Manski, C. (1995), Identification Problems in the Social Sciences, Harvard University Press.

2

The Selection Problem

2.1. The Nature of the Problem

Nonresponse is a perennial concern in the collection of survey data. Some persons are not interviewed and some who are interviewed do not answer some of the questions posed. For example, the U.S. Bureau of the Census (1991, pp. 387–388) reported that in the March 1990 administration of its quarterly Current Population Survey (CPS), approximately 4.5 percent of the 60,000 households in the targeted sample were not interviewed. Incomplete income data were obtained from approximately 8 percent of the persons in the households that were interviewed.

Longitudinal surveys experience nonresponse in the form of sample attrition. Piliavin and Sosin (1988) interviewed a sample of 137 individuals who were homeless in Minneapolis in late December 1985. Six months later, the researchers attempted to reinterview these respondents but succeeded in locating only 78.

Survey nonresponse raises the problem of inference from censored data. Social scientists confront censoring in many other ways as well. We routinely ask *treatment-effect* questions of the form:

What is the effect of _____ on _____?

For example,

What is the effect of schooling on wages?
What is the effect of welfare recipiency on labor supply?
What is the effect of sentencing policy on crime commission?

All efforts to infer treatment effects must confront the fact that the data are inherently censored. One wants to compare outcomes across different treatments, but each unit of analysis, whether a survey respondent or an experimental subject, experiences only one of the treatments.

Whereas the implications of censoring were not well appreciated twenty years ago, they are much better understood today. In particular, social scientists have devoted substantial attention to the *selection problem*. This is the problem of identifying conditional probability distributions from random sample data in which the realizations of the conditioning variables are always observed but the realizations of outcomes are censored.

For example, a classic problem in labor economics is to learn how market wages vary with schooling, work experience, and demographic background. Available surveys such as the CPS provide background data for each respondent and wage data for those respondents who work. Even if all subjects respond fully to the questions posed, there remains a censoring problem in that the surveys do not provide market wage data for respondents who do not work. Labor economists confront the selection problem whenever they attempt to use the CPS or similar surveys to estimate wage regressions.

The selection problem is logically separate from the extrapolation problem discussed in Chapter 1. The extrapolation problem examined there arises from the fact that random sampling does not yield observations of y *off* the support of x. The selection problem arises when a censored random sampling process does not fully reveal the behavior of y *on* the support of x. So selection presents new challenges beyond those faced in extrapolation.

To introduce the selection problem formally, we need to expand the description of the population given in Chapter 1 to include a binary variable indicating when outcomes are observed. Let each member of the population be characterized by a triple (y, z, x). As before, y is the outcome to be predicted and the conditioning variables are denoted by x. The new variable z takes the value one if y is observed and the value zero otherwise.

As in Section 1.3, let a random sample be drawn from the population. One does not, however, observe all the realizations of (y, z, x). One observes all realizations of (z, x), but observes y only when $z = 1$.

This censored-sampling process does not identify $P(y \mid x)$ on the support of x. To isolate the difficulty, use the law of total probability to write $P(y \mid x)$ as the sum

$$(2.1) \quad P(y \mid x) = P(y \mid x, z = 1)P(z = 1 \mid x) + P(y \mid x, z = 0)P(z = 0 \mid x).$$

The censored-sampling process identifies the *selection probability* $P(z = 1 \mid x)$, the *censoring probability* $P(z = 0 \mid x)$, and the distribution $P(y \mid x, z = 1)$ of y conditional on selection. These distributions can be estimated in the manner discussed in Section 1.3. But the sampling process is uninformative regarding the distribution $P(y \mid x, z = 0)$ of y conditional on censoring. Hence the censored-sampling process reveals only that

$$(2.2) \quad P(y \mid x) = P(y \mid x, z = 1)P(z = 1 \mid x) + \gamma P(z = 0 \mid x),$$

for some unknown probability distribution γ.

The logical starting point for investigation of the selection problem is to characterize the problem in the absence of prior information about the distribution of (y, z, x); that is, to learn what restrictions on $P(y \mid x)$ are implied by (2.2) alone. Section 2.2 analyzes this *worst-case* scenario and Section 2.3 provides an empirical illustration. Section 2.4 describes the identifying power of various forms of prior information. Sections 2.5 through 2.8 examine in some depth the problem of identifying treatment effects.[1]

2.2. Identification from Censored Samples Alone

What can and cannot be learned about a conditional distribution from censored data alone, in the absence of prior information restricting the form of this distribution or the censoring rule? I first present two negative facts and then develop a set of positive findings.

Two Negative Facts

It is often assumed in empirical studies that the censored outcomes have the same distribution as the observed ones, conditional on x. That is,

$$P(y \mid x, z = 0) = P(y \mid x, z = 1). \tag{2.3}$$

This assumption, variously described as *exogenous* or *ignorable* selection, identifies $P(y \mid x)$ when combined with the empirical evidence.[2] In particular, it implies that $P(y \mid x)$ coincides with the observable distribution $P(y \mid x, z = 1)$.

Suppose a researcher asserts assumption (2.3). Can this assumption be refuted empirically? The answer is negative in the absence of prior information about the form of $P(y \mid x)$. Censored data reveal nothing about $P(y \mid x, z = 0)$, so assumption (2.3) is necessarily consistent with the empirical evidence.

Exogenous selection is an empirically testable hypothesis only if one maintains assumptions restricting the form of $P(y \mid x)$. In that case, setting $\gamma = P(y \mid x, z = 1)$ in equation (2.2) may yield an infeasible value for $P(y \mid x)$. If so, then one can conclude that assumption (2.3) must be incorrect.

The second negative fact is that, in the absence of prior information, censoring makes it impossible to learn anything about the expected value $E(y \mid x)$ of y conditional on x. To see this, write $E(y \mid x)$ as the sum

$$\begin{aligned} E(y \mid x) = {} & E(y \mid x, z = 1)\, P(z = 1 \mid x) \\ & + E(y \mid x, z = 0)\, P(z = 0 \mid x). \end{aligned} \tag{2.4}$$

The censored-sampling process identifies $E(y \mid x, z = 1)$, $P(z = 1 \mid x)$, and $P(z = 0 \mid x)$, but provides no information on $E(y \mid x, z = 0)$, which might take any value between minus and plus infinity. Hence, whenever the censoring probability $P(z = 0 \mid x)$ is positive, the available empirical evidence does not restrict the value of $E(y \mid x)$.

Bounds on Conditional Probabilities

These negative results do not imply that the selection problem is fatal in the absence of prior information. In fact, censored data do imply informative and easily interpretable bounds on important features of $P(y \mid x)$.

