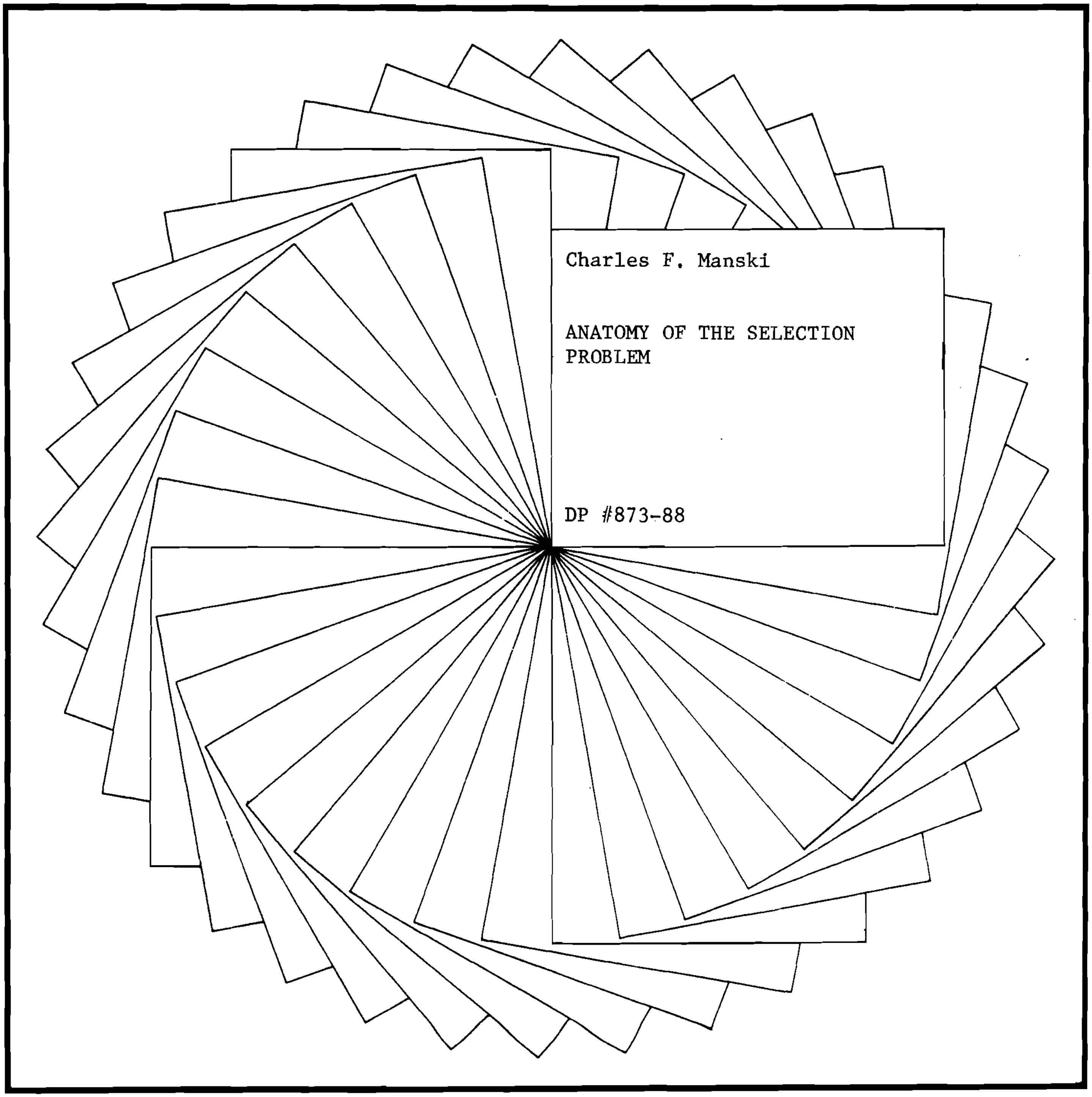
University of Wisconsin-Madison

Institute for Research on Poverty

Discussion Papers



Charles F. Manski

ANATOMY OF THE SELECTION PROBLEM

DP #873-88

Institute for Research on Poverty
Discussion Paper no. 873-88

ANATOMY OF THE SELECTION PROBLEM

Charles F. Manski

Department of Economics
and
Institute for Research on Poverty
University of Wisconsin-Madison

