be blocked by conditioning on another observed variable in the graph other than D. This distinction clarifies that one must be concerned only about the dependence of mechanistic variables on components of back-door paths that cannot be blocked by conditioning on observed variables.

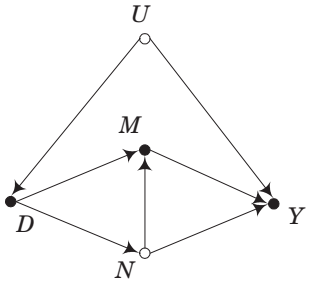
The implication of the necessity of assuming isolation is that a very good understanding of the back-door paths between the causal variable and the outcome variable is needed to justify the assumption that none of the components of unblockable back-door paths have direct effects on the mechanism. If the isolation assumption cannot be maintained, then the mechanistic variables are similarly affected by the same set of dependencies that invalidate basic back-door conditioning as a strategy to identify the causal effect.

Now consider the requirement that the mechanism is exhaustive. For the graph in panel (a) of Figure 8.5, suppose, again, that there is an unblocked back-door path and also two distinct causal pathways from D to Y, which are represented by the two variables M and N. Finally, unlike for Figure 8.3, suppose that N is unobserved.

In this case, the causal pathway from D to Y via N cannot be estimated, and thus the full causal effect of D on Y cannot be estimated by front-door conditioning. If one were to assert a mistaken assumption that M is the full causal mechanism (for example, by substituting a back-door path $D \leftarrow N \rightarrow Y$ for the genuine causal pathway $D \rightarrow N \rightarrow Y$), then one can obtain a causal effect estimate. But the causal effect of D on Y will then be underestimated (assuming all causal effects are positive, etc.) because the part of the causal effect generated by the pathway $D \rightarrow N \rightarrow Y$ is attributed to a mistakenly asserted back-door path $D \leftarrow N \rightarrow Y$.



(a) The causal pathway via M is identified even though N is unobserved



(b) The causal pathway via M is not identified

Figure 8.5: Causal diagrams in which one pathway in an isolated and exhaustive mechanism is unobserved.

Relaxing the Assumption That the Mechanism is Exhaustive

For the example depicted in panel (a) of Figure 8.5, a mechanism-based analysis may still be useful because an isolated piece of the causal effect can still be estimated. Even though N is unobserved, the causal effect of D on M is identified by their observed association because there are no back-door paths from D to M. And, as before, the causal effect of M on Y is identified because all back-door paths from M to Y are blocked by D. One can thereby obtain a consistent estimate of the causal effect of M on Y by conditioning on D, which guarantees that the part of the causal effect that travels through the pathway $D \rightarrow M \rightarrow Y$ can be estimated consistently. This is an important result, especially for practice, because identifying and consistently estimating a distinct causal pathway can be very useful, even if one cannot in the end offer a causal effect estimate of the full effect of the causal variable on the outcome variable.[5]

But, even though this partial estimation result is very useful, there is an important implicit assumption that is hidden by this example: The unobserved causal pathways cannot interact with the observed causal pathways. To see the complications that such an interaction could produce, consider the graph in panel (b) of Figure 8.5. For this graph, the mechanism is isolated, but the unobservable variable N has a causal effect on M. As a result, the causal pathway from D through N to Y is "mixed in" with the causal pathway from D through M to Y. In this situation, the effect of M on Y is not identified. Two back-door paths between M and Y are still blocked by D: $M \leftarrow D \leftarrow U \rightarrow Y$ and $M \leftarrow D \rightarrow N \rightarrow Y$. However, there is now a third back-door path between

[5]Note, further, that one can then assess the portion of the total variation in D and Y that can be attributed to the pathway through M. This calculation can be useful, but it does not identify the amount of the causal effect that is explained by M. The problem is that the association between D and Y that is not attributable to M cannot be apportioned across the two unblocked paths $D \leftarrow U \rightarrow Y$ and $D \rightarrow N \rightarrow Y$.

M and Y that cannot be blocked by conditioning on D or any other observed variable. This new path is $M \leftarrow N \rightarrow Y$, and it remains unblocked because its only intermediate variable N is unobserved (and is not a collider). Not only is it impossible to estimate the effect of D on the unobserved variable N, but the variable N may transmit to both M and Y its own exogenous variation that is completely unrelated to D and U. Without recognition of this outside source of dependence, one could mistakenly infer that the causal pathway from D to Y via M is much more powerful than it is (assuming all causal effects are positive, etc.).

As these simple diagrams show, if one wants to use a mechanism-based strategy to point identify the full effect of a causal variable on an outcome variable, one must put forward an isolated mechanism that is also exhaustive. The exhaustiveness requirement can be relaxed if one is satisfied with identifying only part of the causal effect of interest. But, to secure partial estimation of this form, one must be willing to assume that the observed portion of the mechanism is completely separate from the unobserved portion of the mechanism. And, to assert this assumption, one typically needs to have a theoretical model of the complete and exhaustive mechanism, even though parts of it remain unobserved.

In sum, Pearl's front-door criterion represents a powerful and original set of ideas that clarifies the extent to which a mechanism can be used to identify and estimate a causal effect. And, by approaching the causal inference predicament from the inside–out, the approach helps to shape the question of whether a causal claim qualifies as a sufficiently deep explanation. Rather than estimating a causal effect by some form of conditioning (or with a naturally occurring experiment) and then wondering whether a mechanism can be elaborated to show how the effect comes about, Pearl instead shows that, if we can agree on the variables that constitute an isolated and exhaustive mechanism (in some configuration), then we can estimate the causal effect from its component causal pathways.

The approach cannot be taken out of context, though, lest one claim too much explanatory power for a nonisolated and/or nonexhaustive mechanism. And, at the same time, it must be acknowledged that even a front-door-identified causal effect may not be explained deeply enough by the isolated and exhaustive mechanism that identifies it, for any such mechanism may still be too shallow for a particular substantive area. To consider these issues further, we now turn to recent social scientific writing on mechanisms, wherein the connections among theory, generative mechanisms, and modes of explanation have been fruitfully engaged.

8.3 The Appeal for Generative Mechanisms

To some extent inspired by the work of Jon Elster (e.g., Elster 1989), but more generally by their reading of past successes and failures in explanatory inquiry in the social sciences, Peter Hedström and Richard Swedberg convened a group of

leading scholars to discuss a reorientation of explanatory practices in the social sciences. A collection of papers was then published in 1998 as *Social Mechanisms: An Analytical Approach to Social Theory.* At the same time, John Goldthorpe developed his case for grounding all causal modeling in the social sciences on the elaboration of generative mechanisms. His proposal was published in 2000 as *On Sociology: Numbers, Narratives, and the Integration of Research and Theory.*[6] Most recently, Peter Hedström has laid out in considerable detail his full program of a mechanism-based social science in his 2005 book, *Dissecting the Social: On the Principles of Analytical Sociology.* To orient the reader to this alternative mechanism-based perspective, we will present the slightly different perspectives of both Goldthorpe (2000, 2001) and Hedström (2005).

Goldthorpe develops his proposal by first attempting to drive a wedge between the counterfactual model, which he labels "causation as consequential manipulation," and his alternative mechanistic model of causality. To develop his argument, he advances the claim that the counterfactual model is irretrievably tied to actual experimental manipulations of causal variables. Goldthorpe (2001:6) writes, "... the crux of the matter is of course the insistence of Rubin, Holland, and others that causes must be manipulable, and their unwillingness to allow causal significance to be accorded to variables that are not manipulable, at least in principle."[7] [See our discussion of this issue in the final chapter, in which we present the position of Woodward (2003) that this argument is incorrect.] Goldthorpe also argues, citing the work of statisticians Cox and Wermuth (1996; also Cox 1992), that the counterfactual model too easily settles for causal claims of insufficient depth that do not account for how causes brings about their effects (a point that is clearly supported by our presentation of the natural experiment literature discussed earlier in this chapter).

With this criticism of the counterfactual model in mind, Goldthorpe then proposes his alternative recommendation:

> The approach to causal analysis that is here proposed ... is presented in the form of a three phase sequence: (i) establishing the phenomena that form the *explananda*; (ii) hypothesizing generative processes at the level of social action; and (iii) testing the hypotheses. (Goldthorpe 2001:10)

[6] A chapter from his book was also published in 2001 in the *European Sociological Review* as "Causation, Statistics, and Sociology," which then received comments from a variety of scholars, including the statisticians David Cox and Nanny Wermuth (whose response we will discuss later). Our citations are to the 2001 piece.

[7] It is unclear from this passage what Goldthorpe means by "in principle." However, on the next page of the article, he offers an example that reveals that his conception of a manipulation can be very narrow. Following an example of Holland (1986), he introduces mechanistic variables and argues that in so doing he has produced a causal narrative that has nothing to do with hypothetical manipulations/interventions and hence cannot be represented in a counterfactual framework. The argument seems to be that the mechanistic variables are nonmanipulable because they are typically voluntary decisions of an individual.

Goldthorpe then proceeds to lay out the basic contours of the first two steps, arguing that the most useful mechanisms in the social sciences are those that are based on rational choice theory (because these are sufficiently microscopic and focus on the actions and beliefs of individuals). He then explicates the third step of his approach, which is his requirement that one must test hypotheses suggested by the generative mechanism that one has proposed.[8] And here, he notes that generative causal accounts of the sort that he prefers can never be definitively verified, only indirectly supported through repeated validation of entailed hypotheses.

Consider now the more recent appeal for mechanisms offered by Peter Hedström (2005). Unlike Goldthorpe, Hedström does not confront the counterfactual causality literature directly. He instead develops an argument that suggests the following orienting principle: "The core idea behind the mechanism approach is that we explain not by evoking universal laws, or by identifying statistically relevant factors, but by specifying mechanisms that show how phenomena are brought about" (Hedström 2005:24). Hedström arrives at this position from two different directions. First, he signs on to the dominant position in the philosophy of science that an explanation of how a phenomenon is brought about must necessarily be a causal account of some form that does not rely on unrealistic presuppositions of the existence of invariant general (or covering) laws (see Hedström 2005:15–20). But he also justifies the principle by criticizing regression-based causal models in social science (see Hedström 2005:20–3):

> Such statistical analysis is often described as a form of 'causal analysis'. If a factor appears to be systematically related to the expected value or the conditional probability of the outcome, then the factor is often referred to as a (probabilistic) 'cause' of the outcome. Although it makes little sense to quibble over words, I would like to reserve the word *cause* for a less casual notion of causality. (Hedström 2005:23)

Like Goldthorpe, Hedström also lays out a script for the construction of mechanisms, which is similar in its appeal to the principles of methodological individualism (although he steps back from Goldthorpe's strict reliance on subjective expected utility forms of rational choice theory, noting the promise of agent-based models, and similar approaches). Most importantly, Hedström does not advocate the indirect testing of mechanisms in the way that Goldthorpe does. Instead, he defends the explanatory utility of mechanisms in a much more general way:

> Mechanisms should be seen as theoretical propositions about causal tendencies, not as statements about actualities. An explanation may

[8] By our reading, Goldthorpe ends up back at the counterfactual model at this point. His defense against such a reading is "... while it might seem that, at this stage, attention does after all come to focus on the effects of – given – causes rather than on the causes of effects, this is within the context not of randomized experimental design but of (what should be) a theoretically informed account of a generative process that is subject to ongoing evaluation." (Goldthorpe 2000:13)

> be perfectly correct if understood as a proposition about a causal tendency, and yet it may be inadequate for predicting actual outcomes if other processes are also at work. . . . Since it is the rule rather than the exception that concretely observed phenomena are influenced by several different processes, testing a theory by examining the accuracy of its predictions is likely to conflate the truth or relevance of the postulated mechanism with the importance of other processes, and this may lead us to mistakenly reject perfectly appropriate causal accounts. (Hedström 2005:108)

What, then, is the explanatory status of mechanisms that are "theoretical propositions about causal tendencies"? Here, Hedström seems to argue three basic points: (1) An explanation must be a causal account; (2) causal accounts must include mechanisms; (3) mechanisms may or may not explain actualities. The direct implication of this position is that a valid explanation does not necessarily have to explain actualities, only causal tendencies.[9]

In our view, Goldthorpe and Hedström have developed a wonderful appeal for the need to construct sufficiently deep theories of social processes. Their work is filled with much insight on a variety of promising ways to formulate plausible theoretical mechanisms. And they have helped to further convince many social scientists that specifying the social processes that account for how causes bring about their effects is a central goal of analysis. It is therefore unsurprising that this new wave of scholarship has inspired a movement of sorts in sociology and political science, which we and many others regard as both promising and very healthy for theory construction in the social sciences.[10]

But, to be seen as blueprints for explanatory causal analysis, both of these particular perspectives need to be augmented. And we would argue that they can be usefully extended by embracing the counterfactual model of causality more directly. We explain this position in the remainder of this section, first by offering a slightly different reading of recent developments in the philosophy of science and then by raising the issue of how social scientists can retain the capacity to adjudicate between competing mechanistic accounts of causal effects.

[9] It is unclear from the foregoing passage what the problematic "other processes" are meant to be. Clearly, they must come in two different forms: (1) other mechanisms that are completely independent of the postulated mechanism of primary interest or (2) other mechanisms that interact with the postulated mechanism of primary interest. If one is estimating the effects of a cause, it is only the latter that are problematic for the evaluation of mechanisms, and hence for the evaluation of causal accounts based on mechanisms. See our earlier discussion of partial identification by the front-door criterion.

[10] But we are somewhat more traditional in favoring mechanisms that are formal models, as we are less convinced of the utility of many simulation-based methods of theory construction (see Hedström 2005:76–87, 131–6, 149 on the appeal of such techniques). We agree with Humphreys' (2004:132) caution: "Agent-based models are a powerful addition to the armory of social scientists, but as with any black-box computational procedures, the illusion of understanding is all too easy to generate."

Philosophy of Science and Explanation by Mechanisms

As we noted at the beginning of this chapter, the appeal to mechanisms as elaborations of causal claims has been entrenched in the sociological literature for many decades, and we suspect that this emphasis is true of all of the social sciences.[11] Where perhaps the focus on generative mechanisms is new is in the proposed elevation of their status to the object of primary investigation. This elevation is not unrelated to shifting currents in the philosophy of science.

Some proponents of a mechanism-based social science have drawn inspiration from the demise of covering law models of explanation in the philosophy of science (e.g., Gorski 2004, Hedström 2005). The covering law model, as most famously explicated by Carl Hempel, maintains that all valid explanations can be formulated as logico-deductive entailment from invariant general laws.[12] Since 1970 at least, the model has received near-continuous challenges to its basic premises (see Godfrey-Smith 2003, Salmon 1989, and Woodward 2003). The presupposition that general and exceptionless laws exist and that can be used to warrant all valid causal claims cannot be sustained in most scientific disciplines, and certainly not in the social sciences.

In response, a variety of alternative models of scientific explanation have arisen, with realist models attracting the largest number of adherents (see Psillos 1999). In general, realist models grant ontological status to unobserved quantities (sometimes conferring provisional truth status on them). Most realist models reject causal nihilism and affirm that valid scientific explanation must necessarily be grounded in causal accounts of some form. Especially when invoked for the social sciences, realist models admit the possibility of inherent heterogeneity of causal relationships – in time, space, and within populations. But the depth of the appeal to unobservables varies across types of realist models, as does the level of provisional truth status conferred upon unobservables.

What we wish to emphasize in this section is how uneasily the focus on generative mechanisms sits within this new terrain. As we have just noted, the

[11] Hedström and Swedberg (1998) take a different reading of the literature that followed from the Lazarsfeldian tradition, focusing critical attention on the work of Duncan and his colleagues (see their quotation on page 9). Then, in his 2005 book, Hedström criticized Duncan's models of socioeconomic status (see pages 102-3). As is clear in our main text, our reading of Duncan's record on this issue is somewhat less negative, although we noted in Chapter 1 that it is hard to pin down exactly what Duncan recommended in the theory and the practice of causal modeling. Hedström's negative opinion, however, is certainly applicable to much scholarship that followed Duncan's lead, a position that we suspect Duncan too would have agreed with and which we note in Chapter 1.

Goldthorpe (2001) also has a different reading of Lazarsfeld's method of elaboration. He argues that the "robust dependence" tradition of causal analysis, which he regards as exemplified by Lazarsfeld's work, was based almost entirely on the search for prior causes that can explain away correlations (see Goldthorpe 2001:2–3). This seems inconsistent with what Kendall and Lazarsfeld (1950) refer to as interpretation, as we noted earlier.

[12] We do not attempt to give a full summary of the fall of covering law models, as many accessible texts exist in philosophy (see citations in the main text) and others that are written for social scientists alone (e.g., Gorski 2004 and Hedström 2005). Nor do we give a full explication of the variety of positions that have replaced the covering law model, as these too are well summarized elsewhere (see citations in the main text for realist models in particular, as well as the variety of naturalist positions that contend with them).

appeal for mechanistic explanation in the social sciences is attractive partly because it is claimed that it does not rest on the existence of general laws. But some philosophers who endorse the mechanisms position disagree with this claim, arguing that laws are still essential, perhaps even constituting the defining characteristic of what a mechanism is. For example, having written on the role of mechanisms in scientific explanation for several decades, Mario Bunge (2004:207) concludes in a recent article, "No law, no possible mechanism; and no mechanism, no explanation."[13]

Hedström (2005) considers this complication carefully, and he emphasizes that his mechanism-based approach to explanation is less reliant on invariant laws than the covering law models that he rejects. This position is certainly correct, but one must still consider how to grapple with what seem to be three fairly common declarative statements on the relationships between laws and mechanisms: (1) invariant covering laws do not exist, (2) mechanisms depend to some degree on the existence of laws that are weaker than general, invariant laws, (3) mechanisms are nested within each other. Accepting all three statements seems to delimit laws to the smallest possible configuration within each nested mechanism, such that the laws of a mechanism are dispensed with each time a new black box is opened.

If a mechanism is designed to explain how an actuality comes about, then all seems to be fine with this perspective.[14] But if one instead maintains that mechanisms are the key to the explanation of causal tendencies only – such that the validity of a mechanism cannot be undermined by its inability to explain anything in particular – then this line of thought leads all too easily to the critical realist perspective on mechanisms. We suspect that most empirical social scientists would find critical realism uninspiring. Critical realism's pioneer, Roy Bhaskar, writes with regard to mechanisms:

> The world consists of mechanisms not events. Such mechanisms combine to generate the flux of phenomena that constitute the actual states and happenings of the world. They may be said to be real, though it is rarely that they are actually manifest and rarer still that they are empirically identified by men. They are the intransitive objects of scientific theory. They are quite independent of men – as thinkers, causal agents and perceivers. They are not unknowable though knowledge of them depends upon a rare blending of intellectual, practico-technical and perceptual skills. They are not artificial constructs. But neither are they Platonic forms. For

[13] Given Bunge's position, one then necessarily wonders: How does one discover these crucial explanatory mechanisms? Bunge's (2004:200) guidance is: "There is no method, let alone logic, for conjecturing mechanisms. True, Peirce wrote about the 'method of abduction,' but 'abduction' is synonymous with 'conjecturing', and this – as Peirce himself warned – is an art, not a technique. One reason is that, typically, mechanisms are unobservable, and therefore their description is bound to contain concepts that do not occur in empirical data."

[14] As best we can tell, Woodward (2003; Section 4.6) takes this position, noting the need for only a rather limited "backing" relationship between causal claims and laws.

> they can become manifest to men in experience. Thus we are not imprisoned in caves, either of our own or of nature's making. We are not doomed to ignorance. But neither are we spontaneously free. This is the arduous task of science: the production of the knowledge of those enduring and continually active mechanisms of nature that produce the phenomena of the world. (Bhaskar 1998[1997]:34–5)

If the social sciences sign on to the idea that mechanisms are general and transcendentally valid explanations that may not explain any particularities, we will be led inevitably to a fundamental premise of critical realism: The mechanisms that constitute causal explanations are irreducible to each other, even if they are nested in each other. This position is summarized by Andrew Collier (2005:335): "Critical realism defends the idea that reality is many-layered, and each level has its own kind of laws, irreducible to those of any other layer."

For the social sciences, one might argue that such an irreducibility presumption could be unifying in offering protection against incursions from biology and physics. But, if accepted, it would undermine the work of social scientists who have at least some interest in developing causal claims that unite levels of analysis. If irreducibility were accepted, how then could methodological individualists such as Goldthorpe criticize those who seek to develop macrolevel causal claims with only minimally sufficient reliance on proximate actors and institutions (e.g., Alexander 2003)?[15] Gorski (2004), for example, lays out a constructive realist model of explanation, built up from the causal process perspective of Wesley Salmon, that is completely at odds with the perspective of Goldthorpe (2000).[16] But Goldthorpe (2000) questions the explanatory utility of all secondhand historical analysis, in essence rejecting the capacity of historical analysis, as practiced in sociology, to sustain causal claims of any form. If Goldthorpe were to sign on to irreducibility, which we doubt he would, he could not thereby criticize macrosocial claims of causal relationships.

Given that these extreme positions on mechanisms in the philosophy of science are likely to be unhelpful to practicing social scientists, and given that we suspect Goldthorpe, Hedström, and others in the generative mechanisms movement would agree, which type of philosophy of science gives the appropriate backing for causal analysis? If anything, it is the philosophical writing on the counterfactual model that provides a solid and pragmatic foundation, as best represented by Woodward (2003). The key to understanding why this is the case is to consider alternative ways to adjudicate between the rival mechanisms proposed by alternative investigators, which we turn to next.

[15] Methodological individualism is the basic position of Goldthorpe (2000) and Hedström (2005), as influenced heavily by the scholarship of Raymond Boudon (see Boudon 1998 and citations therein).

[16] Gorski (2004) endorses the causal process model of Wesley Salmon, as developed in Salmon's work from the 1970s and early 1980s (see Salmon 1984). Given the ways in which Salmon's work has developed since then, turning completely toward causal mechanical ideas based on the notion of conserved quantities, his ideas now seem completely at odds with Gorski's statement that "Social science *is* 'nothing but history.' The real error was ever to think it could be anything more" (Gorski 2004:30).

Adjudication Between Rival Mechanisms

Imagine that social scientist A and social scientist B have proposed distinct mechanisms for how X brings about Y. How do they determine whose mechanism is supported? According to Goldthorpe (2000, 2001), each scholar is expected to derive entailed hypotheses and test them with data. One might hope that scholars A and B will be able to agree on a clear critical test that could tip the scales in favor of one mechanism or the other. Unfortunately, the mechanisms of the two scholars may be so different that no such critical test can be derived and agreed on (as would often be the case if scholar A is a sociologist and scholar B is an economist, for example). If no such agreement can be found, the two scholars may end up expending effort seeking to affirm their own entailed indirect hypotheses.