Let B denote any set of possible outcomes and consider the probability $P(y \in B \mid x)$ that y falls in the set B. Write this probability as the sum

$$(2.5) \quad P(y \in B \mid x) = P(y \in B \mid x, z = 1)\,P(z = 1 \mid x) + P(y \in B \mid x, z = 0)\,P(z = 0 \mid x).$$

The censored-sampling process identifies $P(y \in B \mid x, z = 1)$, $P(z = 1 \mid x)$, and $P(z = 0 \mid x)$, but provides no information on $P(y \in B \mid x, z = 0)$. The last quantity, however, necessarily lies between zero and one. This yields the following bound on $P(y \in B \mid x)$:

$$(2.6) \quad \begin{aligned} &P(y \in B \mid x, z = 1)\,P(z = 1 \mid x) \\ &\quad \leq P(y \in B \mid x) \\ &\quad \leq P(y \in B \mid x, z = 1)\;P(z = 1 \mid x) + P(z = 0 \mid x). \end{aligned}$$

The lower bound is the value $P(y \in B \mid x)$ takes if the censored values of y never fall in B, while the upper bound is the value $P(y \in B \mid x)$ takes if all the censored y fall in B.

The bound (2.6) is *sharp*. That is, nothing further can be learned about $P(y \in B \mid x)$ from censored data, in the absence of prior information about the distribution of (y, z, x). The width of the bound is the censoring probability $P(z = 0 \mid x)$. So the bound is informative unless y is always censored. Observe that the bound width may vary with x but not with the set B.[3]

It is often convenient to characterize the probability distribution of a real-valued outcome by its distribution function; that is, by the probability that y falls below different cutoff points. Make B the set

of numbers less than or equal to a specified cutoff point t. Then the bound (2.6) becomes[4]

$$(2.7)\quad \begin{aligned} & P(y \leq t \mid x, z = 1)P(z = 1 \mid x) \\ & \quad \leq P(y \leq t \mid x) \\ & \quad \leq P(y \leq t \mid x, z = 1)\ P(z = 1 \mid x) + P(z = 0 \mid x). \end{aligned}$$

It may seem surprising that one should be able to bound the distribution function of y but not its mean. The explanation is a fact central to the field of robust statistics: the mean of a random variable is not a continuous function of its distribution function. Hence small perturbations in a distribution function can generate large movements in the mean. See Huber (1981).[5]

Statistical Inference

The selection problem is a failure of identification. To keep attention focused on identification, I have treated as known the conditional distributions identified by the censored-sampling process. So the bounds have been expressed as functions of $P(y \mid x, z = 1)$ and $P(z \mid x)$. In practice, one would estimate the relevant features of these distributions, thereby obtaining estimates of the bounds. For example, to estimate the bound (2.6) on a conditional probability $P(y \in B \mid x)$, one could estimate $P(y \in B \mid x, z = 1)$ and $P(z = 1 \mid x)$ in the manner described in Section 1.3.

The precision of an estimate of the bound can be measured in the usual way by placing a confidence interval around the estimate. It is important to understand the distinction between the bound and a confidence interval around its estimate. The bound on $P(y \in B \mid x)$ is a population concept, expressing what could be learned about $P(y \in B \mid x)$ if one knew $P(y \in B \mid x, z = 1)$ and $P(z \mid x)$. The confidence interval is a sampling concept, expressing the precision with which the bound is estimated when estimates of $P(y \in B \mid x, z = 1)$ and $P(z \mid x)$ are obtained from a sample of fixed size. The confi-

dence interval is typically wider than the bound but narrows to match the bound as the sample size increases.

2.3. Bounding the Probability of Exiting Homelessness

To illustrate the bound on conditional probabilities, consider the attrition problem that arose in the study of homelessness undertaken by Piliavin and Sosin (1988). These researchers wished to learn the probability that an individual who is homeless at a given date has a home six months later. Thus the population of interest is the set of people who are homeless at the initial date. The outcome variable y is binary, with $y = 1$ if the individual has a home six months later and $y = 0$ if the person remains homeless. The covariates x are individual background attributes. The objective is to learn $P(y = 1 \mid x)$. The censoring problem is that only a subset of the people in the original sample could be located six months later. So $z = 1$ if a respondent was located for reinterview, $z = 0$ otherwise.

Manski (1989) estimates the bound conditioning on various covariates. Suppose first that the only conditioning variable is a respondent's sex. Consider the males. Initial interview data were obtained from 106 men, of whom 64 were located six months later. Of the latter group, 21 had exited from homelessness. So the estimate of $P(y = 1 \mid \text{male}, z = 1)$ is $21/64$ and that of $P(z = 1 \mid \text{male})$ is $64/106$. Hence the estimate of the bound on $P(y = 1 \mid \text{male})$ is $[21/106, 63/106]$ or approximately [.20, .59].

Now consider the females. Data were obtained from 31 women, of whom 14 were located six months later. Of these, 3 had exited from homelessness. So the estimate of $P(y = 1 \mid \text{female}, z = 1)$ is $3/14$ and the estimate of $P(z = 1 \mid \text{female})$ is $14/31$. Hence the estimate of the bound on $P(y = 1 \mid \text{female})$ is $[3/31, 20/31]$, or approximately [.10, .65].

Interpretation of these estimates should be cautious, given the small sample sizes. Taking the results at face value, we have a tighter bound on $P(y = 1 \mid \text{male})$ than on $P(y = 1 \mid \text{female})$ because the attrition rate for men is lower than that for women. The attrition rates $P(z = 0 \mid x)$, hence estimated bound widths, are .39 for men and .55 for women. The important point is that both bounds are informative. Having imposed no restrictions on the attrition process, we are never-

theless able to place meaningful bounds on the probability that a person who is homeless on a given date is no longer homeless six months later.

The foregoing illustrates estimation of the bound when the conditioning variable is discrete. To provide an example with a continuous conditioning variable, let x = (sex, income). The income variable is the respondent's response, expressed in dollars per week, to the question "What was the best job you ever had? How much did that job pay?"

Usable responses to the income question were obtained from 89 men and 22 women. The sample of women is too small to allow meaningful nonparametric regression analysis, so I shall restrict attention to the men. To keep the analysis simple, I ignore the additional censoring problem implied by the fact that 17 of the 106 men did not respond to the income question.

Figure 2.1 shows a local frequency estimate of the attrition probability $P(z = 0 \mid x)$, computed at the actual income values appearing in the sample. Observe that the estimated attrition probability increases smoothly over the income range where the data are concentrated but seems to turn downward in the high income range where the data are sparse.[6] Figure 2.2 graphs the estimated bound on $P(y = 1 \mid x)$. The lower bound is the estimate of $P(y = 1 \mid x, z = 1)P(z = 1 \mid x)$, which is flat on the income range where the data are concentrated but seems to turn downward eventually. The upper bound is the sum of the estimates for $P(y = 1 \mid x, z = 1)P(z = 1 \mid x)$ and for $P(z = 0 \mid x)$.