December 1988

This research is supported by National Science Foundation Grant SES-8808276 and by funding to the Institute for Research on Poverty from the U.S. Department of Health and Human Services. I am grateful to Irving Piliavin for posing the empirical issue that stimulated me to take a fresh look at the selection problem. I have benefited from discussions with Arthur Goldberger, Peter Gottschalk, and Robert Moffitt. I have also benefited from the opportunity to present this work in a conference at Wingspread and in seminars at Brown, Cornell, and Harvard universities. Any opinions expressed in this paper are those of the author alone and not of the sponsoring institutions.

ABSTRACT

This article considers anew the problem of estimating a regression E(y|x) when realizations of (y,x) are sampled randomly but y is observed selectively. The central issue is the failure of the sampling process to identify E(y|x). The problem faced by the researcher is to find correct prior restrictions which, when combined with the data, identify the regression.

Two kinds of restrictions are examined here. One, which has not been studied before, is a bound on the support of y. Such a bound implies a simple, useful bound on E(y|x). The other, which has received much attention, is a separability restriction derived from a latent variable model.

The selection problem is sometimes confused with the problem of identifying a treatment effect when persons self-select into treatment. This article clarifies the distinction.

1. INTRODUCTION

This article seeks to expose the essence of a problem that has drawn much attention in the past fifteen years: estimation of a regression from selectively observed random sample data.

Suppose that each member of a population is characterized by a triple (y,z,x), where y is a real number, z is a binary indicator, and x is a real vector. A researcher observes a random sample of realizations of (z,x) and, moreover, observes the realizations of y when z = 1. I shall assume that the researcher wants to learn the regression function $E(y|x)$ on the support of the conditioning variable x.

The central issue is identification. The sampling process identifies the regressions $E(y|x,z=1)$ and $E(z|x) = P(z=1|x)$. Given minimal regularity, these functions of x can be estimated consistently. The literature on nonparametric regression analysis offers numerous approaches.

The sampling process does not identify $E(y|x,z=0)$ nor

$$(1) \quad E(y|x) = E(y|x,z=1)P(z=1|x) + E(y|x,z=0)P(z=0|x).$$

On the other hand, $E(y|x)$ may be identified if one can combine the data with suitable prior restrictions on the population distribution of (y,z) conditional on x. The problem faced by the researcher is to find restrictions which are both correct and useful.

Until the early 1970s, researchers almost universally assumed that, conditional on x, y is mean independent of z. That is,

$$(2) \quad E(y|x) = E(y|x,z=1) = E(y|x,z=0).$$

As the sampling process identifies $E(y|x,z=1)$, restriction (2) identifies $E(y|x)$. The plausibility of (2) has subsequently been questioned sharply, especially by researchers who use latent variable models to explain the determination of (y,z). See Gronau(1974).

I shall examine two alternatives to conditional mean independence. Section 2 poses a weak restriction that has not been studied before, namely a bound on the support of y conditional on x. I show that such a bound implies a simple, useful bound on $E(y|x)$ and present an empirical illustration.

Section 3 examines separability restrictions derived from latent variable models. Leading cases include the familiar normal-linear model and recently developed index models.

Section 4 considers the problem of identifying a treatment effect when persons may self-select into treatment. The problem of identifying a treatment effect is often confused with the selection problem. I clarify the distinction.

Section 5 draws conclusions.

