The Estimation of Causal Effects from Observational Data

Annu. Rev. Sociol. 1999. 25:659–706

THE ESTIMATION OF CAUSAL EFFECTS FROM OBSERVATIONAL DATA

Christopher Winship and Stephen L. Morgan
Harvard University, Department of Sociology, William James Hall, 33 Kirkland Street, Cambridge, Massachusetts 02138; e-mail: winship@wjh.harvard.edu; smorgan@wjh.harvard.edu

KEY WORDS: causal inference, causal analysis, counterfactual, treatment effect, selection bias

ABSTRACT

When experimental designs are infeasible, researchers must resort to the use of observational data from surveys, censuses, and administrative records. Because assignment to the independent variables of observational data is usually nonrandom, the challenge of estimating causal effects with observational data can be formidable. In this chapter, we review the large literature produced primarily by statisticians and econometricians in the past two decades on the estimation of causal effects from observational data. We first review the now widely accepted counterfactual framework for the modeling of causal effects. After examining estimators, both old and new, that can be used to estimate causal effects from cross-sectional data, we present estimators that exploit the additional information furnished by longitudinal data. Because of the size and technical nature of the literature, we cannot offer a fully detailed and comprehensive presentation. Instead, we present only the main features of methods that are accessible and potentially of use to quantitatively oriented sociologists.

INTRODUCTION

Most quantitative empirical analyses are motivated by the desire to estimate the causal effect of an independent variable on a dependent variable. Although the randomized experiment is the most powerful design for this task, in most social science research done outside of psychology, experimental designs are infeasible. Social experiments are often too expensive and may require the

0360-0572/99/0815-0659$08.00

unethical coercion of subjects. Subjects may be unwilling to follow the experimental protocol, and the treatment of interest may not be directly manipulable. For example, without considerable power and a total absence of conscience, a researcher could not randomly assign individuals to different levels of educational attainment in order to assess the effect of education on earnings. For these reasons, sociologists, economists, and political scientists must rely on what is now known as observational data—data that have been generated by something other than a randomized experiment—typically surveys, censuses, or administrative records.

The problems of using observational data to make causal inferences are considerable (Lieberson 1985, LaLonde 1986). In the past two decades, however, statisticians (e.g. Rubin, Rosenbaum) and econometricians (e.g. Heckman, Manski) have made considerable progress in clarifying the issues involved when observational data are used to estimate causal effects. In some cases, this hard-won clarity has permitted the development of new and more powerful methods of analysis. This line of research is distinct from the work of sociologists and others who in the 1970s and 1980s developed path analysis and its generalization, covariance structure analysis. Despite their differences, both areas of research are often labeled causal analysis.

Statisticians and econometricians have adopted a shared conceptual framework that can be used to evaluate the appropriateness of different estimators in specific circumstances. This framework, to be described below, also clarifies the properties of estimators that are needed to obtain consistent estimates of causal effects in particular applications.

Our chapter provides an overview of the work that has been done by statisticians and econometricians on causal analysis. We hope it will provide the reader with a basic appreciation of the conceptual advances that have been made and some of the methods that are now available for estimating causal effects. Because the literature is massive and often technical, we do not attempt to be comprehensive. Rather, we present material that we believe is most accessible and useful to practicing researchers.

As is typical of the literature we are reviewing, we use the language of experiments in describing these methods. This usage is an indication of the advances that have been made; we now have a conceptual framework that allows us to use the traditional experimental language and perspective to discuss and analyze observational data. Throughout this chapter, we write of individuals who are subject to treatment, and we describe individuals as having been assigned to either a treatment or a control group. The reader, however, should not assume that the thinking and methods we review apply only to the limited set of situations in which it is strictly proper to talk about treatment and control groups. In almost any situation where a researcher attempts to estimate a causal effect, the analysis can be described, at least in terms of a thought experiment, as an experiment.

The chapter consists of three major sections. The first presents the conceptual framework and problems associated with using observational data to estimate causal effects. It presents the counterfactual account of causality and its associated definition of a causal effect. We also discuss the basic problems that arise when using observational data to estimate a causal effect, and we show that there are two distinct sources of possible bias: Outcomes for the treatment and control groups may differ even in the absence of treatment; and the potential effect of the treatment may differ for the treatment and control groups. We then present a general framework for analyzing how assignment to the treatment group is related to the estimation of a causal effect.

The second section examines cross-sectional methods for estimating causal effects. It discusses the bounds that data place on the permissible range of a causal effect; it also discusses the use of control variables to eliminate potential differences between the treatment and control groups that are related to the outcome. We review standard regression and matching approaches and discuss methods that condition on the likelihood of being assigned to the treatment. These latter methods include the regression discontinuity design, propensity score techniques, and dummy endogenous variable models. This section also discusses the use of instrumental variables to estimate causal effects, presenting their development as a method to identify parameters in simultaneous equation models and reviewing current research on what instrumental variables identify in the presence of different types of treatment-effect heterogeneity.

The third section discusses methods for estimating causal effects from longitudinal data. We present the interrupted time-series design, then use a relatively general model specification for the structure of unobservables to compare change-score analysis, differential linear growth rate models, and the analysis of covariance. The key lesson here is that no one method is appropriate for all cases. This section also discusses how to use data to help determine which method is appropriate in a particular application.

The paper concludes with a discussion of the general importance of the methods reviewed for improving the quality of quantitative empirical research in sociology. We have more powerful methods available, but more important, we have a framework for examining the plausibility of assumptions behind different methods and thus a way of analyzing the quality and limitations of particular empirical estimates.

BASIC CONCEPTUAL FRAMEWORK

In the past two decades, statisticians and econometricians have adopted a common conceptual framework for thinking about the estimation of causal effects—the counterfactual account of causality. The usefulness of the counterfactual framework is threefold. It provides an explicit framework for understanding

(*a*) the limitations of observational data, (*b*) how the treatment assignment process may be related to the outcome of interest, and (*c*) the type of information that is provided by the data in the absence of any assumptions.

The Counterfactual Account Of Causality

Discussions of causality in the social sciences often degenerate into fruitless philosophical digressions (e.g., see McKim & Turner 1997, Singer & Marini 1987). In contrast, the development of the counterfactual definition of causality has yielded practical value. With its origins in the early work on experimental designs by Fisher (1935), Neyman (1923, 1935), Cochran & Cox (1950), Kempthorne (1952), and Cox (1958a,b), the counterfactual framework has been formalized and extended to nonexperimental designs in a series of papers by Rubin (1974, 1977, 1978, 1980, 1981, 1986, 1990; see also Pratt & Schlaifer 1984). However, it also has roots in the economics literature (Roy 1951, Quandt 1972). The counterfactual account has provided a conceptual and notational framework for analyzing problems of causality that is now dominant in both statistics and econometrics. Holland (1986), Pratt & Schlaifer (1988), and Sobel (1995, 1996) provide detailed exegeses of this work.

Let *Y* be an interval level measure of an outcome of interest, either continuous or discrete or a mixture of the two. Examples are earnings, mathematics aptitude, educational attainment, employment status, and age at death. Assume that individuals can be exposed to only one of two alternative states but that each individual could a priori be exposed to either state. Each state is characterized by a distinct set of conditions, exposure to which potentially affects the outcome of interest *Y*. We refer to the two states as treatment and control.[1]

Assume that one group of individuals is assigned to be observed in the treatment state and that a second group of individuals is assigned to be observed in the control state. The key assumption of the counterfactual framework is that individuals assigned to these treatment and control groups have potential outcomes in both states: the one in which they are observed and the one in which they are not observed. In other words, each individual in the treatment group has an observable outcome in the treatment state and an unobservable counterfactual outcome in the control state. Likewise, each individual in the control group has an observable outcome in the control state and an unobservable counterfactual outcome in the treatment state. Thus, the framework asserts that individuals have potential outcomes in all states, even though they can actually only be observed in one state.

[1] Any two states to which individuals could be assigned or could choose to enter can be considered treatment and control. The potential outcome framework also can be generalized to any number of alternative sets of treatment conditions.

Formalizing this conceptualization, the potential outcomes of each individual unit of analysis are defined as the true values of Y that would result from exposure to the alternative sets of conditions that characterize the two states named treatment and control. More formally, let Y_i^t and Y_i^c equal the potential outcomes for each individual i that would result from exposure to the treatment and control conditions. We assume that both potential outcomes exist in theory for every individual, although at most only one potential outcome can be observed for each individual.

The causal effect of the treatment on the outcome for each individual i is defined as the difference between the two potential outcomes in the treatment and control states:

$$\delta_i = Y_i^t - Y_i^c. \qquad 1.$$

Because both Y_i^t and Y_i^c exist in theory, we can define this individual-level causal effect. However, as detailed below, because we cannot observe both Y_i^t and Y_i^c for any single individual, we cannot observe or thus directly calculate any individual-level causal effects.

First note that this definition of a causal effect, while intuitively appealing, makes several assumptions.[2] The most crucial assumption among these is that a change in treatment status of any individual does not affect the potential outcomes of other individuals. Known as the stable unit treatment value assumption (SUTVA) (see Rubin 1980, 1986, 1990), this assumption is most commonly violated when there is interference across treatments (i.e. when there are interactions between treatments). The classical example is the analysis of treatment effects in agricultural research—rain that surreptitiously carries fertilizer from a treated plot to an adjacent untreated plot. Aside from simple interference, the SUTVA may also be violated in other situations, especially when "macro effects" of the treatment alter potential outcomes (see Garfinkel et al. 1992, Heckman et al. 1998). Consider the case where a large job training program is offered in a metropolitan area with a competitive labor market. As the supply of graduates from the program increases, the wage that employers will be willing to pay graduates of the program will decrease. When such complex effects are present, the powerful simplicity of the counterfactual framework vanishes.

Why can we not observe and calculate individual-level causal effects? In order to observe values of Y, we must assign individuals to be observed in one of the two states. To formalize this observation rule, define T_i as a dummy variable equal to 1 if an individual is assigned to the treatment group and equal

[2]One important assumption that we do not discuss is that the treatment must be manipulable. For example, as Holland (1986) argued, it makes no sense to talk about the causal effect of gender or any other nonmanipulable individual trait alone. One must explicitly model the manipulable mechanism that generates an apparent causal effect of a nonmanipulable attribute.

to 0 if an individual is assigned to the control group. The observed Y_i are equal to $Y_i = Y_i^t$ when $T_i = 1$ and $Y_i = Y_i^c$ when $T_i = 0$. As these definitions reveal, causal inference can be seen as a problem of missing data. The observed Y_i do not contain enough information to identify individual-level causal effects because individuals cannot be observed under both the treatment and the control conditions simultaneously.[3]

The main value of this counterfactual framework is that causal inference can be summarized by a single question: Given that the δ_i cannot be calculated for any individual and therefore that Y_i^t and Y_i^c can be observed only on mutually exclusive subsets of the population, what can be inferred about the distribution of the δ_i from an analysis of Y_i and T_i?

Average Effects And The Standard Estimator

Most of the literature has focused on the estimation of the average causal effect for a population. Let $\overline{Y}^t$ be the average value of Y_i^t for all individuals if they are exposed to the treatment, and let $\overline{Y}^c$ be the average value of Y_i^c for all individuals if they are exposed to the control. More formally, $\overline{Y}^t$ is the expected value of Y_i^t in the population, and $\overline{Y}^c$ is the expected value of Y_i^c in the population. The average treatment effect in the population is

$$\overline{\delta} = \overline{Y}^t - \overline{Y}^c \qquad 2.$$

or, again more formally, the expected value of the difference between $\overline{Y}^t$ and $\overline{Y}^c$ in the population.[4]

Because Y_i^t and Y_i^c are unobservable (or missing) on mutually exclusive subsets of the population, $\overline{Y}^t$ and $\overline{Y}^c$ cannot both be calculated. However, $\overline{Y}^t$ and $\overline{Y}^c$ can potentially be estimated, although not very well or without considerable difficulty except in special circumstances. Most methods discussed in this paper attempt to construct from observational data consistent estimates of $\overline{Y}^t$ and $\overline{Y}^c$ in order to obtain a consistent estimate of $\overline{\delta}$.

For example, consider the most common estimator, which we call the standard estimator for the average treatment effect. Let $\overline{Y}^t_{i\in T}$ be the expected value of Y_i^t for all individuals in the population who would be assigned to the treatment group for observation, and let $\overline{Y}^c_{i\in C}$ be the expected value of Y_i^c for all

[3]When one has longitudinal data, an effective strategy may be to use a person as his own control. This strategy only works if age does not otherwise affect the outcome and there are no exogenous period-specific effects. If change with age or period effects is possible, some type of adjustment is needed. We discuss methods that do this in the section on longitudinal analysis.

[4]In many presentations of the counterfactual framework, formal $E[.]$ notation is used. The average treatment effect of Equation 2 is written as $E[\delta] = E[Y^t - Y^c]$. The standard estimator in Equation 3 is considered an attempt to estimate $E[Y^t \mid T = 1] - E[Y^c \mid T = 0]$.

individuals in the population who would be assigned to the control group for observation. Both of these quantities can be calculated and thus effectively estimated by their sample analogs, the mean of Y_i for those actually assigned to the treatment group and the mean of Y_i for those actually assigned to the control group. The standard estimator for the average treatment effect is the difference between these two estimated means:

$$\hat{\overline{\delta}} = \hat{\overline{Y}}^{\mathrm{t}}_{i\in\mathrm{T}} - \hat{\overline{Y}}^{\mathrm{c}}_{i\in\mathrm{C}}, \qquad 3.$$

where the hats on all three terms signify that they are the sample analog estimators (sample means) of the expectations defined above.

Note the two differences between Equations 2 and 3. Equation 2 is defined for the population, whereas Equation 3 represents an estimator that can be applied to a sample drawn from the population. All individuals in the population contribute to the three terms in Equation 2. However, each sampled individual can be used only once to estimate either $\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}}$ or $\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}$. As a result, the way in which individuals are assigned (or assign themselves) to the treatment and control groups determines how effectively the standard estimator $\hat{\overline{\delta}}$ estimates the true average treatment effect $\overline{\delta}$. As we demonstrate, many estimators are extensions of this standard estimator that seek to eliminate the bias resulting from inherent differences between the treatment and control groups.

