Morgan / Stephen L., Winship / Christopher (2007): Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press. ISBN: 978-0-521-67193-4

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 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

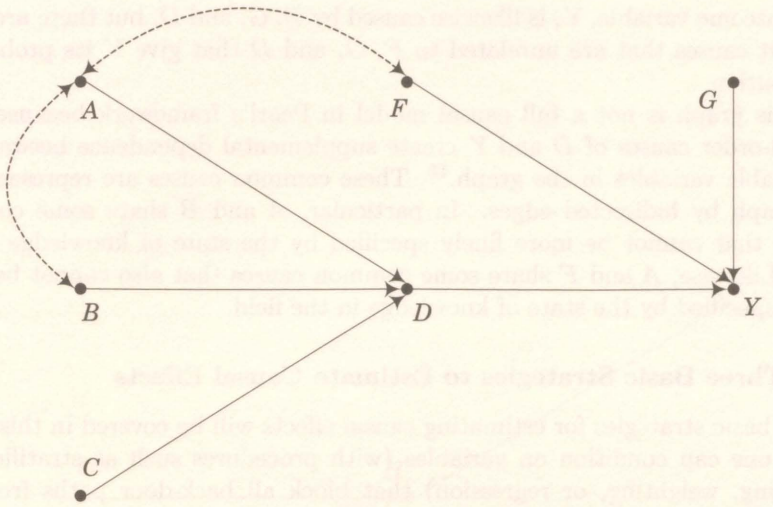

Figure 1.1: A causal diagram in which back-door paths from D to Y can be blocked by observable variables and C is an instrumental variable for D.

(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.

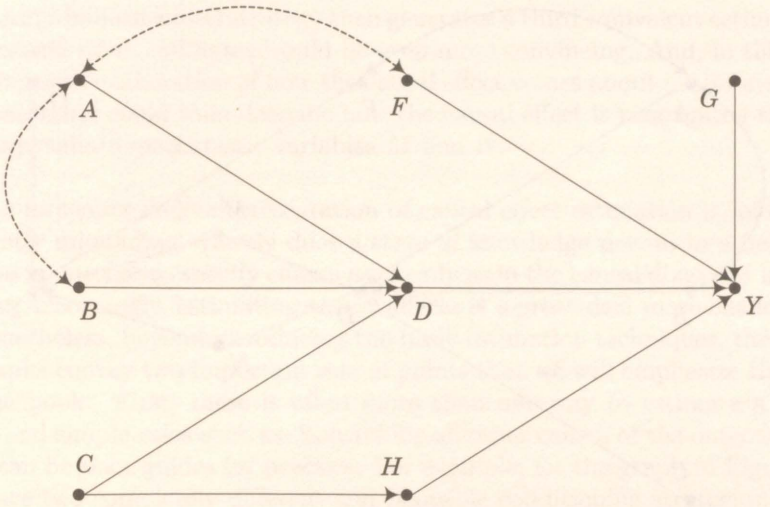

Figure 1.2: A causal diagram in which C is no longer an instrumental variable for D.

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.

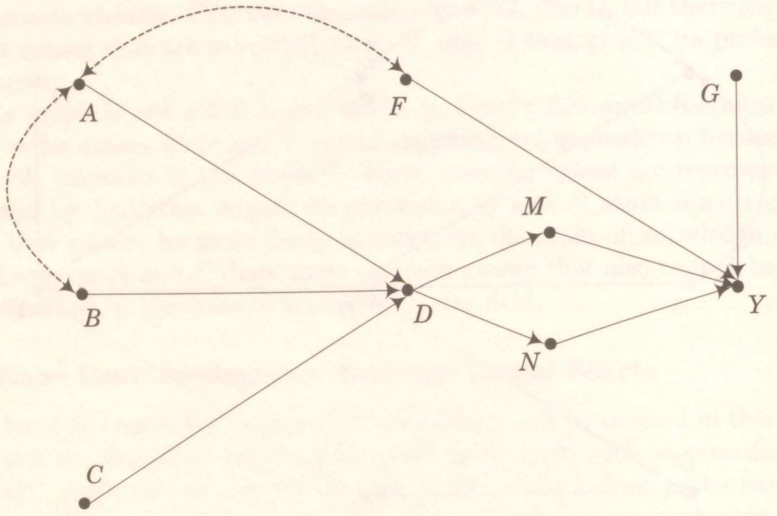

Figure 1.3: A causal diagram in which M and N represent an isolated and exhaustive mechanism for the causal effect of D on Y.

$\{F, C\}$, 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.