2. BOUND ON THE CONDITIONAL SUPPORT OF y

Suppose it is known that, conditional on x and on z = 0, the distribution of y is concentrated in a given interval $[K_{0x}, K_{1x}]$, where $K_{0x} \leq K_{1x}$. That is,

$$(3) \quad P\{y \in [K_{0x}, K_{1x}] | x, z=0\} = 1.$$

Then we may derive an estimable bound on $E(y|x)$. To obtain the bound, observe that

$$(4) \quad P\{y \in [K_{0x}, K_{1x}] | x, z=0\} = 1 \quad \Rightarrow \quad K_{0x} \leq E(y|x, z=0) \leq K_{1x}.$$

Apply this inequality to the right-hand side of equation (1). The result is

$$(5) \quad E(y|x,z=1)P(z=1|x) + K_{0x}P(z=0|x) \leq E(y|x)$$

$$\leq E(y|x,z=1)P(z=1|x) + K_{1x}P(z=0|x).$$

Thus the lower bound is the value $E(y|x)$ takes if, in the non-selected subpopulation, y always equals K_0. The upper bound is the value of $E(y|x)$ if all the non-selected y equal K_1.

This bound on $E(y|x)$ is determined by the bound $[K_{0x}, K_{1x}]$ on y, which is known, and by the regressions $E(y|x,z=1)$ and $P(z|x)$, which are identified by the sampling process. So the bound can be made operational. Methods for estimating the bound from sample data will be provided in Section 2.3.

The bound is informative if $P(z=0|x) < 1$ and if the bound $[K_{0x},K_{1x}]$ on the conditional support of y is nontrivial. Its width is $(K_{1x}-K_{0x})P(z=0|x)$. Thus the width does not depend on $E(y|x,z=1)$. The bound width varies with x in proportion to the two quantities $(K_{1x}-K_{0x})$ and $P(z=0|x)$. This behavior is intuitive. The wider the bound on the conditional support of y, the less prior information one has. The larger is $P(z=0|x)$, the smaller is the probability that y is observed.

It is useful to consider the bound width as a fraction of the width of the original bound on y. The fractional width is $P(z=0|x)$, the probability of not being selected conditional on x. Thus the fractional width does not depend on the variable y whose regression on x is sought. Researchers facing selection problems routinely estimate selection probabilities. So they may easily determine how informative the bound (5) will be for any choice of y and at any value of x.

It is of some historical interest to ask why the literature on selection has not previously recognized the identifying power of a bound on y. The explanation may have several parts.

Timing may have played a role. The literature on selection developed in the 1970s, a period when the frontier of econometrics was nonlinear parametric analysis. At that time, nonparametric regression analysis was just beginning to be formalized by statisticians. Economists were generally unaware that consistent nonparametric regression was possible.

It may be that the historical fixation of econometrics on point identification has inhibited appreciation of the potential usefulness of bounds. Econometricians have occasionally reported useful bounds on quantities that are not point-identified; see, for example, McFadden(1975), Klepper and Leamer(1984), Varian(1985), and Manski(1988a). But the conventional wisdom has been that bounds are hard to estimate and rarely informative. Whatever the validity of this conventional wisdom in other contexts, it does not apply to the bound (5).

Perhaps the preoccupation of researchers with the estimation of wage equations has been a factor. The typical wage regression defines y to be the logarithm of wage. This variable has no obvious upper bound, although minimum wage legislation may enforce a lower bound. Whether or not the logarithm of wage is bounded, wage distributions are always boundable. This is shown in Section 2.1.

2.1. Binary y

When y is a binary indicator variable, the bound takes an especially simple form. Here y is definitionally bounded, with $K_{0x} = 0$ and $K_{1x} = 1$ for all x. Moreover, $E(y|x) = P(y=1|x)$ and $E(y|x,z=1) = P(y=1|x,z=1)$. Hence (5) reduces to

$$(6)\quad P(y=1|x,z=1)P(z=1|x) \leq P(y=1|x)$$
$$\leq P(y=1|x,z=1)P(z=1|x) + P(z=0|x).$$

The binary indicator case may seem special. Actually it has very general application; it provides a bound for any conditional probability. To see this, let w be a random variable taking values in a space W. Let A be any subset of W. Suppose that a researcher wants to bound the probability that w is in A, conditional on x. To do so, one need only observe that

$$(7)\quad P(w\epsilon A|x) = E\{1[w\epsilon A]|x\},$$

where 1[*] is the indicator function taking the value one if the bracketed logical condition holds and zero otherwise. So the bound (6) applies with $y \equiv 1[w\epsilon A]$.