Consider the reaction that Goldthorpe's proposal elicited from the statisticians David Cox and Nanny Wermuth, whose prior work Goldthorpe had used to develop his proposal:

> Goldthorpe (2001) has argued for this . . . view of causality as the appropriate one for sociology with explanation via rational choice theory as an important route for interpretation. To be satisfactory there needs to be evidence, typically arising from studies of different kinds, that such generating processes are not merely hypothesized. Causality is not to be established by merely calling a statistical model causal. (Cox and Wermuth 2001:69)

Cox and Wermuth take the position that generative mechanisms must be directly evaluated, not evaluated only by indirect entailed hypotheses. Anything short of this analysis strategy could result in a flourishing of mechanisms in the social sciences, without an attendant sense of which ones are valid or not. The alternative to mechanism anarchy could be even worse: mechanism warlordism. The mechanisms of the most industrious scholars – those who can dream up the most hypotheses to affirm, who can recruit the largest number of students to do the same, and who can attract the funds to collect the data – could receive the most affirmation. The only defense for out-of-favor mechanisms might then be to appeal to the hidden structures of alternative mechanisms, which one would be tempted to claim cannot be evaluated because of a lack of data.

In sum, the generative mechanisms movement in the social sciences is an admirable call for theory construction.[17] When it does not verge into critical–realist transcendentalism, it is also a very useful call for the pursuit of sufficiently deep causal accounts. But such depth, we would argue, is best secured when it is verified in empirical analysis grounded on the counterfactual model. In the next section, we work our way back to Pearl's front-door criteria, using the language of mechanism sketches and mechanism schemas.

[17] In fact, the first author found it convincing and inspiring when writing Morgan (2005).

8.4 The Pursuit of Explanation with Mechanisms that Bottom Out

Amid the resurgence of writing on mechanisms, we find one statement more helpful than many others, the 2000 article "Thinking About Mechanisms" written by Machamer, Darden, and Craver and published in *Philosophy of Science.* In their article, Machamer, Darden, and Craver develop two particularly helpful lines of thought: (1) the distinctions among a fully articulated mechanism, a mechanism sketch, and a mechanism schema; and (2) the process of "bottoming out" in mechanistic model building. To develop these concepts, Machamer, Darden, and Craver (2000:12) first note that "In a complete description of [a] mechanism, there are no gaps that leave specific steps unintelligible; the process as a whole is rendered intelligible in terms of entities and activities that are acceptable to a field at a time." But they then explain that explanatory inquiry using mechanisms is not an all-or-nothing affair, in which every step is always specified. Variability in the representation of mechanisms is possible because:

> Mechanisms occur in nested hierarchies. . . . The levels in these hierarchies should be thought of as part-whole hierarchies with the additional restriction that lower level entities, properties, and activities are components in mechanisms that produce higher level phenomena . . . ." (Machamer et al. 2000:13)

In spite of such nesting, there is a natural *bottoming out* of mechanism-based explanations:

> Nested hierarchical descriptions of mechanisms typically *bottom out* in lowest level mechanisms. These are the components that are accepted as relatively fundamental or taken to be unproblematic for the purposes of a given scientist, research group, or field. Bottoming out is relative: Different types of entities and activities are where a given field stops when constructing mechanisms. The explanation comes to an end, and description of lower-level mechanisms would be irrelevant to their interests. (Machamer et al. 2000:13)

Then, by thinking through the complexity of the nesting of mechanisms, and how scholars represent mechanisms to each other, they develop two related concepts. A *mechanism schema* is a representation of a set of mechanisms in which some known details (or known nested levels) are suppressed for the sake of simplicity. Or, as Machamer et al. (2000:15) state, "a *mechanism schema* is a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities." In contrast, a *mechanism sketch* is quite different:

> A sketch is an abstraction for which bottom out entities and activities cannot (yet) be supplied or which contains gaps in its stages. The productive continuity from one stage to the next has missing

> pieces, black boxes, which we do not yet know how to fill in. A sketch thus serves to indicate what further work needs to be done in order to have a mechanism schema. Sometimes a sketch has to be abandoned in the light of new findings. In other cases it may become a schema, serving as an abstraction that can be instantiated as needed. (Machamer et al. 2000:18)

Within this framework, one can conceive of causal analysis in the social sciences as the pursuit of explanations that bottom out.[18] Although there will inevitably be differences of opinion on how deep an explanation must be to bottom out, it would seem uncontroversial to state that a valid explanation that invokes a mechanism must bottom out at least as far as the observables in the data at hand.[19]

Now, consider the application of this framework to the sort of causal diagrams we have considered so far. Suppose that one uses some form of back-door conditioning to identify a casual effect of D on Y. Suppose, furthermore, that one has done so from within the counterfactual framework, settling on the average causal effect as the parameter of first-order interest. One then necessarily confronts the question that we raised in the beginning of this chapter: Does a counterfactually defined and consistent estimate of the causal effect of D on Y by itself meet the standard of an explanation that bottoms out?

The answer to this question is clear: It depends on what D and Y are, who is conducting the study, and for what purposes. If one wishes to know only how a hypothetical intervention on D would shift Y, and hence has no interest in anything else whatsoever, then bottoming out has been achieved in some minimalist way. The analysis yields up a consistent estimate of the average causal effect of D on Y that holds for the population within which both are observed. In this case, the model $D \rightarrow Y$ is then regarded as merely a mechanism sketch, which suffices to be treated as a sufficiently deep explanation for the purposes at hand.

However, as we have noted from the beginning of this chapter onward, it will often be the case that an estimate of a warranted causal effect of D on Y, grounded in the potential outcomes framework, is properly considered to be

[18] See also the discussion of modularity in Woodward (2003, Chapter 7).

[19] A critical realist could escape from this position in a variety of ways: asserting irreducibility and transcendentalism and then, more specifically, by arguing in the end that the data that one is forced to consider are but a poor reflection of the phenomena that the mechanism truly does explain. For all of these reasons, the lack of observed explanatory power for observed events would then be argued to be untroubling. This position, however, then becomes a variant of an appeal to a hidden but presupposed valid underlying structure, which Woodward convincingly argues cannot be an acceptable explanatory strategy for any field that hopes to resolve its explanatory controversies because "... the appeal to hidden structure makes it too easy to protect one's favored theory of explanation from genuine counterexamples" (Woodward 2003:175). Moreover, if the particularities in the data are merely a poor reflection of the phenomenon that the mechanism is supposed to explain, then presumably whatever generates the mismatch can be encoded in the mechanism that explains both the genuine phenomenon of interest and the process that generates the misleading data.

an insufficiently deep explanation of how D brings about Y. Such a judgment would be appropriate when the interest of the field is in understanding both how interventions on D and interventions on the mechanistic intervening variables that link D to Y would shift Y. In this case, the model $D \rightarrow Y$ is a mechanism sketch that, for the purposes at hand, cannot be regarded as a sufficiently deep explanation. The arrow in the sketch $D \rightarrow Y$ is a black box that must be filled in through further analysis.

If a warranted claim of a counterfactually defined causal effect is properly regarded as insufficiently deep, the recourse is not to abandon the counterfactual model, but rather to investigate the nested mechanism that intercepts the effect of D on Y. Such analysis may initially take the form of a set of alternative conjectured mechanisms. But, ultimately, any such analysis must return to the particular observed relationship between D and Y in the population of interest (or more broadly in multiple well-defined populations for which the question of interest is relevant). Thus, although theoretical creativity may be required, opening up black boxes is, at least in part, an empirical pursuit. At some point, one must specify the counterfactual states for the variables that constitute the mechanisms, and each link thereby specified must be submitted to its own causal analysis.

Consider this process abstractly with reference to a causal diagram, after which we will introduce a real-world example. Suppose that one pursues further theoretical conjecturing and subsequent empirical analysis, and one then determines that the variables A, B, and C constitute an isolated and exhaustive mechanism that identifies the causal effect of D on Y by Pearl's front-door criterion. At this point, one may be tempted to declare that a sufficiently deep causal explanation of the effect of D on Y has been secured. Such a claim may well be true, but it is not guaranteed. It could be that there are further nested mechanistic variables, such that, for example, three additional variables M, N, and O (each of which is a concept of considerable interest to one's peers) are then found to mediate the casual pathway $D \rightarrow A \rightarrow Y$. In this case, the casual pathway $D \rightarrow A \rightarrow Y$ is then itself best regarded in hindsight as merely a component of a mechanism sketch. When M, N, and O are then observed, $D \rightarrow A \rightarrow Y$ is replaced in the mechanism sketch with, for example, one or more related causal pathways, such as $D \rightarrow M \rightarrow A \rightarrow N \rightarrow Y$ and $D \rightarrow A \rightarrow O \rightarrow Y$. In this example, A, B, and C may well identify the causal effect by the front-door criterion, but they do not qualify as a sufficiently deep causal account of how D brings about Y.

As we noted in the first part of this chapter, the progressive deepening of causal explanation through the modeling of intervening processes is entirely consistent with social science tradition. And yet we also claimed that Pearl's front-door criterion can help guide sharpened analysis practices in this regard. To see what we mean, consider the example of Duncan's research on status attainment processes again. As we noted earlier in Chapter 1, Blau and Duncan (1967) deepened the causal account of how parental social status determines offsprings' social status by specifying what most scholars now regard as the most important link: levels of educational attainment.

Thereafter, Duncan and his colleagues then supported and encouraged further work on the process of educational attainment, most importantly the Wisconsin model of status attainment that we introduced earlier in Subsection 1.3.1. This model is a direct extension of Blau and Duncan's research, in which the causal pathways between parental status and offspring's attainment were elaborated by the introduction of intervening variables for significant others' influence and educational aspirations. In fact, in the most important article in this tradition, Sewell et al. (1969) use mechanistic language to introduce the contribution of their study:

> . . . we present theory and data regarding what we believe to be a logically consistent social psychological model. This provides a plausible causal argument to link stratification and mental ability inputs through a set of social psychological and behavioral mechanisms to educational and occupational attainments. One compelling feature of the model is that some of the inputs may be manipulated through experimental or other purposive interventions. This means that parts of it can be experimentally tested in future research and that practical policy agents can reasonably hope to use it in order to change educational and occupational attainments. (Sewell et al. 1969:84)

The Wisconsin model was very favorably received in sociology, as it was considered to be consistent with the basic features of the model of Blau and Duncan (1967) and yet had a claim to greater causal depth.

Even so, as we also noted in Subsection 1.3.1, critics emerged immediately (see also Morgan 2005, Chapter 2, for a more extended summary). The basic argument was that, even if significant others' influence and educational aspirations have causal effects on educational attainment, they are both grounded in part in sources outside of parental status and mental ability (a point the authors of the Wisconsin model recognized). Thus, although significant others' influence and educational aspirations may be helpful to some extent in offering an interpretation of some of the causal process that generates intergenerational correlations of educational attainment, these intervening variables do not qualify as an isolated and exhaustive mechanism that fully accounts for the effects of parental status on offsprings' status.

In this regard, the Blau and Duncan model can be regarded as a mechanism sketch for the status attainment process, and the Wisconsin model can then be regarded as a mechanism-based attempt to deepen its implied explanation. The Wisconsin model was therefore an important step forward, but it was not conclusive and did not settle all further research.

Nearly 40 years later, it is now clear that the Wisconsin model itself is a mechanism sketch. The research community of inequality scholars in sociology seems to have concluded that its pathways have not bottomed out, and much research continues on the processes that generate educational aspirations (as well as whether or not the relationship between aspirations and attainment is

sufficiently explanatory to be useful). Moreover, some scholars (e.g., Goldthorpe 2000) have produced entirely different mechanism sketches for the relationship between parental status and educational attainment. The future of this research tradition is clearly careful empirical analysis that can adjudicate between these rival mechanism sketches, which will be decisive only when alternative mechanism sketches are pushed down to lower-level entities on which critical tests can then be performed.

8.5 Conclusion

Pearl's front-door criterion for the identification of a causal effect is a powerful and illuminating perspective on the explanatory power of mechanisms. It clearly shows that the identification of a causal effect by a mechanism requires that the mechanism be isolated and exhaustive and that its variables be observed. For such a mechanism to count as a sufficiently deep explanation, its causal pathways must be finely enough articulated that it meets whatever standard of bottoming out is maintained in the relevant field of study. If such a standard is not reached, then the causal effect is identified even though it is not accompanied by a sufficiently deep explanation. Instead, the identifying causal pathways represent a mechanism sketch that demands further analysis.

Considering this chapter and the strategies for causal effect estimation from prior chapters, we have come back full circle to our initial presentation of causal modeling options in Chapter 1. We noted there, with reference to Figure 1.3, that a causal effect that is identified by both the back-door criterion and an IV is best explained when it is also identified by an isolated and exhaustive mechanism. This is the gold standard for an explanatory causal analysis, at least until a field decides that a crucial linkage within a mechanism must then be opened up and subjected to its own analysis.

In the next chapter, we turn in a different direction to consider the extent to which over-time data on an outcome variable can be used to identify and estimate a causal effect. One often hears presentations in which scholars remark "I cannot get at causality because I do not have longitudinal data." We will argue in the next chapter that longitudinal data, although very helpful in many cases, are not the panacea that such statements seem to imply. Moreover, we will show that some long-standing techniques that are thought to reveal causal effects are strongly dependent on assumptions that are often entirely inappropriate and sometimes completely unrecognized.

Chapter 9

Repeated Observations and the Estimation of Causal Effects

As discussed in previous chapters, the fundamental problem of causal inference is that an individual cannot be simultaneously observed in both the treatment and control states. In some situations, however, it is possible to observe the same individual or unit of observation in the treatment and control states *at different points in time*. If time has no effect, then the causal effect of a treatment can be estimated as the difference between an individual's outcome under the control at time 1 and under the treatment at time 2. The assumption that time (and thus age for individuals) has no effect is often heroic. If, however, individuals' outcomes evolve in a predictable way, then it may be possible to use the longitudinal structure of the data to predict the counterfactual outcomes of each individual.

In this chapter, we will assume that the time at which treatment occurs is fixed. We will again focus on the consequences of nonrandom selection of individuals into the treatment and control groups. We will not, however, consider scenarios in which the specific timing of treatment is endogenous. This situation is considerably more complex because a treatment indicator must be modeled for every time period, recognizing that selection of the treatment in any single time period is not only a function of individual characteristics but also of previous decisions and expectations of future decisions.[1]

We begin our discussion with the interrupted time series (ITS) model. This model is the simplest design in which over-time data are utilized, as the data consist of a single person/unit observed at multiple points in time. The goal is to determine the degree to which a treatment shifts the underlying trajectory

[1] For a discussion of the recent literature in this active area, see van der Laan and Robins (2003), as well as citations therein to the important methodological and substantive work of Robins.

of an individual's values for an outcome. Thereafter, we consider the regression discontinuity design. Although not a case in which we have repeated observations on a single individual or unit, the structure of the regression discontinuity (RD) design is sufficiently similar to that of the ITS that it is useful to present it here as well. For the RD design, we also consider the case of fuzzy assignment, which amounts to using IV methods to correct for possible imprecision in the treatment assignment criteria.

Thereafter, we consider panel data: multiple observations over time on multiple individuals or units. We first examine the adequacy of traditional two-period pretreatment/posttreatment adjustment strategies. We show that such methods, despite their considerable popularity, are in general inadequate for making causal inferences, unless the researcher is willing to make strong and generally untestable assumptions about how the outcome evolves over time across individuals. Then, we consider a more comprehensive model-based approach. The key here is to be explicit about the evolutionary dynamics of the outcome and how selection into the treatment depends on these dynamics. This type of strategy typically requires multiple pretreatment waves of data. With data over a sufficient number of time periods, it is possible to test the appropriateness of different models.

9.1 Interrupted Time Series Models

To assess the causal effect of a treatment within an ITS model, a time series model is typically estimated:

$$Y_t = f(T) + D_t b + e_t, \tag{9.1}$$

where Y_t is some function in time [which is represented by $f(T)$ on the right-hand side], D_t is a dummy variable indicating whether the treatment is in effect in time period t, and e_t is time-varying noise. The basic strategy of an ITS analysis is to use the observed trajectory of Y prior to the treatment to forecast the future trajectory of Y in the absence of the treatment (see the introductions in Marcantonio and Cook 1994; McDowall, McCleary, Meidinger, and Hay 1980; Shadish, Cook, and Campbell 2001).[2]

The primary weakness of an ITS model is that the evolution of Y prior to the treatment may not be a sufficiently good predictor of how Y would evolve in the absence of treatment. Consider the trajectory of Y in the hypothetical example depicted in Figure 9.1. The solid line represents the observed data on Y, and the time of the introduction of the treatment is indicated on the horizontal axis. Here, the future (counterfactual) evolution of Y in the absence

[2]Potential outcome notation may be useful here, but it is not needed for understanding the ITS design. Before the treatment is introduced, Y_t is equal to Y_t^0. After the treatment is introduced, Y_t then switches over to Y_t^1. Under this structure, the causal effect is then the difference between Y_t and Y_t^0 in the treatment period. But, because Y_t^0 cannot be observed in the treatment period, an ITS design forecasts the values of Y_t^0 in the treatment period by extrapolating from the trajectory of Y_t in the pretreatment period.

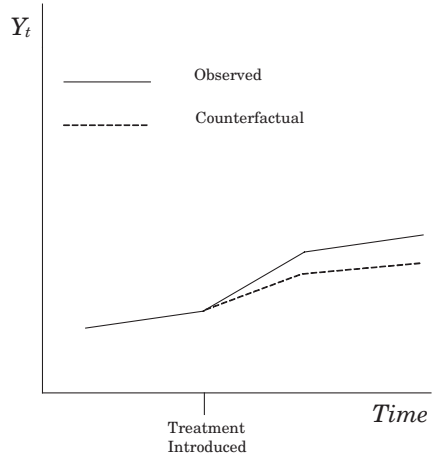


Figure 9.1: Trajectories of the observed and counterfactual outcomes for an ITS model.

of treatment is represented by the dashed line. Clearly, this counterfactual trajectory would be poorly predicted by a straightforward linear extrapolation from the observed data before the treatment. In fact, in this case, assuming that the counterfactual trajectory followed such a linear trajectory would result in substantial overestimation of the treatment effect. In general, if the future in the absence of treatment is quite different from the past, then it may be impossible to successfully estimate the causal effect with an ITS model.

In terms of estimation, an ITS model involves no issues beyond those found in a standard time series analysis. There, the key is that the errors e_t are likely to be correlated over time. If we use least squares regression, the parameter estimates will be consistent, but the standard errors and any hypothesis tests based on them will be wrong. This problem can be especially acute in many cases, as the number of data points in time series datasets can be small. We will not go into all of the issues involved in estimating time series models, as there are many books that cover the topic in depth (e.g., Hamilton 1994 and Hendry 1995).

Instead, we will illustrate the basic thinking behind an ITS analysis with an example from Braga, Kennedy, Waring, and Piehl (2001), which is presented in Figure 9.2. The data presented there are the trend in the monthly youth homicide rate in Boston between June 1991 and May 1998. Braga and his colleagues were interested in evaluating whether an innovative program, "Operation Ceasefire," initiated by the Boston Police Department in June of 1996, reduced youth homicides.

Operation Ceasefire involved meetings with gang-involved youth who were engaged in gang conflict. Gang members were offered educational, employment, and other social services if they committed to refraining from gang-related

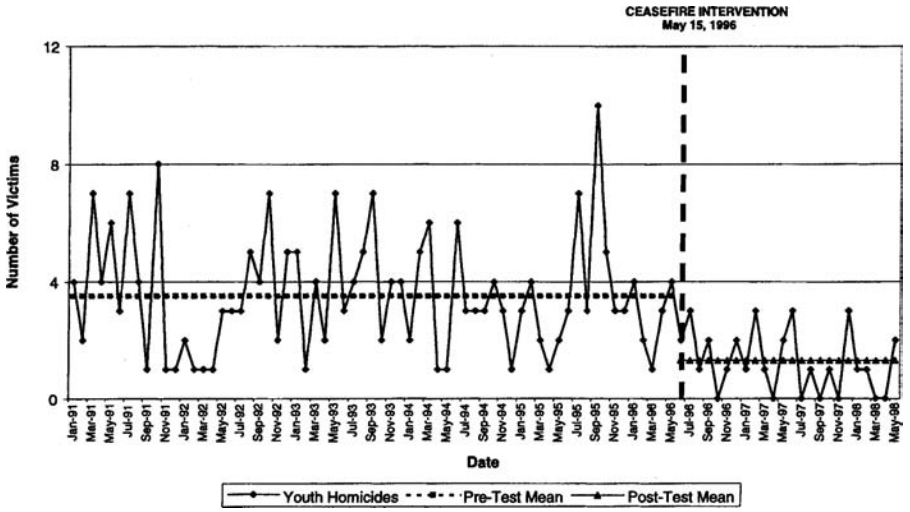


Figure 9.2: Monthly youth homicide rates in Boston, 1991–9. (*Source*: Figure 2, page 205, of Braga, Anthony A., David M. Kennedy, Elin J. Waring, and Anne Morrison Piehl. 2001. "Problem-Oriented Policing, Deterrence, and Youth Violence: An Evaluation of Boston's Operation Ceasefire." *Journal of Research in Crime and Delinquency* 38:195–225. Reprinted by permission of Sage Publications, Inc.)

deviance. At the same time, the police made it clear that they would use every legal means available to see that those who continued to be involved in violent behavior were sent to prison (see Kennedy 1997 for a more detailed description).

The vertical line in Figure 9.2 marks June 1996, the date at which Operation Ceasefire was initiated. The two horizontal lines indicate, respectively, the mean level of youth homicides before and after June 1996. As can be seen in Figure 9.2, there appears to be an abrupt and large drop in the level of youth homicide in Boston immediately after the implementation of Operation Ceasefire.

Braga and his colleagues carry out a more formal analysis using standard time series techniques (although because their dependent variable is a count variable – number of youth homicides in a month – they use a Poisson regression model). In their first model, they adjust only for seasonal effects by using dummy variables for month and a linear term for time. Inclusion of the time trend is particularly important. Although it is not clear in Figure 9.2, there is a slight downward time trend in the homicide rate. For this model, Braga and his colleagues report a large negative and statistically significant effect of Operation Ceasefire on youth homicides.

In general, a researcher would usually prefer a situation in which the underlying time trend and the hypothesized program effect are in the opposite directions. In such a case, disentangling the program effect and the time trend is far easier, strengthening any warrant for a claimed causal effect. However,

for this example, the underlying time trend and the expected program impact move in the same direction. Thus, a researcher should be concerned about the ability to accurately estimate the size of the program effect in the presence of the underlying trend. Recall that our earlier Figure 9.1 illustrated a similar but possibly even more problematic situation. Not only does the time trend move in the same direction as the effect in that hypothetical example, but there is a change in the time trend in the same direction as the treatment, making accurate estimation of the treatment impossible.