To understand when the standard estimator consistently estimates the true average treatment effect for the population, let $\overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}}$ and $\overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}}$ be defined analogously to $\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}}$ and $\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}$ above, and let π equal the proportion of the population that would be assigned to the treatment group. Decompose the average treatment effect in the population into a weighted average of the average treatment effect for those in the treatment group and the average treatment effect for those in the control group and then decompose the resulting terms into differences in average potential outcomes:

$$\begin{aligned}
\overline{\delta} &= \pi\overline{\delta}_{i\in\mathrm{T}} + (1-\pi)\overline{\delta}_{i\in\mathrm{C}} \\
&= \pi\left(\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} - \overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}}\right) + (1-\pi)\left(\overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}} - \overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}\right) \\
&= \left[\pi\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} + (1-\pi)\overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}}\right] - \left[\pi\overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}} + (1-\pi)\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}\right] \\
&= \overline{Y}^{\mathrm{t}} - \overline{Y}^{\mathrm{c}}.
\end{aligned} \qquad 4.$$

The quantities $\overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}}$ and $\overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}}$ that appear explicitly in the second and third lines of Equation 4 cannot be directly calculated because they are based on unobservable values of Y. If we assume that $\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} = \overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}}$ and $\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}} = \overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}}$,

then through substitution starting in the third line of (4):

$$\begin{aligned}\overline{\delta} &= \left[\pi\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} + (1-\pi)\overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}}\right] - \left[\pi\overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}} + (1-\pi)\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}\right] \\ &= \left[\pi\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} + (1-\pi)\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}}\right] - \left[\pi\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}} + (1-\pi)\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}\right] \qquad 5. \\ &= \overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} - \overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}.\end{aligned}$$

Thus, a sufficient condition for the standard estimator to consistently estimate the true average treatment effect in the population is that $\overline{Y}^{\mathrm{t}}_{i\in\mathrm{T}} = \overline{Y}^{\mathrm{t}}_{i\in\mathrm{C}}$ and $\overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}} = \overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}}$. In this situation, the average outcome under the treatment and the average outcome under the control do not differ between the treatment and control groups. In order to satisfy these equality conditions, a sufficient condition is that treatment assignment T_i be uncorrelated with the potential outcome distributions of Y_i^{t} and Y_i^{c}. The principal way to achieve this uncorrelatedness is through random assignment to the treatment.

By definition, observational data are data that have not been generated by an explicit randomization scheme. In most cases, treatment assignment will be correlated with the potential outcome variables. As a result, the standard estimator will usually yield inconsistent estimates of the true average treatment effect in the population when applied to observational data.

An important caveat is that the average treatment effect $\overline{\delta}$ is not always the quantity of theoretical interest. Heckman (1992, 1996, 1997) and Heckman et al. (1997b) have argued that in a variety of policy contexts, it is the average treatment effect for the treated that is of substantive interest. The essence of their argument is that in deciding whether a policy is beneficial, our interest is not whether on average the program is beneficial for all individuals but whether it is beneficial for those individuals who are either assigned or who would assign themselves to the treatment.

For example, if we are interested in determining whether a particular vocational education program in a high school is beneficial, it makes little sense to ask whether its effect is positive for all high school students. For college-bound students, the effects of the program may be negative. Even for non–college-bound students, the program may have positive effects only for some students. To the degree that students can estimate their likely benefit of enrolling in the program before actually doing so, we would expect that those students for whom the expected benefits are positive will be more likely to enroll in the program. The appropriate policy question is whether the program effects for this group of "self-selecting" students are positive and sufficiently large to justify the program costs. The policy–relevant piece of information in need of estimation is the size of the treatment effect for the treated. The average treatment effect for all students in the school is of little or no policy relevance.

As discussed below, it is also the case that in many contexts the average treatment effect is not identified separately from the average treatment effect for the treated. In most circumstances, there is simply no information available on how those in the control group would have reacted if they had instead received the treatment. This is the basis for an important insight into the potential biases of the standard estimator.

Define the baseline difference between the treatment and control groups as $(\overline{Y}^{c}_{i\in T} - \overline{Y}^{c}_{i\in C})$. This quantity can be thought of as the difference in outcomes between the treatment and control groups in the absence of treatment. With a little algebra, it can be shown that Standard estimator $=$ True average treatment effect $+$ (Difference in baseline Y) $+ (1-\pi)$ (Difference in the average treatment effect for the treatment and control groups), or in mathematical notation:

$$\overline{Y}^{t}_{i\in T} - \overline{Y}^{c}_{i\in C} = \overline{\delta} + \left(\overline{Y}^{c}_{i\in T} - \overline{Y}^{c}_{i\in C}\right) + (1-\pi)(\overline{\delta}_{i\in T} - \overline{\delta}_{i\in C}). \quad 6.$$

Equation 6 shows the two possible sources of bias in the standard estimator. The baseline difference, $(\overline{Y}^{c}_{i\in T} - \overline{Y}^{c}_{i\in C})$, is equal to the difference between the treatment and control groups in the absence of treatment. The second source of bias $(\overline{\delta}_{i\in T} - \overline{\delta}_{i\in C})$, the difference in the treatment effect for those in the treatment and control groups, is often not considered, even though it is likely to be present when there are recognized incentives for individuals (or their agents) to select into the treatment group. Instead, many researchers (or, more accurately, the methods that they use) simply assume that the treatment effect is constant in the population, even when common sense dictates that the assumption is clearly implausible (Heckman 1997, Heckman et al. 1997b, Heckman & Robb 1985, 1986, 1988; JJ Heckman, unpublished paper).

To clarify this decomposition, 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 $\overline{\delta}$ in Equation 6. Second, individuals who attend college might have been smarter in the first place. This source of bias is the baseline difference represented by $(\overline{Y}^{c}_{i\in T} - \overline{Y}^{c}_{i\in C})$ in Equation 6. Third, the mental ability of those who attend college may increase more than would the mental ability of those who did not attend college had they in fact attended college. This source of bias is the differential effect of treatment, represented by $(\overline{\delta}_{i\in T} - \overline{\delta}_{i\in C})$ in Equation 6.

To further clarify this last term in the decomposition, assume that those who have attended college and those who have not attended college had the same (average) initial mental ability. Assume further that only those who then

attended college would have benefitted from doing so. If the treatment and control groups are of equal size, the standard estimator would overestimate the true average treatment effect by a factor of two. In this example, and in many other situations, the standard estimator yields a consistent estimate of the average treatment effect for the treated, not the average treatment effect for the entire population.

Equation 6 specifies the two sources of bias that need to be eliminated from estimates of causal effects from observational data. The remainder of the paper examines how this goal can be accomplished. Most of the discussion focuses on the elimination of the baseline difference ($\overline{Y}^{\mathrm{c}}_{i\in\mathrm{T}} - \overline{Y}^{\mathrm{c}}_{i\in\mathrm{C}}$). Fewer techniques are available to adjust for the differential treatment effects component of the bias ($\overline{\delta}_{i\in\mathrm{T}} - \overline{\delta}_{i\in\mathrm{C}}$).

Treatment Assignment Model

To proceed further, we need to develop a basic model for the assignment mechanism that generates the treatment and control groups. Our presentation of the assignment model follows Heckman & Robb (1985, 1986, 1988). Above, we specified that each individual has two potential outcomes, Y_i^{t} and Y_i^{c}, corresponding to potential exposure to the treatment and control. We noted that, in general, for any one individual only one of these two potential outcomes can be observed.

To develop an assignment model, we first write the potential outcomes Y_i^{t} and Y_i^{c} as deviations from their means:

$$Y_i^{\mathrm{c}} = \overline{Y}^{\mathrm{c}} + u_i^{\mathrm{c}},$$

$$Y_i^{\mathrm{t}} = \overline{Y}^{\mathrm{t}} + u_i^{\mathrm{t}}.$$

Combining these two expressions with the observation rule given by the definition of the treatment assignment dummy variable T_i, the equation for any Y_i is

$$\begin{aligned} Y_i &= \overline{Y}^{\mathrm{c}} + T_i(\overline{Y}^{\mathrm{t}} - \overline{Y}^{\mathrm{c}}) + u_i^{\mathrm{c}} + T_i\left(u_i^{\mathrm{t}} - u_i^{\mathrm{c}}\right) \\ &= \overline{Y}^{\mathrm{c}} + T_i\overline{\delta} + u_i, \end{aligned} \qquad 7.$$

where $u_i = u_i^{\mathrm{c}} + T_i(u_i^{\mathrm{t}} - u_i^{\mathrm{c}})$. Equation 7 is known as the structural equation. This equation provides another way of thinking about the problem of consistently estimating the treatment effect. For the standard estimator—which is equivalent to the coefficient on T_i when Equation 7 is estimated by ordinary least squares (OLS)—to be a consistent estimate of the true average treatment effect, T_i and u_i must be uncorrelated.

Consider a supplemental equation, known as the assignment or selection equation, that determines T_i and is written in what is known as an index structure.

Let T_i^* be a latent continuous variable:

$$T_i^* = Z_i a + v_i, \qquad 8.$$

where $T_i = 1$ if $T_i^* \geq 0$ and $T_i = 0$ if $T_i^* < 0$, and where Z_i is a row vector of values on various exogenous observed variables that affect the assignment process, a is a vector of parameters that typically needs to be estimated, and v_i is an error term that captures unobserved factors that affect assignment.

Equations 7 and 8 are general. Additional covariates X_i can be included in Equation 7, as shown below in Equation 10, and X_i and Z_i may have variables in common. Both Z_i and v_i may be functions of an individual's potential outcome after exposure to the treatment (Y_i^{t}), an individual's potential outcome after exposure to the control (Y_i^{c}), or any function of the two potential outcomes, such as their difference ($Y_i^{\mathrm{t}} - Y_i^{\mathrm{c}}$).

We can distinguish between two different ways that T_i and the error term in Equation 7, u_i, can be correlated (Heckman & Robb 1986, 1988; Heckman & Hotz 1989). When Z_i and u_i are correlated, but u_i and v_i are uncorrelated, we have "selection on the observables." In this case, some observed set of factors in Z_i is related to Y_i^{c} and/or Y_i^{t}. This form of selection results in data that are sometimes characterized as having ignorable treatment assignment—the probability of being assigned to the treatment condition is only a function of the observed variables (Rosenbaum & Rubin 1983, Rosenbaum 1984a,b). The second case is where u_i is correlated with v_i, resulting in "selection on the unobservables." Known as nonignorable treatment assignment, in this case the probability of assignment is a function of unobserved variables (and possibly observed variables as well). In the following sections, we examine methods that attempt to deal with both types of selection bias. Not surprisingly, remedies for bias from selection on the observables are easier to implement than are remedies for selection on the unobservables.

CROSS-SECTIONAL METHODS

Bounds For Treatment Effects

In a series of articles that have culminated in a book, Manski has investigated the bounds that are consistent with the data when weak assumptions alone are maintained (Manski 1995; see also Robins 1989). In this section, we point to the fact that in some circumstances the data, without any auxiliary assumptions, provide some information on the size of the treatment effect. Our discussion follows Manski (1994, 1995).

To see that the data can potentially bound a treatment effect, consider a case with a dichotomous zero-one outcome. The average treatment effect, $\overline{\delta}$, cannot exceed 1. The maximum treatment effect occurs when $\overline{Y}^{\mathrm{t}}_{i\in \mathrm{T}} = \overline{Y}^{\mathrm{t}}_{i\in \mathrm{C}} = 1$

Table 1 Hypothetical example illustrating the calculation of bounds on treatment effects

Groups	Mean Outcome Y_i^c	Mean Outcome Y_i^t
Observed mean outcomes[a]		
Control	$\overline{Y}^c_{i\in C} = 0.3$	$\overline{Y}^t_{i\in C} =?$
Treatment	$\overline{Y}^c_{i\in T} =?$	$\overline{Y}^t_{i\in T} = 0.7$
Largest possible treatment effect[b]		
Control	$\overline{Y}^c_{i\in C} = 0.3$	$\overline{Y}^t_{i\in C} = 1$
Treatment	$\overline{Y}^c_{i\in T} = 0$	$\overline{Y}^t_{i\in T} = 0.7$
Small possible treatment effect[c]		
Control	$\overline{Y}^c_{i\in C} = 0.3$	$\overline{Y}^t_{i\in C} = 0$
Treatment	$\overline{Y}^c_{i\in T} = 1$	$\overline{Y}^t_{i\in T} = 0.7$

[a]Standard estimator of treatment effect is 0.4.
[b]Implied upper bound of average treatment effect is 0.7.
[c]Implied lower bound of average treatment effect is −0.3.

and $\overline{Y}^c_{i\in T} = \overline{Y}^c_{i\in C} = 0$. Similarly, the average treatment effect cannot be less than −1. The minimum treatment effect occurs when $\overline{Y}^t_{i\in T} = \overline{Y}^t_{i\in C} = 0$ and $\overline{Y}^c_{i\in T} = \overline{Y}^c_{i\in C} = 1$. Thus, $\overline{\delta}$ is contained in an interval of length 2; more specifically, $\overline{\delta} \in [-1, 1]$.

Now assume that $\overline{Y}^t_{i\in T} = 0.7$ and $\overline{Y}^c_{i\in C} = 0.3$, as is shown in the hypothetical example in Table 1. Both quantities could be estimated from the data, and we do not consider the problem of sampling error. The standard estimator for the treatment effect in this case is $\overline{Y}^t_{i\in T} - \overline{Y}^c_{i\in C} = 0.4$. The largest possible treatment effect (Table 1) indicates the values of $\overline{Y}^t_{i\in C}$ and $\overline{Y}^c_{i\in T}$ that would produce the largest estimate of $\overline{\delta}$, 0.7. The smallest possible treatment effect (Table 1) indicates the values that would produce the smallest estimate of $\overline{\delta}$, −0.3. Thus, the constraints implied by the data guarantee that $\overline{\delta} \in [-0.3, 0.7]$, an interval of length 1, which is half the length of the maximum interval calculated before values for $\overline{Y}^t_{i\in T}$ and $\overline{Y}^c_{i\in C}$ were obtained from the data. Manski calls this interval the no-assumptions bound. Although this bound is still wide, it has substantially reduced our uncertainly about the range of $\overline{\delta}$. Manski (1995) shows that with a zero-one outcome variable, the no-assumptions bound will always be of length 1.

In general (see Manski 1994), the treatment effect will only be bounded when the outcome variable itself is bounded or when one is analyzing a function of the distribution of the dependent variable that is bounded. Because $\overline{Y}^t_{i\in C}$ and $\overline{Y}^c_{i\in T}$ are both unobserved, in the absence of any restriction they can take on any

value from minus infinity to plus infinity. Thus, in the absence of any known restriction on $\overline{Y}^{t}_{i\in C}$ and $\overline{Y}^{c}_{i\in T}$, $\overline{\delta}$ can take on any value from minus infinity to plus infinity.

The goal of Manski's research is to analyze how additional assumptions narrow the bound for the estimated treatment effect while recognizing that the more assumptions an analysis entails, the less credible it is. He argues that researchers should first attempt to learn as much as possible about a treatment effect maintaining the weakest possible assumptions. Manski shows that weak and often plausible assumptions can substantially narrow the no-assumptions bound. For example, in many situations it may be reasonable to assume that the treatment effect cannot be negative (or alternatively positive) for any individual. Manski (1997) labels this assumption the monotone treatment response assumption. Under this assumption, the lower bound for the treatment effect is 0. Thus, for the example presented in Table 1, the bound for the treatment effect would be [0, 0.7].