For example, let w be the logarithm of a worker's wage. Suppose that the support of w conditional on x is unbounded. Then the bound (5) on $E(w|x)$ is the trivial $(-\infty,\infty)$. But one can obtain an informative bound on the conditional probability $P(w\leq\tau|x)$ for any real number τ. Just define $y \equiv 1[w\leq\tau]$ and apply (6). Varying τ, one may bound the distribution function of w conditional on x.

It may seem surprising that one should be able to bound the distribution function of a random variable but not its mean. The explanation is a fact that is widely appreciated by researchers in 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).

To obtain some intuition for this fact, consider the following thought experiment. Let w be a random variable with $1-\varepsilon$ of its probability mass in the interval $(-\infty,T]$ and ε mass at some point $S > T$. Suppose w is perturbed by moving the mass at S to some $S_1 > S$. Then $P(w\leq\tau)$ remains unchanged for $\tau < S$ and falls by at most ε for $\tau \geq S$. But E(w) increases by the amount $\varepsilon(S_1-S)$. Now let $S_1 \to \infty$. The perturbed distribution function remains within an ε-bound of the original one but the mean of the perturbed random variable converges to infinity.

2.2. Bounding the Effect of a Change in x

The objective of a regression analysis is sometimes not to learn $E(y|x)$ at a given value of x but rather to learn how $E(y|x)$ moves as x changes. One can use (5) to bound the magnitude of this movement. In some cases one can bound its direction.

Suppose that one wants to learn $E(y|x=\xi) - E(y|x=\rho)$, where ξ and ρ are given points in the support of x. The bound (5) implies that $E(y|x=\xi) - E(y|x=\rho)$ is bounded from below by the difference between the lower bound on $E(y|x=\xi)$ and the upper bound on $E(y|x=\rho)$. Similarly, $E(y|x=\xi) - E(y|x=\rho)$ is bounded from above by the difference between the upper bound on $E(y|x=\xi)$ and the lower bound on $E(y|x=\rho)$. That is,

$$
\begin{aligned}
(8)\quad & E(y|x{=}\xi,z{=}1)P(z{=}1|x{=}\xi) \;-\; E(y|x{=}\rho,z{=}1)P(z{=}1|x{=}\rho) \\
& + K_{0\xi}P(z{=}0|x{=}\xi) \;-\; K_{1\rho}P(z{=}0|x{=}\rho) \\
& \qquad\qquad \leq \quad E(y|x{=}\xi) \;-\; E(y|x{=}\rho) \quad \leq \\
& E(y|x{=}\xi,z{=}1)P(z{=}1|x{=}\xi) \;-\; E(y|x{=}\rho,z{=}1)P(z{=}1|x{=}\rho) \\
& + K_{1\xi}P(z{=}0|x{=}\xi) \;-\; K_{0\rho}P(z{=}0|x{=}\rho).
\end{aligned}
$$

The width of this bound is the sum of the widths of the bounds on $E(y|x{=}\xi)$ and on $E(y|x{=}\rho)$. Depending on the case, the bound may or may not lie entirely to one side of the origin.

The foregoing concerns a finite change in x. On occasion one would like to learn the derivative $\partial E(y|x)/\partial x$. A bound on y does not by itself restrict this derivative. A bound on y combined with one on $\partial E(y|x,z{=}0)/\partial x$ does.