Braga et al. (2001) is a very high-quality example of how to carry out an ITS analysis, in part because they consider these issues carefully. They offer four types of supplemental analysis, each of which is broadly applicable to all ITS designs, and each of which can be used to strengthen the warrant for a causal assertion.

First, they considered additional dependent variables that Operation Ceasefire should have affected. Specifically, they analyzed whether the program affected the number of gun assaults and reports of gun shots fired. For these dependent variables, they find an even larger program impact. Their analysis would have been strengthened further if they also had considered dependent variables that Operation Ceasefire should not have affected as much (e.g., number of robberies and incidents of domestic violence). Here, evidence of an impact would suggest that factors other than the program itself at least partially accounted for the observed drop in youth homicides.

Second, they focused their hypothesis and considered it within meaningful subgroups. In particular, they analyzed the level of gun assaults in police district B-2, the district where gang violence was the highest in the early 1990s and Operation Ceasefire was most active. As they predicted, the program effect was larger there than in the city as a whole. If there had been districts with high levels of youth violence where Operation Ceasefire was inactive, it would have been useful to have tested for a program impact. If evidence were found that the program had an impact in these districts, it would suggest that something other than the program was responsible for the observed decline in youth homicides, gun assaults, and gun shots fired. Unfortunately, at least for the analyst, almost all youth violence in Boston occurs in three adjacent police districts, all districts in which Operation Ceasefire was active. As a result, such an analysis was not possible.

Third, Braga and his colleagues included additional adjustment variables in their time series models in order to capture the underlying time trend as well as the year-by-year variability before and after the introduction of the treatment. These time-varying covariates included unemployment rates, the size of the youth population, the robbery rate, the homicide rate for older victims, and the drug-related arrest rate. The advisability of adjusting for the latter three variables, given that they are certainly endogenous, is questionable. Nonetheless, conditioning on these additional variables produced little change in their estimate of the program impact on any of their dependent variables.

As a final method to assess the robustness of their results, Braga and his colleagues compared the time trend in homicides in Boston with the time trend

in 41 other cities where no interventions were implemented. Their goal was to determine whether other cities experienced declines in rates as abrupt as the one observed in Boston. The explanation of Braga and his colleagues for the abruptness of the decline in Boston – in fact, a decline that emerged in only two months – was that word got out on the street quickly that the police meant business.

For the other cities that they considered, homicide rates fell throughout the 1990s in many of them. But the declines were, with the exception of those of New York, substantially smaller than the decline in Boston. Braga and his colleagues then showed that abrupt declines did occur in five other cities, but the exact timing of these abrupt declines was different than in Boston. This evidence raises perhaps the greatest doubt about the assertion that Operation Ceasefire causally reduced youth homicide rates in Boston because there is no clear explanation for why these abrupt declines occurred elsewhere either. And, because it may be implausible that Operation Ceasefire's effect could have fully taken hold in as short as two months, the possibility exists that the decline in homicide rates and the introduction of Operation Ceasefire were coincidental.

If Braga and his colleagues had carried out their evaluation a number of years later, they could have implemented one additional strategy. In 1999, Operation Ceasefire was cut back and then terminated. If Braga and his colleagues had performed their analysis through 2006, they could have examined whether the termination of the program resulted in an increase in homicide rates. In fact, since 1999, youth homicide rates have risen such that, by the summer of 2006, they were at nearly the same level as in the early 1990s. Although no recent formal analysis has been carried out with time-varying covariates (at least of which we are aware), the recent increase in the youth homicide rate certainly provides *prima facie* evidence for Operation Ceasefire's impact.

This example nicely illustrates the variety of general strategies that are potentially available to strengthen an ITS analysis:

1. Assess the effect of the cause on multiple outcomes that should be affected by the cause.

2. Assess the effect of the cause on outcomes that should not be affected by the cause.

3. Assess the effect of the cause within subgroups across which the causal effect should vary in predictable ways.

4. Adjust for trends in other variables that may affect or be related to the underlying time series of interest.

5. Compare the time trend with the time trend for other units or populations to determine whether breaks in the time series are likely to occur in the absence of the cause.

6. Assess the impact of termination of the cause in addition to its initiation.

These strategies are often available for other types of analysis, and they are also widely applicable to all forms of data analysis that attempt to infer causation from over-time relationships.

9.2 Regression Discontinuity Designs

An RD design is very similar to an ITS design, except that the treatment is a function of a variable other than time. RD applies in situations in which treatment assignment is a discontinuous variable (or highly nonlinear function of a continuous variable) that is believed to also directly affect the outcome. RD has been applied to a variety of problems: the effect of student scholarships on career aspirations (Thistlewaite and Campbell 1960), the effect of unemployment benefits for former prisoners on recidivism (Berk and Rauma 1983), the effect of financial aid on attendance at a particular college (Van der Klaauw 2002), the effect of class size on student test scores (Angrist and Lavy 1999), and the willingness of parents to pay for better schools (Black 1999).

Campbell was the first to propose the RD method (see Trochim 1984 and Hahn, Todd, and Van der Klaauw 2001). It is most easily understood with an example. Here we consider the example of Mark and Mellor (1991), which is discussed also by Shadish et al. (2001). Mark and Mellor were concerned with the effect that an event of high personal salience may have on hindsight bias – the claim that an event was foreseeable after it occurred. In their study, they examine the effects of being laid off in a large manufacturing plant. Individuals with less than 20 years of seniority were laid off; those with 20 years or more were not. Figure 9.3 shows the relationship between seniority and hindsight bias.

As shown in Figure 9.3, there is an abrupt discontinuity in the relationship between foreseeability and seniority at the point in seniority where individuals were laid off. Those who were not laid off (i.e., individuals with higher levels of seniority) were more likely to see the event as foreseen. Note here that the effect of seniority within the group of those who were laid off and those who were not is in the opposite direction of the treatment effect: laid-off individuals with the highest seniority were less likely to believe that the layoffs were foreseeable. Because these effects are in the opposite direction, Mark and Mellor have greater confidence that being laid off increases the likelihood that an individual thinks an event is foreseeable.

In general, an RD design can be estimated in the same way as an ITS model because most of the same issues apply. One key and helpful difference is that in most RD designs individuals are sampled independently. As a result, the problem of correlated errors in an ITS design is absent in an RD design.

A generalization of the RD design, known as the fuzzy RD design, has received recent attention. Here, treatment is a function of an assignment process in which there is some error that may be associated with Y. Consider the causal diagram in Figure 9.4. Here, $f(Z)$ represents the assignment rule to D as a function of Z. Note that, if this assignment rule were perfect, then we

Figure 9.3: Foreseeability of a layoff as an example of an RD design. (*Source*: Figure 1 of Mark, Melvin M. and Steven Mellor. 1991. "Effect of Self-Relevance of an Event on Hindsight Bias: The Foreseeability of a Layoff." *Journal of Applied Psychology* 76:569–77. Reprinted with permission.)

could simply condition on Z, which would allow us to block the back-door path $D \leftarrow f(Z) \leftarrow Z \rightarrow Y$.[3]

In the case of fuzzy RD, the assignment rule is not deterministic because D is a function of both $f(Z)$ and U_D, and because U_D and U_Y may be correlated. The situation is directly analogous to an experiment with noncompliance in which noncompliance is thought to be nonrandom. By conditioning on Z, $f(Z)$ can be used as an instrument for D because it is uncorrelated with both U_D and U_Y and affects Y only through D. Conditioning on Z is necessary to eliminate the back-door path $f(Z) \leftarrow Z \rightarrow Y$. This will work, however, only if $f(Z)$ is some nonlinear function of Z, so that $f(Z)$ and Z are not linearly dependent.[4]

Angrist and Lavy (1999) use the fuzzy RD design to study the effects of class size on student test performance in the Israeli public school system. In the Israeli school system, an additional class is added to a grade within a school when the existing classes contain more than 40 students. Within any school, this intervention creates a discontinuous relationship between enrollment and class size. This discontinuity allowed Angrist and Lavy to create a nonlinear

[3]In other words, if assignment were perfect, there would be no potential for a back-door path through U_D and U_Y (i.e., the bidirected edge between U_D and U_Y would not be present).

[4]As Imbens and Rubin (1997) show, one can deal with noncompliance by treating the intention to treat indicator as an instrument, Z, for actual treatment.

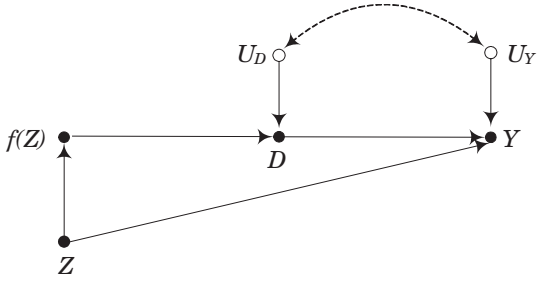


Figure 9.4: An example of a fuzzy RD design.

function of enrollment that can then be used as an instrument for class size. They find that class size has a substantial effect on test performance for fourth and fifth graders, but not for third graders.

As shown by this example, RD designs are simple and can be convincing. But the same weaknesses of ITS designs are present. Counterfactual values must be extrapolated from observed data below/before and above/after the application of the treatment. If the assumptions built into the chosen method of extrapolation are unreasonable, then causal effect estimates will be incorrect.

9.3 Panel Data

A severe limitation of time series data is that we have data on only a single unit over time. This limitation is problematic for two reasons. First, because of the relatively small amount of data, it is very common to obtain large standard errors, making inference about the presence of a treatment effect difficult. Second, because we do not observe the treated unit of analysis in the control state after the treatment is introduced, the only way to impute the counterfactual outcome is to assume that the future can be predicted accurately from the past. This assumption is generally untestable.

Panel data, where multiple individuals or units are observed over time, may solve both problems. Assuming that each individual's time series is relatively long, separate ITS analyses could be conducted for each individual and then pooled to form an average causal effect estimate. Moreover, because individuals receive the treatment at different times (and ages) or do not receive the treatment at all, it is possible to observe how Y_t^0 changes over time for some individuals after others have received the treatment. To the degree that time affects Y_t^0 for individuals in the treatment and control groups similarly, it may be possible to make reasonable predictions about how the counterfactual values of Y_t^0 would have evolved over time for individuals in the treatment group during the posttreatment period. If these predictions are reasonable, inferences about the causal effect of the treatment, D, can be made by comparison of the observed Y_t^1 for those who are treated ($D = 1$) with the predictions of counterfactual values of Y_t^0 for those who are treated ($D = 1$). The crux of the

matter, of course, is how to use observed values of Y_t^0 among the control group to make reasonable predictions about future counterfactual values of Y_t^0 for the treatment group.

For the remainder of this chapter, in which we must deal now with some quantities that vary only over individuals, others that vary only over time, and others that vary over both, we will subscript with i for individuals and with t for time. In some cases, such subscripting is redundant. For example, in prior chapters, we have represented the causal effect as δ, recognizing that this varies over individuals. Now, however, we will represent the individual-level causal effect as δ_i, so that it is clear that in this form we are assuming that it does not vary with time. For a time-varying causal effect, we would instead need to subscript it as δ_{it}.

Moreover, we will also distinguish between two different treatment indicator variables. D_{it} is a time-varying dummy variable that indicates whether individual i receives the treatment in time period t. In contrast, D_i^* is a time-constant dummy variable that indicates whether individual i ever receives the treatment at any point in the time span under study. D_{it} is best thought of as a treatment exposure indicator variable, and D_i^* is best thought of as a treatment group indicator variable. In some sense, this distinction reveals the potential value of panel data. For a cross-sectional study in which all observation occurs in a single time period, $D_{it} = D_i^*$. However, a panel dataset over multiple time periods allows a researcher to consider how treatment exposure (D_{it}) can be separated from treatment group membership (D_i^*) and then exploit this difference to estimate the causal effect.

This distinction has consequences for the relationship between observed and potential outcomes. For the control group, $Y_{it} = Y_{it}^0$. But, for the treatment group, $Y_{it} = Y_{it}^0$ before treatment exposure, and $Y_{it} = Y_{it}^1$ after treatment exposure. More formally, Y_i is defined with reference to D_{it}, such that $Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$ for all t. As a result, if treatment exposure occurs between time period t' and t'', then $Y_{it'} = Y_{it'}^0$ for those in the control group ($D_i^* = 0$) and for those in the treatment group ($D_i^* = 1$). But, in time period t'', D_{it} and D_i^* diverge, so that $Y_{it''} = Y_{it''}^0$ for those in the control group ($D_i^* = 0$) and $Y_{it''} = Y_{it''}^1$ for those in the treatment group ($D_i^* = 1$).

Finally, as will become clear, we consider only treatment effects where all heterogeneity of δ_i across the population is random. This restriction allows us to avoid the complications (discussed extensively in Chapters 4 and 5) that arise when heterogeneity of a causal effect exists across levels of conditioning variables. All of those discussions apply here as well, but for now we consider the specific complications of panel data models.

9.3.1 Traditional Adjustment Strategies

The most common situation in panel data analysis consists of nonequivalent treatment and control groups and two periods of data, where the first wave of data is from pretreatment time period t and the second wave of data is from posttreatment time period $t + 1$. Such two-period, pretreatment/posttreatment

panel data are sometimes thought to be a panacea for not having a randomized experiment. Typically, it is assumed that changes over time in the control group can be used to adjust the changes observed for the treatment group, with the net change then representing a consistent estimate of the causal effect of the treatment.

Unfortunately, the situation is far more complicated. There are an infinite number of ways to adjust for initial differences, and different methods of adjustment give estimates that sometimes differ dramatically. Specifically, as we will show in this section, by choosing a particular adjustment technique, any estimate that a researcher may want can be obtained.

Consider the two most common methods used in the analysis of pretreatment/posttreatment data, usually referred to as *change score* and *analysis of covariance* models. These two models are equivalent to estimating the following two equations for the observed Y_{it} with OLS regression:

$$\text{change score:} \qquad Y_{it+1} - Y_{it} = a + D_i^* c + e_i, \tag{9.2}$$

$$\text{analysis of covariance:} \qquad Y_{it+1} = a + Y_{it} b + D_i^* c + e_i. \tag{9.3}$$

These two equations provide different means of adjustment for Y_{it}. In the change score model, one adjusts Y_{it+1} by subtracting out Y_{it}. For the analysis of covariance model, one adjusts Y_{it+1} by regressing it on Y_{it}.[5]

Consider a real-world example that shows how these two models can yield different results. After decades of studying the environmental and genetic determinants of intelligence, considerable controversy remains over their relative effects on life outcomes. As discussed in Devlin, Fienberg, Resnick, and Roeder (1997) and other similar collections, these debates resurfaced after the publication of *The Bell Curve: Intelligence and Class Structure in American Life* by Herrnstein and Murray in 1994. Even though existing reviews of the literature emphasized the malleability of IQ (see Ceci 1991), Herrnstein and Murray concluded in their widely read book that:

> Taken together, the story of attempts to raise intelligence is one of high hopes, flamboyant claims, and disappointing results. For the foreseeable future, the problems of low cognitive ability are not going to be solved by outside interventions to make children smarter. (Herrnstein and Murray 1994:389)

As discussed in Winship and Korenman (1997), the weight of evidence supports the claim that education determines measured intelligence to some degree, even though debate remains on how best to measure intelligence.

[5] Also, it bears mentioning that when we present alternative equations such as Equations (9.2) and (9.3) in this chapter, we give generic notation – such as a, b, and c – to standard regression coefficients such as intercepts and treatment effect estimates. We do the same, in general, for regression residuals, and so on. We do not mean to imply that such quantities are equal across equations, but it is cumbersome to introduce distinct notation for each coefficient across equations to make sure that we never imply equality by reusing generic characters such as a, b, c, and e.

Consider now a very specific question associated with this controversy: What is the causal effect of a twelfth year of education on measured IQ? The following results, based on data from the National Longitudinal Survey of Youth, show how different the results from change score and analysis of covariance models can be. For both models, IQ is measured in the twelfth grade for all individuals who meet various analysis-sample criteria ($N = 1354$). IQ is then measured again two years later, and the treatment variable is high school completion. A change score analysis yields a treatment effect estimate of 1.318 (with a standard error of .241) and an analysis of covariance model yields a treatment effect estimate of 2.323 (with a standard error of .217).[6] These estimates are quite different: The analysis of covariance model suggests that the effect of a twelfth year of schooling on IQ is 76 percent larger than that of the change score model (that is, $[2.323 - 1.318]/1.318$). Which estimate should one use? Before we discuss how (and if) one can choose between these two types of traditional adjustment, consider a more general, but still simple, hypothetical example.

Panel Data Demonstration 1

For this hypothetical example, there are three points in time, $t = \{1, 2, 3\}$, and treatment selection is assumed to occur between $t = 2$ and $t = 3$. We present three alternative treatment selection rules. For the first variant, treatment selection is a random variable D^* with the probability that D^* is equal to 1 specified as a logistic distribution in A and B:

$$\Pr[D_i^* = 1|A_i, B_i] = \frac{\exp(-2 + 2A_i + 2B_i)}{1 + \exp(-2 + 2A_i + 2B_i)}, \tag{9.4}$$

where A and B are bivariate normal with

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & .3 \\ .3 & 1 \end{bmatrix}. \tag{9.5}$$

For the second variant, the treatment selection probability is a function in A, B, and Y_2^0:

$$\Pr[D_i^* = 1|A_i, B_i, Y_{i2}^0] = \frac{\exp\{-2 + 2A_i + 2B_i + .1(Y_{i2}^0 - E_N[Y_{i2}^0])\}}{1 + \exp\{-2 + 2A_i + 2B_i + .1(Y_{i2}^0 - E_N[Y_{i2}^0])\}}, \tag{9.6}$$

where A and B are defined as for the first variant. And for the third variant, the treatment selection probability is now a function in A, B, and δ:

$$\Pr[D_i^* = 1|A_i, B_i, \delta_i] = \frac{\exp\{-2 + 2A_i + 2B_i + 1(\delta_i - E_N[\delta_i])\}}{1 + \exp\{-2 + 2A_i + 2B_i + 1(\delta_i - E_N[\delta_i])\}}, \tag{9.7}$$

[6] For completeness, we report additional features of these models here. Each was estimated with three other covariates: age, year of the test, and a standardized measure of socioeconomic status. The coefficients on these three variables were −.184, −.755, and −.429 for the change score model and −.966, −.495, and 2.047 for the analysis of covariance model. The R^2 was .06 for the change score model and .68 for the analysis of covariance model, in which the lag coefficient on IQ in the twelfth grade was .623.

where A and B are, again, defined as for the first variant.

Finally, the potential outcomes y_{it}^0 are constructed as

$$\begin{aligned} y_{i1}^0 &= 98 + 2a_i + 2b_i + 5v_{i1}^0, \\ y_{i2}^0 &= 99 + 2a_i + 2b_i + 5v_{i2}^0, \\ y_{i3}^0 &= 100 + 2a_i + 2b_i + 5v_{i3}^0, \end{aligned} \tag{9.8}$$

where the v_t^0 are multivariate normal with

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}. \tag{9.9}$$

Because of the timing of treatment, the values y_{i1}^0 and y_{i2}^0 become the observed values y_{i1} and y_{i2} for all individuals. But because treatment exposure occurs between $t = 2$ and $t = 3$, the values y_{i3}^0 become observed values y_{i3} for those in the control group whereas the values of y_{i3}^0 become counterfactual for those in the treatment group.

The potential outcomes y_{it}^1 of individuals are then constructed as $y_{it}^0 + \delta_i$:

$$\begin{aligned} y_{i1}^1 &= y_{i1}^0 + \delta_i, \\ y_{i2}^1 &= y_{i2}^0 + \delta_i, \\ y_{i3}^1 &= y_{i3}^0 + \delta_i, \end{aligned} \tag{9.10}$$

where δ is normally distributed with mean 10 and variance 1. The values y_{i1}^1 and y_{i2}^1 are counterfactual, in the sense that no one is exposed to the treatment in time periods 1 and 2, and thus no one receives the boost δ_i that is attributable to the treatment. But, for $t = 3$, the values y_{i3}^1 become the observed values y_{i3} for those in the treatment group and remain counterfactual values for those in the control group.

For this example, Table 9.1 shows that the analysis of covariance and change score models often give very different results. Consider for now only the first column of the results, where treatment assignment is based on A and B only. The true average treatment effect is the same as the average treatment effects for the treated and for the untreated because treatment selection is not based on expectations of the causal effect. However, the naive estimate of 14.69 is upwardly biased because A and B are positively associated with both the outcome and treatment assignment. We also offer two other cross-sectional estimators just below the naive estimator. Note that if we observe both A and B, then we can estimate the average causal effect consistently as 10.00 because in this case selection is on the relevant observables only. However, if we do not observe B, then a regression of Y_3 (our posttreatment observed outcome) only on D^* and A will overestimate the true average treatment effect (in this case with a value of 12.30, which is between the true average treatment effect and the naive estimate). Here, B is regarded as an omitted variable, which is correlated with both Y_3 and D^*. It is precisely in these situations when researchers turn to

Table 9.1: Hypothetical Treatment Effect Estimates from Analysis of Covariance and Change Score Models

	Treatment selection		
	Eq. (9.4)	Eq. (9.6)	Eq. (9.7)
	True average treatment effects		
Average treatment effect, $E[\delta]$	10.00	10.00	10.00
Average treatment effect for the treated, $E[\delta \mid D^* = 1]$	10.00	10.00	10.30
Average treatment effect for the untreated, $E[\delta \mid D^* = 0]$	10.00	10.00	9.87
	Estimators of the coefficient on D		
Cross-sectional estimators:			
Naive estimator	14.69	15.23	14.82
Regression of Y_3 on A and D^*	12.30	13.08	12.52
Regression of Y_3 on A, B, and D^*	10.00	10.95	10.42
Analysis of covariance estimators:			
Regression of Y_3 on Y_2 and D^*	11.90	11.89	12.14
Regression of Y_3 on Y_2 , A, and D^*	11.03	11.00	11.33
Regression of Y_3 on Y_2 , A, B, and D^*	10.00	10.00	10.42
Change score estimator:			
Regression of $(Y_3 - Y_2)$ and D^*	10.00	9.48	10.30

a pretreatment outcome measure (or pretest) in order to solve such omitted-variable-bias problems.