Another possible assumption is that those who actually receive the treatment have higher average outcomes under potential exposure to both the treatment and control (i.e. $\overline{Y}^{t}_{i\in T} \geq \overline{Y}^{t}_{i\in C}$ and $\overline{Y}^{c}_{i\in T} \geq \overline{Y}^{c}_{i\in C}$). Manski & Pepper (1998) present this monotone treatment selection assumption with the example of the effect of education on wages. This case is equivalent to assuming that individuals with higher educational attainments would on average receive higher wages than would individuals with lower educational attainments, even if counterfactually the two groups had the same levels of educational attainment. For the example presented in Table 1, the monotone treatment selection assumption implies that the standard estimator would be an upper bound for the average treatment effect. Therefore, if we invoke the monotone treatment response and selection assumptions together, the bound on the treatment effect is [0, 0.4], which is considerably more narrow than the no-assumptions bound. Applications of Manski's approach can be found in Manski & Nagin (1998) and in Manski et al. (1992). We discuss Manski's work further below.

Regression Methods

The basic strategy behind regression analysis and related methods is to find a set of control variables that can be included in the regression equation in order to remove the correlation between the treatment variable and the error term. In order to understand the relationship between regression and other cross-sectional methods, it is worth formalizing this idea. Assume that we are interested in estimating Equation 8 above and that we believe the treatment indicator, T_i, is correlated with the error term, u_i, because treatment assignment is not random. We could attempt to deal with this problem by controlling for

various observed Xs, estimating a regression equation of the form

$$Y_i = b_0 + T_i\overline{\delta} + X_i b + w_i. \quad 9.$$

Estimating Equation 9 by OLS is equivalent to following the double residual regression procedure (Malinvaud 1970, Goldberger 1991): (*a*) Regress Y_i on X_i and calculate $Y_i^* = Y_i - \hat{Y}_i$; (*b*) regress T_i on X_i and calculate $T_i^* = T_i - \hat{T}_i$; and (*c*) estimate $Y_i^* = T_i^*\overline{\delta} + w_i^*$, where $w_i^* = w_i - X_i b$. This three step procedure will yield the same estimate of $\overline{\delta}$ as OLS on Equation 9. Thus, OLS regression is equivalent to estimating the relationship between residualized versions of Y_i and T_i from which their common dependence on other variables has been subtracted out.

A number of techniques, all falling under what Heckman & Robb (1985) label control function estimators, can be understood as variants of this strategy. We discuss only a few such methods where a control function (i.e. some function of one or more variables) is entered into a regression equation in an attempt to eliminate the correlation between the treatment indicator variable and the error term. As is discussed below, instrumental variable techniques are based on a strategy that is the mirror image of the control function approach.

ANALYSIS OF COVARIANCE AND MATCHING The analysis of covariance is probably the most common technique used to adjust for possible differences between treatment and control groups. Although it was originally developed to adjust for chance differences in observed Xs across treatment and control groups in randomized designs, it is now routinely used to attempt to control for differences between treatment and control groups in observational studies. Technically, the analysis of covariance is just a specific application of regression analysis. We consider a model somewhat more general than the standard model.

If we had a large data set and believed that either Y_i^c or δ_i varied as a function of the Xs, then one approach would be to stratify the sample on the Xs and carry out the analysis separately within each stratum. We could then estimate separate average treatment effects, $\overline{\delta}_x$, for each stratum. If a single treatment effect estimate was desired, we could then average these estimated effects across the strata, weighting each estimated treatment effect by the relative size of its stratum.

An analogous set of analyses could be mounted in a regression framework. Let the potential outcomes Y_i^t and Y_i^c depend on some set of variables X_i:

$$Y_i^c = b_0^c + X_i b + e_i^c \quad 10a.$$

and

$$Y_i^t = b_0^t + X_i(b + c) + e_i^t. \quad 10b.$$

The observed data can be written as a combination of these two equations:

$$Y_i = b_0^c + T_i\left(b_0^t - b_0^c\right) + X_i b + T_i(X_i c) + e_i. \qquad 11.$$

For individuals for whom $X_i = 0$, the treatment effect in Equation 11 is equal to $(b_0^t - b_0^c)$. The $X_i b$ term represents how the baseline level of Y_i, the Y_i^c, varies with the observed X_i. The hope is that by including the $X_i b$ term, we eliminate the baseline difference between the treatment and control groups, $(\overline{Y}^c_{i\in T} - \overline{Y}^c_{i\in C})$.

The $X_i c$ term represents how the treatment effect, δ_i, varies with X_i. This term is not typically included in a standard analysis of covariance model. The hope is that by including the $X_i c$ term, we eliminate the difference in the treatment effects between the treatment and control groups, $(\overline{\delta}_{i\in T} - \overline{\delta}_{i\in C})$. This may often be an unrealistic assumption, because it implies that the researcher can forecast an individual's treatment effect just as accurately as the individual himself can. If individuals have pertinent information that is unavailable to the researcher (i.e. information that is not contained in the *X*s), then it is likely that there will be differences in the treatment effects between the treatment and control groups that are not captured by observed *X*s (Heckman 1989, 1992, 1996, 1997). Note that the treatment effect in Equation 11 is equal to $(b_0^t - b_0^c) + X_i c$. Obviously, this is not the treatment effect for the entire population but rather for individuals with characteristics X_i.

One problem with the regression approach is that it imposes a linearity constraint. Nonlinear terms can be added, but it is often difficult to know how the nonlinearity should be approximated. As White (1981) has shown, polynomial and related expansions may inadequately model nonlinearity and lead to biased estimates.

An alternative technique that avoids this problem is matching. Common in biomedical research but not in social scientific research, matching is closely related to the stratification procedure described above. Smith (1997) provides an excellent introduction for social scientists. Matching has several advantages. First, it makes no assumption about the functional form of the dependence between the outcome of interest and the other *X*s. Second, matching ensures that only those portions of the distribution of the *X*s in the observed data that contain individuals in both the treatment and control groups enter the estimation of the treatment effect.[5] Third, because fewer parameters are estimated than

[5]In two important empirical papers, Heckman et al (1997, 1998a) show that the bias due to selection on the unobservables, although significant and large relative to the size of the treatment effect, is small relative to the bias that results from having different ranges of *X*s for the treatment and control groups and different distributions of the *X*s across their common range. Matching solves both of the latter problems, although the average effect is not for the total population but only for that portion of the population where the treatment and control groups have common *X* values.

in a regression model, matching is more efficient. Efficiency can be important with small samples. A major problem with the traditional matching approach is that unless an enormous sample of data is available and there are more than a few *X*s, it may be difficult to find both treatment and control cases that match. [See below for the ingenious solution to this problem developed by Rosenbaum & Rubin (1983)].

REGRESSION DISCONTINUITY DESIGN A key limitation of the analysis of covariance and related designs is that they do not directly conceptualize how the *X*s are related to the likelihood of being assigned to the treatment group. Rather, the approach is to model the determinants of Y_i, thereby including *X*s that are believed to affect the outcome and that may also be associated with assignment to the treatment group. By including many determinants of Y_i, one hopes to eliminate all differences between the treatment and control groups that are related to the outcome but that are not due to the treatment itself.

The philosophy behind regression discontinuity designs and propensity score methods is quite different from the strategy behind analysis of covariance. The strategy is to attempt to control for observed variables, Z_i, that affect whether an individual is assigned to the treatment group or the control group. By controlling for *Z*s that affect the treatment assignment, one hopes to eliminate any correlation between T_i and u_i in Equation 7.

The regression discontinuity design (Cook & Campbell 1979, Judd & Kenny 1981, Marcantonio & Cook 1994) is the simplest way of relating an observed variable, Z_i, to the assignment to a treatment group. The basic strategy is to find a Z_i that is related to the assignment of treatment in a sharply discontinuous way, as in Figure 1. The jump on the vertical axis at the point of treatment on the horizontal axis is the estimate of the main treatment effect. In Figure 1, the treatment effect is even more complex. The treatment also affects the slope of the relationship between Z and Y. Thus, the size of the treatment effect varies with Z.

The strength of the regression discontinuity design is determined by the accuracy of the estimate of the conditional relationship between Y and Z in the absence of treatment over the range of Z that receives the treatment. If the relationship between Z and Y is nonlinear, this can be highly problematic. Figure 2 provides an example. As can be seen from Figure 2, if we poorly estimate the values of Y that would be observed in the absence of treatment, we poorly estimate the effect of the treatment. The problem here is directly related to matching. One of the strengths of matching is that it ensures that we have both control and treatment cases over the range of Z that is relevant to the analysis. In the regression discontinuity design, the opposite is the case. There are no values of Z that contain both treatment and control cases. The power of the design hinges solely on the ability to extrapolate accurately.

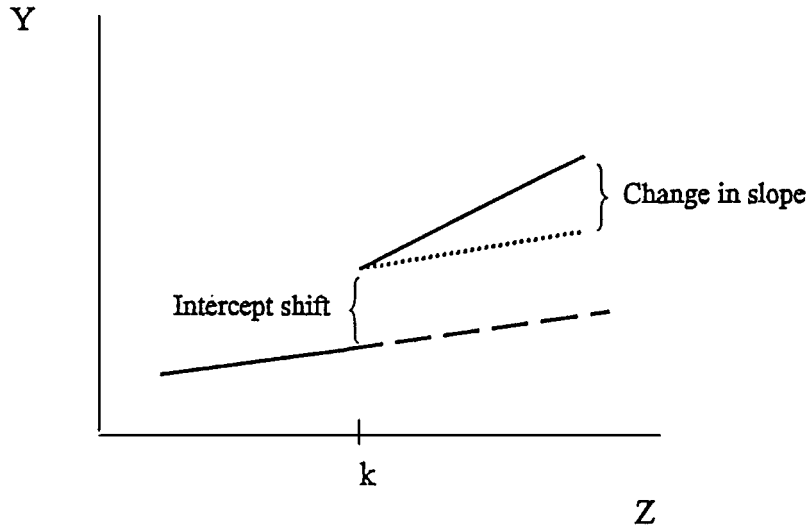

Figure 1 The regression discontinuity design. Note: If $Z \geq k$, the individual receives the treatment. If $Z < k$, the individual does not receive the treatment. (Solid line) Observed outcome; (dashed line) the assumed outcome in the absence of treatment.

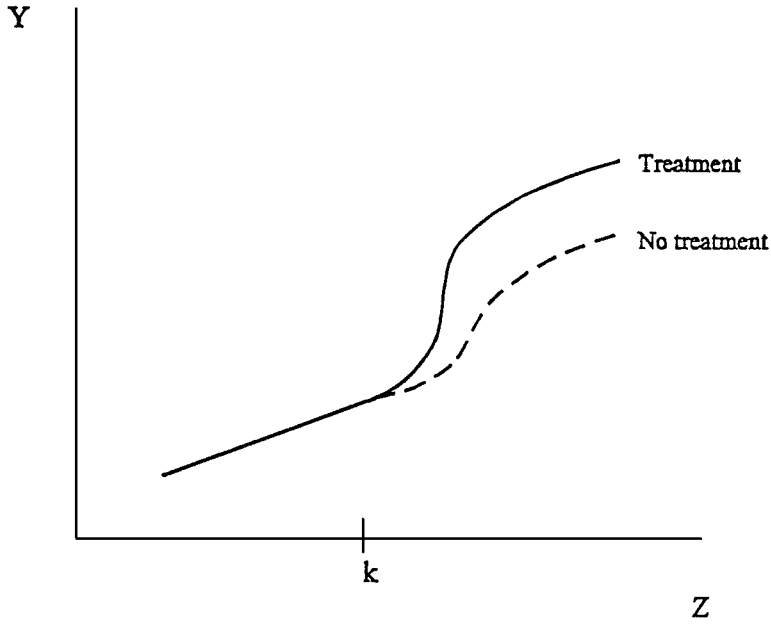

Figure 2 The regression discontinuity design with unrecognized nonlinearity.

PROPENSITY SCORES The essence of the regression discontinuity design is the direct tie between the treatment assignment and an observed variable Z. The propensity score method (Rosenbaum & Rubin 1983, 1984, 1985; Rosenbaum 1984a,b, 1995; Rubin 1991; Rubin & Thomas 1996) provides a much more general approach that is nonetheless based on the same strategy as the regression discontinuity design. The propensity score for an individual is simply the probability that an individual, with a set of observed characteristics Z_i, is assigned to the treatment group instead of the control group, or

$$P(Z_i) = \text{Prob}(T = 1 \mid Z_i). \qquad 12.$$

If treatment assignment is purely a function of the observed Zs (or in the language used above, selection is only on the observables), then conditional on the Zs, assignment is random with respect to the outcomes.[6] The importance of this result is that the analysis can then safely proceed after either matching or stratifying on the propensity score, $P(Z_i)$. In general, the propensity score will not be known but can be estimated using standard methods such as a logit or probit model.

Rosenbaum & Rubin (1983) show that there is nothing to be gained by matching (or stratifying) in a more refined way on the variables in Z than on just the propensity scores alone that are a function of the variables in Z. The propensity score contains all the information that is needed to create a balanced design—a design where the treatment and control groups do not differ with respect to Z in any way that is also related to treatment assignment T_i. This fact is of enormous importance because it means that matching can be done on a single dimension. As a result, even when there are many variables in Z that determine treatment assignment, matching is still feasible. Stratification on the propensity score is typically feasible only with large data sets.

A variety of matching schemes are possible. Nearest available matching on the estimated propensity score is the most common and one of the simplest (see Rosenbaum & Rubin 1985). First, the propensity scores for all individuals are estimated with a standard logit or probit model. Individuals in the treatment are then listed in random order.[7] The first treatment case is selected, and its propensity score is noted, and then matched to the control case with the closest propensity score. Both cases are then removed from their respective lists,

[6]See Rosenbaum & Rubin (1983) for a proof. Heckman et al (1997, 1998b) point out that this proof involves the true propensity score and that in most applications the propensity score needs to be estimated. It is unclear whether this is consequential.

[7]In most empirical applications of matching techniques, the treatment group is considerably smaller than the control group. This need not be the case in all applications, and if the reverse is true, the nearest available matching scheme described here runs in the opposite direction. Treatment cases would be matched to the smaller subset of control cases.

the second treatment case is matched to the remaining control case with the closest propensity score, and so on, until all treatment cases have received a matched control case. Other matching techniques that use propensity scores are implemented by (*a*) using different methods and different sets of *Z*s to estimate propensity scores, (*b*) matching on some important *Z*s first and then on propensity scores second, (*c*) defining the closeness of propensity scores and *Z*s in different ways, and/or (*d*) matching multiple control cases to each treatment case (see Rosenbaum 1995, Rubin & Thomas 1996, Smith 1997).

In principle, the propensity score can also be entered as a control variable in a regression model in a fashion similar to the inclusion of X_i in Equation 9 or 11. Rubin & Rosenbaum have advocated matching because it implicitly deals with the problem of nonlinearity and uses fewer degrees of freedom, making it more efficient. To better understand the propensity-score method, it is useful, however, to consider the approach within a regression framework.