The argument extends that leading to (5). It follows from (1) that

$$
\begin{aligned}
(9)\quad \partial E(y|x)/\partial x \quad = & \\
& P(z{=}1|x)[\partial E(y|x,z{=}1)/\partial x] \;+\; E(y|x,z{=}1)[\partial P(z{=}1|x)/\partial x] \\
+ & \; P(z{=}0|x)[\partial E(y|x,z{=}0)/\partial x] \;+\; E(y|x,z{=}0)[\partial P(z{=}0|x)/\partial x],
\end{aligned}
$$

provided that these derivatives exist. Of the quantities on the right-hand side of (9), all but $E(y|x,z{=}0)$ and $\partial E(y|x,z{=}0)/\partial x$ are identified by the sampling process. Suppose that (3) holds. Moreover, let it be known that, for a given $[D_{0x}, D_{1x}]$,

$$(10) \qquad \partial E(y|x,z=0)/\partial x \quad \epsilon \quad [D_{0x}, D_{1x}].$$

Then the unidentified quantities are both bounded. The result is a bound on $\partial E(y|x)/\partial x$, namely

$$\begin{aligned}(11) \quad & P(z=1|x)[\partial E(y|x,z=1)/\partial x] + E(y|x,z=1)[\partial P(z=1|x)/\partial x] \\ & \qquad + P(z=0|x)D_{0x} + K_{0x}[\partial P(z=0|x)/\partial x] \\ & \qquad\qquad \leq \quad \partial E(y|x)/\partial x \quad \leq \\ & P(z=1|x)[\partial E(y|x,z=1)/\partial x] + E(y|x,z=1)[\partial P(z=1|x)/\partial x] \\ & \qquad + P(z=0|x)D_{1x} + K_{1x}[\partial P(z=0|x)/\partial x].\end{aligned}$$

I shall not discuss this bound further. The knowledge needed to obtain it is much less readily available than that which suffices to bound the finite difference $E(y|x=\xi) - E(y|x=\rho)$. The support of y is often definitionally bounded. The derivative $\partial E(y|x,z=0)/\partial x$ is rarely so.

2.3. Estimation of the Bound

A simple way to estimate the bound on $E(y|x)$ is to estimate $E(y|x,z=1)$ and $P(z|x)$, both of which are identified by the sampling process. I shall present an equivalent approach whose statistical properties are a bit easier to derive.

First rewrite (5) in an equivalent form. Observe that

$$(12) \quad E(yz|x) = E(yz|x,z=1)P(z=1|x) + E(yz|x,z=0)P(z=0|x)$$
$$= E(y|x,z=1)P(z=1|x).$$

Also observe that

$$(13) \quad P(z=0|x) = E(1-z|x).$$

It follows from (12) and (13) that (5) may be rewritten as

$$(5') \quad E(yz|x) + K_{0x}E(1-z|x) \leq E(y|x) \leq E(yz|x) + K_{1x}E(1-z|x).$$

The above shows that to estimate the bound, it suffices to estimate two linear combinations of $E(yz|x)$ and $E(1-z|x)$. I shall first pose a simple method that works at values of x having positive probability in the population. I shall then extend the method to make it work at any point in the support of x.

Consider a point ξ such that $P(x=\xi) > 0$. Let N denote the sample size. The natural estimates for $E(yz|x=\xi)$ and $E(1-z|x=\xi)$

are the sample averages of yz and 1-z across those observations for which x = ξ, namely

$$(14)\quad b_{N\xi} \equiv \frac{\sum_{i=1}^{N} y_i z_i \, 1[x_i=\xi]}{\sum_{j=1}^{N} 1[x_j=\xi]}$$

and

$$(15)\quad c_{N\xi} \equiv \frac{\sum_{i=1}^{N} (1-z_i) \, 1[x_i=\xi]}{\sum_{j=1}^{N} 1[x_j=\xi]} .$$

Note that $b_{N\xi}$ is computable even though y_i is not always observed; if $z_i = 0$, then $y_i z_i = 0$. Given (5'), (14), and (15), we may estimate the bound at ξ by

$$(16)\quad [b_{N\xi} + K_{0\xi}c_{N\xi}, b_{N\xi} + K_{1\xi}c_{\xi N}].$$

This estimate is consistent. The strong law of large numbers implies that as N -> ∞,

$$(17)\qquad (b_{N\xi}, c_{N\xi}) \;\; -> \;\; [E(yz|x=\xi), E(1-z|x=\xi)]$$

almost surely. Hence (16) converges almost surely to the true bound.