Now, consider the final four rows of Table 9.1. For the first two analysis of covariance models, the treatment effect estimates are upwardly biased at 11.90 and 11.03. The third estimate is not, but this result arises because the regression equation includes both A and B, such that no omitted-variable bias is present. In contrast, the change score model yields an unbiased estimate of the outcome of 10.00. Thus, it is easy to see here that a researcher should favor the change score model over the analysis of covariance model. But such a conclusion is clear here only because we know by construction that 10.00 is the correct answer. Typically, a researcher would be left with two estimates – such as 11.90 and 10.00 – with no clear sense of which is correct.

For these simulated data, the change score model gives the correct answer because the Y^0 outcomes for the treatment group and the control group change at the same rate over time [see the terms 98 through 100 in the lines of Equation (9.8)] and the treatment effect does not vary with time [see the same term

δ_i in all lines of Equation (9.10)]. Such uniformity of time trends cannot be assumed to prevail in general. Finally, to see what is coming in later sections of this chapter, note also that the situation is even more complex when treatment selection is on Y_2^0 or δ, as is shown in the second and third columns of Table 9.1.

Return to the question that motivates this hypothetical example, where we are not in the fortunate situation of knowing what the true average treatment effect is. All we have are alternative treatment effect estimates suggested by a change score model and an analysis of covariance model. How should one choose between them? There are at least three possible ways to decide:

1. Choose the method that gives us the results we want.
2. Choose based on the nature of problem. As Allison (1990) suggests: If selection is based on fixed characteristics, use change score analysis. If selection is based on the dependent variable, use an analysis of covariance.
3. Use the data to determine which model, if either, is appropriate (as in Heckman and Hotz 1989).

Hopefully, the first decision method is not a serious consideration. The second is a better option, in that it at least suggests that one should begin to think through the specific nature of the problem of interest. The third decision method appears most promising, at least at face value. However, in some cases (perhaps most where these two designs are utilized), we have data from only two points in time. This is the situation for the estimate of the causal effect of a twelfth year of schooling on IQ. It would also be the case for the hypothetical example if we had data only from time periods 2 and 3.

Unfortunately, with only two time periods, both the change score and analysis of covariance models are just identified. As such, the data cannot be used to test whether one of the two methods is more appropriate. As we will explain when we discuss model-based approaches later, we need at least two periods of pretreatment data to carry out a test of model fit. We would be able to perform such tests on the hypothetical example, if we used the data from time period 1 as well.

For now, consider the case in which we continue to have data from only one pretreatment time point, t, and one posttreatment time point, $t+1$ (i.e., time periods 2 and 3 for the hypothetical example). Consider the implicit assumptions that are made if it is asserted that either the change score model or the analysis of covariance model represents a consistent estimator of the treatment effect of interest:

- The change score model assumes that, *in the absence of treatment*, any difference between the expectations of Y for those in the treatment and control groups remains constant over time. In the counterfactual framework, the assumption is that $E[Y_{it}^0|D_i^* = 1]$ and $E[Y_{it}^0|D_i^* = 0]$ differ by the same constant k in every time period t.

- The analysis of covariance model assumes that, *in the absence of treatment*, any difference between the expectations of Y for those in the the treatment and control groups shrinks by a multiplicative factor r between each time period. The implication here is that, after enough time, the analysis of covariance model assumes that there would be no difference in the expected outcomes for the treatment and control groups if the treatment is not introduced. In the counterfactual framework, the assumption is that the difference between $E[Y^0_{it}|D^*_i = 1]$ and $E[Y^0_{it}|D^*_i = 0]$ declines by the same amount in every period of time, such that, by time period $t = \infty$, $E[Y^0_{it=\infty}|D^*_i = 1]$ will be equal to $E[Y^0_{it=\infty}|D^*_i = 0]$.

Now, consider a general equation that represents the value of the average treatment effect:

$$\begin{aligned} E[\delta_{it+1}] &= \big(E[Y^1_{it+1}|D^*_i = 1] - E[Y^0_{it+1}|D^*_i = 0]\big) \\ &\quad - \alpha\left(E[Y^0_{it}|D^*_i = 1] - E[Y^0_{it}|D^*_i = 0]\right) \end{aligned} \tag{9.11}$$

for some unknown value α. The term in the first set of parentheses is equal to the naive estimator in the posttreatment period $t + 1$ [see Equation (2.7)]. It is therefore equal to $E[Y_{it+1}|D^*_i = 1] - E[Y_{it+1}|D^*_i = 0]$. The second term is an adjustment factor: How much of the initial difference in time period t of the expectation of Y^0 for those in the treatment and for those in the control group should be subtracted out? The piece of this term inside the second set of parentheses is equal to $E[Y_{it}|D^*_i = 1] - E[Y_{it}|D^*_i = 0]$ because Y^0 is observed as Y for both the treatment group and the control group in time period t. Thus Equation (9.11) can be rewritten as

$$\begin{aligned} E[\delta_{it+1}] = (&E[Y_{it+1}|D^*_i = 1] - E[Y_{it+1}|D^*_i = 0]) \\ &- \alpha(E[Y_{it}|D^*_i = 1] - E[Y_{it}|D^*_i = 0]), \end{aligned} \tag{9.12}$$

and its right-hand side can then be written even more simply with words as

$$(\text{posttreatment difference in } Y) \;-\; \alpha\,(\text{pretreatment difference in } Y)\,. \tag{9.13}$$

The change score model and analysis of covariance model can be seen as alternative methods that make very different and very rigid assumptions about the value of α in Equations (9.11)–(9.13). Change score models implicitly assume that $\alpha = 1$. In contrast, analysis of covariance models implicitly assume that $\alpha = r$, where r is the intraclass correlation between Y_{t+1} and Y_t (i.e., r is the correlation for individuals' Y_{it+1} and Y_{it}). In other contexts, this correlation r is known as Y's reliability. If other covariates are included in the model, then analysis of covariance models estimate a coefficient on Y_t that can be thought of as the intraclass correlation between residualized variants of Y_{t+1} and Y_t, where their common linear dependence on the covariates has been purged. In this case, the basic intuition still holds, but the coefficient on Y_t is not the reliability of Y anymore, but rather a conditional variant of it.

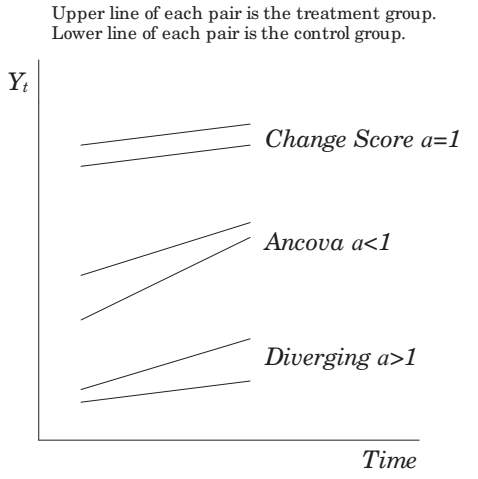


Figure 9.5: Three examples of possible trajectories for the treatment and control groups of the outcome in the absence of treatment.

Researchers often believe that, because r is estimated from the data, an analysis of covariance model is superior to a change score model. This position is wrong. To be sure, an analysis of covariance model does estimate r from the data on Y. But, when using the results from an analysis of covariance model, a researcher is assuming implicitly that α in Equations (9.11)–(9.13) should be equal to r (or some conditional variant of it). As we noted earlier, this is an assumption that, in the absence of treatment, the difference between the expectation of Y^0 for those in the treatment group and those in the control group would shrink by the factor r between t and $t+1$ for every value of t.

Consider the graph presented in Figure 9.5, which is adapted from Judd and Kenny (1981; Figure 6.4). It presents three scenarios for changes over time in the expected value of Y for the treatment and control groups in the absence of treatment (i.e., Y^0). For each pair of lines, the upper line is for the treatment group and the lower line is for the control group. Most importantly, this graph shows a possibility (the bottom pair of lines) that is not consistent with either a change score analysis or an analysis of covariance. Here, in the absence of treatment, the expectations of Y^0 for the treatment group and for the control group diverge over time. There are many possible examples of this situation, but the most famous is represented by situations described in the gospel of Matthew, where it is written "Unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath" (quoted in Merton 1968, who is credited with introducing this version of the idea into the social sciences).

The key point is that different methods of estimation make different implicit assumptions about how the difference in the expectations between the treatment

and control groups would change with time in the absence of treatment (which, in the counterfactual tradition, are different assumptions about treatment and control group differences in the evolution of Y^0). Researchers typically use either change score models or analysis of covariance models without taking note of these assumptions. Nevertheless, these assumptions can be very consequential, as we now show in a general way.

Our claim that any assumption about α is potentially reasonable can be justified by consideration of the following model for the generation of the potential outcome variables:

$$Y_{it}^0 = \lambda_i + T\tau_i, \tag{9.14}$$
$$Y_{it}^1 = Y_{it}^0 + \delta_i, \tag{9.15}$$

where λ_i is an individual-varying intercept, the variable T identifies the time period, and τ_i is an individual-varying coefficient on time T. For this model, the following equality holds:

$$\begin{aligned} E[Y_{it}^0|D_i^* = 1] - E[Y_{it}^0|D_i^* = 0] = &(E[\lambda_i|D_i^* = 1] - E[\lambda_i|D_i^* = 0]) \\ &+ T\left(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0]\right), \end{aligned} \tag{9.16}$$

where the T on the right-hand side is set equal to the value of t in the subscript on the left-hand side. Without loss of generality, assume for the moment that $(E[\lambda_i|D_i^* = 1] - E[\lambda_i|D_i^* = 0]) > 0$. In this case, note that whether the initial difference between Y^0 between those in the treatment group and those in the control group remains the same, grows, or shrinks is, respectively, a function of whether $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])$ is equal to 0, is greater than 0, or is less than 0.

If we assume that $(E[\lambda_i|D_i^* = 1] - E[\lambda_i|D_i^* = 0]) = 0$, then the appropriate adjustment factor, α, equals

$$\alpha = 1 + \left(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0]\right). \tag{9.17}$$

If $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0]) = 0$ (i.e., Y^0 changes on average over time at the same rate for individuals in the treatment and control group), then $\alpha = 1$ in Equation (9.17), and the assumptions of the change score model are appropriate. If $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0]) = (r - 1)$, which is necessarily nonpositive (because $0 < r < 1$), then $\alpha = r$ in Equation (9.17), and the assumptions of the analysis of covariance model are appropriate instead.

Of course, there is no reason that $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])$ should necessarily be equal to either 0 or $r-1$. Thus, it is possible that neither the change score model nor the analysis of covariance model provides the correct adjustment. To bring this point home, consider Table 9.2, in which the stipulated true causal effect is 1. The table reports different estimates of the causal effect of the treatment for different combinations of true and assumed adjustment factors. Equivalently, because $\alpha = 1 + (E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])$, the table reports

Table 9.2: Estimated Treatment Effects (With True Effect Equal to 1) for Different Combinations of True and Assumed Adjustment Factors

	Assumed α				
True α	2	1.5	1	.5	0
2	1	1.5	2	2.5	3
1.5	.5	1	1.5	2	2.5
1	0	.5	1	1.5	2
.5	−.5	0	.5	1	1.5
0	−1	−.5	0	.5	1

estimates for different actual and assumed values of the difference in slopes for the treatment and control groups.

Note first that all of the diagonal elements of Table 9.2 are equal to 1. If the assumed adjustment factor equals the correct adjustment factor, we get the correct estimate of the causal effect, 1. Below the diagonal are cases where we have overadjusted (that is, in which the assumed adjustment factor is greater than the correct one). As a result, we get estimates of the causal effect that are too low, ranging from .5 to −1, including an estimate of no effect at all. Above the diagonal, we have cases where we have underadjusted (that is, the assumed adjustment factor is smaller than the correct one). As a result, our estimates of the causal effect are too high, ranging from 1.5 to 3.

For this example, the true causal effect is 1 by construction. Across Table 9.2, we have estimates ranging from −1 to 3. We could easily expand the range of these estimates by considering a broader range of true and/or assumed adjustment factors. And alternative examples could be developed in which the true causal effect equals alternative values, and in which the range of estimates varies just as widely. Although the calculations behind Table 9.2 are simple, the point of the table is to show that one can obtain any estimate of a causal effect by making different assumptions about the appropriate adjustment factor.

In view of this problem, what should a researcher do? If there are strong theoretical reasons for arguing that a particular adjustment factor is correct, and others agree, then analysis is straightforward. If not, which we suspect is generally the case, then it may be possible to argue for a range of adjustment factors. In this case, a researcher may be able to bound the causal effect to a region on which all researchers can agree.

In general, the assumption that a particular correction factor is correct must be based on a set of assumptions about how the expectation of Y^0 for those in the treatment and control groups evolves over time. This leads naturally to a more explicit model-based approach. As we will also see, with data from more than one pretreatment time period, it may be possible to test the adequacy of the assumptions and thus the appropriateness of a particular adjustment factor.

In fact, when panel data are used, there may be no feasible alternative to using an explicit model-based approach.

9.3.2 Model-Based Approaches

In discussing panel data models we have until now considered only traditional methods of estimating a causal effect. There is, however, much merit to considering explicit models of the evolution of Y^1 and Y^0 and then to ask "Under what model assumptions do different methods give consistent estimates?" In this section, we take this approach and address four questions:

1. What is the dynamic structure of the outcome? That is, how are future values of the outcome related to previous values of the outcome? Answering this question is critical if our goal is to estimate counterfactual values. In the counterfactual framework for the sort of treatment examples we will consider, we are interested primarily in the dynamic structure of Y^0, which is the potential outcome random variable under the control state.

2. How is assignment to the treatment determined? As in cross-sectional attempts to estimate causal effects, this is crucial if a researcher hopes to generate consistent estimates of a particular causal effect.

3. What are the alternative methods of estimation that can be used to consistently estimate the average causal effect, given a valid set of assumptions?

4. How can the estimated model be tested against the data?

We will consider these four questions in this order.

Dynamic Structure

As shown in any advanced time series textbook, the dynamic structure of the outcome can be infinitely complex. In the context of panel data models, researchers have typically considered fairly simple structures, often because of the limited number of waves of data that are available. Rather than trying to provide an exhaustive account – which would take a book in and of itself – we primarily focus on conceptual issues.

The broad statistics and econometric literature on panel data models is quite distinct from the estimation of treatment effects from a counterfactual perspective. Implicit in much of this literature is the assumption that causal effects are constant across individuals, such that causes/treatments simply shift the outcome by fixed amounts. From a counterfactual perspective, such assumptions are overly rigid. A necessary component of estimating a treatment effect is the consideration of the hypothetical evolution of Y^0 for the treatment group after the treatment occurs. If treatment effects are heterogeneous and selection is on the treatment effect itself, then the treatment effect for the treated is usually the parameter of interest, as it is often the only one that can be identified by any model (and, fortunately, it is also often of inherent substantive interest).

Consider the following possible two equations for the generation of Y^0:

$$Y^0_{it+1} = \lambda_i + e_{it+1}, \tag{9.18}$$

$$e_{it+1} = \rho e_{it} + v_{it}, \tag{9.19}$$

where λ_i is a time-constant, individual-varying fixed effect, v_{it} is pure random noise (that is, uncorrelated with everything), and ρ is the correlation between e_{it} over time (not the correlation between Y^0_{it} over time, which we labeled as r earlier). Equation (9.19) specifies an autoregressive process of order (1). It is order (1) because the current e_{t+1} is dependent on only the last e_t, not e_{t-1} or any errors from prior time periods. There are many possible ways that the current error could be dependent on past errors. These define what are known as the class of autoregressive moving average (ARMA) models.

Within the current context [i.e., assuming that we know that Equations (9.18) and (9.19) are capable of representing the full dynamic structure of Y^0], determining the dynamic portion of the model for Y^0 amounts to asking whether $\text{Var}(\lambda) = 0$, $\rho = 0$, or both. Multiple tests are available to evaluate these restrictions (see, again, texts such as Hamilton 1994 and Hendry 1995 for comprehensive details). Most standard data analysis programs allow a researcher to estimate a full model on the pretreatment values of Y, assuming that neither $\text{Var}(\lambda) = 0$ nor $\rho = 0$, and then to reestimate various constrained versions. Thereafter, a researcher can then use standard likelihood ratio tests of these model constraints. But such tests on the pretreatment data are not full tests of how Y^0 evolves for the treatment group in the absence of treatment (here again, we are back to the issue in ITS analysis in Figure 9.1). Thus, a researcher most likely will need to make some untestable assumptions.

Consider the following scenarios. If both $\text{Var}(\lambda)$ and ρ are nonzero (and, furthermore, that selection into the treatment is on λ), how then do the values of Y^0 in the treatment and control group behave? When asking this question, we are implicitly considering how $E[Y^0_{it}|D^*_i = 1]$ and $E[Y^0_{it}|D^*_i = 0]$ regress over time toward one or more values. Consider a summary of these different situations in Table 9.3, which are then depicted as pairs of lines in Figure 9.6.

Note that Model A is consistent with the assumptions of the change score model. Model B is consistent with the assumptions of the analysis of covariance model. Model C is consistent with neither, but we suspect that it is the most common scenario in empirical research.

Many of the most common models in panel data analysis assume that there is virtually no dynamic structure to the process that generates the outcome. In fact, most versions of the model that generally goes by the name of a "fixed effects" model are based on the following implicit model for the generation of Y^0 and Y^1:

$$Y^0_{it} = \lambda_i + T\tau + v_{it}, \tag{9.20}$$

$$Y^1_{it} = Y^0_{it} + \delta_i, \tag{9.21}$$

where λ_i is a fixed time constant, individual-level determinant of the outcome

Table 9.3: Alternative Trajectories of the Outcome Under the Control State for Different Assumptions About its Dynamic Structure

Model	Assumed Constraints		Evolution of Y^0
A	$\rho = 0$	$\text{Var}(\lambda) \neq 0$	Immediate regression of individual values to separate group expectations
B	$\rho \neq 0$	$\text{Var}(\lambda) = 0$	Regression over time of individual values to a common expectation
C	$\rho \neq 0$	$\text{Var}(\lambda) \neq 0$	Regression over time of individual values to separate group expectations

in the absence of treatment, $T\tau$ is a time trend common to all (because T is a variable measuring time, and τ is a constant coefficient that does not vary over individuals or time), v_{it} is random noise, and δ_i is an individual-specific additive causal effect that is assumed to be independent of λ_i and v_{it}. The assumed data generation process that motivates the most standard form of a fixed effect model is equivalent to the assumption that each individual has his or her own intercept but there is neither serial correlation in v_{it} nor individual-specific trajectories in time.

The motivation for the standard fixed effects model can be generalized by allowing each individual to have his or her own slope with respect to time, which is indicated by subscripting τ by i in the assumed data generation model in Equations (9.20) and (9.21). This more general model is then

$$Y_{it}^0 = \lambda_i + T\tau_i + v_{it}, \tag{9.22}$$
$$Y_{it}^1 = Y_{it}^0 + \delta_i, \tag{9.23}$$

which, apart from the stochastic term v_{it}, was considered already in the previous section [see Equations (9.14) and (9.15)]. There, we showed that allowing for differences between $E[\tau_i|D_i^* = 1]$ and $E[\tau_i|D_i^* = 0]$ could lead to the necessity of adjustment factors ranging from negative to positive infinity. The attractiveness of this model, of course, is that it allows the expectation of Y^0 for the treatment and control groups to evolve in parallel, diverge, or converge. This will depend, respectively, on whether the difference in the expected slopes for the treatment and control groups is zero, positive, or negative. But substantial amounts of data are needed to estimate it, and certainly from more than just one pretreatment time period and one posttreatment time period.

Determining the Assignment Process

As we have argued in previous chapters, the key to estimating a treatment effect is understanding the process of treatment assignment/selection. One of

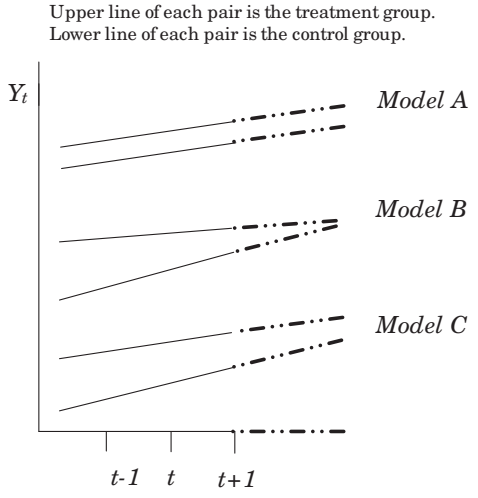


Figure 9.6: Three examples of the possible trajectories summarized in Table 9.3.

the advantages of conceptualizing and then analyzing the dynamic process of the outcome is that it provides evidence of factors in the assignment process. Specifically, knowing the dynamic structure of Y indicates how Y might be predicted, which is important in situations where selection is likely to be on the predicted Y.

Two general cases are of particular interest and lead to quite different estimation strategies. The issue, however, is potentially tricky in that we may want to condition on one or more endogenous variables. As we have discussed at several points throughout this book, conditioning on an endogenous variable can often unblock already blocked back-door paths, thus creating new problems.

Consider first the case in which assignment is directly a function of previous values of Y, as in Figure 9.7. In this diagram, the association between Y_{t+1} and D does not identify the causal effect of D on Y_{t+1} because they are connected by a back-door path through Y_t: $D \leftarrow Y_t \rightarrow Y_{t+1}$ However, this back-door path can be blocked by conditioning on Y_t.

Note that Y_t is a collider on the path that connects e_{t-1} and e_t via Y_{t-1} and Y_t. Thus, conditioning on Y_t will induce associations between both e_{t-1} and Y_{t-1} with e_t. These new associations are unproblematic, however, because they do not create a new back-door path between D and Y_{t+1}. Note that if we thought that D depended on earlier values of Y, we could condition on these Y's without creating problematic back-door paths.

Consider an alternative and much more complex model, presented in Figure 9.8, where D is determined by λ as opposed to Y_t. For this model, there is an unblocked back-door path connecting Y_{t+1} and D: $D \leftarrow \lambda \rightarrow Y_{t+1}$. What happens if we condition on Y_t? Obviously, the unblocked back-door path $D \leftarrow$

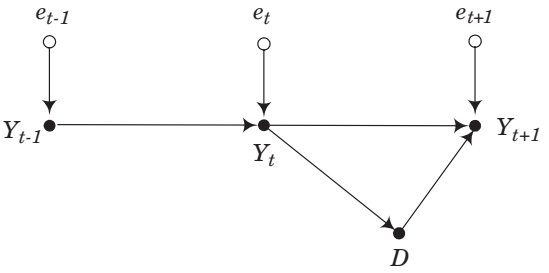


Figure 9.7: A model of endogenous treatment assignment in which selection is on the pretreatment outcome.