Observe that the estimated bound is tightest at the low end of the income domain and spreads as income increases. The interval is [.24, .55] at income \$50 and [.23, .66] at income \$600. This spreading reflects the fact, shown in Figure 2.1, that the estimated probability of attrition increases with income.

Is the Cup Part Empty or Part Full?

Consider the bound estimate for the probability that a male exits homelessness: that is, [.20, .59]. Even ignoring sampling variability in the estimate, this seems a modest finding. After all, $P(y = 1 \mid \text{male})$

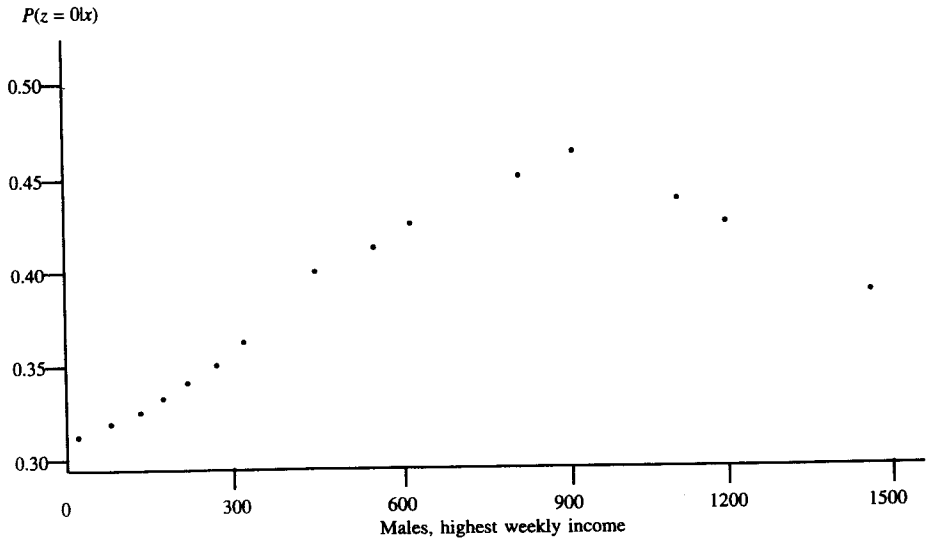

Figure 2.1 Attrition probabilities. (Source: Manski, 1989, fig. 1. Reprinted by permission of the University of Wisconsin Press.)

could take any value in an interval of width .39. Surely we would like to pin the value down more tightly than that. One might be tempted to use the midpoint of the interval [.20, .59] as a point estimate of $P(y = 1 \mid \text{male})$ but, in the absence of prior information, there is no justification for doing so.

The bound appears more useful when one focuses on the fact that it establishes a domain of consensus about the value of $P(y = 1 \mid \text{male})$. Researchers making different assumptions about the attrition process may logically reach different conclusions about the position of the exit probability within the interval [.20, .59]. But all researchers who accept the Piliavin and Sosin data as a censored random sample must agree that, abstracting from sampling error, the exit probability is neither less than .20 nor greater than .59. It is valuable to be able to narrow the region of potential dispute from the interval [0, 1] to the interval [.20, .59].

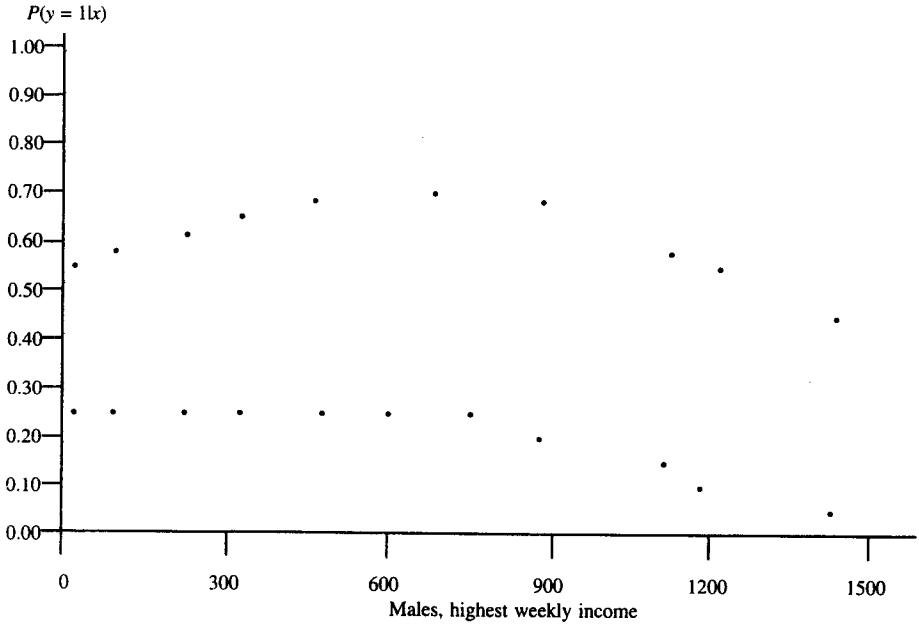

Figure 2.2 Bound on exit probabilities. (Source: Manski, 1989, fig. 2. Reprinted by permission of the University of Wisconsin Press.)

I believe that the historical fixation of social scientists on point identification has inhibited appreciation of the usefulness of bounds. Estimable bounds on quantities that are not identified have been reported from time to time. For example, a bound developed by Frechet (1951) will be exploited in Chapter 3. But the conventional wisdom has been that bounds are uninformative.

It is relevant to note that forty years ago, in a study of the statistical problems of the Kinsey report on sexual behavior, Cochran, Mosteller, and Tukey (1954, pp. 274–282) used bounds of the form (2.6) to express the possible effects of nonresponse to the Kinsey survey. Unfortunately, the subsequent literature did not pursue the idea. In fact, Cochran (1977) dismissed bounding the effects of survey nonresponse. Using the symbol W_2 to denote the censoring probability, he stated (p. 362): "The limits are distressingly wide unless W_2 is very

small." Cochran appears not to have recognized the value of worst-case bounds in establishing a domain of consensus among researchers.

2.4. Prior Distributional Information

Estimation of worst-case bounds should be the starting point for empirical analysis but ordinarily will not be the ending point. Having determined what can be learned in the absence of prior information about the distribution of (y, z, x), one should then ask what more can be learned if plausible assumptions are imposed.

Ideally, we would like to learn the identifying power of all distributional restrictions, so as to characterize the entire spectrum of inferential possibilities. But there does not appear to be any effective way to conduct an exhaustive identification analysis. I shall therefore focus on forms of prior information that are often asserted in empirical studies and that have identifying power.