Consider Equations 7 and 8 again. The assumption behind these two equations is that Z_i directly affects treatment assignment but does not directly affect either Y_i^{t} or Y_i^{c}. Z_i, however, is potentially correlated with u_i, which may include both observed and unobserved components. In some cases, the Z_i may overlap with observed components of u_i. However, we do not think of either the Z_i or the propensity score $P(Z_i)$ as being determinants of the outcome. Thus, Z_i does not belong in the structural Equation 7. Z_i determines assignment, not the outcome.

What are we doing if we enter the propensity score, or some nonlinear transformation of it, into Equation like 9 or 10, as if it were an X? Heckman & Robb (1986, 1988) have pointed out that Rosenbaum and Rubin's propensity-score method is one example of a control function. As discussed above, the goal when a control variable, in this case the propensity score, is entered into Equation 7 as a regressor is to make the treatment assignment variable uncorrelated with the new error term. Above, we noted that conditional on the propensity score, assignment to the treatment group is random by construction. This means that by entering the propensity score, or some nonlinear transformation of it, into regression Equation 9, for example, we are "subtracting out" of Y_i and T_i that component of their correlation that is due to the assignment process.

To understand what we are doing further, consider Figure 3 where we are interested in estimating the effect of T_i on Y_i, but we are concerned that T_i and u_i might be correlated. There are two reasons they might be correlated. First, u_i and T_i might be correlated because the Z_i or equivalently the propensity score, $P(Z_i)$, and T_i are correlated. This is selection on the observables. Second, there is a possibility that T_i and u_i are correlated because u_i and v_i are correlated. This is selection on the unobservables. The propensity-score method, however, assumes that all the selection is on the observables. Thus there is no

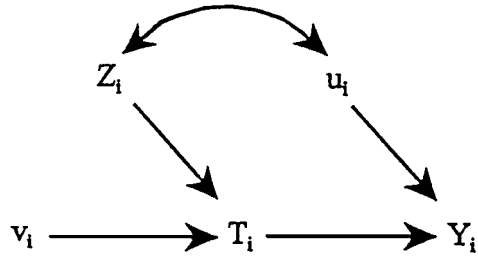

Figure 3 The propensity score strategy for the estimation of a treatment effect when selection is nonrandom.

arrow connecting u_i and v_i in Figure 3. This is a very strong assumption. It implies that there are no common omitted variables that determine both treatment assignment and the outcome. Estimation of the propensity-score model amounts to estimating the effect of T_i on Y_i where both variables have been residualized with respect to $P(Z_i)$. As Figure 3 indicates, conditional on Z_i, or equivalently the propensity score $P(Z_i)$, T_i and u_i are assumed to be uncorrelated. As a result, estimation by OLS using residualized Y_i and T_i consistently estimates the treatment effect.

SELECTION MODELS Heckman's early work in the late 1970s on selection bias, particularly his lambda method, has received some attention in sociology. Since that time, considerable new research has appeared, primarily in the econometrics literature. Winship & Mare (1992) provide a review of much of this literature. Heckman's closely related work on dummy endogenous variables, pursued at the same time as his well-known selection-bias research, has received less attention (Heckman 1978). Although his terminology is different, his work also addresses the estimation of treatment effects when assignment is nonrandom.

The selection and nonrandom assignment problems are intimately connected. In essence, the nonrandom assignment problem is two selection problems in one. If the focus is only on Y_i^c, we have a selection problem because Y_i^c is only observed for individuals who are exposed to the control. Similarly, we have a selection problem if we focus solely on Y_i^t because it is only observed for individuals who are exposed to the treatment. In both cases, we are concerned that individuals have selected (or been selected) on the dependent variable and thus that treatment exposure is a function of Y_i^t, Y_i^c, or some function of the two. When this occurs, standard regression techniques yield an inconsistent estimate of the treatment effect.

Although completed prior to most of Rubin's and Rosenbaum's work on propensity scores, Heckman's work on the dummy endogenous variable

problem can be understood as a generalization of the propensity-score approach. It is also another example of a control function estimator.[8] As with Rosenbaum's and Rubin's propensity score method, Heckman focuses on the selection Equation 9. Heckman, however, is interested in the conditional mean of T_i^*, the latent continuous variable, rather than the probability that $T_i = 1$. Specifically, using the linearity of Equation 8, he is interested in

$$E[T_i^* \mid Z_i a, T_i] = Z_i a + E[v_i \mid Z_i a, T_i]. \qquad 13.$$

Note that the expected value here of T_i^* is a function of both $Z_i a$ and T_i. This allows Heckman to take account of selection that may be a function of both the observables Z_i and the unobservables v_i. As shown in Figure 3, we now assume that u_i and v_i may be correlated. This correlation would occur if respondents know more about their potential outcomes under the treatment and control than the researcher and use their private information when "selecting" themselves into the treatment or control group.

If v_i is only correlated with observed components of u_i (i.e. the Xs in our notation), then the selection problem is easily solved. We can adjust for nonrandom assignment by simply controlling for these Xs when estimating Equation 8, as in the analysis of covariance and its extensions that are discussed above. However, if v_i is correlated with unobserved components of u_i, a more complicated solution is required.

If we could observe v_i, we could enter it into Equation 7 or 11 as a control variable, adopting a strategy similar in spirit to Rosenbaum's and Rubin's propensity score method. In so doing, we would control for a function of the assignment process in order to create residualized Y_i and T_i so that the residualized T_i would no longer be correlated with the new error term. The brilliance of Heckman's research was his recognition that although one could not observe v_i directly, one could calculate its expected value from Equation 13 and that this expected value of v_i could be used as a control variable (function) to consistently estimate Equation 7.

In order to calculate the expected value of v_i in Equation 13, one needs to make an assumption about the distribution of v_i. Typically, the distribution is assumed to be normal. If $f(.)$ is the normal density function and $F(.)$ is the corresponding cumulative distribution function, then

$$E[v_i \mid Z_i a, T_i] = \frac{f(Z_i a)}{[1 - F(Z_i a)]} \quad \text{when} \quad T_i = 1 \qquad 14a.$$

[8]The general selection model considered by Heckman (1979) can also be estimated by maximum likelihood or nonlinear least squares, although this involves stronger distributional assumptions than does the lambda method discussed here (see Winship & Mare 1992 for a brief discussion).

and

$$E[v_i \mid Z_i a, T_i] = \frac{-f(Z_i a)}{F(Z_i a)} \quad \text{when} \quad T_i = 0. \qquad 14b.$$

Equation 14a simply gives the formula for lambda in a standard sample selection problem. In the treatment context, one would calculate a lambda for those in the treatment condition ($T_i = 1$) using Equation 14a and a lambda for those in the control equation using Equation 14b. These lambdas would then be entered into Equation 7 or, similarly, Equation 11 as controls, analogous to the inclusion of two more *X*s. Thus, the procedure here is identical to Heckman's lambda method for correcting for selection bias, except that two distinct lambdas, one for the treatment and one for the control group, are utilized.

As Heckman and many others have come to recognize, estimates from his method can be sensitive to assumptions about the distribution of v_i. This issue is discussed in Winship & Mare (1992). Typically, if one is estimating, for example, Equation 11, there should be *Z*s in the selection equation that are not also *X*s. Recently, Heckman and his colleagues (1998a) have suggested that one might, in the spirit of Rubin's and Rosenbaum's propensity score method, match on lambda. This strategy is similar to methods proposed by Powell (1987) and Honore & Powell (1994) for dealing with sample selection.

Instrumental Variables

When an independent variable in a regression model is endogenous (i.e. correlated with the error term), the traditional approach in econometrics is to use instrumental variables. In our context, if there is some variable (or set of variables) that affects assignment but does not affect the outcome, then this variable (or set of variables) can be used as an instrument to deal with the possibility that assignment to treatment is nonrandom. The power of the instrumental variable approach is derived solely from the assumption that the instrument only affects the outcome indirectly through the independent variables in the model. In general, this assumption cannot be tested.

Instrumental variable techniques were first developed by economists to estimate simultaneous equation models with jointly determined supply and demand equations from a set of competitive markets (Hood & Koopmans 1953). For any one market, only one point is observed—the competitive equilibrium price and quantity at the intersection of the supply and demand curves. In order to estimate the demand curve, a variable is needed that shifts the supply curve. One can then observe different points of intersection between the demand curve and the shifted supply curve. Similarly, in order to estimate the supply curve, a second variable is needed that shifts the demand curve so that one can observe

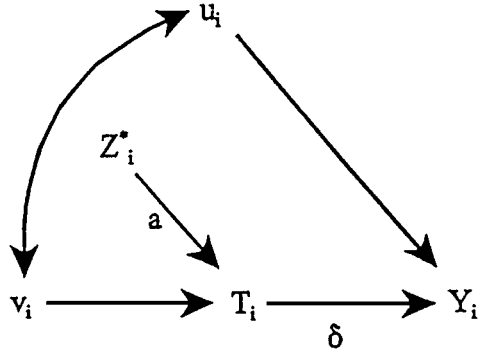

Figure 4 The instrumental variable strategy for the estimation of a treatment effect when selection is nonrandom.

different points of intersection between the supply curve and the shifted demand curve. Variables that fulfill these functions are instrumental variables for quantity supplied and quantity demanded, respectively.

TRADITIONAL INSTRUMENTAL VARIABLES In the counterfactual causality framework outlined above, an instrument is a variable that affects assignment but does not directly affect either Y_i^t or Y_i^c. Consider the simple path model in Figure 4, analogous to Figure 3, for the observed Y_i. In order to simplify the discussion, assume that u_i only contains unobserved determinants of Y_i (i.e. there are no Xs). Equivalently, assume that the effects of any Xs have already been conditioned out. In Figure 4, the potential instrument Z_i^* is assumed to be uncorrelated with u_i. Contrast this with Figure 3, where Z_i or more accurately the propensity score, $P(Z_i)$, is assumed to be (strongly) correlated with u_i. In Figure 3, the strength of the correlation between Z_i and u_i is sufficiently strong so that u_i and v_i, and thus u_i and T_i, are assumed to be uncorrelated. Figures 3 and 4 show that Z_i and Z_i^* relate to u_i in totally opposite ways.

Ignoring scale factors, or similarly assuming that all variables have been standardized, we can see that in Figure 4 the covariance between Z_i^* and Y_i is δa if the covariance between Z_i^* and T_i is a and δ is the direct effect of T_i on Y_i (i.e. the effect of T_i on Y_i not including its indirect effect through u_i). Thus, we can estimate the treatment effect as $\hat{\delta} = \text{Cov}(Y_i, Z_i^*)/\text{Cov}(T_i, Z_i^*)$.

One way of understanding instrumental variables is in terms of an exclusion restriction—the instrumental variable only affects the outcome indirectly through the treatment variable. In the previous section, we discussed the use of control functions—residualizing Y_i and T_i with respect to some X_i and Z_i (or

set of *X*s and *Z*s) such that the residualized T_i is no longer correlated with the resulting error term. Instrumental variable techniques attempt to achieve the same goal of creating a new T_i that is uncorrelated with the resulting error term but do so in the opposite way. Instead of constructing residualized variables, instrumental variables construct predicted Y_i and T_i where the predicted T_i is uncorrelated with the resulting error term. Analogous to steps (*a*) through (*c*) above, instrumental variable estimates can be obtained by the following two-stage procedure: (*a*) Regress Y_i on Z_i^* and calculate $\hat{Y}_i$; (*b*) regress T_i on Z_i^* and calculate $\hat{T}_i$; and (*c*) estimate $\hat{Y}_i = \hat{T}_i\overline{\delta} + w_i$. This three-step procedure illustrates another way of thinking about instrumental variables. We express both our dependent variable Y_i and independent variable T_i as functions of a third variable Z_i^* that is uncorrelated with the error term. Because Z_i^* is uncorrelated with the error term, the new predicted T_i, $\hat{T}_i$, is uncorrelated with the error term. We can then regress the new predicted Y_i, $\hat{Y}_i$, on $\hat{T}_i$ to obtain a consistent estimate of the treatment effect.

A comparison of alternative strategies based on instrumental variables and control functions is instructive. When using a propensity score, or more generally a control function strategy, we look for control variables, as in Figure 3, that are highly correlated with the error term in the structural equation so that after conditioning on these variables, the treatment indicator variable is no longer correlated with any portion of the error term that remains. When using an instrumental variables strategy, we look for a variable or set of variables, as in Figure 4, that is uncorrelated with the error term. If we can then express the outcome and treatment variables as functions of this variable or set of variables, we can calculate the treatment effect with a simple regression using the new variables that have been predicted from the instrument(s).

A third way of thinking about instrumental variables is as naturally occurring randomization (Angrist et al. 1996, Heckman 1996). This perspective is easiest to appreciate when the instrument is binary, because the standard instrumental variable estimator takes the simple form

$$\hat{\beta}_{\text{IV}} = \frac{(\overline{Y_i} \mid Z_i^* = 1) - (\overline{Y_i} \mid Z_i^* = 0)}{(\overline{T_i} \mid Z_i^* = 1) - (\overline{T_i} \mid Z_i^* = 0)}. \qquad 15.$$

known in econometrics as the Wald estimator. The numerator is the standard estimator for the treatment effect of Z_i^* on Y_i, and the denominator is the standard estimator for the treatment effect of Z_i^* on T_i. If Z_i^* is randomly assigned, as in the case of a natural experiment, then both estimates are consistent. Because we assume that Z_i^* only affects Y_i through T_i, the ratio of these two effects consistently estimates the effect of T_i on Y_i.

WEAKNESSES OF CONVENTIONAL IV TECHNIQUES Instrumental variable (IV) techniques have three main weaknesses (see Heckman 1997 for a detailed discussion). First, assumptions that exclusion restrictions are valid are generally untestable and sometimes unbelievable. Second, the standard errors of IV estimates can be large if the instrument is weak or the sample size is not large. Third, IVs only consistently estimate the true average treatment effect when the treatment effect is constant for all individuals, an assumption that is often unreasonable. We discuss each of these problems in turn.

Even within economics, the assumed validity of an exclusion restriction is often controversial. Consider one of the most celebrated example of IVs—the draft lottery number as an instrument for veteran status in estimating the effect of military service on earnings (Angrist 1990). The case for excluding lottery number from the earnings equation that is the primary interest of the study is the randomization of the draft lottery (numbers assigned by day of birth). However, differential mortality patterns may lead to sample selection that spoils the randomization (Moffitt 1996). In addition, employers may behave differently with respect to individuals with different lottery numbers, investing more heavily in individuals who are less likely to be drafted. As a result, lottery number may be a direct, though probably weak, determinant of future earnings (Heckman 1995, 1997).