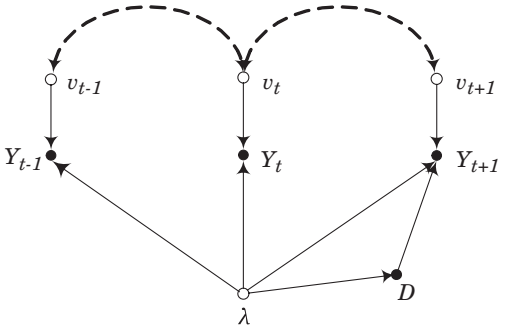


Figure 9.8: A model of endogenous treatment assignment in which selection is on a fixed effect that also determines the outcome.

$\lambda \rightarrow Y_{t+1}$ remains unblocked because Y_t does not lie along it. In addition, conditioning on Y_t unblocks another back-door path because Y_t is a collider on the previously blocked back-door path $D \leftarrow \lambda \rightarrow Y_t \leftarrow v_t \dashleftarrow\dashrightarrow v_{t+1} \rightarrow Y_{t+1}$. In combination, conditioning on Y_t has only made things worse. We failed to block the original back-door path that was of concern, and we unblocked an already blocked back-door path.

Consider this problem from another perspective. Suppose that the standard motivation of a fixed effect model is in place, such that one has reason to believe that Y^0 and Y^1 are generated as in Equations (9.20) and (9.21). Suppose, therefore, that we wish to estimate the following regression equation:

$$Y_{it+1} = l_i + D_{it+1}c + e_i \tag{9.24}$$

where the l_i are individual-specific intercepts, and where we then hope that the estimated coefficient c on the treatment exposure variable in time period $t+1$ will then be equal to the average causal effect, $E[\delta]$.

If we had a measure of λ_i, then estimating the model in Equation (9.24) would be straightforward. We could put in a dummy variable specification to parameterize the l_i intercepts. In the causal diagram in Figure 9.8, and from a standard structural equation perspective, Y_{it} can be thought of as a measure of λ_i. This then suggests that the following regression equation can be estimated instead of the one in Equation (9.24):

$$Y_{it+1} = a + Y_{it}b + D_{it+1}c + e_i \ . \tag{9.25}$$

If we think of Y_{it} as a measure of λ_i, then it contains measurement error because it is also a function of v_{it}. As a result, the coefficient on Y_{it} will be downwardly biased. In fact, the estimate of b will be a consistent estimate of r, which is the reliability of Y. This is not surprising because Equation (9.25) is the analysis of covariance model. Thus, the analysis of covariance model can be interpreted as a fixed effect model in which we have used a noisy measure of λ_i as a right-hand side variable.

In the last section, we saw that the choice of either the analysis of covariance or the change score model can be consequential. Accordingly, it is critical to determine whether assignment to D is function of Y_{it} or λ_i. If we have two or more pretreatment observations, this is easy to do. The first step is to determine whether $b = c$ or $c = 0$ in the following model:

$$\text{Logit}(D_i) = a + Y_{it}b + Y_{it-1}c \ . \tag{9.26}$$

In Figure 9.7, D is dependent only on Y_t. Thus, c should equal 0. In Figure 9.8, D is associated with only Y_t and Y_{t-1} through their joint dependence on λ. As a result, $b = c$.[7] Obviously, in this test, we may also include observed X's that we believe also determine D. And the test generalizes in the obvious way when we observe Y at more than two pretreatment time periods.

Effect Estimation

Extensive discussions exist in the econometrics literature on the estimation of the fixed effects model and its generalizations. Basically, there are two different

[7] These assumptions can more easily be tested by estimating the model

$$\text{Logit}(D_i) = a + Y_{it}b + (Y_{it} + Y_{it-1})c$$

and testing whether $b = 0$ or $c = 0$. Here, $(Y_{it} + Y_{it-1})$ is essentially acting as a measure of λ_i. This strategy is based on the following trick often used to test for the equality of coefficients for two variables X and Z. Let the coefficient on X be m and on Z be $m + n$. Run the following regression equation:

$$\begin{aligned} Y &= Xm + Z(m + n) + u \\ &= (X + Z)m + Zn + u \end{aligned}$$

and use a standard statistical test to evaluate the null hypothesis that $n = 0$.

estimation strategies. For the differencing approach, the specification is

$$\begin{aligned} \text{differencing: } Y_{it+1} - Y_{it} &= (\lambda_i - \lambda_i) + (D_{it+1} - D_{it})d \\ &\quad + (X_{it+1} - X_{it})b + (e_{it+1} - e_{it}) \\ &= (D_{it+1} - D_{it})d + (X_{it+1} - X_{it})b + (e_{it+1} - e_{it}), \end{aligned} \tag{9.27}$$

where treatment exposure occurs between time period t and $t+1$. In contrast, the dummy variable approach is

$$\text{individual dummies: } Y_{it} = P_i l_i + D_{it}d + X_{it}b + e_{it}, \tag{9.28}$$

where P_i is a dummy variable for person i, and l_i is its associated coefficient. This second method amounts to estimating a separate intercept for each individual. It is also equivalent to differencing Y and X from their respective individual-level means. If one wants to estimate the generalized fixed effect model in which there is an interaction between an individual effect and time, one can do this by either differencing additional times or by interacting the individual dummies, P_i, with time and estimating a separate interaction term for each dummy.

To understand the differences between these two approaches, evolving conventions in data analysis must be noted. For the representation of the change score model and the analysis of covariance model in Equations (9.2) and (9.3), we conformed to the convention in the literature wherein the models are written out so that they can be estimated easily as a cross-sectional dataset. In other words, time is encoded in the variables, so that time-varying outcome variables – Y_{it} and Y_{it+1}– are regressed on a cross-sectional treatment group dummy variable D_i^* in an implicit dataset with one record for each of N individuals. This data setup is also the implicit background for the differencing specification of the fixed effect estimator in Equation (9.27).

As can be seen in virtually any panel data textbook (e.g., Baltagi 2005), the convention is to now structure one's dataset in person–time records, which results in $N \times T$ records rather than N records. Forcing time to be constant within each data record allows for variables such as D_{it} and D_i^* to be cleanly parameterized in subsequent analysis. This data setup is the implicit background for the individual dummy specification of the fixed effect estimator in Equation (9.28), and that is why there is no reference to time t versus time $t+1$ in Equation (9.28). Throughout the remainder of this section, we will write with such an $N \times T$ dataset in mind. The ideas, however, do not depend on such an implicit structuring, as one can switch back and forth between both setups based on analogs to Equations (9.27) and (9.28).

As for the alternative fixed effect specifications in Equations (9.27) and (9.28), the two methods give identical answers when there are only two points in time. When the errors are correlated across time and there are more than two time periods, the two methods will give somewhat different estimates, though both estimators are consistent. Which estimator is preferable depends on the nature of the correlation structure. We need not be concerned about this issue here (see Baltagi 2005, Hsiao 2003, and Wooldridge 2002 for discussion).

Traditional fixed effect and differencing methods are generally inefficient, however, if the goal is only to estimate the effect of a treatment. These methods simply eliminate unobserved individual effects from the data. Doing so powerfully eliminates all associations between the treatment variable D_{it} and unobserved time-constant, individual-level variables. If the coefficients of all observed variables are of interest, then this is appropriate.

In the present case, however, our focus is simply on estimating the effect of treatment exposure, D_{it}. As pointed out repeatedly in previous chapters, one approach to consistently estimate the effect of D_{it} is to balance the data with respect to all systematic determinants of D_{it}. As discussed in the last section, our interest in the case of linear models (the situation with nonlinear models being more complicated) is in differences in the expected trajectories of Y^0 for those in the treatment and control groups. If we want to use the control group's Y^0 to predict what the treatment group's Y^0 would have been in the absence of treatment, then it is essential that differences in the two groups' trajectories be modeled.

To be more concrete, suppose that we have three time points, as for the demonstration reported in Table 9.1. If the only difference in the expected trajectories of Y^0 for the two groups is in their levels (that is, their intercepts), then all we need to do is allow for differences in group-level intercepts by estimating

$$Y_{it} = a + D_i^* b + D_{it} d + e_{it}. \tag{9.29}$$

Here, the coefficient b captures differences in the intercept in the expected trajectories for the treatment and control groups, such that the intercept for the control group is a and the intercept for the treatment group is $a + b$.[8]

If the expected trajectories also differ in their slopes, then we need to include a term for time and an interaction term between group membership and time, as in

$$Y_{it} = a + D_i^* b + Tc + (D_i^* \times T)c' + D_{it} d + e_{it}. \tag{9.30}$$

The coefficient b again captures differences in the intercepts, and c' now captures differences in the slopes of the expected trajectories. Interactions between D^* and higher-order polynomials of T (or any other function of time) can also be introduced, assuming sufficient pretreatment data are available.

Estimating the model in Equation (9.30) is equivalent to differencing out the treatment/control group expectations of λ_i and τ_i in the following assumed data generation model for Y^0 and Y^1 [based on an augmentation of Equations (9.22) and (9.23)]. Expanding a standard fixed effect model separately for the treatment and control groups using the time-constant indicator of the treatment

[8] Notice that we can include both D_{it} and D_i^* in the same regression equation because we have multiple records for each individual over time in our dataset. In posttreatment time periods $D_i^* = D_{it}$ for all individuals (assuming that no one leaves the treatment state before the end of the study), but in pretreatment time periods $D_i^* = D_{it}$ only for individuals in the control group.

group D^* yields

$$\text{for } D_i^* = 0: \tag{9.31}$$
$$Y_{it,D^*=0}^0 = (\mu_{\lambda,D^*=0} + v_{i,D^*=0}) + T(\mu_{\tau,D^*=0} + \tau'_{i,D^*=0}),$$
$$\text{for } D_i^* = 1:$$
$$Y_{it,D^*=1}^0 = (\mu_{\lambda,D^*=1} + v_{i,D^*=1}) + T(\mu_{\tau,D^*=1} + \tau'_{i,D^*=1}),$$
$$\text{for } D_i^* = 0 \text{ and } D_i^* = 1:$$
$$Y_{it}^1 = Y_{it}^0 + \delta_i \ . \tag{9.32}$$

Here, $\mu_{\lambda,D^*=0}$ is the expectation of λ_i in the control group, and $\lambda_i = \mu_{\lambda,D^*=0} + v_{i,D^*=0}$ for those in the control group. Likewise, $\mu_{\tau,D^*=0}$ is the expectation of τ_i in the control group, and $\tau_i = \mu_{\tau,D^*=0} + \tau'_{i\lambda,D^*=0}$ for those in the control group. The terms for the treatment group are defined analogously.

In this setup, the terms $v_{i,D^*=0}$, $v_{i,D^*=1}$, $T\tau'_{i,D^*=0}$, and $T\tau'_{i,D^*=1}$ become components of the error term e_{it} of Equation (9.30) as constant individual-level differences v_i and time-varying individual differences $T\tau'_i$. Because $v_{i,D^*=0}$, $v_{i,D^*=1}$, $T\tau'_{i,D^*=0}$, and $T\tau'_{i,D^*=1}$ are all by construction uncorrelated with D_i^*, e_{it} is uncorrelated with D_i^* assuming any extra individual or time-varying noise embedded within e_{it} is completely random. Furthermore, the coefficient a in Equation (9.30) is equal to $\mu_{\lambda,D^*=0}$, and the coefficient b is equal to $\mu_{\lambda,D^*=1} - \mu_{\lambda,D^*=0}$. Thus, b captures the difference in the expected intercept for individuals in the treatment and control groups. Likewise, the coefficient c in Equation (9.30) is equal to $\mu_{\tau,D^*=0}$, and the coefficient c' is then equal to $\mu_{\tau,D^*=1} - \mu_{\tau,D^*=0}$. And thus c' captures the difference in expected slope of the time trends for individuals in the treatment and control groups.

The coefficient d on D_{it} is a consistent estimate of average treatment effect because the expectations of λ_i and τ_i are balanced across the treatment and control groups. All of their systematic components related to treatment group membership are parameterized completely by a, b, c, and c'. This leaves all remaining components of the distributions of λ_i and τ_i safely relegated to the error term.[9]

There are two advantages of estimating Equation (9.30), as opposed to using traditional methods for estimating fixed effect models and their generalizations. First, conceptually, the model makes clear that if the goal is consistent estimation, then the expected trajectories of Y^0 for the treatment and control groups must be correctly modeled (not necessarily all individual-specific trajectories). Later, we show how this principle leads to a general specification test. Second, there are potential efficiency gains. For example, in a standard fixed effect model, half of the overall degrees of freedom are lost by specifying individual-specific fixed effects (when there are only two time periods of data). In estimating Equation (9.30), only one degree of freedom is lost to an estimate of the difference in the intercept for the mean of Y^0. We should note, however,

[9]Moreover, because this model is set up as a linear specification of the treatment effect, a lack of balance in higher-order (centered) moments of λ_i and τ_i does not affect the estimation of d.

that this minimization in the loss of degrees of freedom is moderated by the fact that the errors in Equation (9.30) are likely to be highly correlated within individuals (because the errors within individuals include a common fixed effect). An estimation procedure should be used for the standard errors that accounts for this correlation.

In the situation in which Y_{t+1} is directly affected by Y_t, estimation is far simpler. As already discussed with reference to Figure 9.7, conditioning on Y_t is sufficient to block all back-door paths connecting Y_{t+1} and D. Conditioning could be done by matching, as in the analysis of Dehejia and Wahba (1999) for the National Supported Work data (although there is no evidence that they attempt to test the suitability of this specification as opposed to a fixed effect specification).[10] Alternatively, Y_t could be conditioned on by a regression model, as in an analysis of covariance model. More complicated specifications in which Y_{t+1} is a function of both Y_t and unobserved individual-level variables are also possible. In general, IVs are needed to estimate these models effectively. Halaby (2004) provides a clear introduction to these methods.

Model Testing

Hopefully, by now, we have convinced the reader that maintained modeling assumptions can have large consequences. Given this dependence, it is critical that researchers be explicit about the assumptions that they have made and be able to defend those assumptions. Assumptions can be defended either theoretically or on empirical grounds. Often neither is done. In fact, they are made often without any explicit recognition. Fortunately, if pretreatment observations are available for multiple time periods, it is possible in many circumstances to test the assumptions against the data. Here, we describe two conceptually different, but mathematically closely related, approaches.

In discussing strategies to increase confidence in a causal effect estimate from an ITS model, we suggested that a researcher could either use a dependent variable for which no effect should occur or estimate the effect for a group for which no treatment effect should occur. Evidence of a treatment effect in either case is evidence that the model is misspecified.

Heckman and Hotz (1989) suggest applying this same principle to panel data when two or more pretreatment waves of data are available. Specifically, they suggest taking one of the pretreatment Y's and analyzing it as if it occurred posttreatment. A researcher then simply applies the estimation method to the new pseudo-posttreatment data and tests for whether there is evidence of a "treatment effect." Because neither the treatment nor the control group has experienced a treatment in the data under consideration, evidence of a treatment effect is evidence that the model is misspecified (i.e., that the model has failed to fully adjust for differences between the treatment and control groups).

In the analysis of covariance model, care must be taken. Here, it is implicitly assumed that selection is on Y_{it}. For example, for a logit specification of the

[10] For detailed discussions of the appropriate model specification for these data, see Smith and Todd (2005) and associated comment and reply.

probability of treatment selection, it is implicitly assumed:

$$\text{Logit}(D_i^*) = a + Y_{it}b. \tag{9.33}$$

In this model, D_i^* is a function of Y_{it}. This is mathematically equivalent to maintaining that D_i^* is a function of $Y_{it-1} + (Y_{it} - Y_{it-1})$. Generally, Y_{it} will be correlated with $(Y_{it} - Y_{it-1})$. Consider the following model:

$$Y_{it} = a + Y_{it-1}r + D_i^*c + u_i. \tag{9.34}$$

Because D_i^* is a function of both Y_{it} and $(Y_{it} - Y_{it-1})$, and Y_{it} is correlated with the latter term, in general c will not equal 0. The basic point is that D_i^* is partially a function of a component of Y_{it} that is not contained in Y_{it-1}. In general, the coefficient c on D_i^* is a function of this dependence.

We can, however, run time backwards. Accordingly, we can estimate

$$Y_{it-1} = a + Y_{it}r + v_i. \tag{9.35}$$

If we are going to use the testing strategy of Heckman and Hotz (1989) to evaluate an analysis of covariance model, we should then test whether $c = 0$ in the following related model:

$$Y_{it-1} = a + Y_{it}r + D_i^*c + e_i. \tag{9.36}$$

Because there is no component of D_i^* that depends on Y_{it-1} conditional on Y_{it}, c should equal 0 in this model if the analysis of covariance model has correctly adjusted for treatment group differences in Y_{it-1}.

Heckman and Hotz's test also indicates how two-period pretreatment/posttreatment data can be used. What we should do is fit a cross-sectional model. We should then treat the pretreatment Y as if it were a posttreatment Y and then test for a treatment effect. Evidence of a treatment effect is evidence that our cross-sectional model has failed to fully adjust for pretreatment differences between the treatment and control groups.

To better understand these procedures, consider a more general specification of this type of test. Recall that, net of our adjustments for various functions of time and other variables, we seek to evaluate whether the trajectories of Y^0 are equivalent in the pretreatment data for the treatment and control groups. A variety of different models can be assessed. A fixed effect model allows for differences in the intercepts for two groups. A model with individual- or group-specific time coefficients allows for differences in slopes. If we have enough pretest data, we can add additional functions of time into our model and thereby allow for even more heterogeneity of trajectories for Y^0.

The most general specification would be to use one time period as the base and create a dummy variable indicator for all other time periods and to allow these dummy variables to fully interact with our treatment group indicator D^*. This is what is known as the saturated model in this tradition. It is a completely flexible function form and allows for completely separate time trajectories for

Table 9.4: Specification Tests from the Analysis of Heckman and Hotz (1989) of the Effect of the National Supported Work Program on the Earnings of High School Dropouts

		p values for specification tests	
	Estimated effect	Preprogram 1975 earnings	Postprogram 1978 earnings
Experiment	−48 (144)		
Regression	−1884 (247)	.000	.000
Fixed effect model (pre-1972)	−1886 (242)	.000	.000
Random-growth model (pre-1972 and 1973)	−231 (414)	.375	.329

Note: Results are from Tables 3 and 5 of Heckman and Hotz (1989), for the B1 specification of basic demographic characteristics.

Y^0 for the treatment and control groups. It is of little use in estimating the true causal effect.

Using just the pretest data, we can, however, compare the saturated model with any more restrictive model – such as a fixed effects model – using an F-test or likelihood ratio test. Evidence that the more restrictive model does not fit is evidence that the more restrictive model fails to fully model the differences between the treatment and control groups in the trajectories of Y^0.

Consider the results of Heckman and Hotz (1989), a portion of which is presented in Table 9.4.[11] For their analysis, Heckman and Hotz (1989) estimated a wide range of alternative models of the effect of the National Supported Work program on the 1978 earnings of participants who were high school dropouts. The first column reports selected estimated effects from their study. The experimental estimate indicates that the program has no effect (that is, −48 dollars with a standard error of 144 dollars). The regression and fixed effect models show large negative effects, which are statistically significant by conventional standards. Their random-growth models, which allow for individual slope coefficients for the trajectories of earnings, suggest a modest but still nonsignificant negative effect.

[11] Although Heckman and Hotz (1989) is an exemplary early example of this sort of analysis, the basic specification test approach is used in one form or another in other work as well (e.g., Petersen, Penner, and Høgsnes 2006, Table 4).

The second column of Table 9.4 reports the p values for tests of no treatment effect, in which the preprogram 1975 earnings are used as if they were in fact posttreatment earnings. If the models that are tested adequately adjust for underlying differences between the treatment and control groups in the trajectories of earnings, one would expect the treatment effect estimate to be nonsignificant (i.e., have a high p value). In the case of the regression and fixed effects models, the faux-treatment-effect estimate is highly significant, indicating a lack of model fit. In the case of the random-growth model, however, it appears to fit the data.

The third column reports results from a similar test, where Heckman and Hotz analyze the valid 1978 posttreatment data as if one time period was in fact pretreatment data. Again, they test for whether there is evidence of a treatment effect. As with the tests reported in the second column, if the model is properly specified, then there should be no evidence of a treatment effect. But here also the p values for the regression and fixed effects models suggest a significant treatment effect, indicating that these models do not fit the data. And, as before, the p value for the random-growth model indicates that it fits the data.[12]

The National Supported Work data have been analyzed by many different social scientists, and they are perhaps the most widely used data to assess the relative explanatory power of alternative types of panel data estimators. There has been considerable debate about whether or not researchers need methods that take account of unobservables or whether adjusting only for observables is sufficient. Smith and Todd (2005; see also the associated comment and reply) offer the most up-to-date analysis. Their analyses show clearly how sensitive estimates can be to the sample that is chosen. They support Heckman's position that there are important situations for which treatment selection is likely to be a function of unobserved variables.

9.4 Conclusions

Longitudinal data may be helpful for estimating causal effects, but longitudinal data do not constitute a magic bullet. Defendable assumptions about the treatment assignment process must be specified. And, to use longitudinal data to its maximum potential, researchers must carefully consider the dynamic process that generates the outcome, clearly define the causal effect of interest, and then use constrained models only when there is reason to believe that they fit the underlying data.

[12] A note of caution is warranted here. As in all situations, the power of these tests is a function of sample size. In this case, there are only 566 treatment cases. From the data, it is impossible to tell whether the random-growth model fits the data because it is a sufficiently correct model of the underlying individual-specific trajectories or rather because the sample is small.

Part 4: Conclusions

Chapter 10

Counterfactual Causality and Future Empirical Research in the Social Sciences

What role should counterfactual models play in causal analysis in observational social science? Some claim that it is the only correct way to think about causality while others claim that it is of limited value. We take an intermediate position. We see the methods of counterfactual causal modeling as constituting a useful set of tools that can help to improve the investigation of causal relationships within the social sciences. We believe that counterfactual methods both complement and extend existing approaches to causal analysis.

The strength of counterfactual modeling is that it demands that the researcher specify precisely how changing the treatment state, holding other relevant conditions constant, would change the expected outcome for a relevant unit of analysis. Counterfactual models also reveal the essential requirement of explanations that appeal to mechanisms: One must document how a treatment effect is propagated by a mechanism to an outcome, considering each distinctive causal pathway along the way.