It is important to understand that some kinds of prior information have no identifying power. The distribution $P(y \mid x, z = 1)$ of observed outcomes and the distribution $P(z \mid x)$ of selected and censored cases are identified by the censored-sampling process. Assumptions restricting these distributions are thus superfluous from the perspective of identification. Such assumptions may be useful from the perspective of statistical inference, as they may enable more precise estimation of $P(y \mid x, z = 1)$ and $P(z \mid x)$. But that is not our concern here.

Exogenous Selection and Self-Selection

Each year, the U.S. Bureau of the Census publishes statistics on the distribution of income in its report *Money Income of Households, Families, and Persons in the United States*. The raw data for this report are drawn from the March CPS. As noted at the beginning of the chapter, some households in the sampling frame are not interviewed and incomplete income data are obtained from some of the persons in the interviewed households. The Census Bureau copes with this nonresponse problem by assuming that, conditional on whatever variables x are observed for a given person, nonreported incomes have the same distribution as reported incomes. That is, the Census Bureau imposes the exogenous selection assumption given in equation (2.3).

Until the early 1970s, social scientists almost universally assumed exogenous selection. The perspective of many changed dramatically after economists began a sustained effort to understand the role of self-selection in determining when behavioral outcomes are observed. The result was that the assumption of exogenous selection diminished sharply in credibility.

The work of labor economists studying market wage determination has been particularly influential. Consider the process determining whether wage data are available in surveys such as the CPS. Labor economists reasoned as follows (see Gronau, 1974):

(1) Wage data are available only for respondents who work.
(2) Respondents work when they choose to do so.
(3) The wage one would be paid influences the decision to work.
(4) Hence the distributions of observed and unobserved wages may differ.

A simple model often used to express this reasoning supposes that each individual knows the wage y that would be paid if he or she were to work. The individual chooses to work if y is greater than some lowest acceptable wage R, called the person's *reservation wage,* and chooses not to work if y is below the reservation wage. So wages are observed when $y > R$ and are censored when $y < R$.

The reservation-wage model does not predict whether a person works if $y = R$, but it is conventionally assumed that this event occurs with probability zero in the population. Hence the borderline case may be ignored. With this caveat, the reservation-wage model implies that

$$P(y \mid x, z = 1) = P(y \mid x, y > R), \tag{2.8}$$

$$P(y \mid x, z = 0) = P(y \mid x, y < R), \tag{2.9}$$

and

$$P(z = 1 \mid x) = P(y > R \mid x). \tag{2.10}$$

If the reservation-wage model accurately describes the decision to work, then it is correct to assume exogenous selection if and only if

$$P(y \mid x, y > R) = P(y \mid x, y < R), \tag{2.11}$$

or equivalently, if

$$P(y \mid x, r > 0) = P(y \mid x, r < 0), \tag{2.12}$$

where $r \equiv y - R$ denotes the difference between market and reservation wages. Condition (2.12) holds if persons with high market wages also tend to have high reservation wages. In particular, it holds if the difference r between market and reservation wages is statistically independent of the market wage y. But the condition generally does not hold otherwise.

For example, suppose all persons with attributes x have the same reservation wage $R(x)$. This *fixed-threshold* assumption implies that censored wages tend to be lower than observed wages, conditional on x. In particular,

$$P[y \leq t \mid x, y < R(x)] \geq P[y \leq t \mid x, y > R(x)], \tag{2.13}$$

for any cutoff point t.

The reservation-wage model of labor supply provides a good example of observability determined by self-selection. There are many others that could equally well be cited. As another example, suppose one wishes to predict the outcomes a high school graduate would experience if he or she were to enroll in college (see Willis and Rosen, 1979). The outcomes of college enrollment are observable only for those high school graduates who actually enroll. The persons who enroll presumably are those who anticipate that college will have favorable outcomes for them, relative to nonenrollment. If anticipated outcomes are related to realized ones, then the observable distribution of outcomes experienced by those who actually enroll may differ from the censored distribution of outcomes that nonenrollees would have experienced had they enrolled.

The message is that self-selection may make the observability of a behavioral outcome depend partly on the value of the outcome. Self-selection does not imply that the assumption of exogenous selection is necessarily wrong, but it does cast much doubt on the assumption. Anyone concerned with prediction of y conditional on x must take notice. Using the observed distribution of outcomes to predict y conditional on x leads one astray to the extent that $P(y \mid x, z = 1)$ differs from $P(y \mid x)$.

Latent-Variable Models

It is easier to show why selection may not be exogenous than to find a credible alternative assumption that identifies $P(y \mid x)$. To illustrate, let us examine further the problem of predicting market wages. Labor economists have widely used the reservation-wage model to explain labor supply and, hence, the observability of market wages. Suppose that the reservation-wage model is correct. What is its identifying power?

The startling answer is that the reservation-wage model has no identifying power at all. The model states in equation (2.9) that the distribution of censored wages has the form $P(y \mid x, y < R)$. But the model imposes no restrictions on the form of $P(y \mid x, y < R)$. So the model implies no restrictions on the distribution of censored wages.

Facing this fact, labor economists could decide to limit themselves to estimating worst-case bounds on the distribution of market wages. Instead, the prevailing practice has been to augment the reservation-wage model by imposing assumptions on the distribution $P(y, R \mid x)$ of market and reservation wages. Because reservation wages are never observed and market wages are only sometimes observed, the reservation-wage model accompanied by restrictions on $P(y, R \mid x)$ is often referred to as a *latent-variable model.*

The identifying power of a latent-variable model depends on the assumptions imposed on $P(y, R \mid x)$. The most common practice has been to restrict $P(y, R \mid x)$ to a family of distributions specified up to some parameters. One specification, the *normal-linear model,* has received by far the most attention (see Heckman, 1976, and Maddala, 1983).

The normal-linear model assumes that, conditional on x, the distribution of $(\log y, \log R)$ is bivariate normal with mean $(x\beta_1, x\beta_2)$ and variance matrix Σ. This assumption reduces the problem of identifying $P(y \mid x)$ to that of identifying the unknown parameters $(\beta_1, \beta_2, \Sigma)$. These parameters are identified if, under the maintained log-normality assumption, there is exactly one value of $(\beta_1, \beta_2, \Sigma)$ that implies the observable distributions $P(y \mid x, z = 1)$ and $P(z \mid x)$.[7]

Some studies make the fixed-threshold assumption that all persons with attributes x have the same reservation wage $R(x)$. The magnitude of $R(x)$ need not be known a priori as it is identified by the sampling process. The reservation-wage model states that y is observed if and only if $y > R(x)$. So $R(x)$ is the lower bound of the support of the distribution $P(y \mid x, z = 1)$ of observed market wages.