IV point estimates of treatment effects are often accompanied by wide confidence intervals.[9] The variance of the IV estimator for a bivariate regression with a single instrument is

$$\mathrm{Var}(\hat{\beta}_{\mathrm{IV}}) = \frac{\sigma_e^2 \,\mathrm{Var}(Z_i^*)}{n\,\mathrm{Cov}(T_i, Z_i^*)^2}, \qquad 16.$$

where n is the sample size and σ_e^2 is the variance of the error term. The standard error of an IV estimate is inversely proportional to both the covariance between T_i and Z_i^* and the sample size. To obtain precise estimates, either the sample size must be unusually large and/or T_i and Z_i^* must be strongly correlated. The latter case has led researchers to describe the perfect instrument as an apparent contradiction. A valid instrument must be uncorrelated with the error term but highly correlated with the treatment variable T_i. However, because T_i is correlated with the error term, motivating the use of instrumental variables in the first place, any variable that is highly correlated with T_i is likely also to be correlated with the error term, even though this is not necessarily so.

Angrist & Krueger (1991, 1992) have capitalized on the large size of census datasets, using quarter of birth as an instrument for education when estimating

[9]IV estimates always have larger variance than OLS estimates. Thus, even if it is known that OLS estimates are biased, they may be preferred to apparently unbiased IV estimates if the mean-squared error of OLS estimates is smaller.

the effect of education on earnings. Because of education laws regarding minimum ages for school entry and later voluntary dropping out, individuals born just before and just after specific cutoff dates (e.g. January 1) are likely to differ in their levels of educational attainment. Angrist & Krueger (1991, 1992) find that quarter of birth is indeed (weakly) correlated with educational attainment and assume that it has no direct effect on earnings. Because of the large size of the census samples they utilize, they are able to obtain precise IV estimates of the effect of education on earnings. Their work, however, has received considerable criticism (Bound et al. 1995; also see Winship 1998). All these critics point out that the covariance of earnings with quarter of birth and the covariance of educational attainment with quarter of birth are both weak. In this case, the instrumental variable estimator is essentially the ratio of two very small numbers, the covariance between quarter of birth and education and the covariance between quarter of birth and earnings. As a result, the IV estimate may potentially be unstable. Even if the direct effect of quarter of birth on earnings is small, it will make a substantial contribution to the covariance between these two variables. As a result, large biases in the IV estimate will occur. Bound et al (1995) discuss a variety of reasons that quarter of birth might have a small but non-zero direct effect on earnings. If this direct effect is non-zero, as they argue, then Angrist's and Krueger's IV estimates are likely to be substantially biased.

As already noted, the instrumental variable estimator only estimates the average treatment effect when the treatment effect is constant. What does it estimate when the treatment effect is heterogenous? Recent work by Imbens & Angrist (1994), Angrist & Imbens (1995), Angrist et al (1996), and Imbens & Rubin (1997) investigates this issue by extending the potential outcome framework discussed at the beginning of this paper. This extension is accomplished by assuming that treatment assignment is a function of an exogenous instrument Z_i^*.

For simplicity, assume that both the treatment and the instrument are binary, and that the instrument Z_i^* is a randomly assigned incentive to enroll in the treatment program (e.g. a cash subsidy). When both the treatment and incentive are binary, individuals eligible to receive the treatment can be categorized into four mutually exclusive groups. Individuals who would only enroll in the program if offered the incentive and thus who would not enroll in the program if not offered the incentive are labeled compliers [i.e. individuals for whom $T_i(Z_i^* = 0) = 0$ and $T_i(Z_i^* = 1) = 1$]. Likewise, individuals who would only enroll in the program if not offered the incentive are called defiers [i.e. individuals for whom $T_i(Z_i^* = 0) = 1$ and $T_i(Z_i^* = 1) = 0$]. Individuals who would always enroll in the program, regardless of the incentive, are called always-takers [i.e. individuals for whom $T_i(Z_i^* = 0) = T_i(Z_i^* = 1) = 1$]. Finally, individuals who would never enroll in the program, regardless of the incentive, are called never-takers [i.e. individuals for whom $T_i(Z_i^* = 0) = T_i(Z_i^* = 1) = 0$].

Based on the potential treatment assignment function, Imbens & Angrist (1994) define a monotonicity condition. In the binary-treatment-binary-instrument case, their condition requires that either $T_i(Z_i^* = 1) \geq T_i(Z_i^* = 0)$ or $T_i(Z_i^* = 1) \leq T_i(Z_i^* = 0)$ for all i. In words, the instrument must affect the treatment assignment of all individuals in the same direction and thus in a monotone fashion. For all individuals, an increase (decrease) in their Z_i^* must either leave their treatment condition the same or, among individuals who change, change them in the same way. There may be either defiers or compliers but not both among those eligible to receive the treatment. Conventional IV methods make no assumptions about the coexistence of compliers and defiers.[10]

When an exclusion restriction is satisfied and when the treatment assignment process satisfies the monotonicity condition, the conventional IV estimate is an estimate of what is known as the local average treatment effect (LATE), the average treatment effect for either compliers alone or for defiers alone, depending on which group exists in the population.[11] LATE is the average effect for the subset of the population whose treatment assignment is affected by the instrument. The individual-level treatment effects of always-takers and never-takers are excluded in the calculation of LATE. When the monotonicity condition is not satisfied and treatment effect heterogeneity seems likely, the conventional IV estimator yields a parameter estimate that has no clear interpretation.

LATE has three problems: (*a*) It is defined by the instrument, and thus different instruments define different average treatment effects for the same group of individuals eligible to receive the treatment; (*b*) it is an average treatment effect for a subset of individuals that is inherently unobservable no matter what the instrument; (*c*) it is hard to interpret when the instrument measures something other than an incentive to which individuals can consciously respond by complying or defying.

BOUNDS WITH INSTRUMENTAL VARIABLES If IV techniques generally do not provide an estimate of the average treatment effect when there is treatment effect heterogeneity, then can IVs tell us anything at all about the average treatment effect? In a recent paper, Manski & Pepper (1998) investigate this question in some depth showing what can be learned when standard and when weaker IV assumptions are maintained.

[10]Note that when an instrument is valid, there must be at least some compliers or some defiers, otherwise the sample would be composed of only always-takers and never-takers. In this case, Z_i^* would not be a valid instrument because it would be uncorrelated with treatment assignment.

[11]The exclusion restriction that defines LATE is stronger than the conventional exclusion restriction that the instrument be mean-independent of the error term. Instead, Imbens & Angrist (1994) require that the instrument be fully independent of the error term. Imbens & Rubin (1997) argue that the strong independence restriction is more realistic because it continues to hold under transformations of the outcome variable. An assumption about the distribution of the outcome is thereby avoided.

Manski & Pepper (1998) define the traditional IV assumption in terms of mean independence. Specifically, in our notation, for arbitrary values s and s'

$$E\left[Y_i^{\mathrm{t}} \mid X, Z_i^* = s\right] = E\left[Y_i^{\mathrm{t}} \mid X, Z_i^* = s'\right] \quad 17a.$$

and

$$E\left[Y_i^{\mathrm{c}} \mid X, Z_i^* = s\right] = E\left[Y_i^{\mathrm{c}} \mid X, Z_i^* = s'\right]. \quad 17b.$$

In words, Equations 17a and 17b require that the mean values of the outcomes in each subpopulation defined by values of Z_i^* be equivalent to those in the population as a whole. The implication of this assumption is that the bounds assumption analysis, discussed earlier, and the monotone treatment response assumption alone also apply within each subpopulation defined by Z_i^*. As a result, the bound on the treatment effect can be defined as the intersection of the bounds across subpopulations defined by Z_i^* (see Manski 1994,1995; Manski & Pepper 1998). The common bound can only be narrowed with the aid of an IV if the bounds differ across subpopulations. Because the monotone treatment selection assumption, discussed briefly above, is an assumption about how treatment is assigned, it may or may not make sense to assume that it holds within subpopulations defined by the instrument.

As we and many others have noted, the standard IV assumption is a strong condition. Manski & Pepper consider a weaker assumption, the monotone IV assumption (MIV). It states that for $s \geq s'$,

$$E\left[Y_i^{\mathrm{t}} \mid X, Z_i^* = s\right] \geq E\left[Y_i^{\mathrm{t}} \mid X, Z_i^* = s'\right] \quad 18a.$$

and

$$E\left[Y_i^{\mathrm{c}} \mid X, Z_i^* = s\right] \geq E\left[Y_i^{c} \mid X, Z_i^* = s'\right]. \quad 18b.$$

Thus, in Equations 18a and 18b, the mean values of both potential outcomes are weakly increasing functions in Z_i^*.

It is easier to demonstrate how the MIV condition bounds the mean of each outcome than it is to demonstrate directly how the MIV condition bounds the average treatment effect that is a function of these means. Without loss of generality, consider the mean of Y_i^{t} in the population. Under the standard IV assumption, the upper bound for this mean will be equal to the smallest upper bound across the different subpopulations defined by the instrument. Under the MIV assumption, the upper bound of the conditional mean within the subpopulation defined by a particular value, s', of the instrument will be equal to the smallest upper bound for all subpopulations defined by values of the instrument greater than or equal to s'. The upper bound for the overall mean of Y_i^{t} will simply be the weighted average of the subpopulation upper bounds

where the weights are equal to the proportions of the sample in the various subpopulations defined by Z_i^*. The determination of the analysis for the lower bound of Y_i^t is analogous, as are the determination of the bounds on Y_i^c.

Manski & Pepper (1998) use the assumptions of monotone treatment response, monotone treatment selection, and MIV to determine the bounds on the effect of education on the logged wages of respondents to the National Longitudinal Survey of Youth. When they invoke monotone treatment response and selection assumptions, they find that the bound for the effect of a twelfth year of schooling is [0, 0.199], that the bound for the effect of a fifteenth year of schooling is [0, 0.255], and that the bound for the effect of a sixteenth year of schooling is [0, 0.256]. When they use the Armed Forces Qualifying Test as a monotone instrumental variable while still maintaining the monotone treatment response and selection assumptions, they obtain narrower bounds respectively of [0, 0.126], [0, 0.162], and [0, 0.167]. Although these bounds are somewhat broader than one might wish, they are consistent with the range of estimates typically found in the literature.

LONGITUDINAL METHODS

The use of longitudinal data to estimate treatment effects has a long history. Longitudinal data are useful because they allow individuals to serve as their own controls. The treatment effect for an individual can then be estimated as the change in the pretest and the posttest measurements of their outcome. Of course, any such estimator implicitly assumes that the outcome would have remained unchanged in the absence of treatment. As this is often an unrealistic assumption, we need to be able to estimate for those individuals in the treatment group how their outcomes would have evolved in the absence of treatment.

There are two possible sources of information for constructing this counterfactual trajectory. First, if there are multiple pretest observations, it may be possible to extrapolate from these observations and estimate what the outcome would have been in the absence of treatment, assuming that the future is similar to the past. Second, if there is a control group, then the evolution of its outcome may be used to model what the outcome would have been in the absence of treatment, assuming that the treatment and control groups are similar in key respects.

In the past two decades, many new techniques have been developed to utilize longitudinal data to estimate causal effects. Five important insights have emerged from this research: (*a*) in many circumstances, aggregate cohort-level data contain sufficient information to consistently estimate a causal effect (Heckman & Robb 1985, 1986, 1988); (*b*) whenever possible, the data should be used to test the appropriateness of alternative models; (*c*) multiple measurements of the outcome before and after the treatment are essential

both for estimating sophisticated models and for testing the appropriateness of alternative specifications; (*d*) understanding the underlying behavior that generates assignment to the treatment and control groups is critical to the proper modeling of suspected unobservable effects; and (*e*) it is only possible to estimate the average treatment effect for the treated in most longitudinal models because the average treatment effect for the entire population is typically unidentified.

Heckman & Robb (1985, 1986, 1988) provide an extensive, although challenging, review of alternative methods for estimating causal effects using longitudinal (as well as cross-sectional) data. Space does not permit us to provide a similar review here. Moreover, we are confident that many readers would find a full exposition of the technical details of these models more overwhelming than illuminating. Our aim in this section, rather, is to provide an overview of commonly used methods, both old and new, and an assessment of their utility. In so doing, we hope to provide insight into the types of information that are available in longitudinal data to aid in the estimation of a causal effect. We discuss five basic models: interrupted time series models, fixed effect models, differential linear growth rate models, analysis of covariance models, and covariance stationary models.

Interrupted Time Series Design

Perhaps the simplest data structure for estimating causal effects, the interrupted time series (ITS) design uses standard time series methods on multiple observations over time for a single unit in order to estimate a causal effect of a variable. The core of the method involves the specification and estimation of the error structure (i.e. the nature of the interdependence of the period-specific error terms over time). A variety of textbooks provide comprehensive treatments of time series methods (e.g. Harvey 1990, Hamilton 1994, Judge et al 1985). We do not review them here.

The logic of the ITS design parallels that of the regression discontinuity design discussed earlier. In an ITS analysis, time plays the role of Z, and there are now multiple measures over time for a single unit of analysis. The unit might be a country, city, cohort of individuals, or a single person. It is assumed that the treatment is introduced at a specific time and has an immediate impact. The goal is then to estimate how the dependent variable would evolve over time in both the presence and absence of a treatment effect.

We now change notation slightly. Let Y_t be the outcome at time t. For an ITS analysis we do not need an "i" subscript because we are only analyzing data for a single unit of analysis. We continue to denote treatment by the dummy variable T.

We can formally represent the ITS model as

$$Y_t = b_{0_t} + T_t b_{1_t} + e_t. \quad 19.$$

Note that both the intercept, b_{0_t}, and the treatment effect, b_{1_t}, potentially vary over time. This model is not identified without imposing further structure on how these two parameters are related to time. Return to Figure 1, which presents the basic intuition behind both the regression discontinuity design and the ITS design. For the ITS model, this figure assumes that Y_t, under both treatment and control conditions, grows linearly with time. This implies that for all t, the differences $b_{0_{t+1}} - b_{0_t}$ and $b_{1_{t+1}} - b_{1_t}$ are constants. The dashed line shows the predicted evolution for Y_t in the absence of the treatment. As shown in Figure 1, in this particular example, the treatment has caused a shift in Y_t and a change in the slope.

Equation 19 could be augmented by the inclusion of covariates, X_t. A frequent problem with time series analyses (unlike most cross-sectional analyses) is that the number of parameters in the model may be large relative to the number of observations. As a result, the amount of information available to estimate the parameters may be small. This problem can be especially acute when there is strong dependence among the period-specific error terms, e_t.

The ITS design has the same potential problems as the regression discontinuity design. An ITS analysis assumes that the future is sufficiently like the past that the past can be used to estimate how Y_t would have evolved in the absence of treatment. As with the regression discontinuity design, Figure 2 illustrates the bias in the estimate of the treatment effect that can result when this assumption does not hold.