Nonetheless, counterfactual models require much less than some versions of structural equation modeling. Counterfactual models do not require a full specification of all causes that produce an outcome, and they do not assume (or require) that a causal effect estimate be population invariant. In many cases, their parameters can be estimated under far weaker assumptions. For many average causal effects of interest, for example, the data merely need to be balanced with respect to the determinants of treatment assignment. Causal claims do not rest on assumptions that require full independence between the observed causes of the outcome and all other unobserved causes. Thus, counterfactual

modeling allows us to go beyond traditional analyses that focus on conditional associations and the problem of omitted variables, pushing us to develop richer and more specific causal accounts. These models, however, do not require that we specify full structural models of all of the determinants of an outcome when neither theory nor data can sustain them.

In this chapter, we first shore up our presentation of the counterfactual model by considering several critical perspectives on its utility. We weigh in with the arguments that we find most compelling, and it will not be surprising to the reader that we find these objections less serious than do those who have formulated them. However, we use quotations from the published literature extensively in order to demonstrate the richness of these debates and to encourage a full reading of the material that we are able to summarize only briefly here.

We then conclude the chapter with a discussion of causal modeling practices, noting the alternative modes of causal inquiry that exist in observational social science. We argue that counterfactual-based modeling is useful because it encourages the explicit analysis of well-defined causal relationships. As such, the model facilitates movement between modes of causal inquiry, guiding analysis to the deepest level that theory and data can enable at any given point in time.

10.1 Objections to Features of the Counterfactual Model

Although attractive in many respects, the counterfactual model is not without serious objections from thoughtful critics. In this section, we present some of the concerns from the published literature.[1] If any of these objections are accepted, then a justification for eschewing much of what we have presented in prior chapters is available. Such a decision, however, would presumably require the adoption of the counterposition(s) of those who raise the following objections. We will not lay out these alternative approaches to causal analysis here, but their main features are implicit in the critiques of the counterfactual model.

Objection 1: Nonmanipulable Causes Cannot be Analyzed

The counterfactual model is most natural when one can conceive of a specific procedure for manipulating the cause of interest, or in the words of Berk

[1] There is one common objection that we do not present or discuss in the main text. Some argue that the counterfactual model encourages undue obsession with causal inference that can lead to poor statistical practice as a by-product. It is often asked, Is it a good idea to focus so much on unbiasedness and consistency of estimates? An emphasis on consistency, for example, could lead one to favor a consistent but very inefficient IV estimate over an inconsistent but very efficient OLS estimate. We do not address this objection in the main text because it is an objection to all approaches to causal modeling. Accordingly, we do not advocate downgrading attention to matters of statistical inference, only upgrading of the care with which we pursue causal inference.

(2004:91), "in the sense that conscious human action can change things." If the effects of a cause that cannot be manipulated by "human action" are of interest, then the contribution of the framework loses some of its transparency (see Holland 1986). There are two reactions to this objection.

First, it is not as limiting as is often claimed. Woodward (2003) has argued this position more completely than anyone else (see, in particular, his section 3.4 "In What Sense Must Interventions Be Possible?"). In developing his conception of an intervention, Woodward concludes:

> The sorts of counterfactuals that cannot be legitimately used to elucidate the meaning of causal claims will be those for which we cannot coherently describe what it would be like for the relevant intervention to occur at all or for which there is no conceivable basis for assessing claims about what would happen under such interventions because we have no basis for disentangling, even conceptually, the effect of changing the cause variable alone from the effects of other sorts of changes that accompany changes in the cause variable. (Woodward 2003:132)

In this regard, what matters is not the ability for humans to manipulate the cause through some form of actual physical intervention but rather that we be able, as observational analysts, to conceive of the conditions that would follow from a hypothetical (but perhaps physically impossible) intervention.[2] Moreover, the manipulability criterion is not as strong as is sometimes supposed. For example, if racial discrimination is the topic of study, race itself does not need to be manipulated, only the perception of race by the potential discriminator. Deception can then be used to gain leverage on the causal effect of interest, even if race itself remains fixed.

Even if Woodward's position is rejected [and we concede that some discussions, such as those of Berk (2004), do remain convincing to a large extent], and thus if a researcher does not feel comfortable using the counterfactual framework to define the causal effects for hard-to-conceive-of-actually-manipulating attributes, the counterfactual framework can still be used in an as-if mode in order to sharpen the goals of an analysis. This is the position of Glymour (1986), who responds to the position that nonmanipulable attributes cannot be causes by arguing that counterfactuals can be used to elucidate effects that are presumed to have been caused by a set of potentially manipulable factors but are merely referred to collectively by nominal labels. The goal of research is to push the analysis further down from the nominal labels to the separable manipulable factors. In pursuit of this goal, it can be efficient to first define the counterfactuals associated with nonmanipulable attributes as if they could be easily and naturally manipulated by human action.

[2] Woodward (2003:122) also discusses how the focus on conceivable interventions allows one to avoid having to contemplate preposterously unreasonable counterfactuals.

Consider studying the effect of race on socioeconomic outcomes, one of the most common topics of study in the literature on social inequality. Although race cannot be manipulated easily, the framework can still be used to construct thought experiments that clarify the questions of interest. For example, the counterfactual model could be used to motivate an attempt to estimate the average gain an employed black male working full time, full year would expect to capture if all prospective employers believed him to be white. This would be a quite different effort than an attempt to estimate what the black–white gap would be if black men had grown up in families with as much family income as whites, and hence would have had the available resources with which to purchase higher-quality secondary and postsecondary education. By helping to frame such fine distinctions between alternative states, the counterfactual model helps to sharpen research questions and then shape reasonable interpretations of whatever data are at hand. Reskin (2003) notes, correctly to be sure, that such a basic framing of differentials is but a first step. But it is a necessary first step, for only thereafter can one begin to systematically investigate the intervening mechanisms that generate the causal relationships and the differences across attributes that the extant literature defines as important.

In this regard, stating merely that an attribute D causes an outcome Y is simply the crudest form of a mechanism sketch (as we discussed in Chapter 8). To further investigate this mechanism sketch, counterfactuals must then be defined for whatever presupposed process is thought to generate the causal effect of D on Y. Reskin (2003), for example, provides a clear exposition of how the investigation of such effects of attributes should proceed. She argues that researchers should lay out alternative mechanisms that generate ascriptive inequality – primarily those based on the discriminatory motives of gatekeepers versus those based on unequal access in structures of opportunity – and then evaluate the relative of importance of each mechanism in targeted empirical analysis.

Objection 2: The Counterfactual Approach Is Ill Suited to the Discovery of the Causes of Effects

If an investigator is interested in estimating the effects of causes, then the counterfactual model is quite natural. However, if an investigator is interested in all of the causes of an observed effect, then the counterfactual model is less helpful. This limitation is most obvious when the effect of interest is an important event that demands explanation.

Although we would not want to argue that causes-of-effects questions (or causes-of-events questions) should be of no concern, such global questions of causation are often ill posed and can encourage sloppy analysis. Michael Sobel has provided perhaps the most vigorous indictment of the common practice of attempting to estimate simultaneously a large numbers of causes of an effect of interest from survey data. To cite one example of his perspective, he writes in the overview essay "Causal Inference in the Social Sciences" for a millennium issue of the *Journal of the American Statistical Association*:

> ... much of quantitative political science and sociology may be characterized as a highly stylized search for new causes of effects. Researchers typically begin with one or more outcomes and a list of causes identified by previous workers. Potentially new causes are then listed; if these account ("explain") for additional variability of the response, then the new causes are held to affect the outcome, and the significant coefficients in the new model are endowed with a causal interpretation. The process is repeated by subsequent workers, resulting in further "progress." When researchers realize that they are merely adding more and more variables to a predictive conditioning set, one wonders what will take the place of the thousands of purported (causal) effects that currently fill the journals. (Sobel 2000:650)

If Sobel is correct (and we are inclined to agree with him), then the social sciences give too much attention to the identification of the causes of effects and too little attention to the simpler and more tractable goal of estimating the effects of particular causes. But even if one disagrees with this position, maintaining instead that causes-of-effects questions should always be in the foreground, it is hard to deny that the quest for a full account of the causes of any effect is aided (and perhaps, at times, best advanced) by the pursuit of well-defined questions that focus narrowly on the effects of particular causes. As knowledge of these effects accumulates, we can then attempt to build full causal accounts of outcomes and events. If counterfactual modeling is then justified only as an instrumental path toward the ultimate goal of sustaining full causal accounts, little of its appeal is diminished in our view.

The philosopher Paul Humphreys offers an elegant appeal to this pragmatic realist–empiricist position in the concluding paragraph of his 1989 book, *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences*. After reflecting on the demise of the covering law model of explanation and some of its failed alternatives, Humphreys writes:

> What, then, do we have to replace the traditional account of explanatory knowledge? It is not propositional knowledge .... Nor is it *de dicto* knowledge .... It is *de re* knowledge of the causes that contributed to the effect, gained in a cumulative manner, coupled with the discovery of previously unknown structures that pushes the moving boundary of the observable even further from ordinary experience and allows us to become acquainted with finer and finer details of how the world works. When we know of all of the causes, then we shall know all there is to know of the kind with which we are concerned, for all that we will be left is pure chance, and chance is, as I have said, literally nothing. (Humphreys 1989: 140–1)

The idea here is that we should strive to explain phenomena by estimating the effects of putative causes on particular outcomes. And, in practice for observational data analysis, all such attempts require a consideration of at least two

types of other causes of the outcome: (1) those that may lie along unblocked back-door paths in a causal diagram and (2) those that constitute the mechanisms that link the putative causal variable to the outcome variable. Over time, we collect more data and estimate the effects of the same causes on the same outcomes in a variety of conditions, which deepens our understanding of both forms of variables that were considered unobservables initially.

Objection 3: Causal Inference Should Not Depend on Metaphysical Quantities

In a wide-ranging article titled "Causal Inference Without Counterfactuals," the statistician A. P. Dawid (2000) argues that potential outcomes are inherently metaphysical, and thus the counterfactual model for causal inference is "generally unhelpful and frequently misleading" (Dawid 2000:409). He argues that causal claims must rest only on inherently testable ideas, and he presents a decision-theoretic alternative that he argues succeeds in realizing the goal of estimating the effects of causes.[3]

Although some aspects of Dawid's critique are rather involved, some of its central features are quite simple. He argues that counterfactual modeling embraces "fatalism" because it implicitly assumes that the potential outcomes of individuals (i.e., y_i^1 and y_i^0 for a binary cause) are fixed values that are regarded as "predetermined attributes" of individuals "waiting only to be uncovered by suitable experimentation" (Dawid 2000:412).[4] He regards this implicit assumption as inherently untestable, and yet it is necessary in order to make sense of other concepts that are central to counterfactual modeling (such as SUTVA and the existence of compliers and defiers for LATE analysis):

> [This untestable assumption] ... leaves no scope for introducing realistic stochastic effects of external influences acting between the times of application of treatment and of the response. Any account of causation that requires one to jettison all of the familiar statistical framework and machinery should be treated with the utmost suspicion, unless and until it has shown itself completely indispensable for its purpose. (Dawid 2000:413)

For his alternative, Dawid considers a basic Bayesian decision model that, without reference to potential outcomes, allows a statistician to (1) design a randomized experiment to compare the effects of a treatment on an outcome (in comparison with a base state of either no treatment or an alternative treatment) and (2) convey the results of this experiment to inform a relevant decision maker of the expected effect of applying the treatment to an additional subject similar

[3] For brevity, we do not cover Dawid's position on modeling the causes of effects here. We just addressed the causes-of-effects versus effects-of-causes position in the last section.

[4] He also then objects to the arbitrariness of selecting the linear difference $y_i^1 - y_i^0$ to represent the individual-level causal effect, which is a bit of a red herring, given that virtually everyone has conceded that other possibilities exist but cannot always be profitably analyzed given the constraints of data.

to those on whom the experiment was conducted.[5] After contrasting his model with the counterfactual model, he concludes:

> I have argued that the counterfactual approach to causal inference is essentially metaphysical, and full of temptation to make "inferences" that cannot be justified on the basis of empirical data and are thus unscientific. An alternative approach based on decision analysis, naturally appealing and fully scientific, has been presented. This approach is completely satisfactory for addressing the problem of inference about the effects of causes, and the familiar "black box" approach of experimental statistics is perfectly adequate for this purpose. (Dawid 2000:423)

The response to Dawid's critique of the counterfactual model has been largely negative, as revealed in the comments published with it. The crux of the counterargument is the following. If perfect experiments for every causal question of interest could be designed and then implemented, then potential outcomes are unnecessary (even though they might still be useful for some purposes, such as to think through the results that might have emerged from alternative experimental protocols). Furthermore, no one seems to disagree with the claim that we should use study designs that can rescue us from having to apply assumptions to what-if quantities. In this regard, it would of course be preferable to be able to use a crossover study in all situations, which Rothman and Greenland (1998) describe as follows:

> The classic crossover study is a type of intervention study in which two (or more) interventions are compared, as in any experimental study. In a crossover study, however, each subject receives both interventions, with one following the other in a randomized sequence. Enough time is allocated between administration of the interventions so that the effect on the subject of each intervention can be measured before the other intervention is given. Obviously, the crossover study is only feasible for studying interventions whose effects occur promptly and do not persist, so that the effect of the second intervention is not intermingled with the effect of the first. (Rothman and Greenland 1998:111)

The appeal of this sort of a study design, in view of Dawid's critique, is that each ostensibly metaphysical potential outcome would become an observed outcome. But, unfortunately, crossover research designs work effectively only when one has control over the allocation of the treatments and only when treatment effects are

[5] Following in the Bayesian tradition, the decision maker (who could also be the statistician) can then assess the consequences of applying the treatment to the subject, with reference to cost considerations and other determinants of her loss function. But these additional points are not stressed by Dawid, in part because they are not incompatible with the counterfactual model.

sufficiently ephemeral. These conditions sometimes exist for the causal questions that concern social scientists, but rarely is this the case.

For observational data analysis, which Dawid barely mentions in his article, we are stuck with having to assert what-if assumptions about potential outcomes in order to move forward. Robins and Greenland (2000), when commenting on the Dawid critique, consider many of the applied examples for which counterfactual models have been used. They conclude:

> By narrowly concentrating on randomized experiments with complete compliance, Dawid, in our opinion, incorrectly concludes that an approach to causal inference based on "decision analysis" and free of counterfactuals is completely satisfactory for addressing the problem of inference about the effects of causes. We argue that when attempting to estimate the effects of causes in observational studies or in randomized experiments with noncompliance ... [a] reliance on counterfactuals or their logical equivalent cannot be avoided. (Robins and Greenland 2000:431)

This basic point is echoed by most of the other commentators on the article. Cox (2000:424) asks this question: "And yet: has the philosophical coherence [of Dawid's position], if not thrown the baby out with the bathwater, at least left the baby seriously bruised in some vital organs?" Casella and Schwartz (2000:425–6) conclude, "Dawid insists that such choices [about how to conduct causal analysis] ... must be based on strict principles that can be verified empirically. We believe that such a program is so overly rigid that, in the end, science is not served."

Even so, Dawid's objection to the metaphysical nature of potential outcomes does bring up a deeper question: What is the epistemological status of the counterfactual model?[6] The potential outcomes framework is certainly not positivist, as it breaks radically from the positivist–empiricist prescription that analysis must consider only observable quantities. The reliance on what-if potential outcomes and the consideration of unobservable characteristics of the treatment assignment/selection process consigns the counterfactual model to the postpositivist model of science generally labeled realism (see Psillos 1999, Putnam 1975; see also Godfrey-Smith 2003 for an overview). But, because each unobserved potential outcome could have become an observed outcome if the unobservables in the treatment assignment process had been configured in an alternative way, the entire framework aspires to help us become reductive empiricists.

[6]When wading into the philosophy literature, one point leaps out in regard to Dawid's charge. The literature on causality is considered by most philosophers to be properly situated in the subfield of metaphysics. Thus, Dawid's charge that the potential outcome framework is metaphysical is, from the perspective of philosophers, both accurate and untroubling. See also the position of Pearl (2000:33–4), who argues that, even if counterfactuals are metaphysical, individuals are all too happy to use them in their daily lives. For this reason alone, it seems natural to build notions of causality around them.

In this regard, we suspect that most social scientists who use the counterfactual model would find the pragmatic position of Donald Campbell similar to their own, as he laid it out in 1977:

> I am a fallibilist and antifoundationalist. No part of our system of knowledge is immune to correction. There are no firm building blocks, either as indubitable and therefore valid axioms, or in any set of posits that are unequivocal once made. Nor are there even any unequivocal experiences or explicit operations on which to found certainty of communication in lieu of a certainty of knowledge.
>
> I am some kind of a realist, some kind of a critical, hypothetical, corrigible, scientific realist. But I am against direct realism, naive realism, and epistemological complacency. (Campbell 1988 [1977]:444–5)

We read this passage as indicating that Campbell recognizes, grudgingly to be sure, that the pursuit of valid knowledge necessitates the invocation of unobservable quantities, such as potential outcomes. This tips one into the realm of realism, wherein such unobservables are given provisional truth status for pragmatic purposes. But, for any corrigible realist, the ultimate aspiration is to bring conjectured unobservables out in the open, drawing analysis down to its most generic form as an as-if positivist endeavor.

For scholars who find the counterfactual approach to observational data analysis unappealing, perhaps because of the persuasiveness of one of these objections, a variety of responses exists. Some scholars argue, simply, that we should devote much more attention to descriptive modeling because causal analysis can be so intractable. We begin the next section with this position, after which we then move on to consider four alternative modes of causal inquiry.

10.2 Modes of Causal Inquiry in the Social Sciences

Some scholars argue that many long-standing questions in the social sciences can and should be pursued without any reference to causality. When the sole interest is a parsimonious account of "what the data show" or "what the historical record reveals," then all strategies for causal analysis are largely irrelevant, including the counterfactual framework. But, there are also good reasons to consider descriptive analysis more generally. Berk (2004:218) notes that "... good description is the bread and butter of good science and good policy research." Sobel (1996:376) declares that "... many sociological questions neither require nor benefit from the introduction of causal considerations, and the tendency to treat such questions as if they are causal only leads to confusion."

We agree with this position to a great extent. But the appeal for descriptive research cannot be taken too far. Deep epistemological questions still arise in descriptive research, as not all descriptive accounts can be considered equally

worthy of our attention. Perhaps more importantly, social scientists may not want to backpedal from causal inquiry, lest journalists, partisan think-tank scholars, and politicians take it over entirely. We would most prefer to see both judicious descriptive analysis and careful causal analysis jointly crowd out poor causal analysis in the social sciences.

As we have noted at several points, when considering the causal controversies that pervade the literature, many social scientists have echoed the point summarized most clearly for sociology by John Goldthorpe, who writes, "sociologists have to find their own ways of thinking about causation, proper to the kinds of research that they can realistically carry out and the problems that they can realistically address" (Goldthorpe 2001:8). With this appeal in mind, we conclude this book with a discussion of the four primary modes of causal inquiry that exist in observational data analysis.

We have two goals for this concluding presentation of complementary modes of causal inquiry: (1) to affirm that all four modes of inquiry are valuable and (2) to argue that the counterfactual model is consistent with all of them and can help us to move productively between them.

Even so, we would not want to argue for the hegemony of counterfactual thinking. The specter of widespread mechanical adoption of the counterfactual model is truly frightening to us: No one wishes to have to review journal submissions in which scholars define the treatment effect for the treated, offer up a fairly arbitrary set of matching estimates, and then a disingenuous sensitivity analysis that purportedly bolsters the results. This prospect alone is sufficient for us to recommend considerable caution.[7]

But, perhaps more important, we cannot envision the pursuit of empirical analysis in our own research areas without engaging in analysis that has no explicit connection with the counterfactual model. In other words, we can see many future profitable lines of inquiry in our own research in which counterfactual thinking will play no role other than as a background tool that encourages clear thinking.

With that backing, we offer the following (necessarily oversimplified) representation of complementary modes of causal inquiry in observational social science:

Mode 1: Associational Analysis. In practice, most causal inquiry begins with an assessment, sometimes unreported, of whether a putative causal variable and outcome variable are associated in some way. It is often stated that establishing such an association is a precondition for subsequent causal analysis, as reflected in the aphorism "no causation without association."[8]

[7] When reading the penultimate draft of this book, one colleague asked, "Why is this worse than the senseless regression estimates that we have to review now?" It perhaps bears mentioning that this colleague was coming to the end of a tour of duty on the editorial board of a major journal.

[8] Formally, of course, there are cases for which this may not be true, such as when individual-varying causal effects perfectly cancel out each other or when suppression effects exist. We ignore exceptions such as these here, as they are rare. The sorts of physical equilibria that

Mode 2: Conditional Associational Analysis. After an association has been established, it is customary to then reestimate the association after conditioning on values of other observed variables. Although quite general, the typical approach usually follows one of two implementations: (1) conditioning on other variables that are thought to determine the outcome and that may also be related to the cause and (2) conditioning on variables that determine the cause and that may also be related to the outcome. Although conceptually quite different (as we noted in our discussion in Section 3.3), the goal of such conditioning is to eliminate prior causes of both the putative causal variable and the outcome variable, so as eliminate obvious sources of spuriousness.

Mode 3: Mechanism-Based Analysis. Perhaps after obvious forms of spuriousness have been eliminated through conditioning, a common practice in causal inquiry is to then introduce intervening variables between the putative causal variable and the outcome variable in an effort to provide a mechanistic explanation of the process that generates the causal effect. Such a form of causal inquiry can proceed, as is often the case, even though it may be unclear whether or not all forms of spurious association have been eliminated.

Mode 4: All-Cause Structural Analysis. Finally, at its most ambitious level, causal inquiry is pursued as an attempt to identify all causes in a chain of causality from the putative causal variable to the outcome variable, eliminating all spurious forms of association as a by-product. This approach is best represented in the forms of structural equation modeling that prevail in economics, wherein all specifications are justified with appeals to microeconomic theory and purport to explain the entire "who, when, where, and how" of all of the causes of the outcome of interest.

Consider now how the counterfactual model of causality can be related to these four modes of inquiry and how it facilitates movement between them as advances in theory construction and data collection are achieved. For mode 1 (associational analysis), the counterfactual model encourages the consideration of causes for which one can conceive of reasonable and theoretically possible conditions associated with an intervention that changes the causal variable from one value to another. When causal inquiry is stuck by convention on the consideration of the effects of attributes, for which the conditions of an as-if intervention are unclear, the model encourages additional effort to break the associational analysis into pieces that are more amenable to analysis by consideration of specific counterfactual dependencies. The hope of such a refined and expansive strategy is that the analysis can be brought down to a level where manipulations are more easily conceivable, which may then allow for a redirection of the conventional forms of associational causal inquiry. Such redirection may be necessary in order to open up the possibility of advancement beyond modes 1 and 2.

rigidly generate balancing of responses to causal shocks have no clear analogs in the social sciences that would generate similar perfect cancellation of unit-specific causal effects. And, although suppression effects exist in the social sciences, we cannot think of examples for which they have been shown to completely erase bivariate associations that are at least partly causal.