The fixed-threshold assumption implies that $P(y \leq t \mid x)$ is identified at all cutoff points t greater than $R(x)$. To see this, observe that when $t > R(x)$,

$$(2.14) \quad P(y \leq t \mid x, z = 0) = P[y \leq t \mid x, y < R(x)] = 1.$$

It follows that

$$\begin{aligned} (2.15) \quad P(y \leq t \mid x) &= P(y \leq t \mid x, z = 1)P(z = 1 \mid x) \\ &\quad + P(y \leq t \mid x, z = 0)P(z = 0 \mid x) \\ &= P(y \leq t \mid x, z = 1)P(z = 1 \mid x) \\ &\quad + P(z = 0 \mid x). \end{aligned}$$

The quantities on the right side of (2.15) are all identified.

Latent-variable models with the same formal structure as the reservation-wage model have been applied to infer conditional distributions in many settings where self-selection is thought to determine the observability of outcomes. For example, economists studying the income returns to college enrollment often assume that a person enrolls in college if the income y the person would receive following enrollment exceeds the opportunity cost R of enrollment.

In the mid-1970s, the development of parametric latent-variable models of self-selection was greeted with widespread enthusiasm

among economists. The initial enthusiasm was partly based on a misconception that specification of a latent-variable model offers a universally applicable "solution" to the selection problem. Initially, it was not appreciated that solutions to the selection problem are only as good as the assumptions imposed.

Enthusiasm waned in the early 1980s after methodological studies showed that seemingly minor changes in the distributional assumptions placed on reservation wages and similar latent variables could generate large changes in the implied value of $P(y \mid x)$. See Hurd (1979), Arabmazar and Schmidt (1982), and Goldberger (1983). Empirical researchers typically were unable to provide solid arguments substantiating their distributional assumptions. Hence they were forced to face up to the fragility of their inferences.

Once aware of the limitations of latent-variable models, economists reacted in at least three distinct ways. Some have continued to use the models, but with a more skeptical attitude than previously. Some have returned to the earlier convention of assuming exogenous selection. Some, asserting that no useful inference is possible in the presence of censoring, have argued that controlled experimentation provides the only trustworthy basis for empirical research. The experimental movement will be discussed in Chapter 3.

Exclusion Restrictions

Whether they have assumed exogenous selection or used latent-variable models, researchers have generally sought to impose prior information that fully identifies the conditional distribution $P(y \mid x)$. There is a vast, mostly unexplored middle ground between the worst-case analysis of Section 2.2 and the strong assumptions considered earlier in this section. This middle ground contains assumptions having some identifying power, but not enough to identify $P(y \mid x)$ fully.

One part of the middle ground that is now understood concerns the identifying power of *exclusion restrictions*. Social scientists often assume that some component of the regressor vector x does not affect the distribution of y, but does affect whether y is observed. For example, a labor economist studying wage determination might assume that a person's nonlabor income (for example, stock dividends and

interest on savings) influences the decision to work but does not affect the wage a person would be offered by a firm.

This idea may be formalized by decomposing the conditioning variables x into two components w and v; so $x = (w, v)$. The assumption is that, holding w fixed, the outcome distribution $P(y \mid w, v)$ does not vary with v, while the selection probability $P(z = 1 \mid w, v)$ does vary with v. Thus v is excluded from the determination of y, conditional on w. The excluded component v is sometimes referred to as an *instrumental variable*.

It turns out to be easy to characterize the identifying power of an exclusion restriction. The simple result (see Manski, 1990a, 1994a) is that an exclusion restriction allows one to replace the bounds available in the absence of prior information with the intersection of these bounds across all values of v.

For example, consider the probability $P(y \in B \mid w, v)$ that y falls in some set B, conditional on (w, v). The worst-case bound on $P(y \in B \mid w, v)$ was given in equation (2.6). Now assume that, holding w fixed, $P(y \in B \mid w, v)$ does not vary with v; so $P(y \in B \mid w, v) = P(y \in B \mid w)$ for all values of v. Then $P(y \in B \mid w)$ must lie within the intersection of the bounds (2.6) on $P(y \in B \mid w, v)$ across all values of v. Thus if v can take the two values v_1 and v_2, we have

$$
\begin{aligned}
(2.16)\quad & \max_{j=1,2} P[y \in B \mid (w, v_j), z = 1]P(z = 1 \mid w, v_j) \\
& \leq P(y \in B \mid w) \\
& \leq \min_{j=1,2} P[y \in B \mid (w, v_j), z = 1]P(z = 1 \mid w, v_j) \\
& \quad + P(z = 0 \mid w, v_j).
\end{aligned}
$$

The new bound (2.16) typically improves on (2.6), but not enough to identify $P(y \in B \mid x)$.[8]

2.5. Identification of Treatment Effects

I noted earlier that we routinely ask questions of the form: What is the effect of _____ on _____? These questions aim to compare the distributions of outcomes that would be realized if alternative

treatments were applied to a population. A central problem of empirical research is to learn these distributions when some members of the population are observed to receive one treatment and the rest are observed to receive another. The remainder of the chapter examines this inferential problem.

Suppose, for example, that patients ill with a specified disease might be treated by drugs or by surgery. The relevant outcome might be life span. Then we may wish to determine the distribution of life spans that would be realized if patients with specified risk factors were all treated by drugs. And we may wish to compare this with the distribution of life spans that would be realized if these same patients were instead treated by surgery. The problem is to infer these outcome distributions from observations of the life spans of patients some of whom actually were treated by drugs and the rest by surgery.

Or, in the realm of economic policy, suppose that workers displaced by a plant closing might be retrained or given assistance in searching for a new job. The outcome of interest might be earned income. Then we may wish to determine the distribution of income that would be realized if all workers with specified backgrounds were retrained, and compare this with the distribution that would be realized if these same workers were instead assisted in job search. The problem is to infer these distributions from observations of the incomes earned by workers some of whom were retrained and some of whom were given job search assistance.

Switching Processes

To formalize the problem, let the treatments being compared be labeled 1 and 0, and let the associated outcomes be y_1 and y_0. Thus y_1 is the outcome that a person would realize if he or she were to receive treatment 1, and y_0 is the outcome that would be realized if the person were to receive treatment 0. Let $P(y_1 \mid x)$ denote the distribution of outcomes that would be realized if all persons with covariates x were to receive treatment 1, and let $P(y_0 \mid x)$ denote the analogous distribution of outcomes under treatment 0. Then the objective is to compare the distributions $P(y_1 \mid x)$ and $P(y_0 \mid x)$.[9]

We suppose that each member of the population actually receives one or the other of the two treatments. Let the binary variable z indicate which treatment a person receives. A random sample is drawn and, for each sampled person, one observes the covariate value x, the treatment z received, and the outcome under that treatment. One thus observes y_1 when $z = 1$ and y_0 when $z = 0$.