At the beginning of this section, we noted that the availability of aggregated cohort-level data alone is sometimes sufficient for estimating a treatment effect. This conclusion can be presented in the framework of an ITS model where we assume that Y_t measures the average value for a cohort of individuals on some dependent variable (e.g. wages). Equation 19 is consistent with a specification in which all individuals receive the treatment. In this case, b_{1_t} represents the contemporaneous increase in wages caused by the treatment (e.g. training), and variation in b_1 over time represents the changes in wage growth caused by the treatment. What if only some known portion of the cohort, π, received training? As shown by Heckman & Robb (1985, 1986, 1988), we can still consistently estimate the average treatment effect for those who received training. In the situation where b_{1_t} does not vary with time, the average treatment effect for the treated equals (b_{1_t}/π).

The time series literature provides a host of sophisticated ways of modeling data. The core material in this literature is typically covered in a one- or even

two-semester advanced graduate-level econometrics course. Time and space limitations prevent us from providing even a brief overview of these models. The time series literature also contains alternative conceptions of causality to those considered here. The key ideas are those of Granger causality and cointegration (see Harris 1995 and Hendry 1995 for definitions and further discussion; see Holland 1986 and Sobel 1995 for connections with the counterfactual framework). Robins (1986, 1987, 1997) provides a full analysis of the estimation of causal effects when a treatment may be applied repeatedly and at multiple times.

General Model Specification

The methods that we want to consider in the remainder of this section all assume that we have individual-level data with pretest and posttest values on the outcome for both treatment and control groups. The goal is to use the control group (as well as possible multiple pretest measures on the treatment group) to forecast what the values of the dependent variable would have been for the treatment group in the absence of treatment. This goal can only be accomplished if we know or can effectively estimate what the relationship would have been in the absence of treatment between the pretest and posttest values of the treatment and control groups.

Consider the simplest but by far the most common situation, where we have a single pretest and posttest value for the two groups. As Judd & Kenny (1981) demonstrated, even in a linear world there are at least three possibilities. These are shown in Figures 5*a*, *b*, and *c*. In all three figures, the observed values are identical. The estimate of the treatment effect, however, differs substantially, depending on what we assume would have happened to the treatment group if they had not been exposed to the treatment.

As is discussed below, Figures 5*a*, *b*, and *c* characterize three traditional models for estimating a causal effect with pretest and posttest data. To understand the assumptions behind each of these models, we first build a general model of which the three models are special cases. Consider the following model:

$$\begin{aligned} Y_{it} = {} & b_{0_t} + T_{it}b_1 + && \text{(Basic structural parameters)} \\ & X_{it}b_{2_t} + T_{it}X_{it}b_3 + && \text{(Observed heterogeneity)} \\ & \lambda_{it} + T_{it}\alpha_i + e_{it} && \text{(Unobserved heterogeneity),} \end{aligned} \qquad 20.$$

where $e_{it} = \rho e_{it-1} + v_{it}$. The first term is b_{0_t}, the intercept that varies with t in order to capture the general effects of time; b_1 is the treatment effect that we assume is time invariant. This assumption is not essential. Because we want to allow for the possibility that the treatment effect may vary across individuals, we assume that b_1 is the average treatment effect for the population of interest or the group for whom $X_{it} = 0$.

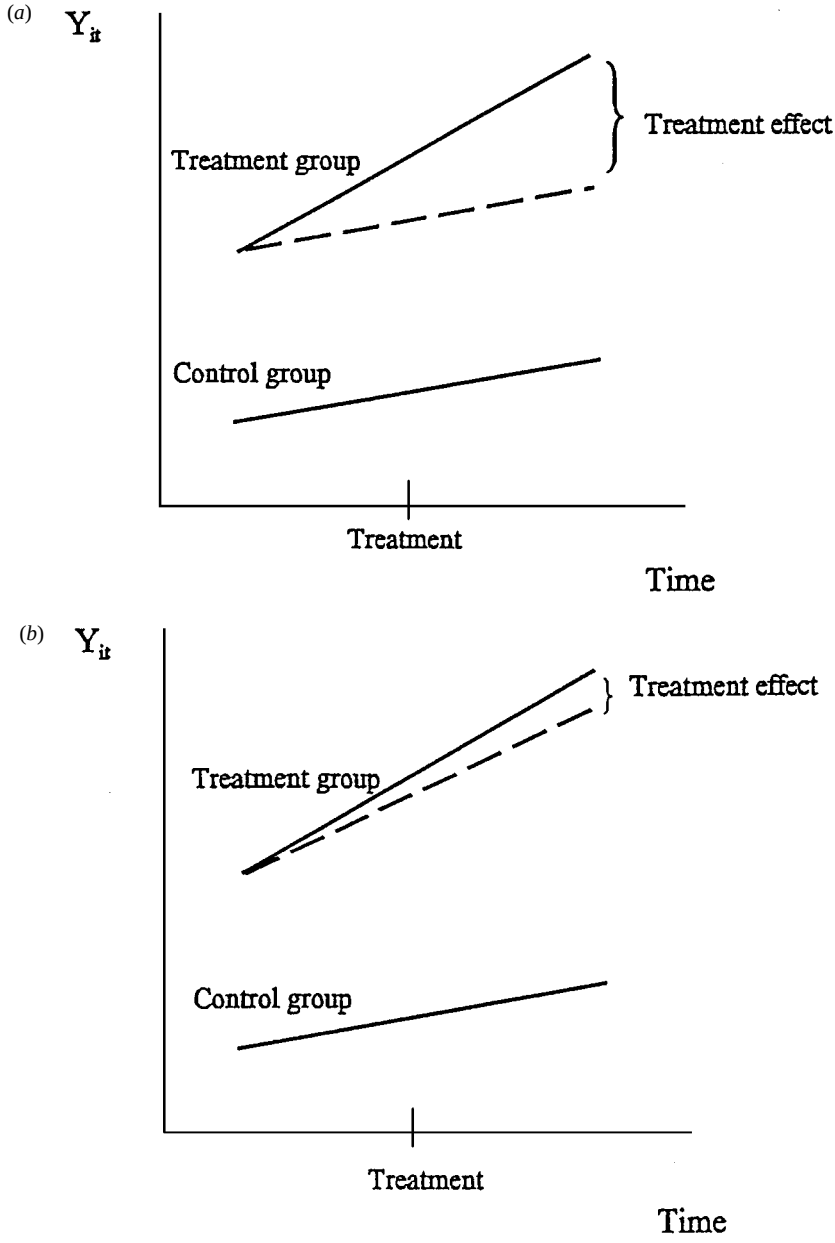

Figure 5 The true average treatment effect when unobserved heterogeneity does not differentially affect the rate of growth for both groups.

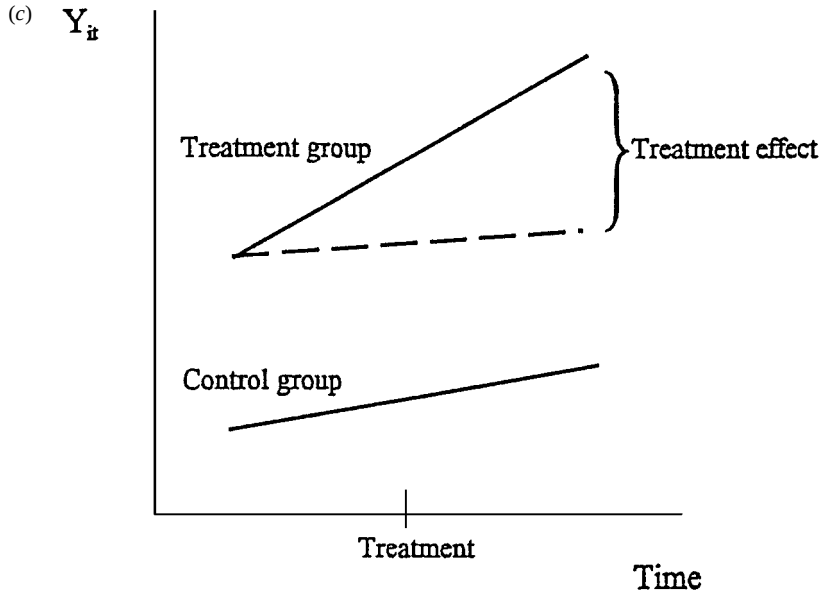

Figure 5 (*Continued*)

X_{it} consists of the values of fixed, current, past, and future values of exogenous control variables that are relevant at time t. The two terms in the second line of Equation 20, which are interactions of the two terms of the first line with X_{it}, represent observed heterogeneity. The coefficients in b_{2_t} represent (possibly time-varying) shifts in the intercept that are a function of X_{it}. The coefficients in b_3 represent interactions between X_{it} and the treatment effect. We assume that these interactions are time invariant, but this assumption is not essential.

The next three terms constitute the different components of the error term for Equation 20. The first two terms are measures of unobserved individual heterogeneity that are analogous to the two terms for observed individual heterogeneity in the second line. The first term, λ_{it}, represents an individual specific intercept that we allow to vary with time. Thus, the components of λ_{it} capture both constant (or fixed) differences between individuals as well as the possibility that Y_{it} may grow at different rates across individuals i in ways that are not captured in the Xs.

The second term, $T_{it}\alpha_i$, measures the degree to which the treatment differentially impacts each individual i. As with the previous treatment effect terms, we assume that this treatment effect is time invariant. In general, we would expect that individuals who are most likely to benefit from the treatment (i.e.

individuals who have high α_is) would self-select into the treatment. Unfortunately, longitudinal data typically do not provide a way to control for these differences. In most cases, it is only possible to estimate the average treatment effect for the treated, not the average treatment effect for the population as a whole.

The final component of Equation 20, e_{it}, represents an individual-and-period-specific component of the error term. We want to allow for interdependence among the e_{it} over time to capture what is known in the time series literature as transitory effects. We assume an autoregressive structure of order one (AR1), that is $e_{it} = \rho e_{it-1}$ where ρ is the correlation between e_{it} and e_{it-1}. More complicated dependence in the form of an autoregressive moving-average structure could be assumed. These are reviewed in the standard time series textbooks cited above.

Finally, v_{it} is a pure time-specific error that is uncorrelated with anything else. To simplify the exposition, we assume that v_{it} has constant variance across individuals. In the time series literature, v_{it} is often referred to as the innovation in the process. Because it is purely random, it cannot be forecast. Typically, it is assumed to be a priori unknown to both the individual and the analyst.

How does Equation 20 relate to Figures 5*a–c*? Assume for the moment (and for most of our discussion of longitudinal methods below) that all of the heterogeneity is represented by the unobservables. Standard techniques can be used to eliminate differences that are a function of observed *X*s. When estimating the treatment effect, we would like the treatment and control groups to be identical, at least conditional on X.[12] Our concern is that the two groups may also differ in terms of the unobservable components found in Equation 20 because assignment to the treatment group may be a function of unobservable characteristics. The question then is whether there are techniques that can eliminate potential differences between the treatment and control groups that are a function of the unobservable components in Equation 20. After eliminating differences due to X, can we "control out" the effects of the unobservables that are potentially related to treatment assignment?

In Figure 5*a*, in the absence of a treatment, the parallel lines for the two groups indicate that the differences between Y_{it} for individuals in the treatment and control groups (on average) remain constant over time. This case is consistent with a model in which unobserved differences between the treatment and control group are a function solely of a λ_i that is time invariant or fixed, as would occur if we had omitted an X that was constant over time and correlated with the treatment variable. For example, if we were estimating the effects of additional schooling on wages, there might be unmeasured and thus unobserved components of family background that need to be controlled.

[12]In the linear models used in our exposition, it is only necessary that the expected values of the components on the right-hand side of Equation 20 be identical for the treatment and control groups.

Figure 5*b* illustrates a situation in which, in the absence of treatment, the growth rate for Y_{it} differs for individuals in the treatment and the control groups. Assume that the λ_{it} change linearly with time such that $\lambda_{it} - \lambda_{it-1} = \tau_i$, where τ_i is a constant for each individual i for all t. In Figure 5*b* the rate of increase (in the absence of the treatment) in λ_{it}, τ_i, is higher for those in the treatment group than for those in the control group. As an example, more intelligent individuals may be more likely to continue in school, but whether they are in school or not (that is, whether they receive the treatment or not), they may still learn faster than individuals with less intelligence.

Figure 5*c* might result in two different ways. First, as in Figure 5*b*, the rate of increase in λ_{it}, τ_i, may differ between the treatment and control group. Here, however, the rate of increase is greater for those in the control group than for those in the treatment groups. For example, an increase in the incidence rate of disease might be greater in a control group if willingness to take other unobservable preventive measures (unrelated to the treatment regime itself) is greater on average for those in the treatment group.

A second circumstance when Equation 20 is consistent with Figure 5*c* is when ρ in Equation 20 is positive. In this case, if assignment to the treatment group instead of the control group is a function of e_{it}, then assignment is a function of transitory components of the unobservables. If assignment is only a function of transitory components, then over-time differences between the two groups would shrink to zero in the absence of treatment.

A number of empirical examples exist where assignment to the treatment group is a function of transitory components of the unobservables. The classic case, which reverses the labeling of the treatment and control groups in Figure 5*c*, is where individuals who are experiencing low wages in the near term are more likely to enroll in a job-training program because they experience lower opportunity costs (Ashenfelter 1978, Ashenfelter & Card 1985). Regression toward the mean in this case produces an apparent training effect because the wages of individuals who received training would have increased on average (regressed toward the mean) even in the absence of training. This case differs from the one presented in Figure 5*c* in that the treatment program is compensatory—individuals with the lower wages are in the treatment group.

Several approaches are available to consistently estimate the treatment effect, b_1, in Equation 20. The most obvious, but often difficult method, would be to use panel data to fully estimate a complicated model specification such as Equation 20. If the model is properly specified, then the treatment effect b_1 can be consistently estimated. In general, it will only be possible to estimate models of this type of complexity if one has multiple pretest and post test measures of Y_{it}. The technical issues involved in estimating models of this type vary across different model specifications.

Alternative Methods

Over the years, a number of simple methods for estimating treatment effects have been proposed. These include charge score analysis, the analysis of covariance, and their generalizations. We now consider these models and the conditions under which they give consistent estimates of the treatment effect, b_1. An extensive literature has argued the relative merits of these different approaches (see Judd & Kenny 1981; Holland & Rubin 1983; Allison 1990).

The appropriateness of a model depends on whether it provides a consistent estimate of the treatment effect in a particular context. This depends on whether the model's assumptions are congruent with the underlying process that generates the data. The appropriateness of a model for a specific situation can only be determined through theoretical and empirical analysis. No one statistical model is a panacea.

The problem of whether a model is consistent with an underlying process that generates Y_{it} is potentially complicated. As we discuss below, if one is to have confidence in one's results, it is essential to test the appropriateness of one's model specification. But what constitutes a proper model specification? As discussed above, one approach to consistently estimating the treatment effect in a longitudinal model is to attempt to specify and estimate a full model specification for Y_{it}. To accomplish this, it may be necessary to have multiple pre- and posttest observations on Y_{it}.