For the transition from mode 1 to mode 2 (that is, from associational to conditional associational analysis), the counterfactual model encourages separate but comparative consideration of the determinants of the cause and the determinants of the outcome. Determinants of causal variables lost some deserved attention as regression methods became dominant in observational social science in the 1980s and 1990s, but the matching literature associated with the counterfactual model has restored some of this attention by focusing analysis on the goal of balancing the data with respect to the determinants of the cause. Moreover, the joint consideration of both types of variables then allows for a determination, by carefully defined assumptions grounded in theory, of which sufficient subset of such determinants may be used in a conditioning strategy to achieve identification of the causal effect by eliminating all spurious associations. If the assumptions cannot be sustained for any set of observed variables, then point identification of the causal effect by back-door conditioning is impossible. If no other research design is possible, such as those that exemplify modes 3 and 4, then the counterfactual model suggests clearly why analysis should then expand beyond efforts to point estimate causal effects to alternative forms of partial identification and sensitivity analysis.

In addition, the counterfactual model also encourages especially careful examination of all back-door paths from the causal variable to the outcome variable (rather than simply a one-by-one consideration of all variables that lie along any back-door path). This form of systematic consideration of model identification can prevent researchers from mistakenly generating new spurious associations between the causal variable and the outcome variable, as occurs when conditioning on a collider variable that lies along an already blocked back-door path.

For the transition from mode 2 to mode 3, the counterfactual model shows that such a transition is substantive, not methodological. No new techniques beyond those used for mode 2 are required. All that is required to pursue mode 3 is a theory that suggests which variables (and in which configuration) represent the intervening causal mechanism that brings about the causal effect.

Moreover, the model shows that typical mechanism-based analyses must be clearly separated into two very different varieties: those that merely point to possible causal pathways and those that fully identify the causal effect of interest. For analysis of the former type, some explanation can be achieved. But, to realize the more ambitious goals of the latter, the mechanistic variables that are specified must be isolated and exhaustive (or made so by suitable conditioning). Mechanistic variables that are not independent of unblocked back-door paths after conditioning on observed variables cannot be used to identify a causal effect.

For transitions from modes 2 and 3 to mode 4, the counterfactual model sensitizes researchers to the stringency of the assumptions that usually must be maintained for all-cause structural analysis. Pearl (2000) is surely correct that our ambition should always be to get as close as possible to all-cause structural models. But, the counterfactual model clarifies the dubious nature of the maintained assumptions that pervade the published literature where all-cause models are offered.

For example, most such models are based on combinations of two identification strategies – basic conditioning and instrumental variable techniques – under maintained assumptions that causal effects are homogeneous (or at least conditionally random). As the matching literature associated with the counterfactual model has shown, such homogeneity is rarely justified empirically, and most parametric forms of conditioning average systematic causal effect heterogeneity in arcane ways. Moreover, the IV literature associated with the counterfactual model has shown that, in the presence of such heterogeneity, IVs estimate marginal causal effects that cannot then be extrapolated to other segments of the population of interest without introducing unsupportable assumptions of homogeneity of effects.

By making these issues clear, the counterfactual model shows how high the demands on theory and on data must be in order to sustain the all-cause mode of causal inquiry: Answers to all of the "who, when, where, and how" questions for the causal relationship of interest must be provided. Dodging these questions by introducing homogeneity assumptions that have only a dubious grounding in theory (or a solid grounding in dubious theory) can undermine causal claims, at least in the eyes of a fair critic. In this sense, the counterfactual model encourages modesty of causal inquiry.

Finally, such modesty can be pursued outside of these four modes of inquiry, after which analysis can then shift into whichever of these four modes seems most appropriate. As shown perhaps most clearly in Manski's partial identification perspective (see our discussion in Chapter 6), the range that a causal effect may take on can be defined and then examined within the counterfactual framework before any data are considered. Thereafter, an assessment can be undertaken of the data that are available to estimate average causal effects of various forms. For these steps, assumptions about why a causal relationship may exist need not be introduced, nor in fact assumptions that there is any systematic relationship between the processes that generate causal relationships at the individual level. Such flexibility leaves analysis wide open at the outset, focusing attention on clear definitions of effects of interest, independent of data availability issues.

But, once an analysis has begun, a theoretical position must be adopted in order to provide answers to at least some of the "who, when, where, and how" questions for the causal relationship of interest. Given provisional answers to these questions, the counterfactual model then reveals very precisely which modes of causal inquiry are feasible. Analysis may have to stop at the associational level, yielding nothing other than analogs to the naive estimator. If some of the determinants of the cause are systematic and observed, then analysis can move down to conditional variants of associational analysis, followed by an analysis of bounds (and perhaps a sensitivity analysis). If theory and data are available to examine what brings about the effect of the cause, then analysis can move down toward a mechanism-based mode of causal inquiry. Finally, if theory is finely articulated and supported by past substantive scholarship, and if all data requirements are met, then full structural modeling can be attempted. Such all-cause models represent a standard that is rarely achieved but

properly valued, for they provide complete explanations not only of the causes of an outcome but of every linkage in the causal chain between them. Counterfactual modeling guides us as close to this standard as is appropriate, given the constraints of theory and of data that prevail at any given time.

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *Journal of the American Statistical Association* 97:284–92.

Abadie, Alberto, David Drukker, Jane L. Herr, and Guido W. Imbens. 2004. "Implementing Matching Estimators for Average Treatment Effects in Stata." *The Stata Journal* 4:290–311.

Abadie, Alberto and Guido W. Imbens. 2004. "On the Failure of the Bootstrap for Matching Estimators." Working Paper, John F. Kennedy School of Government, Harvard University.

—. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74:235–67.

Abbott, Andrew D. 2001. *Time Matters: On Theory and Method*. Chicago: University of Chicago Press.

Agresti, Alan. 2002. *Categorical Data Analysis*. New York: Wiley.

Alexander, Jeffrey C. 2003. *The Meanings of Social Life: A Cultural Sociology*. New York: Oxford University Press.

Alexander, Karl L. and Aaron M. Pallas. 1983. "Private Schools and Public Policy: New Evidence on Cognitive Achievement in Public and Private Schools." *Sociology of Education* 56:170–82.

—. 1985. "School Sector and Cognitive Performance: When Is a Little a Little?" *Sociology of Education* 58:115–28.

Allison, Paul D. 1990. "Change Scores as Dependent Variables in Regression Analysis." *Sociological Methodology* 20:93–114.

Althauser, Robert P. and Donald B. Rubin. 1970. "The Computerized Construction of a Matched Sample." *American Journal of Sociology* 76:325–46.

—. 1971. "Measurement Error and Regression to the Mean in Matched Samples." *Social Forces* 50:206–14.

Altonji, Joseph G., Todd E. Elder, and Christopher Taber. 2005a. "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling." *Journal of Human Resources* 40:791–821.

—. 2005b. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113:151–83.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80:313–36.

—. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66:249–88.

Angrist, Joshua D. and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90:431–42.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 87:328–36.

Angrist, Joshua D. and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106:979–1014.

—. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association* 87:328–36.

—. 1994. "Why Do World War II Veterans Earn More Than Nonveterans?" *Journal of Labor Economics* 12:74–97.

—. 1999. "Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, edited by O. C. Ashenfelter and D. Card, Vol. 3, pp. 1277–1366. Amsterdam: Elsevier.

—. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15:65–83.

Angrist, Joshua D. and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114:533–75.

Arminger, Gerhard, Clifford C. Clogg, and Michael E. Sobel, Eds. 1995. *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Plenum.

Armor, David J. 1995. *Forced Justice: School Desegregation and the Law*. New York: Oxford University Press.

Ashenfelter, Orley C. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60:47–57.

Ashenfelter, Orley C. and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67:648–60.

Baltagi, Badi H. 2005. *Econometric Analysis of Panel Data*. Chichester, U.K.: Wiley.

Bang, Heejung and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61:962–72.

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." *Evaluation Studies Review Annual* 5:43–59.

Becker, Gary S. 1993[1964]. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. Chicago: University of Chicago Press.

Becker, Sascha O. and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal* 2:358–77.

Behrens, Angela, Christopher Uggen, and Jeff Manza. 2003. "Ballot Manipulation and the 'Menace of Negro Domination': Racial Threat and Felon Disenfranchisement in the United States, 1850–2002." *American Journal of Sociology* 109:559–605.

Berger, Mark C. and Barry T. Hirsch. 1983. "The Civilian Earnings Experience of Vietnam-Era Veterans." *Journal of Human Resources* 18:455–79.

Berk, Richard A. 1988. "Causal Inference for Sociological Data." In *Handbook of Sociology*, edited by N. J. Smelser, pp. 155–72. Newbury Park: Sage.

—. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks: Sage.

—. 2006. "An Introduction to Ensemble Methods for Data Analysis." *Sociological Methods and Research* 34:263–95.

—. 2007. *Statistical Learning from a Regression Perspective*. New York: Springer.

Berk, Richard A. and Jan de Leeuw. 1999. "An Evaluation of California's Inmate Classification System Using a Generalized Discontinuity Design." *Journal of the American Statistical Association* 94:1045–52.

Berk, Richard A. and Phyllis J. Newton. 1985. "Does Arrest Really Deter Wife Battery? An Effort to Replicate the Findings of the Minneapolis Spouse Abuse Experiment." *American Sociological Review* 50:253–62.

Berk, Richard A., Phyllis J. Newton, and Sarah Fenstermaker Berk. 1986. "What a Difference a Day Makes: An Empirical Study of the Impact of Shelters for Battered Women." *Journal of Marriage and the Family* 48:481–90.

Berk, Richard A. and David Rauma. 1983. "Capitalizing on Nonrandom Assignment to Treatments: A Regression-Discontinuity Evaluation of a Crime-Control Program." *Journal of the American Statistical Association* 78:21–7.

Bhaskar, Roy. 1998[1997]. "Philosophy and Scientific Realism." In *Critical Realism: Essential Readings*, edited by M. S. Archer, R. Bhaskar, A. Collier, T. Lawson, and A. Norrie, pp. 16–47. London, New York: Routledge.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114:577–99.

Blalock, Hubert M. 1964. *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press.

Blau, Peter M. and Otis Dudley Duncan. 1967. *The American Occupational Structure*. New York: Wiley.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

—. 1995. "Structural Equation Models That Are Nonlinear in Latent Variables: A Least Squares Estimator." *Sociological Methodology* 25:223–51.

—. 1996a. "An Alternative Two-Stage Least Squares (2SLS) Estimator for Latent Variable Equations." *Psychometrika* 61:109–21.

—. 1996b. "A Limited-Information Estimator for Lisrel Models With or Without Heteroscedastic Errors." In *Advanced Structural Equation Modeling: Issues and Techniques*, edited by G. A. Marcoulides and R. E. Schumacker, pp. 227–41. Mahwah, NJ: Erlbaum.

—. 2001. "Two-Stage Least Squares and Latent Variable Models: Simultaneous Estimation and Robustness to Misspecifications." In *Structural Equation Modeling: Present and Future – a Festschrift in Honor of Karl Jöreskog*, edited by R. Cudeck, S. du Toit, and D. Sörbom, pp. 119–38. Lincolnwood, IL: Scientific Software International.

Boudon, Raymond. 1998. "Social Mechanisms without Black Boxes." In *Social Mechanisms: An Analytical Approach to Social Theory*, edited by P. Hedström, and R. Swedberg, pp. 172–203. Cambridge: Cambridge University Press.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90:443–50.

Bourdieu, Pierre. 1973. "Cultural Reproduction and Social Reproduction." In *Knowledge, Education, and Cultural Change: Papers in the Sociology of Education*, edited by R. K. Brown, pp. 71–112. London: Tavistock.

Bowden, Roger J. and Darrell A. Turkington. 1984. *Instrumental Variables*. Cambridge: Cambridge University Press.

Brady, Henry E. and David Collier, Eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD.: Rowman & Littlefield.

Braga, Anthony A., David M. Kennedy, Elin J. Waring, and Anne Morrison Piehl. 2001. "Problem-Oriented Policing, Deterrence, and Youth Violence: An Evaluation of Boston's Operation Ceasefire." *Journal of Research in Crime and Delinquency* 38:195–225.

Brand, Jennie E. and Charles N. Halaby. 2006. "Regression and Matching Estimates of the Effect of Elite College Attendance on Educational and Career Achievement." *Social Science Research* 35:749–70.

Browning, Harley L., Sally C. Lopreato, and Dudley L. Poston Jr. 1973. "Income and Veteran Status: Variations Among Mexican Americans, Blacks and Anglos." *American Sociological Review* 38:74–85.

Bunge, Mario. 2004. "How Does It Work? The Search for Explanatory Mechanisms." *Philosophy of Social Science* 34:182–210.

Bush, Robert R. and Frederick Mosteller. 1955. *Stochastic Models for Learning*. New York: Wiley.

—. 1959. "A Comparison of Eight Models." In *Studies in Mathematical Learning Theory*, edited by R. R. Bush, and W. K. Estes, pp. 335–49. Stanford: Stanford University Press.

Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54:397–312.

—. 1988[1977]. *Methodology and Epistemology for Social Science: Selected Papers*. Chicago: University of Chicago Press.

Campbell, Donald T. and Julian C. Stanley. 1966[1963]. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, edited by O. C. Ashenfelter and D. Card, Vol. 3A, pp. 1801–63. Amsterdam: Elsevier.

Casella, George and Stephen P. Schwartz. 2000. "Comment on 'Causal Inference without Counterfactuals' by A. P. Dawid." *Journal of the American Statistical Association* 95:425–7.

Ceci, Stephen J. 1991. "How Much Does Schooling Influence General Intelligence and Its Cognitive Components? A Reassessment of the Evidence." *Developmental Psychology* 27:703–22.

Chapin, F. Stuart. 1932. "The Advantages of Experimental Sociology in the Study of Family Group Patterns." *Social Forces* 11:200–7.

—. 1947. *Experimental Designs in Sociological Research*. New York: Harper.

Chubb, John E. and Terry M. Moe. 1990. *Politics, Markets, and America's Schools*. Washington, D.C.: Brookings Institution.

Clotfelter, Charles T. 2004. *After Brown: The Rise and Retreat of School Desegregation*. Princeton: Princeton University Press.

Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24:295–313.

Cochran, William G. and Gertrude M. Cox. 1950. *Experimental Designs*. New York: Wiley.

Cochran, William G. and Donald B. Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhya* 35:417–46.

Coleman, James S. 1961. *The Adolescent Society: The Social Life of the Teenager and Its Impact on Education*. New York: Free Press.

—. 1964. *Introduction to Mathematical Sociology*. New York: Free Press.

—. 1981. *Longitudinal Data Analysis*. New York: Basic Books.

Coleman, James S. and Thomas Hoffer. 1987. *Public and Private Schools: The Impact of Communities*. New York: Basic Books.

Coleman, James S., Thomas Hoffer, and Sally Kilgore. 1982. *High School Achievement: Public, Catholic, and Private Schools Compared*. New York: Basic Books.

Collier, Andrew. 2005. "Philosophy and Critical Realism." In *The Politics of Method in the Human Sciences: Positivism and Its Epistemological Others*, edited by G. Steinmetz, pp. 327–45. Durham: Duke University Press.

Collins, John, Ned Hall, and L. A. Paul, Eds. 2004. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Cook, Michael D. and William N. Evans. 2000. "Families or Schools? Explaining the Convergence in White and Black Academic Performance." *Journal of Labor Economics* 18:729–54.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cox, D. R. 2000. "Comment on 'Causal Inference without Counterfactuals' by A. P. Dawid." *Journal of the American Statistical Association* 95:424–5.

Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.

—. 1992. *Planning of Experiments*. New York: Wiley.

Cox, David R. and Nancy Reid. 2000. *The Theory of the Design of Experiments*. Boca Raton: Chapman & Hall/CRC.

Cox, David R. and Nanny Wermuth. 1996. *Multivariate Dependencies: Models, Analysis and Interpretation*. New York: Chapman & Hall.

—. 2001. "Some Statistical Aspects of Causality." *European Sociological Review* 17:65–74.

Crain, Robert L. and Rita E. Mahard. 1983. "The Effect of Research Methodology on Desegregation-Achievement Studies: A Meta-Analysis." *American Journal of Sociology* 88:839–54.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand." Working Paper, Department of Economics, UC Berkeley.

Cutright, Phillips. 1974. "The Civilian Earnings of White and Black Draftees and Nonveterans." *American Sociological Review* 39:317–27.

Dawid, A. P. 2000. "Causal Inference without Counterfactuals." *Journal of the American Statistical Association* 95:407–24.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*:1053–62.

—. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84:151–161.

De Tray, Dennis. 1982. "Veteran Status as a Screening Device." *American Economic Review* 72:133–42.

Devlin, Bernie, Stephen E. Fienberg, Daniel P. Resnick, and Kathryn Roeder, Eds. 1997. *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve*. New York: Springer.

Diamond, Alexis and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper, Travers Department of Political Science, UC Berkeley.

DiPrete, Thomas A. and Markus Gangl. 2004. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." *Sociological Methodology* 34:271–310.

Draper, Norman R. and Harry Smith. 1998. *Applied Regression Analysis*. New York: Wiley.

Duncan, Otis Dudley. 1966. "Path Analysis: Sociological Examples." *American Journal of Sociology* 72:1–16.

—. 1975. *Introduction to Structural Equation Models*. New York: Academic.

—. 1984. *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation.

Duncan, Otis Dudley, David L. Featherman, and Beverly Duncan. 1972. *Socioeconomic Background and Achievement*. New York: Seminar.

Elster, Jon. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Evans, William and Robert M. Schwab. 1995. "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics* 110:41–74.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyde.

Fox, John. 1984. *Linear Statistical Models and Related Methods: With Applications to Social Research*. New York: Wiley.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58:21–9.

Frank, Kenneth A. 2000. "Impact of a Confounding Variable on the Inference of a Regression Coefficient." *Sociological Methods and Research* 29:147–94.

Freedman, David A. 1983. "A Note on Screening Regression Equations." *American Statistician* 37:152–5.

—. 2005. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Freedman, Ronald and Amos H. Hawley. 1949. "Unemployment and Migration in the Depression." *Journal of the American Statistical Association* 44:260–72.

Fu, Vincent Kang, Christopher Winship, and Robert D. Mare. 2004. "Sample Selection Bias Models." In *Handbook of Data Analysis*, edited by M. A. Hardy, and A. Bryman, pp. 409–30. Thousand Oaks: Sage.

Fuller, Bruce and Richard F. Elmore. 1996. *Who Chooses? Who Loses? Culture, Institutions, and the Unequal Effects of School Choice*. New York: Teachers College Press.

Garen, John. 1984. "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable." *Econometrica* 52:1199–1218.

Garfinkel, Irwin, Charles F. Manski, and C. Michalopoulos. 1992. "Micro Experiments and Macro Effects." In *Evaluating Welfare and Training Programs*, edited by C. F. Manski and I. Garfinkel, pp. 253–73. Cambridge, MA: Harvard University Press.

Gastwirth, Joseph L., Abba M. Krieger, and Paul R. Rosenbaum. 1998. "Dual and Simultaneous Sensitivity Analysis for Matched Pairs." *Biometrika* 85:907–20.

—. 2000. "Asymptotic Separability in Sensitivity Analysis." *Journal of the Royal Statistical Society, Series B* 62:545–55.

Gelman, Andrew and Jennifer Hill. 2007. *Applied Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew and Gary King. 1990. "Estimating Incumbency Advantage without Bias." *American Journal of Political Science* 34:1142–64.

Gelman, Andrew and Xiao-Li Meng, Eds. 2004. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*. New York: Wiley.

Glymour, Clark. 1986. "Statistics and Metaphysics: Comment on Holland's 'Statistics and Causal Inference'." *Journal of the American Statistical Association* 81:964–6.

Godfrey-Smith, Peter. 2003. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: University of Chicago Press.

Goldberger, Arthur S. 1972. "Structural Equation Methods in the Social Sciences." *Econometrica* 40:979–1001.

—. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Goldberger, Arthur S. and Glen G. Cain. 1982. "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report." *Sociology of Education* 55:103–22.

Goldthorpe, John H. 2000. *On Sociology: Numbers, Narratives, and the Integration of Research and Theory*. Oxford: Oxford University Press.

—. 2001. "Causation, Statistics, and Sociology." *European Sociological Review* 17:1–20.

Gorski, Philip S. 2004. "The Poverty of Deductivism: A Constructive Realist Model of Sociological Explanation." *Sociological Methodology* 34:1–33.

Greene, William H. 2000. *Econometric Analysis*. Upper Saddle River: Prentice-Hall.

Greenland, Sander and Babette Brumback. 2002. "An Overview of Relations Among Causal Modelling Methods." *International Journal of Epidemiology* 31:1030-7.

Greenwood, Ernest. 1945. *Experimental Sociology: A Study in Method*. New York: King's Crown Press.

Haavelmo, Trygve. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 11:1–12.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66:315–31.

Hahn, Jinyong and Jerry Hausman. 2003. "Weak Instruments: Diagnosis and Cures in Empirical Economics." *American Economic Review* 93:118–25.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69:201–9.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory Into Practice." *Annual Review of Sociology* 30:507–44.

Ham, J. C., X. Li, and P. B. Reagan. 2003. "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men." Working Paper, Department of Sociology and Center for Human Resource Research, Ohio State University.

Hamilton, James D. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Hansen, Ben B. 2004a. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99:609–18.

—. 2004b. "Optmatch, an Add-on Package for R." Department of Statistics, University of Michigan.

Harding, David J. 2003. "Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping out and Teenage Pregnancy." *American Journal of Sociology* 109:676–719.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hauser, Philip M. 1962. "On Stouffer's Social Research to Test Ideas." *Public Opinion Quarterly* 26:329–34.

Hauser, Robert M., John Robert Warren, Min-Hsiung Huang, and Wendy Y. Carter. 2000. "Occupational Status, Education, and Social Mobility in the Meritocracy." In *Meritocracy and Economic Inequality*, edited by K. J. Arrow, S. Bowles, and S. N. Durlauf, pp. 179–229. Princeton: Princeton University Press.

Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.

Heckman, James J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica* 42:679–94.

—. 1978. "Dummy Endogenous Variables in a Simultaneous Equation." *Econometrica* 46:931–61.

—. 1979. "Selection Bias as a Specification Error." *Econometrica* 47:153-61.

—. 1989. "Causal Inference and Nonrandom Samples." *Journal of Educational Statistics* 14:159–68.

—. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, edited by C. F. Manski and I. Garfinkel, pp. 201–30. Cambridge, MA: Harvard University Press.