The estimate has a limiting normal distribution if, conditional on $x = \xi$, the bivariate random variable $(yz, 1-z)$ has finite variance matrix. Let Σ_ξ denote the variance matrix. The central limit theorem implies that as $N \to \infty$,

$$(18) \quad \sqrt{N}[(b_{N\xi}, c_{N\xi}) - \{E(yz|x=\xi), E(1-z|x=\xi)\}] \quad \to \quad N(0,\Sigma_\xi)$$

in distribution. Hence $\sqrt{N}$ times the difference between $(b_{N\xi}+K_{0\xi}c_{N\xi}, b_{N\xi}+K_{1\xi}c_{\xi N})$ and the true endpoints of the bound has a limiting normal distribution with mean zero and variance matrix

$$\begin{bmatrix} 1 & K_{0\xi} \\ 1 & K_{1\xi} \end{bmatrix} \Sigma_\xi \begin{bmatrix} 1 & 1 \\ K_{0\xi} & K_{1\xi} \end{bmatrix} .$$

Now consider the problem of estimating the bound at values of x that have probability zero in the population but are in the support of x. That is, let $\| \ \|$ denote a norm and consider ξ such that $P(x=\xi) = 0$ but $P[\|x-\xi\|<\delta] > 0$ for all $\delta > 0$.

Estimating $E(yz|x=\xi)$ and $E(1-z|x=\xi)$ by (14) and (15) clearly will not work; with probability one there are no sample observations for which $x = \xi$. On the other hand, it seems reasonable to estimate these quantities by the sample averages of yz and 1-z across those observations for which x is "close" to ξ, provided that one tightens the criterion of closeness appropriately as the sample size increases. This intuitive idea does work; it is the basis of nonparametric regression analysis.

To formalize the idea, let $W_{Ni}(\xi)$, $i = 1,\ldots,N$ be chosen weights that sum to one and redefine $(b_{N\xi}, c_{N\xi})$ to be estimates of the form

$$(19) \quad \begin{bmatrix} b_{N\xi} \\ c_{N\xi} \end{bmatrix} \equiv \sum_{i=1}^{N} W_{Ni}(\xi) \begin{bmatrix} y_i z_i \\ 1-z_i \end{bmatrix} .$$

The earlier definitions of $(b_{N\xi}, c_{N\xi})$ are subsumed by (19); they are the special case in which

$$(20) \quad W_{Ni}(\xi) \equiv \frac{1[x_i=\xi]}{\sum_{j=1}^{N} 1[x_j=\xi]} .$$

A large menu of nonparametric regression estimates having the form (19) are available for application. Perhaps the simplest is the "histogram" method. Here the researcher selects a "bandwidth" $\delta > 0$ and lets

$$(21) \quad W_{Ni}(\xi) \equiv \frac{1[\|x_i-\xi\|<\delta]}{\sum_{j=1}^{N} 1[\|x_j-\xi\|<\delta]} .$$

If the weights are chosen as in (21) and if δ is fixed, we have a consistent estimate of $[E(yz|\|x-\xi\|<\delta), E(1-z|\|x-\xi\|<\delta)]$. On the other hand, if the researcher lets δ vary with the sample in a way that makes $\delta \rightarrow 0$ as $N \rightarrow \infty$, it is plausible that we obtain a consistent estimate of $[E(yz|x),E(1-z|x)]$. This turns out to be so, provided that the rule used to choose δ makes $\delta \to 0$ sufficiently slowly as $N \to \infty$.

Two classes of nonparametric regression methods that have drawn much attention are the "kernel" estimators and the "nearest-neighbor" estimators. Both classes have the form (19); they choose the weights W in different ways. The histogram estimator is a member of the kernel class. Bierens(1987) and Hardle(1988) provide excellent expositions of kernel regression, complete with numerical examples. Prakasa Rao(1983) and Hardle(1988) cover the nearest-neighbor approach. Manski(1988b) introduces a close cousin of the kernel and nearest-neighbor methods, called "smallest neighborhood" estimation.