This *switching process* identifies the treatment distribution $P(z \mid x)$, the distribution $P(y_1 \mid x, z = 1)$ of y_1 conditional on receiving treatment 1, and the distribution $P(y_0 \mid x, z = 0)$ of y_0 conditional on receiving treatment 0. Thus the inferential question formalizes as:

> What does knowledge of $P(z \mid x)$, $P(y_1 \mid x, z = 1)$ and $P(y_0 \mid x, z = 0)$ reveal about $P(y_1 \mid x)$ and $P(y_0 \mid x)$?

The situations of $P(y_1 \mid x)$ and $P(y_0 \mid x)$ are symmetric, so it is enough to consider the problem of inference on $P(y_1 \mid x)$. A switching process yields richer data than did the censored-sampling process examined in Sections 2.1 through 2.4. Whereas we earlier assumed that no outcome is observed when $z = 0$, a switching process yields an observation of y_0. Whereas censored sampling identifies $P(z \mid x)$ and $P(y_1 \mid x, z = 1)$, a switching process also identifies $P(y_0 \mid x, z = 0)$. Thus a switching process differs from censored sampling to the extent that knowledge of $P(y_0 \mid x, z = 0)$ reveals something about the censored distribution $P(y_1 \mid x, z = 0)$.

Section 2.6 examines several situations in which knowing $P(y_0 \mid x, z = 0)$ does reveal something about $P(y_1 \mid x, z = 0)$. Before that, however, there are important things to say about the comparison of treatments in situations where switching and censoring are equivalent.

Comparing Treatments with No Prior Information: The Cup Is Half Full

Suppose one has no prior information about the distribution of (y_1, y_0, z, x). Then knowing $P(y_0 \mid x, z = 0)$ reveals nothing about

$P(y_1 \mid x, z = 0)$. So a switching process reveals no more about $P(y_1 \mid x)$ than does censored sampling, and the findings of Section 2.2 continue to hold. The same conclusion applies to inference on $P(y_0 \mid x)$, except that selection and censoring are reversed.

Suppose a researcher asserts that treatment has no effect on outcomes. That is, the researcher asserts that

$$P(y_1 \mid x) = P(y_0 \mid x). \tag{2.17}$$

This hypothesis is not refutable in the absence of prior information. To see this, use the law of total probability to write

$$\begin{aligned} P(y_1 \mid x) = {} & P(y_1 \mid x, z = 1)P(z = 1 \mid x) \\ & + P(y_1 \mid x, z = 0)P(z = 0 \mid x) \end{aligned} \tag{2.18a}$$

and

$$\begin{aligned} P(y_0 \mid x) = {} & P(y_0 \mid x, z = 1)P(z = 1 \mid x) \\ & + P(y_0 \mid x, z = 0)P(z = 0 \mid x). \end{aligned} \tag{2.18b}$$

A switching process reveals nothing about $P(y_1 \mid x, z = 0)$ and $P(y_0 \mid x, z = 1)$, so these distributions could equal $P(y_0 \mid x, z = 0)$ and $P(y_1 \mid x, z = 1)$ respectively. In that event, $P(y_1 \mid x)$ and $P(y_0 \mid x)$ are the same.

Although the hypothesis (2.17) is not testable, useful inferences can still be made about the relationship between $P(y_1 \mid x)$ and $P(y_0 \mid x)$. It is common to compare treatments by the difference in the probability that the outcome falls in a specified set, namely,

$$T(B \mid x) \equiv P(y_1 \in B \mid x) - P(y_0 \in B \mid x), \tag{2.19}$$

where B is the specified set of outcome values. When no prior information is available, sharp bounds on $T(B \mid x)$ can be obtained directly

from the bounds on $P(y_1 \in B \mid x)$ and $P(y_0 \in B \mid x)$ given in equation (2.6). The lower bound in $T(B \mid x)$ is the difference between the lower bound on $P(y_1 \in B \mid x)$ and the upper bound on $P(y_0 = 1 \mid x)$. The upper bound on $T(B \mid x)$ is determined similarly. Hence we find that

$$
\begin{aligned}
(2.20)\quad & P(y_1 \in B \mid x, z = 1)P(z = 1 \mid x) \\
& \quad - P(y_0 \in B \mid x, z = 0)P(z = 0 \mid x) - P(z = 1 \mid x) \\
& \leq T(B \mid x) \\
& \leq P(y_1 \in B \mid x, z = 1)P(z = 1 \mid x) + P(z = 0 \mid x) \\
& \quad - P(y_0 \in B \mid x, z = 0)P(z = 0 \mid x).
\end{aligned}
$$

Observe that the bound (2.20) always has width one. This is an important fact with a simple explanation. The width of the bound on $T(B \mid x)$ is the sum of the widths of the bounds on $P(y_1 \in B \mid x)$ and $P(y_0 \in B \mid x)$. The latter widths are $P(z = 0 \mid x)$ and $P(z = 1 \mid x)$, respectively, as the censoring of y_1 coincides with the selection of y_0.

If no data were available, $T(B \mid x)$ could lie anywhere in the interval $[-1, 1]$, an interval of width two. So observing the outcomes of a switching process allows one to confine $T(B \mid x)$ to half of its logically possible range. In this sense, a researcher wishing to compare treatments in the absence of prior information finds that the cup is exactly half full.

Treatment Independent of Outcomes

It is often assumed that the treatment z received by each person with covariates x is statistically independent of the person's outcomes (y_1, y_0). The purpose of experimentation with randomized selection of treatment is to justify this assumption. Treatment independent of outcomes is also commonly assumed in empirical studies using nonexperimental data, where it may be referred to as *exogenous switching*

(Maddala, 1983) or as *strongly ignorable* treatment assignment (Rosenbaum and Rubin, 1983). The assumption implies that

$$(2.21a)\quad P(y_1 \mid x) = P(y_1 \mid x, z = 1) = P(y_1 \mid x, z = 0)$$

and

$$(2.21b)\quad P(y_0 \mid x) = P(y_0 \mid x, z = 1) = P(y_0 \mid x, z = 0).$$

So $P(y_1 \mid x)$ and $P(y_0 \mid x)$ are identified.

When treatment is independent of outcomes, the outcome distributions $P(y_1 \mid x)$ and $P(y_0 \mid x)$ can be expressed in a manner that makes no reference to switching. Let

$$(2.22)\quad y \equiv y_1 z + y_0(1 - z)$$

denote the outcome actually realized by a member of the population, namely, y_1 when $z = 1$ and y_0 otherwise. Observe that

$$(2.23a)\quad P(y \mid x, z = 1) = P(y_1 \mid x, z = 1)$$

and

$$(2.23b)\quad P(y \mid x, z = 0) = P(y_0 \mid x, z = 0).$$

When treatment is independent of outcomes, (2.21) and (2.23) combine to yield

$$(2.24a)\quad P(y_1 \mid x) = P(y \mid x, z = 1)$$

and

$$(2.24b)\quad P(y_0 \mid x) = P(y \mid x, z = 0).$$

Thus inference on $P(y_1 \mid x)$ and $P(y_0 \mid x)$ when the data are generated by a switching process is the same as inference on $P(y \mid x, z = 1)$ and $P(y \mid x, z = 0)$ under random sampling of (y, x, z).