Our above discussion of Rubin's and Rosenbaum's propensity score, however, suggests that estimation of the full model for Y_{it} may not be necessary to correct for the effects of assignment. We demonstrated that if we could condition on the probability of assignment (or at least on those factors that determine assignment), the treatment effect could be consistently estimated even in the presence of omitted variables.

Does this mean that we can get away with not estimating the full model for Y_{it}? As Heckman has argued repeatedly over the years, in many situations this is not likely to be possible. If individual choice is involved in the assignment process, it is likely that individuals will choose to be in the treatment and control group based on the consequences of treatment for their future Y_{it}. In this case, individuals (at least crudely) use the previous history of their Y_{it}, plus the total history of Y_{it} for others, both pre- and posttreatment, to project the future values of their Y_{it} under both the treatment and control. If so, the assignment process will be a function of the parameters of the model that the individual uses to predict future Y_{it}.

The question then is what model is the individual using to predict their future Y_{it}? If it is simpler than the full model, then it may well be possible to condition on only those components of the model that determine assignment and consistently estimate the treatment effect. In many situations, it is unclear

why an individual would not use something like the full model to predict future Y_{it}. Thus, we may be stuck with having to try to specify and estimate the full model that generates Y_{it}. Of course, the greatest concern is that the individual may be using a prediction model that is more complicated or more accurate than the one used by the analyst. This might be due to the fact that the analyst has used too simple a model and/or that the individual has access to information that the analyst has no way of incorporating into her model—either directly through observed *X*s or indirectly through a particular specification of the structure of the unobservables. In this situation, it may simply be impossible to consistently estimate the treatment effect.

CHANGE SCORE OR FIXED EFFECTS MODELS Change score or fixed-effect models are a common and simple method for estimating causal effects when pretest and posttest data are available for separate treatment and control groups. The basic model can be formalized in two ways. The standard change-score model is

$$(Y_{it} - Y_{it-1}) = c_0 + T_i c_1 + (X_{it-1} - X_{it-1})c_2 + u_{it}, \qquad 21.$$

where $T_i = 1$ if the individual received the treatment (and $T_i = 0$ otherwise), $c_0 = b_{0_t} - b_{0_{t-1}}$, $c_1 = b_1$, $c_2 = b_2$, and $u_{it} = e_{it} - e_{it-1}$. This model can also be formalized as a fixed effect model making its relation to Equation 20 more transparent:

$$Y_{it} = b_{0_t} + T_{it} b_1 + X_{it} b_2 + \lambda_i + e_{it}, \qquad 22.$$

where λ_i is a time invariant or fixed individual specific effect and the e_{it} for individual i and across time are assumed to be uncorrelated. The fixed-effect formulation allows for the possibility that multiple pretest and posttest outcomes may be observed on each individual. The model implies that there are permanent fixed differences between individuals in their Y_{it}. As a result, as the process evolves from time $t-1$ to time t there will be regression toward the mean in Y_{it}, but the regression will be toward the individual specific mean of Y_{it} not the overall population mean of Y_{it}.

Because the λ_i terms represent all fixed, time-invariant differences between individuals, the effects of constant *X*s are absorbed into λ_i. This is most apparent in Equation 21, where we see that only the effects of *X*s that change over time are estimated. The fixed-effect model is equivalent to a standard regression model where a separate dummy variable has been included for each individual, which is then estimated by OLS. Alternatively, Equation 21 can be estimated by OLS. Heckman & Robb (1985, 1986, 1988) show that if we know the identity of individuals who will receive the treatment, then the fixed-effect model can be estimated from cohorts based on repeated cross sections.

As can be seen from Equation 22, the fixed-effect model is a constrained version of the general model in Equation 20 because it assumes there is no transitory component to the error term ($\rho = 0$) and the effect of the X_{it} are invariant with respect to time ($b_{2_t} = b_2$). The first constraint implies that any unobserved differences between the treatment and control groups must be constant over time, as shown in Figure 5*a*. As with all the longitudinal models we consider, the fixed-effects model also assumes that the effect of the treatment is constant across individuals ($b_3 = \alpha_i = 0$). If this is not the case, then the treatment effect estimate is a consistent estimate only of the average treatment effect for the treated, not the average treatment effect for the entire population.

The fixed-effect model will only provide consistent estimates of the treatment effect if Equation 22 correctly models the time series structure of Y_{it} or if the fixed effects, λ_i, are the only unobservables that determine assignment to the treatment group. Framed in terms of Heckman's concern above about the consequences of assignment due to individual choice, the fixed-effect model will provide consistent estimates of the treatment effect only if assignment is a function of the fixed effects in Equation 20. However, it only makes sense for an individual to make choices this way if in fact Equation 22, the pure fixed-effects specification, is the correct model for Y_{it}.

DIFFERENTIAL RATE OF GROWTH MODELS In many situations it may be the case that not only are there fixed unobserved individual differences, λ_i, but that there are differences across individuals in the rate of change in Y_{it}. We allow for this possibility by permitting λ_{it} to vary with time. The simplest case is where we assume that the λ_{it} grow linearly but at different rates across individuals (i.e. $\lambda_{it} - \lambda_{it-1} = \tau_i$, a constant growth rate for individual i across all t). Figures 5*b* and *c* are illustrative of this type of process. For example, consistent with Figure 5*b*, we might believe that some individuals learn faster than others, or that because of previous education and training some individuals' wages would grow faster than others.

The differential growth rate model can be estimated as a standard regression model using OLS by including a dummy variable for each individual entered in the equation by themselves and also interacted with time. Alternatively, the model can be estimated by applying OLS to the double difference of both the right- and left-hand sides of Equation 20. If λ_{it} grows quadratically or as a function of even a higher-order polynomial in time, this can be dealt with by differencing further.[13] The differential growth rate model will consistently

[13]In these models, the variance of the outcome or equivalently of the error term may grow without bound. As a result, these models do not have a typical autoregressive moving-average structure. We know of no methods for estimating the differential growth rate model when it includes a transitory auto-regressive component.

estimate the treatment effect only if it accurately models the process generating Y_{it} or assignment is only a function of an individual's fixed effect and individual growth parameter (Heckman & Robb 1985, 1986, 1988).

ANALYSIS OF COVARIANCE MODELS The most common model used to estimate causal effects when both pretest and posttest data are available is the analysis-of-covariance model. In its simplest form, the model is

$$Y_{it} = b_0 + Y_{it-1}\gamma + T_{it}b_1 + u_{it}, \qquad 23.$$

where b_1 is an estimate of the treatment effect and Equation 23 is estimated by OLS.[14] The coefficient γ is equal to the pooled within-treatment group regression of Y_{it} on Y_{it-1}. If u_{it} has constant variance (which is generally assumed and which we also assume), then in the absence of treatment, γ is equal to the correlation between Y_{it} and Y_{it-1}, (that is, the intracluster correlation with each individual considered a separate cluster). As a result, when u_{it} has constant variance, γ must be less than or equal to one. It measures the degree to which each individual's Y_{it} regresses between times $t-1$ and t toward the overall mean of Y_{it}. This regression toward the mean differs from that in the fixed effects model where the individual Y_{it} regress toward individual specific means.

To simplify the exposition, consider the properties of the analysis-of-covariance model in the absence of treatment for all individuals. If we generalize to allow for multiple time periods, then the analysis-of-covariance model is equivalent to an autoregressive model of degree 1:

$$Y_{it} = b_{0_t} + e_{it}, \qquad 24.$$

where $e_{it} = \rho e_{it-1} + v_{it}$. Here ρ is the correlation between temporally adjacent e_{it}, and v_{it} is pure random error that is assumed to be independent of everything. b_{0_t} is a time-varying intercept that follows the generating equation $b_{0_{t+1}} - b_{0_t} = \rho(b_{0_t} - b_{0_{t-1}})$. This model is a constrained version of Equation 20. It makes the strong assumption that all differences in Y_{it} across individuals are transitory. There are no fixed or permanent differences or differences across individuals in the growth rates of their Y_{it}. Thus, Equation 24 implies that between Y_{it-1} and Y_{it} there will be regression toward the mean of a very strong form. Y_{it} across all individuals regress toward the same grand mean.

[14] As written, econometricians would typically interpret Equation 23 as indicating that i is determined in part by its lagged value Y_{it-1}. Under this interpretation, Equation 23 should be estimated using instrumental variables, because under almost any reasonable assumption about the error structure, Y_{it-1} will be correlated with u_{it}, invalidating OLS. Heckman & Robb (1985, 1986, 1988) point out that equations with lagged Y_{it}s can be dealt with by putting them in reduced form. This strategy then yields equations similar in form to Equation 20 which can be dealt with by the techniques discussed here and in their papers.

If Y_{it} is in fact generated by Equation 24, then $\rho = \gamma$. In most situations, however, we would expect that Y_{it} would have both fixed and transitory components. A simple specification that captures this idea is

$$Y_{it} = b_{0_t} + \lambda_i + e_{it}, \qquad 25.$$

where $e_{it} = \rho e_{it-1} + v_{it}$ and where ρ and v_{it} are as in Equation 24. In this case, $\gamma = [\text{var}(\lambda_i) + \rho\text{var}(e_{it})]/[\text{var}(\lambda_i) + \text{var}(e_{it})]$ which is necessarily greater than or equal to the correlation ρ. If there is no transitory component, $\text{var}(e_{it}) = 0$, and γ is still less than one because regression toward the mean is due to the pure random component, v_{it}. If there is no permanent component, then $\gamma = \rho$.

The key to understanding the analysis-of-covariance model is to rewrite Equation 23 as

$$(Y_{it} - Y_{it-1}\gamma) = b_0 + T_{it}b_1 + u_{it}. \qquad 26.$$

Equation 26 shows that γ is a measure of the degree to which Y_{it} should be adjusted by its previous pretreatment value, Y_{it-1}. Specifically, γ measures the degree to which the pretest difference in the treatment and control group Y_{it-1} should be used to correct the post-treatment difference in Y_{it} in estimating the treatment effect:

$$\text{Treatment effect} = b_1 = \left(\overline{Y}^{\text{t}}_{it} - \overline{Y}^{\text{c}}_{it}\right) - \gamma\left(\overline{Y}^{\text{t}}_{it-1} - \overline{Y}^{\text{c}}_{it-1}\right). \qquad 27.$$

If $\gamma = 0$, then no adjustment is needed. The treatment effect is simply equal to the average difference between the treatment and control group in Y_{it}. This would be appropriate only if the Y_{it} were a function of the pure random component of the unobservables, v_{it}.

If $\gamma = 1$, then the Y_{it} are fully adjusted. The treatment effect is then equal to the difference in Y_{it} between the treatment and control groups net of their initial difference. In the latter case, the analysis of covariance model is equivalent to the change-score/fixed-effect model discussed above. This would be appropriate if there is no transitory component in Equation 26.

Assume that Equation 25—which models Y_{it} as a function of fixed, transitory, and random effects—holds and that we estimate γ from the data. In this case $1 > \hat{\gamma} > \rho$. $\hat{\gamma} = 1$ only if there is no transitory component or random component, that is, v_{it} in Equation 25. $\hat{\gamma} = \rho$ only if there is no fixed effect term in Equation 25.[15]

[15]If there is measurement error in Y_{it}s, the measurement error will bias downward the estimate of γ, resulting in an underadjustment for pretreatment differences between the treatment and control groups. This underadjustment will bias the estimate of the treatment effect.

Consider the assignment of individuals to the treatment and control groups. If assignment is a function of the fixed effects λ_i, then no adjustment is needed to the pretreatment difference in Y_{it}. Using the estimated γ in Equation 27 will overstate the treatment effect because the correct adjustment factor is $\gamma = 1$. The analysis-of-covariance model will only give a consistent estimate of the treatment effect in this situation if the estimated $\gamma = 1$. This will occur, however, only when there is no transitory term, e_{it}, or random component, v_{it} in Equation 25. In this case, the prediction of Y_{it} from previous values is trivial because Y_{it} is a constant.

If assignment were a function of only the transitory component, e_{it}, and Y_{it} depends on a fixed component, using the estimated γ would result in an understatement of the treatment effect because the correct adjustment factor is $\gamma = \rho$, which is necessarily less than the estimated γ. In general, the estimated γ will be the correct adjustment factor only if selection is on Y_{it-1}. In this case, in the absence of treatment, the expected shrinkage in the difference in the pretreatment means of Y_{it} for the treatment and control groups and their posttreatment difference is proportional to γ.

But under what circumstances would it make sense for assignment to be based only on Y_{it-1}? As Heckman has argued, if Y_{it} were generated according to, for instance, Equation 25 or the even more complicated Equation 20, it would be reasonable to assume that an individual would want to use values of Y_{it} prior to Y_{it-1} to predict Y_{it}. In essence, one could imagine an individual (at least crudely) estimating their individual specific fixed effect and growth rate so that they could accurately predict what their Y_{it} would be in the absence of treatment. Assuming that is known, values of Y_{it} prior to Y_{it-1} could be ignored only in the situation where Y_{it} had a simple AR1 structure, i.e. where Y_{it} is generated by Equation 24. This leads us to a strong and negative conclusion about the applicability of the analysis-of-covariance model. An analysis of covariance generally will only properly adjust for the pretreatment difference in outcomes between the treatment and control group if treatment assignment is solely a function of the pretreatment outcome, Y_{it-1}. In general, an individual would only choose his assignment based on Y_{it-1} if prior values of Y_{it} or other relevant information were not available or if Y_{it} followed an AR1 specification, as in Equation 24. The latter condition is an extremely strong assumption because it implies that all unobserved differences between individuals are only transitory.

COVARIANCE STATIONARY MODELS The change-score and analysis-of-covariance models (or, similarly, their specifications respectively as a pure fixed-effect model and a pure transitory-effect model) represent extreme model specifications. These two extremes make strong assumptions about how, in

the absence of treatment, the difference between the pretreatment mean Y_{it} for the treatment and control groups will change during the posttreatment period. In the change-score model, the assumption is that there will be no change. For the analysis-of-covariance model, the assumption is that there will be shrinkage of a specific amount. In particular, the shrinkage will be equivalent to the amount of regression toward the mean observed at the individual level. Thus, although ρ is estimated from the data, the analysis-of-covariance model simply assumes that this is the correct shrinkage factor.

In most instances, we would like to use a method that allows for both fixed and transitory effects in the generation of Y_{it} or at least in the assignment process. Equivalently, we would like to estimate how much adjustment is appropriate when estimating the treatment effect using pretreatment differences between the treatment and control groups. The change score and analysis-of-covariance models simply assume alternative levels of adjustment.