—. 1996. "Randomization as an Instrumental Variable." *Review of Economics and Statistics* 77:336–41.

—. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *The Journal of Human Resources* 32:441–62.

—. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *The Quarterly Journal of Economics* 115:45–97.

—. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35:1–97.

Heckman, James J. and V. Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84:862–74.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017–98.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64:605–54.

—. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65:261–94.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, edited by O. C. Ashenfelter and D. Card, Vol. 3, pp. 1865–2097. Amsterdam: Elsevier.

Heckman, James J. and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, edited by J. J. Heckman and B. Singer, pp. 156–245. Cambridge: Cambridge University Press.

—. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." In *Drawing Inferences from Self-Selected Samples*, edited by H. Wainer, pp. 63–113. New York: Springer-Verlag.

—. 1989. "The Value of Longitudinal Data for Solving the Problem of Selection Bias in Evaluating the Impact of Treatment on Outcomes." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 512–38. New York: Wiley.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64:487–535.

Heckman, James J., Justin L. Tobias, and Edward Vytlacil. 2003. "Simple Estimators for Treatment Parameters in a Latent-Variable Framework." *The Review of Economics and Statistics* 85:748–55.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." Working Paper, Department of Economics, University of Chicago.

Heckman, James J. and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America* 96:4730–4.

—. 2000. "The Relationship between Treatment Parameters within a Latent Variable Framework." *Economics Letters* 66:33–9.

—. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73:669–738.

Hedström, Peter. 2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Hedström, Peter and Richard Swedberg. 1998. *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Hendry, David F. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Hernan, Miguel A., Sonia Hernandez-Diaz, and James M. Robins. 2004. "A Structural Approach to Selection Bias." *Epidemiology* 15:615–25.

Herrnstein, Richard J. and Charles A. Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Herron, Michael C. and Jasjeet S. Sekhon. 2003. "Overvoting and Representation: An Examination of Overvoted Presidential Ballots in Broward and Miami-Dade Counties." *Electoral Studies* 22:21–47.

Hirano, Keisuke and Guido W. Imbens. 2001. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services & Outcomes Research Methodology* 2:259–78.

—. 2004. "The Propensity Score with Continuous Treatments." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, edited by A. Gelman and X.-L. Meng, pp. 73–84. New York: Wiley.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71:1161–89.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2004. "Matchit: Nonparametric Preprocessing for Parametric Causal Inference." Department of Government, Harvard University.

—. 2005. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Working Paper, Department of Government, Harvard University.

Hoffer, Thomas, Andrew M. Greeley, and James S. Coleman. 1985. "Achievement Growth in Public and Catholic Schools." *Sociology of Education* 58:74–97.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–70.

Holland, Paul W. and Donald B. Rubin. 1983. "On Lord's Paradox." In *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, edited by H. Wainer and S. Messick, pp. 3–25. Hillsdale: Erlbaum.

Hong, Guanglei and Stephen W. Raudenbush. 2006. "Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data." *Journal of the American Statistical Association* 101:901–10.

Honoré, Bo E. and James L. Powell. 1994. "Pairwise Difference Estimators of Censored and Truncated Regression Models." *Journal of Econometrics* 64:241–78.

Hood, William C. and Tjalling C. Koopmans, Eds. 1953. *Studies in Econometric Method*. New York: Wiley.

Howell, William G. and Paul E. Peterson. 2002. *The Education Gap: Vouchers and Urban Schools*. Washington, D.C.: Brookings Institution Press.

Hoxby, Caroline M. 1996. "The Effects of Private School Vouchers on Schools and Students." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by H. F. Ladd, pp. 177–208. Washington, D.C.: Brookings Institution Press.

Hoxby, Caroline M, Ed. 2003. *The Economics of School Choice*. Chicago: University of Chicago Press.

Hsiao, Cheng. 2003. *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Humphreys, Paul. 1989. *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences*. Princeton: Princeton University Press.

—. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. New York: Oxford University Press.

Hyman, Herbert H. 1962. "Samuel A. Stouffer and Social Research." *Public Opinion Quarterly* 26:323–8.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2006. "The Balance Test Fallacy in Matching Methods for Causal Inference." Working Paper, Department of Government, Harvard University.

Imai, Kosuke and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99:854–66.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87:706–10.

—. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86:4–29.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467–75.

Imbens, Guido W. and Donald B. Rubin. 1997. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models." *Review of Economic Studies* 64:555–74.

Joffe, Marshall M. and Paul R. Rosenbaum. 1999. "Propensity Scores." *American Journal of Epidemiology* 150:327–31.

Judd, Charles M. and David A. Kenny. 1981. *Estimating the Effects of Social Interventions*. New York: Cambridge University Press.

Kempthorne, Oscar. 1948. "Review of Experimental Designs in Sociological Research by F. Stuart Chapin." *Journal of the American Statistical Association* 43:489–92.

—. 1952. *The Design and Analysis of Experiments*. New York: Wiley.

Kendall, Patricia L. and Paul F. Lazarsfeld. 1950. "Problems of Survey Analysis." In *Continuities in Social Research: Studies in the Scope and Method of "The American Soldier,"* edited by R. K. Merton and P. F. Lazarsfeld, pp. 133–96. Glencoe: Free Press.

Kennedy, David M. 1997. "Pulling Levers: Chronic Offenders, High-Crime Settings, and a Theory of Prevention." *Valparaiso University Law Review* 31:449–84.

Keyfitz, Nathan. 1948. "Review of Experimental Designs in Sociological Research by F. Stuart Chapin." *American Journal of Sociology* 54:259–60.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton. Princeton University Press.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

Koopmans, Tjalling C. and Olav Reiersøl. 1950. "The Identification of Structural Characteristics." *Annals of Mathematical Statistics* 21:165–81.

Krueger, Alan B. and Pei Zhu. 2004. "Another Look at the New York City School Voucher Experiment." *American Behavioral Scientist* 47:658–98.

Krueger, Thomas M. and William F. Kennedy. 1990. "An Examination of the Super Bowl Stock Market Predictor." *Journal of Finance* 45:691–7.

Ladd, Helen F. 2002. "School Vouchers: A Critical View." *Journal of Economic Perspectives* 16:3–24.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76:604–20.

—. 1995. "The Promise of Public Sector-Sponsored Training Programs." *Journal of Economic Perspectives* 9:149–68.

Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet. 1955[1948]. "Political Interest and Voting Behavior." In *The Language of Social Research: A Reader in the Methodology of Social Research*, edited by P. F. Lazarsfeld and M. Rosenberg, pp. 155–8. Glencoe: Free Press.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

—. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73:31–43.

Lechner, Michael. 2002a. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies." *The Review of Economics and Statistics* 84:205–20.

—. 2002b. "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods." *Journal of Royal Statistical Society* 165:59–82.

Leuven, Edwin and Barbara Sianesi. 2003. "Psmatch2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing." Boston College Department of Economics, Statistical Software Components.

Lewis, David. 1973. "Causation." *Journal of Philosophy* 70:556–67.

Lieberson, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Lieberson, Stanley and Freda B. Lynn. 2002. "Barking up the Wrong Branch: Scientific Alternatives to the Current Model of Sociological Science." *Annual Review of Sociology* 28:1–19.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Lu, Bo, Elaine Zanutto, Robert Hornik, and Paul R. Rosenbaum. 2001. "Matching with Doses in an Observational Study of a Media Campaign Against Drug Abuse." *Journal of the American Statistical Association* 96:1245–53.

Lunceford, Jared K. and Marie Davidian. 2004. "Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." *Statistics in Medicine* 23:2937–60.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67:1–25.

Mahoney, James and Gary Goertz. 2006. "A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research." *Political Analysis* 14:227–49.

Manski, Charles F. 1994. "The Selection Problem." In *Advances in Econometrics: Sixth World Congress*, edited by C. A. Sims, Vol. 1, pp. 143–70. Cambridge: Cambridge University Press.

—. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

—. 1997. "Monotone Treatment Response." *Econometrica* 65:1311–34.

—. 1999. "Comment on 'Choice as an Alternative to Control in Observational Studies' by Rosenbaum." *Statistical Science* 14:279–81.

—. 2003. *Partial Identification of Probability Distributions*. New York: Springer.

Manski, Charles F. and Irwin Garfinkel, Eds. 1992. *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.

Manski, Charles F. and Daniel S. Nagin. 1998. "Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism." *Sociological Methodology* 28:99–137.

Manski, Charles F. and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68:997–1010.

Manski, Charles F., Gary D. Sandefur, Sara McLanahan, and Daniel Powers. 1992. "Alternative Estimates of the Effect of Family Structure During Adolescence on High School Graduation." *Journal of the American Statistical Association* 87:25–37.

Manza, Jeff and Christopher Uggen. 2004. "Punishment and Democracy: Disenfranchisement of Nonincarcerated Felons in the United States." *Perspectives on Politics* 2:491–505.

Marcantonio, Richard J. and Thomas D. Cook. 1994. "Convincing Quasi-Experiments: The Interrupted Time Series and Regression-Discontinuity Designs." In *Handbook of Practical Program Evaluation*, edited by J. S. Wholey, H. P. Hatry, and K. E. Newcomer, pp. 133–51. San Francisco: Jossey-Bass.

Marini, Margaret M. and Burton Singer. 1988. "Causality in the Social Sciences." *Sociological Methodology* 18:347–409.

Mark, Melvin M. and Steven Mellor. 1991. "Effect of Self-Relevance of an Event on Hingsight Bias: The Foreseeability of a Layoff." *Journal of Applied Psychology* 76:569–77.

McDowall, David, Richard McCleary, Errol E. Meidinger, and Richard A. Hay Jr. 1980. *Interrupted Time Series Analysis*. Beverly Hills: Sage.

Mebane, Walter R. 2004. "The Wrong Man Is President! Overvotes in the 2000 Presidential Election in Florida." *Perspectives on Politics* 2:525–35.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159:56–63.

Moffitt, Robert A. 1996. "Comment on 'Identification of Causal Effects Using Instrumental Variables' by Angrist, Imbens, and Rubin." *Journal of the American Statistical Association* 91:462–65.

—. 2003. "Causal Analysis in Population Research: An Economist's Perspective." *Population and Development Review* 29:448–58.

Morgan, Stephen L. 2001. "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning." *Sociology of Education* 74:341–74.

—. 2005. *On the Edge of Commitment: Educational Attainment and Race in the United States*. Stanford: Stanford University Press.

Morgan, Stephen L. and David J. Harding. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods and Research* 35:3–60.

Murnane, Richard J., Stephen E. Newstead, and Randall J. Olsen. 1985. "Comparing Public and Private Schools: The Puzzling Role of Selectivity Bias." *Journal of Business and Economic Statistics* 3:23–35.

Neal, Derek. 1997. "The Effects of Catholic Secondary Schooling on Educational Achievement." *Journal of Labor Economics* 14:98–123.

—. 2002. "How Vouchers Could Change the Market for Education." *Journal of Economic Perspectives* 16:25–44.

Neyman, Jerzy Splawa. 1935. "Statistical Problems in Agricultural Experimentation (with Discussion)." *Journal of the Royal Statistical Society, Series B* 2:107–80.

—. 1990[1923]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5:465–80.

Nie, Norman H., Jane Junn, and Kenneth Stehlik-Barry. 1996. *Education and Democratic Citizenship in America*. Chicago: University of Chicago Press.

Noell, Jay. 1982. "Public and Catholic Schools: A Reanalysis of 'Public and Private Schools'." *Sociology of Education* 55:123–32.

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Petersen, Trond, Andrew Penner, and Geir Høgsnes. 2006. "The Male Marital Wage Premium: Sorting Versus Differential Pay." Working Paper, Department of Sociology, University of California, Berkeley.

Peterson, Paul E. and William G. Howell. 2004. "Efficiency, Bias, and Classification Schemes: A Response to Alan B. Krueger and Pei Zhu." *American Behavioral Scientist* 47:699–717.

Powers, Daniel A. and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. San Diego: Academic.

Pratt, John W. and Robert Schlaifer. 1984. "On the Nature and Discovery of Structure." *Journal of the American Statistical Association* 79:9–33.

—. 1988. "On the Interpretation and Observation of Laws." *Journal of Econometrics* 39:23–52.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Putnam, Hilary. 1975. *Mind, Language, and Reality*. New York: Cambridge University Press.

Quandt, Richard E. 1972. "A New Approach to Estimating Switching Regression." *Journal of the American Statistical Association* 67:306–10.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–63.

Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

—. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks: Sage.

Reiersøl, Olav. 1941. "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis." *Econometrica* 9:1–24.

Reskin, Barbara F. 2003. "Including Mechanisms in Our Models of Ascriptive Inequality." *American Sociological Review* 68:1–21.

Robins, James M. and Sander Greenland. 2000. "Comment on 'Causal Inference without Counterfactuals' by A. P. Dawid." *Journal of the American Statistical Association* 95:431–5.

Robins, James M. and Ya'acov Ritov. 1997. "Toward a Curse of Dimensionality Appropriate (Coda) Asymptotic Theory for Semi-Parametric Models." *Statistics in Medicine* 16:285–319.

Robins, James M. and Andrea Rotnitzky. 2001. "Comment on 'Inference for Semiparametric Models: Some Questions and an Answer' by Bickel and Kwon." *Statistica Sinica* 11:920–36.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89:846–66.

Rose, Arnold. 1962. "Review of 'Social Research to Test Ideas: Selected Writings' by Samuel A. Stouffer." *American Sociological Review* 27:720–1.

Rosen, Sherwin and Paul Taubman. 1982. "Changes in Life-Cycle Earnings: What Do Social Security Data Show?" *Journal of Human Resources* 17:321–38.

Rosenbaum, Paul R. 1984a. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147:656–66.

—. 1984b. "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment." *Journal of the American Statistical Association* 79:41–8.

—. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82:387–94.

—. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84:1024–32.

—. 1991. "Sensitivity Analysis for Matched Case Control Studies." *Biometrics* 47:87–100.

—. 1992. "Detecting Bias with Confidence in Observational Studies." *Biometrika* 79:367–74.

—. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14:259–304.

—. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983a. "Assessing Sensitivity to an Unobserved Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society* 45:212–18.

—. 1983b. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

—. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516–24.

—. 1985a. "The Bias Due to Incomplete Matching." *Biometrics* 41:103–16.

—. 1985b. "Constructing a Control Group Using Multivariate Matched Sampling Methods." *American Statistician* 39:33–8.

Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38:827–74.

Rossell, Christine H., David J. Armor, and Herbert J. Walberg, Eds. 2002. *School Desegregation in the 21st Century*. Westport: Praeger.

Rothman, Kenneth J. and Sander Greenland. 1998. *Modern Epidemiology*. Philadelphia: Lippincott.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3:135–46.

Rubin, Donald B. 1973a. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159–83.

—. 1973b. "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29:185–203.

—. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688–701.

—. 1976. "Multivariate Matching Methods That Are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32:109–20.

—. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2:1–26.

—. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6:34–58.

—. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74:318–28.

—. 1980a. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36:293–298.

—. 1980b. "Comment on 'Randomization Analysis of Experimental Data in the Fisher Randomization Test' by Basu." *Journal of the American Statistical Association* 75:591–3.

—. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6:377–400.

—. 1986. "Which Ifs Have Causal Answers (Comment on 'Statistics and Causal Inference' by Paul W. Holland)." *Journal of the American Statistical Association* 81:961–2.

—. 1990. "Formal Modes of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25:279–92.

—. 1991. "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism." *Biometrics* 47:1213–34.

—. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100:322–31.

Rubin, Donald B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52:249–64.

—. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573–85.

Ruppert, David, M. P. Wand, and Raymond J. Carroll. 2003. *Semiparametric Regression*. Cambridge: Cambridge University Press.

Ruud, Paul A. 2000. *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

—. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Saltelli, Andrea, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. 2004. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Hoboken: Wiley.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. "Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association* 94:1096–1120.

Schofield, Janet Ward. 1995. "Review of Research on School Desegregation's Impact on Elementary and Secondary School Students." In *Handbook of Research on Multicultural Education*, edited by J. A. Banks and C. A. M. Banks, pp. 597–617. New York: Macmillan.

Schultz, Henry. 1938. *The Theory and Measurement of Demand*. Chicago: University of Chicago Press.

Schwartz, Saul. 1986. "The Relative Earnings of Vietnam and Korean-Era Veterans." *Industrial and Labor Relations Review* 39:564–572.

Sekhon, Jasjeet S. 2004. "The 2004 Florida Optical Voting Machine Controversy: A Causal Analysis Using Matching." Working Paper, Department of Government, Harvard University.

—. 2005. "Matching: Multivariate and Propensity Score Matching with Balance Optimization." Travers Department of Political Science, UC Berkeley.

Sewell, William H. 1964. "Community of Residence and College Plans." *American Sociological Review* 29:24–38.

Sewell, William H., Archibald O. Haller, and George W. Ohlendorf. 1970. "The Educational and Early Occupational Status Attainment Process: Replication and Revision." *American Sociological Review* 35:1014–1024.

Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. "The Educational and Early Occupational Attainment Process." *American Sociological Review* 34:82–92.

Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Singer, Burton and Margaret M. Marini. 1987. "Advancing Social Research: An Essay Based on Stanley Lieberson's Making It Count." *Sociological Methodology* 17:373–391.

Smith, Herbert L. 1990. "Specification Problems in Experimental and Nonexperimental Social Research." *Sociological Methodology* 20:59–91.

—. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27:325–53.

—. 2003. "Some Thoughts on Causation as It Relates to Demography and Population Studies." *Population and Development Review* 29:459–69.

Smith, Jeffery A. and Petra Todd. 2005. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125: 305–353.

Sobel, Michael E. 1995. "Causal Inference in the Social and Behavioral Sciences." In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel, pp. 1–38. New York: Plenum.

—. 1996. "An Introduction to Causal Inference." *Sociological Methods and Research* 24:353–79.

—. 2000. "Causal Inference in the Social Sciences." *Journal of the American Statistical Association* 95:647–51.

—. 2006. "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference." Working Paper, Department of Sociology, Columbia University.

Sørensen, Aage B. 1998. "Theoretical Mechanisms and the Empirical Study of Social Processes." In *Social Mechanisms: An Analytical Approach to Social Theory*, *Studies in Rationality and Social Change*, edited by P. Hedström and R. Swedberg, pp. 238–66. Cambridge: Cambridge University Press.

Sørensen, Aage B. and Stephen L. Morgan. 2000. "School Effects: Theoretical and Methodological Issues." In *Handbook of the Sociology of Education*, edited by M. T. Hallinan, pp. 137–60. New York: Kluwer/Plenum.

Staiger, Douglas and James H. Stock. 1997. "IV Regression with Weak Instruments." *Econometrica* 65:557–86.

Stolzenberg, Ross M. 2004. "Multiple Regression Analysis." In *Handbook of Data Analysis,* edited by M. A. Hardy and A. Bryman, pp. 165–207. Thousand Oaks: Sage.

Stouffer, Samuel A. 1949. *The American Soldier*. Princeton: Princeton University Press.

—. 1950. "Some Observations on Study Design." *American Journal of Sociology* 55:355–61.

—. 1955. *Communism, Conformity, and Civil Liberties: A Cross-Section of the Nation Speaks Its Mind*. Garden City: Doubleday.

—. 1962. *Social Research to Test Ideas*. Glencoe: Free Press.

Thistlewaite, D. L. and Donald T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51:309–17.

Thompson, Steven K. 2002. *Sampling*. New York: Wiley.

Trochim, William M. K. 1984. *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Beverly Hills: Sage.

Tu, Wanzhu and Xiao-Hua Zhou. 2002. "A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification." *Health Services & Outcomes Research Methodology* 3:135–47.

Uggen, Christopher, Angela Behrens, and Jeff Manza. 2005. "Criminal Disenfranchisement." *Annual Review of Law and Social Science* 1:307–22.

Uggen, Christopher and Jeff Manza. 2002. "Democratic Contraction? Political Consequences of Felon Disenfranchisement in the United States." *American Sociological Review* 67:777–803.

Van der Klaauw, Wilbert. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43:1249–87.

van der Laan, M. J. and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Verba, Sidney and Norman H. Nie. 1972. *Participation in America: Political Democracy and Social Equality*. New York: Harper.

Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Harvard University Press.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70:331–41.

Wald, Abraham. 1940. "The Fitting of Straight Lines If Both Variables Are Subject to Error." *Annals of Mathematical Statistics* 11:284–300.

Wand, Jonathan, Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr., Michael C. Herron, and Henry E. Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95:793–810.

Wells, Amy Stuart and Robert L. Crain. 1994. "Perpetuation Theory and the Long-Term Effects of School Desegregation." *Review of Educational Research* 64:531–55.

Willis, Robert and Sherwin Rosen. 1979. "Education and Self-Selection." *Journal of Political Economy* 87:S7–S35.

Willms, J. Douglas. 1985. "Catholic-School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-Up Study." *Sociology of Education* 58:98–114.

Winship, Christopher and David J. Harding. Forthcoming. "A Mechanism Based Approach to the Identification of Age-Period-Cohort Models." *Sociological Methods and Research*.

Winship, Christopher and Sanders Korenman. 1997. "Does Staying in School Make You Smarter? The Effect of Education on IQ in the Bell Curve." In *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve*, edited by B. Devlin, S. E. Fienberg, D. P. Resnick, and K. Roeder, pp. 215–34. New York: Springer.

Winship, Christopher and Robert D. Mare. 1984. "Regression Models with Ordinal Variables." *American Sociological Review* 49:512–25.

—. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–50.

Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659–706.

Winship, Christopher and Michael E. Sobel. 2004. "Causal Analysis in Sociological Studies." In *Handbook of Data Analysis*, edited by M. A. Hardy and A. Bryman, pp. 481–503. Thousand Oaks: Sage.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Working, E. J. 1927. "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics* 41:212–35.

Working, Holbrook. 1925. "The Statistical Determination of Demand Curves." *Quarterly Journal of Economics* 39:503–45.

Wright, Sewall. 1921. "Correlation and Causation." *Journal of Agricultural Research* 20:557–85.

—. 1925. "Corn and Hog Correlations." U.S. Department of Agriculture, Washington, D.C.

—. 1934. "The Method of Path Coefficients." *Annals of Mathematical Statistics* 5:161-15.

Xie, Yu. 2006. "Otis Dudley Duncan's Legacy: The Demographic Approach to Quantitative Reasoning in Social Science." Population Studies Center, University of Michigan, Ann Arbor.

Yinger, Milton J., Kiyoshi Ikeda, and Frank Laycock. 1967. "Treating Matching as a Variable in a Sociological Experiment." *American Sociological Review* 32:801–12.

Index