Equation (2.24) can be used to rewrite the probability difference $T(B \mid x)$ of (2.19) as

$$T(B \mid x) = P(y \in B \mid x, z = 1) - P(y \in B \mid x, z = 0). \tag{2.25}$$

Empirical studies often refer to estimates of $P(y \in B \mid x, z = 1) - P(y \in B \mid x, z = 0)$ as estimates of treatment effects. This practice is well founded if treatment is independent of outcomes, but not otherwise. When treatment is not independent of outcomes, equation (2.25) generally does not hold.[10]

2.6. Information Linking Outcomes across Treatments

This section examines three distributional assumptions implying that observation of y_0 is informative about y_1, and vice versa. The situations of y_1 and y_0 are symmetric, so I focus on the problem of inferring $P(y_1 \mid x)$.

Shifted Outcomes with an Exclusion Restriction

Empirical studies sometimes assume that y_1 and y_0 always differ by a constant, so that y_1 is a shifted version of y_0. For example, a long-standing concern of labor economics is to determine the effect of union membership on wages. Let y_1 be the wage that a person would earn if he or she were a union member and let y_0 be the wage that person would earn as a nonmember. Labor economists have often assumed that the *union wage differential* $y_1 - y_0$ is the same for all people.

Formally, assume there exists a constant k such that

$$y_1 = y_0 + k. \tag{2.26}$$

This assumption implies that, for all cutoff points t,

$$P(y_1 \leq t \mid x, z = 0) = P(y_0 \leq t - k \mid x, z = 0). \tag{2.27}$$

The distribution $P(y_0 \mid x, z = 0)$ is identified by the switching process. So $P(y_1 \mid x, z = 0)$ is identified if the value of the shift parameter k can be determined. It can be shown that k is identified if $P(y_1 \mid x)$ satisfies an exclusion restriction of the type discussed in Section 2.4.[11] Hence $P(y_1 \mid x)$ is identified if a switching process generates the data, outcomes are shifted, and an exclusion restriction holds.

This is a powerful finding, identifying $P(y_1 \mid x)$ without imposing any assumptions on the switching rule determining whether y_1 or y_0 is observed. The result is achieved at high cost, however. An exclusion restriction may or may not be available in a given application, but the assumption of shifted outcomes strains credibility.

Consider, for example, the union wage differential. Is it plausible to assume that union membership offers the same wage increment for all workers? Union contracts are often thought to tie wages and job security more closely to seniority than to merit. It therefore seems likely that, within a given job category, the less productive workers experience a larger union wage differential than do the more productive ones.

Ordered Outcomes

Similar in spirit to shifted outcomes is the assumption that y_1 and y_0 are ordered with, say, the value of y_1 always at least as large as the value of y_0. For example, suppose that an ill person may be treated by drug therapy ($z = 1$) or by placebo ($z = 0$). Let the outcomes y_1 and y_0 be the life span following each treatment. One might not know the value of drug therapy but might feel confident that it can do no harm. If so, then one can assume that y_1 must be at least as large as y_0.

Formally, the assumption is

$$y_1 \geq y_0. \tag{2.28}$$

This assumption implies that, for all cutoff points t,

$$P(y_1 \leq t \mid x, z = 0) \leq P(y_0 \leq t \mid x, z = 0). \tag{2.29}$$

Hence

$$
\begin{aligned}
(2.30)\quad P(y_1 \leq t \mid x) &= P(y_1 \leq t \mid x, z = 1)P(z = 1 \mid x) \\
&\quad + P(y_1 \leq t \mid x, z = 0)P(z = 0 \mid x) \\
&\leq P(y_1 \leq t \mid x, z = 1)P(z = 1 \mid x) \\
&\quad + P(y_0 \leq t \mid x, z = 0)P(z = 0 \mid x) \\
&= P(y \leq t \mid x).
\end{aligned}
$$

On the one hand, the upper bound on $P(y_1 \leq t \mid x)$ given here improves on the one in equation (2.7), which is the best available under censored sampling. On the other hand, the ordered-outcomes assumption does not tighten the lower bound in (2.7).

Selection of the Treatment with the Larger/Smaller Outcome

The shifted-outcomes and ordered-outcomes assumptions link the outcomes (y_1, y_0) by restricting the form of their joint distribution $P(y_1, y_0 \mid x)$. These outcomes may also be linked by assumptions on the switching rule determining whether y_1 or y_0 is realized. I shall discuss two symmetric cases.

Economic analyses of voluntary treatment policies often assume that the treatment yielding the larger outcome is selected. An example is the Roy (1951) model of occupation choice, where a person selects between two occupations by choosing the occupation with the higher wage. Let $z = 1$ if one occupation is chosen, $z = 0$ if the other, and let y_1 and y_0 be the wages that would be earned in the two occupations. Then the Roy model asserts that

$$(2.31)\quad z = 1 \text{ if } y_1 > y_0 \quad \text{and} \quad z = 0 \text{ if } y_1 < y_0.$$

No prediction is made if $y_1 = y_0$, but it is conventionally assumed that this event occurs with probability zero in the population. So the borderline case may be ignored.

To a researcher studying occupational wage determination, the Roy model is prior information specifying the switching rule de-

termining whether y_1 or y_0 is observed. The situation is the same as in the reservation-wage model except that there, $z = 0$ implied that no outcome was observed, while here the outcome y_0 is observed.

Observability of y_0 makes a crucial difference. Whereas the reservation-wage model was earlier shown to have no identifying power in the absence of distributional assumptions (see Section 2.4), the Roy model does have identifying power. In particular, equation (2.31) implies that for all cutoff points t,

$$(2.32) \quad P(y_1 \leq t \mid x, z = 0) \geq P(y_0 \leq t \mid x, z = 0).$$

It follows that

$$\begin{aligned}(2.33) \quad P(y_1 \leq t \mid x) &= P(y_1 \leq t \mid x, z = 1)P(z = 1 \mid x) \\ &\quad + P(y_1 \leq t \mid x, z = 0)P(z = 0 \mid x) \\ &\geq P(y_1 \leq t \mid x, z = 1)P(z = 1 \mid x) \\ &\quad + P(y_0 \leq t \mid x, z = 0)P(z = 0 \mid x) \\ &= P(y \leq t \mid x).\end{aligned}$$

This lower bound on $P(y_1 \leq t \mid x)$ improves on the one in (2.7), which is the best available under censored sampling. The upper bound on $P(y_1 \leq t \mid x)$, however, remains as in (2.7).