Heckman & Robb (1985) show that it is possible to estimate a model that combines an individual fixed effect along with a transitory AR1 component. In fact, all that needs to be assumed is that the process is covariance stationary.[16] This model is consistent with most autoregressive moving-average specifications, including the change-score/fixed effect and analysis-of-covariance models. Assume that you have at least three equally spaced (in time) measures of Y_{it}, at least two of which occur prior to treatment. Label these times respectively $t-2$, $t-1$, and t, with only t occurring after the treatment. Assuming there are no relevant Xs for the moment, it is easy to show through multiplication and two substitutions that the covariance between Y_{it} and Y_{it-1} is equal to

$$\text{Cov}(Y_{it}, Y_{it-1}) = \text{Cov}(Y_{it-1}, T_i)b_1 + \text{Cov}(Y_{it}, Y_{it-2}), \qquad 28.$$

where T_i is as before a dummy variable treatment indicator and b_1 is the treatment effect. All three of these covariances can be estimated from the data, allowing us to solve out for b_1. If additional time periods are available, the assumption of stationary covariance can be tested. An overall test of the model can be obtained by comparing alternative estimates of the treatment effect using different time-period triplets.

Testing Alternative Models

The point of the above discussion is that traditional methods such as change-score analysis and the analysis of covariance are flawed because they make strong assumptions that are rarely examined and almost never tested. Without

[16] A process is covariance stationary if it has a finite mean and finite variance and the covariance between any two Y over time is only a function of the time elapsed between them.

confidence that their assumptions are valid, resulting estimates of a causal effect have no guaranteed validity.

When only a single pretest measure and a single posttest measure are available, the appropriateness of alternative models cannot be tested. At best one can only make arguments for one specification as opposed to another on theoretical grounds. Hopefully, by embedding these models in Equation 20, we have made it clear to the reader what the nature of these arguments would have to be. With multiple waves of data, however, it becomes possible to determine whether a particular specification is appropriate for the data being analyzed.

One approach would be to use both pre- and posttest data to determine the structure of the unobservables in the data. This is a standard topic in the analysis of panel data. An extensive collection of relevant papers is provided in Maddala (1993). We note that it can often be difficult to determine which among the possible specifications is appropriate because different specifications can produce similar patterns in the data. Also, standard time series methods typically will not work because they assume that the error term consisting of different unobserved components is uncorrelated with any of the observed right-hand-side variables. In this paper, we are interested in situations where the treatment variable may be correlated with unobserved components.

Fortunately, less-sophisticated and more easily applied methods can be used. In order to consistently estimate the treatment effect, it is not necessary that we correctly specify the full structure of the unobservables. Rather, we must only control for those aspects of the unobservables that differ between the treatment and control groups. Heckman & Hotz (1989) discuss an imaginative way of testing this condition that is also simple. One should take all the pretest observations and then analyze them as if they consisted of both pretest and posttest data, testing whether a treatment effect is significant on the pretest observations alone. Because no treatment has yet occurred, no treatment effect should be observed. If a pretreatment effect is found, this is strong evidence that one's model is misspecified. Whatever procedure has been used to control for unobserved differences between the treatment and control group has failed because the significant pretreatment effect indicates that there are still differences between the two groups that have not been accounted for. The posttest data can be used in a similar way. In this case, no treatment effect should be found if the model has been correctly specified, because no additional treatment has occurred. It may, however, be necessary to account for the possibility that the treatment effect dissipates over time. A third possible test is to enter past and future values of the outcome as regressors. If the model is appropriately specified, they should have no effect on the current outcome (Heckman & Hotz 1989).

CONCLUSION

We have tapped only a fraction of the methods and literature that has appeared over the past couple of decades relevant to estimating causal effects. Our intention has been to focus on methods that are relatively accessible and likely to be useful to quantitatively oriented researchers. We hope the reader is impressed by how far the research literature has gone beyond standard regression models.

The appropriateness of alternative models for the estimation of a causal effect depends both on the structure of the data that are available and on the nature of the substantive problem. Given the large number of options, it is critical that researchers, to the degree that it is possible, test for the appropriateness of a chosen specification. Otherwise, a variety of methods should be implemented to determine how robust the treatment effect estimate is to alternative methods.

Besides providing the reader with an introduction to a variety of methods that can be used to estimate causal effects, we hope that we have also presented a conceptual scheme that will be useful to all researchers in trying to think through their own particular analysis problems. In particular, we have shown how a counterfactual interpretation of causality leads to a precise definition of what is meant by a causal effect. Furthermore, this definition points to two important sources of bias in the estimation of treatment effects: (*a*) initial differences between the treatment and control groups in the absence of treatment, and (*b*) the difference between the two groups in the potential effect of the treatment. The latter component is particularly important in situations where there is likely to be selection into the treatment group based on the projected effects of the treatment.

The estimation of causal effects continues to be one of the most active areas of research in both statistics and econometrics. Perhaps one of the most important new developments is the investigation of the quality of estimates that are produced by the different techniques we have discussed. Rubin and Rosenbaum are actively involved in applying matching methods based on the propensity score to different problems. Heckman and his coworkers have been examining matching as well as other methods. It is important to note that they have been extending the methods discussed here to semi-parametric and nonparametric approaches. Their findings (Heckman et al 1997a,b, 1998a,b), using the Job Training Partnership Act data, suggest that at least in some circumstances, the assumption of specific functional forms can be an important source of bias.

ACKNOWLEDGMENT

We thank Gary King, Peter Marsden, William Morgan, Herb Smith, Michael Sobel, Peng He, and Aage Sørensen for helpful comments. The research was supported in part by the National Science Foundation through grant no. SBR

9411875 awarded to Winship and a graduate research fellowship awarded to Morgan.

Literature Cited

Allison PD. 1990. Change scores as dependent variables in regression analysis. In *Sociological Methodology 1990*, ed. CC Clogg, 20:93–114. Washington, DC: Am. Sociol. Assoc.

Angrist JD. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from Social Security administrative records. *Am. Econ. Rev.* 80:313–36

Angrist JD, Imbens GW. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Stat. Assoc.* 90:431–42

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–72

Angrist JD, Krueger AB. 1991. Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* 106:979–1014

Angrist JD, Krueger AB. 1992. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *J. Am. Stat. Assoc.* 87:328–36

Ashenfelter O. 1978. Estimating the effect of training programs on earnings. *Rev. Econ. Stat.* 60:47–57

Ashenfelter O, Card D. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Rev. Econ. Stat.* 67:648–60

Bound J, Jaeger DA, Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90:443–50

Cochran WG, Cox GM. 1950. *Experimental Design*. New York: Wiley. 2nd ed.

Cook TD, Campbell DT. 1979. *Quasi-Experimental: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin

Cox DR. 1958a. *Planning of Experiments*. New York: Wiley

Cox DR. 1958b. The interpretation of the effects of non-additivity in the Latin square. *Biometrika* 45:69–73

Fisher RA. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd

Garfinkel I, Manski CF, Michalopoulos C. 1992. Micro experiments and macro effects. See Manski & Garfinkel 1992, pp. 253–73

Goldberger AS. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard Univ. Press

Hamilton JD. 1994. *Time Series Analysis*. Princeton, NJ: Princeton Univ. Press

Harris RID. 1995. *Using Cointegration Analysis in Econometric Modelling*. New York: Prentice Hall

Harvey A. 1990. *The Econometric Analysis of Time Series*. Cambridge, MA: MIT Press. 2nd ed.

Heckman JJ. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46:931–61

Heckman JJ. 1979. Selection bias as a specification error. *Econometrica* 47:153–61

Heckman JJ. 1989. Causal inference and nonrandom samples. *J. Educ. Stat.* 14:159–68

Heckman JJ. 1992. Randomization and social policy evaluation. See Manski & Garfinkel 1992, pp. 201–30

Heckman JJ. 1995. *Instrumental Variables: A Cautionary Tale*. Cambridge, MA: Natl. Bur. Econ. Res.

Heckman JJ. 1996. Randomization as an instrumental variable. *Rev. Econ. Stat.* 77:336–41

Heckman JJ. 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *J. Hum. Resour.* 32:441–62

Heckman JJ, Hotz VJ. 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *J. Am. Stat. Assoc.* 84:862–80

Heckman JJ, Ichimura H, Smith J, Todd P. 1998a. Characterizing selection bias using experimental data. *Econometrica* 6:1017–99

Heckman JJ, Ichimura H, Todd P. 1997a. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Rev. Econ. Stud.* 64:605–54

Heckman JJ, Ichimura H, Todd P. 1998b. Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* 65:261–94

Heckman JJ, Lochner L, Taber C. 1998c. General-equilibrium treatment effects: a study of tuition policy. *Am. Econ. Rev.* 88(2): 381–92

Heckman JJ, Robb R. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. JJ Heckman, B Singer, pp. 156–245. Cambridge, UK: Cambridge Univ. Press

Heckman JJ, Robb R. 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inferences from Self-Selected Samples*, ed. H Wainer, pp. 63–113. New York: Springer-Verlag

Heckman JJ, Robb R. 1988. The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatment on outcomes. In *Panel Surveys*, ed. G Duncan, G Kalton, pp. 512–38. New York: Wiley

Heckman JJ, Smith J, Clements N. 1997b. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.* 64:487–535

Hendry F. 1995. *Dynamic Econometrics*. New York: Oxford Univ. Press

Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81:945–70

Holland PW, Rubin DB. 1983. On Lord's Paradox. In *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, Ed. H Wainer, S Messick, pp. 3–25. Hillsdale, NJ: Erlbaum

Honore BE, Powell JL. 1994. Pairwise difference estimators of censored and truncated regression models. *J. Econom.* 64:231–78

Hood WC, Koopmans TC, eds. 1953. *Studies in Econometric Method*. New York: Wiley

Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75

Imbens GW, Rubin DB. 1997. Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* 64:555–74

Judd CM, Kenny DA. 1981. *Estimating the Effects of Social Interventions*. New York: Cambridge Univ. Press

Judge G, Hill C, Griffiths W, Lee T. 1985. *The Theory and Practice of Econometrics*. New York: Wiley

Kempthorne O. 1952. *Design and Analysis of Experiments*. New York: Wiley

LaLonde RJ. 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* 76:604–20

Lieberson S. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: Univ. Calif. Press

Maddala GS. 1993. *The Econometrics of Panel Data*, Vols. 1, 2. Hants, UK: Elgar

Malinvaud EB. 1970. *Statistical Methods of Econometrics*. Amsterdam: North-Holland

Manski CF. 1994. The selection problem. In *Advances in Econometrics*, Vol. I, ed. C Sims, pp. 147–70. Cambridge, UK: Cambridge Univ. Press

Manski CF. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard Univ. Press

Manski CF. 1997. Monotone treatment response. *Econometrica* 65:1311–34

Manski CF, Garfinkel I, eds. 1992. *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard Univ. Press

Manski CF, Nagin DS. 1998. Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociol. Methodol.* 28:99–137

Manski CF, Pepper JV. 1998. *Monotone instrumental variables: with an application to the returns to schooling*. Presented at Winter Meet. Am. Sociol. Assoc., Chicago

Manski CF, Sandefur GD, McLanahan S, Powers D. 1992. Alternative estimates of the effect of family structure during adolescence on high school graduation. *J. Am. Stat. Assoc.* 87:25–37

Marcantonio RJ, Cook TD. 1994. Convincing quasi-experiments: the interrupted time series and regression-discontinuity designs. In *Handbook of Practical Program Evaluation*, ed. JS Wholey, HP Hatry, KE Newcomer, pp. 133–54. San Francisco: Jossey-Bass

McKim VR, Turner SP. 1997. *Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. South Bend, IN: Univ. Notre Dame Press

Moffitt RA. 1996. Comment on "Identification of causal effects using instrumental variables" by Angrist, Imbens, and Rubin. *J. Am. Stat. Assoc.* 91:462–65

Neyman JS. 1923. On the application of probability theory to agricultural experiments. Essay on principles. Transl. DM Dabrowska, TP Speed, 1990, in *Stat. Sci*. 5:465–80 (From Polish)

Neyman J. 1935. Statistical problems in agricultural experimentation. *J. R. Stat. Soc.* 2:107–80

Powell JL. 1987. *Semiparametric estimation of bivariate latent variables models. Working pap. no. 8704*. Madison, WI: Univ. WI, Soc. Syst. Res. Inst.

Pratt JW, Schlaifer R. 1984. On the nature and discovery of structure. *J. Am. Stat. Assoc.* 79:9–33

Pratt JW, Schlaifer R. 1988. On the interpretation and observation of laws. *J. Econom.* 39:23–52

Quandt R. 1972. A new approach to estimating switching regression. *J. Am. Stat. Assoc.* 67:306–10

Robins JM. 1986. A new approach to causal

inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math. Model* 7:1393–512
Robins JM. 1987. Addendum to 'A new approach to causal inference in mortality studies with sustained exposure period-application to control of the healthy worker survivor effect.' *Comp. Math. Appl.* 14:923–45
Robins JM. 1989. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, ed. L Sechrest, H Freeman, A Mulley, pp. 113–59. Washington, DC: US Public Health Serv.
Robins JM. 1997. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics* Ed. M Berkane. 120:69–117. New York: Springer-Verlag
Rosenbaum PR. 1984a. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc.* 147:656–66
Rosenbaum PR. 1984b. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. *J. Am. Stat. Assoc.* 79:41–48
Rosenbaum PR. 1995. *Observational Studies*. New York: Springer-Verlag
Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 76:41–55
Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79:516–24
Rosenbaum PR, Rubin DB. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* 39:33–38
Roy AD. 1951. Some thoughts on the distribution of earnings. *Oxford Econ. Pap.* 3:135–46
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701
Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* 2:1–26
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58
Rubin DB. 1980. Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *J. Am. Stat. Assoc.* 75:591–93
Rubin DB. 1981. Estimation in parallel randomized experiments. *J. Educ. Stat.* 6:377–400
Rubin DB. 1986. Which ifs have causal answers? Discussion of "Statistics and causal inference" by Holland. *J. Am. Stat. Assoc.* 83:396
Rubin DB. 1990. Formal modes of statistical inference for causal effects. *J. Stat. Plan. Inference* 25:279–92
Rubin DB. 1991. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 47:1213–34
Rubin DB, Thomas N. 1996. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 52:249–64
Singer B, Marini MM. 1987. Advancing social research: an essay based on Stanley Lieberson's *Making It Count*. In *Sociological Methodology 1987*, ed. CC Clogg, pp. 373–91. Washington, DC: Am. Sociol. Assoc.
Smith HL. 1997. Matching with multiple controls to estimate treatment effects in observational studies. *Sociol. Methodol.* 27:325–53
Sobel ME. 1995. Causal inference in the social and behavioral sciences. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. G Arminger, CC Clogg, ME Sobel, pp. 1–38. New York: Plenum
Sobel ME. 1996. An introduction to causal inference. *Sociol. Methods Res.* 24:353–79
White H. 1981. Consequences and detection of misspecified nonlinear regression models. *J. Am. Stat. Assoc.* 76:419–33
Winship C. 1998. *Multicollinearity and model misspecification: a Bayesian analysis*. Presented at Winter Meet. Am. Sociol. Assoc., Chicago
Winship C, Mare RD. 1992. Models for sample selection bias. *Annu. Rev. Sociol.* 18:327–50

Annual Review of Sociology
Volume 25, 1999

CONTENTS