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

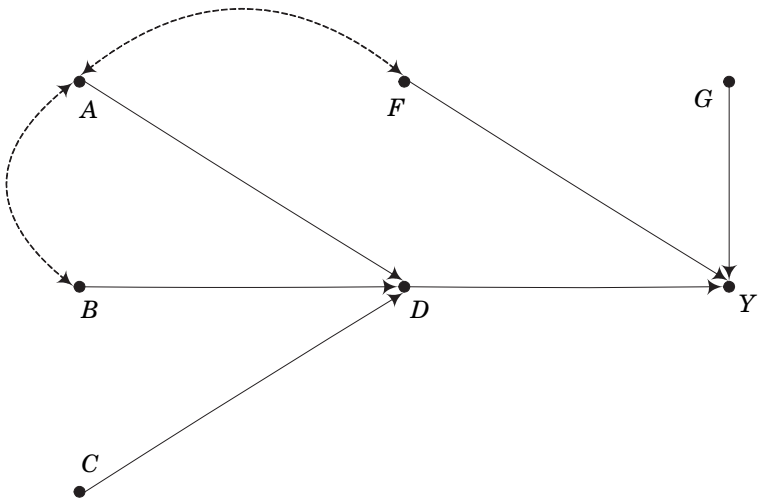

Figure 1.1: A causal diagram in which back-door paths from D to Y can be blocked by observable variables and C is an instrumental variable for D.

(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.

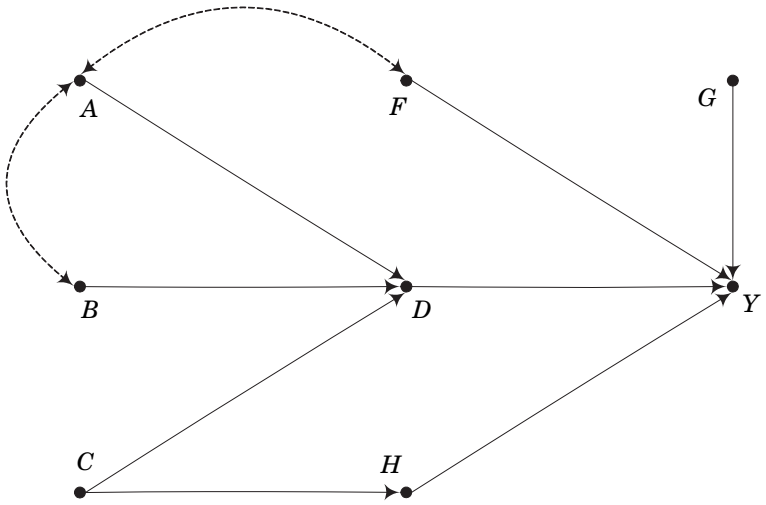

Figure 1.2: A causal diagram in which C is no longer an instrumental variable for D.

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.

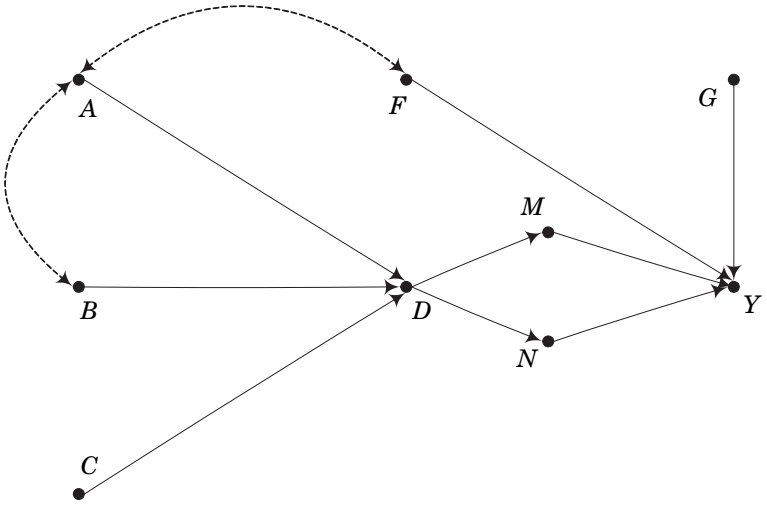

Figure 1.3: A causal diagram in which M and N represent an isolated and exhaustive mechanism for the causal effect of D on Y.

$\{F, C\}$, 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.