The *competing-risks* model of survival analysis assumes that the treatment with the smaller outcome is selected, rather than the larger (see Kalbfleisch and Prentice, 1980). A person with two terminal diseases, for example, dies of the disease that first manifests itself.[12] Let $z = 1$ if one disease causes death, $z = 0$ if the other, and let y_1 and y_0 be the span of time before each disease manifests itself. Then the competing-risks model asserts that

$$(2.34) \quad z = 1 \text{ if } y_1 < y_0 \quad \text{and} \quad z = 0 \text{ if } y_1 > y_0.$$

It follows from (2.34) that

$$(2.35) \quad P(y_1 \leq t \mid x, z = 0) \leq P(y_0 \leq t \mid x, z = 0)$$

or all cutoff points t. This is the same finding as was reported in equation (2.29) under the ordered-outcomes assumption. Again (2.30) gives the upper bound on $P(y_1 \leq t \mid x)$.

2.7. Predicting High School Graduation If All Families Were Intact

To illustrate inference on treatment effects, I shall carry further the analysis of high school graduation begun in Section 1.5. The discussion there concluded by observing that the estimated graduation probabilities presented in Table 1.1 do not suffice to predict outcomes if the American environment of the 1980s were to change in some way. Among other things, these estimates do not permit one to predict what would happen if the fraction of intact families in the population were increased.

The study by Manski et al. (1992) goes beyond the estimates in Table 1.1 and addresses the extreme version of this question, namely: What would high school graduation probabilities be if all families were intact? The study also compares this scenario with the opposite extreme in which all families are nonintact.

We imagine that each child is characterized by two hypothetical high school graduation outcomes, y_1 and y_0. Variable y_1 indicates the outcome if the child were to reside in an intact family, with $y_1 = 1$ if the child would graduate and $y_1 = 0$ otherwise. Analogously, y_0 indicates the outcome if the same child were to reside in a nonintact family. We wish to learn the probability $P(y_1 = 1 \mid x)$ that a child with covariates x would graduate if all such children were to reside in intact families, and to compare $P(y_1 = 1 \mid x)$ with the probability $P(y_0 = 1 \mid x)$ of graduation if all children with covariates x were to reside in nonintact families. The covariates x are race, sex, and parents' schooling. In Section 1.5 family structure was also a covariate, but this variable now appears as a treatment instead.

The inferential problem stems from the fact that each child in the NLSY sample actually realizes only one of the two graduation outcomes; y_1 is realized if a child actually resides in an intact family ($z = 1$) and y_0 is realized otherwise ($z = 0$). The empirical evidence therefore reveals the probability $P(z = 1 \mid x)$ that a child with covari-

Table 2.1 Estimated probabilities of residence in an intact family, for children whose parents have completed twelve years of schooling

White male	White female	Black male	Black female
.82	.82	.54	.46

Source: Computations based on Manski et al. (1992), table 4.

ates x resides in an intact family, the probability $P(y_1 = 1 \mid x, z = 1)$ of graduation conditional on residing in an intact family, and the probability $P(y_0 = 1 \mid x, z = 0)$ of graduation conditional on residing in a nonintact family.

As in Section 1.5, I focus here on children whose parents have both completed twelve years of schooling. Estimates of $P(y_1 = 1 \mid x, z = 1)$ and $P(y_0 = 1 \mid x, z = 0)$ have already been presented in Table 1.1. Estimates of $P(z = 1 \mid x)$ are given in Table 2.1. The striking feature of the table is the difference across races. Holding sex and parents' schooling fixed, we find that a white child is much more likely than a black child to reside in an intact family at age 14.

The data in Tables 1.1 and 2.1 provide all the components needed to infer the graduation probabilities $P(y_1 = 1 \mid x)$ and $P(y_0 = 1 \mid x)$ under the various assumptions considered in Sections 2.5 and 2.6. Table 2.2 presents the estimates for $P(y_1 = 1 \mid x)$. Similar estimates may be computed for $P(y_0 = 1 \mid x)$.

Table 2.2 Estimated bounds on high school graduation probabilities if all families were intact, for children whose parents have completed twelve years of schooling

Prior information	White male	White female	Black male	Black femal
No prior information (equation 2.6)	[.73, .91]	[.77, .95]	[.47, .93]	[.44, .98]
Ordered outcomes (equation 2.30)	[.87, .91]	[.91, .95]	[.83, .93]	[.91, .98]
Treatment with the larger outcome (equation 2.33)	[.73, .87]	[.77, .91]	[.47, .83]	[.44, .91]
Treatment independent of outcomes (equation 2.24)	.89	.94	.87	.95

Source: Computations based on Tables 1.1 and 2.1.

Table 2.2 makes clear how prior information affects the conclusions one may draw about graduation outcomes if all families were intact. In the absence of prior information, one may estimate the bounds in the top row of the table. The width of each bound is the estimated probability of residing in a nonintact family.

Prior information may allow one to tighten the worst-case bounds. Suppose one believes that residing in an intact family can never harm a child's schooling prospects; so outcomes are ordered with $y_1 \geq y_0$ for all children. Then mandating that all families be intact cannot decrease graduation probabilities relative to those actually found among the NLSY respondents. The actual graduation probabilities, namely,

$$(2.36)\quad P(y = 1 \mid x) = P(y_1 = 1 \mid x, z = 1)\,P(z = 1 \mid x) + P(y_0 = 1 \mid x, z = 0)\,P(z = 0 \mid x),$$

are estimated to be .87 for white males, .91 for white females, .83 for black males, and .91 for black females. These values become lower bounds under the ordered-outcomes assumption.

Suppose one believes that realized family structure reflects parents' decisions about what is best for their children's schooling prospects—in other words, that families choose the treatment with the larger outcome. (For example, parents may compare the likely impact on their children of maintaining a marriage characterized by constant fighting and hostility with the impact of raising the children in a single-parent household.) If family structure is determined in this manner, then mandating that all families be intact cannot increase graduation probabilities relative to those actually found among the NLSY respondents, and may decrease them. So the actual graduation probabilities now become upper bounds on $P(y_1 = 1 \mid x)$ rather than lower bounds.

Finally, suppose one believes that family structure is exogenous with respect to children's schooling prospects. Then the graduation probabilities $P(y_1 = 1 \mid x)$ coincide with the probabilities $P(y_1 = 1 \mid x, z = 1)$ reported earlier in the left column of Table 1.1, and now found in the fourth row of Table 2.2.

Each of the rows of Table 2.2 presents a logically valid conclusion

about the graduation outcomes that would be realized if all families were intact. Any social scientist who accepts the NLSY as empirical evidence must agree that the probabilities $P(y_1 = 1 \mid x)$ lie within the worst-case bounds of the first row. Beyond that, the conclusions one draws necessarily depend on the assumptions one is willing to maintain.