It would carry us too far afield to survey here the asymptotic properties and operational characteristics of the many available procedures. A few general remarks will suffice.

First, almost any intuitively reasonable estimator of the form (19) provides a consistent estimate of $[E(yz|x),E(1-z|x)]$. Most estimators have limiting normal distributions, although not always centered at zero. The rate of convergence is generally slower than the $\sqrt{N}$ rate obtainable in classical estimation problems. Bierens(1987) introduces an easily computed kernel type estimate that converges as rapidly as is possible and that has a limiting normal distribution centered at zero.

Second, some researchers find it uncomfortable that so many different choices of the weights $W_{Ni}(\xi)$, i =1,...,N yield estimates with similar asymptotic properties. Simply put, the problem is that the available statistical theory gives the researcher too little guidance on choosing the weights in practice. Many researchers advocate use of "cross-validation" to select an estimator. In cross-validation, one computes alternative estimates on subsamples and uses them to predict y on the complementary subsamples. One selects the estimate which has the best predictive power.

Third, there is consensus among practitioners that nonparametric regression methods usually work well when the regressor variable x has low dimension. On the other hand, it is common to find that, in samples of realistic size, performance is poor when the dimension of x is high. This phenomenon has led some researchers to develop approaches that impose dimension-reducing restrictions on the regression but remain nonparametric in part. Section 3.3 will describe applications to the analysis of selection problems.

2.4. An Empirical Example: Attrition in a Survey of the Homeless

To illustrate the bound, I consider a selection problem that arose in a recent study of exit from 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 variable y is binary, with $y = 1$ if the individual has a home six months later and $y = 0$ if he or she remains homeless. The regressors x are individual background attributes. The objective is to learn $E(y|x) = P(y=1|x)$.

The investigators interviewed a random sample of the people who were homeless in Minneapolis in late December 1985. Six months later they attempted to reinterview the original respondents but succeeded in locating only a subset. So the selection problem is attrition from the sample: $z = 1$ if a respondent is located for reinterview, $z = 0$ otherwise.

Let us first estimate the bound on a very simple regression, that in which x is the respondent's sex. Consider the males. First 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 $E(yz|\text{male})$ is $b_{N\text{male}} = 21/106$ and that of $E(1-z|\text{male})$ is $c_{N\text{male}} = 42/106$. The estimate of the bound on $P(y=1|\text{male})$ is $[21/106,63/106] \approx [.20,.59]$.

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 $b_{Nfemale} = 3/31$ and $c_{Nfemale} = 17/31$. The estimated bound on $P(y=1|female)$ is $[3/31,20/31] \approx [.10,.65]$.

Interpretation of these estimates should be cautious, given the small sample sizes. Taking the results at face value, we have a tighter bound on $P(y=1|male)$ than on $P(y=1|female)$. The attrition frequencies, hence 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 selection process, we are nevertheless 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 regressor is a discrete variable. To provide an example in which x is continuous, I regress y on sex and an income variable. The latter 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 from 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 selection problem implied by the fact that 17 of the 106 men did not respond to the income question.

Let x be the bivariate regressor (male,income). Figure 1 graphs a nonparametric estimate of $P(z=0|x)$ on the sample income data. This and the estimate of $E(yz|x)$ were obtained by cross-validated logistic kernel regression, using the program NPREG described in Manski and Thompson(1987). 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.

Figure 2 graphs the estimate of the bound on $P(y=1|x)$. The lower bound is the estimate of $E(yz|x)$, which is flat on the income range where the data are concentrated but turns downward eventually. The upper bound is the sum of the estimates for $E(yz|x)$ and for $P(z=0|x)$.

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 1, that the estimated probability of attrition increases with income.

FIGURE 1: P[z=0|x]

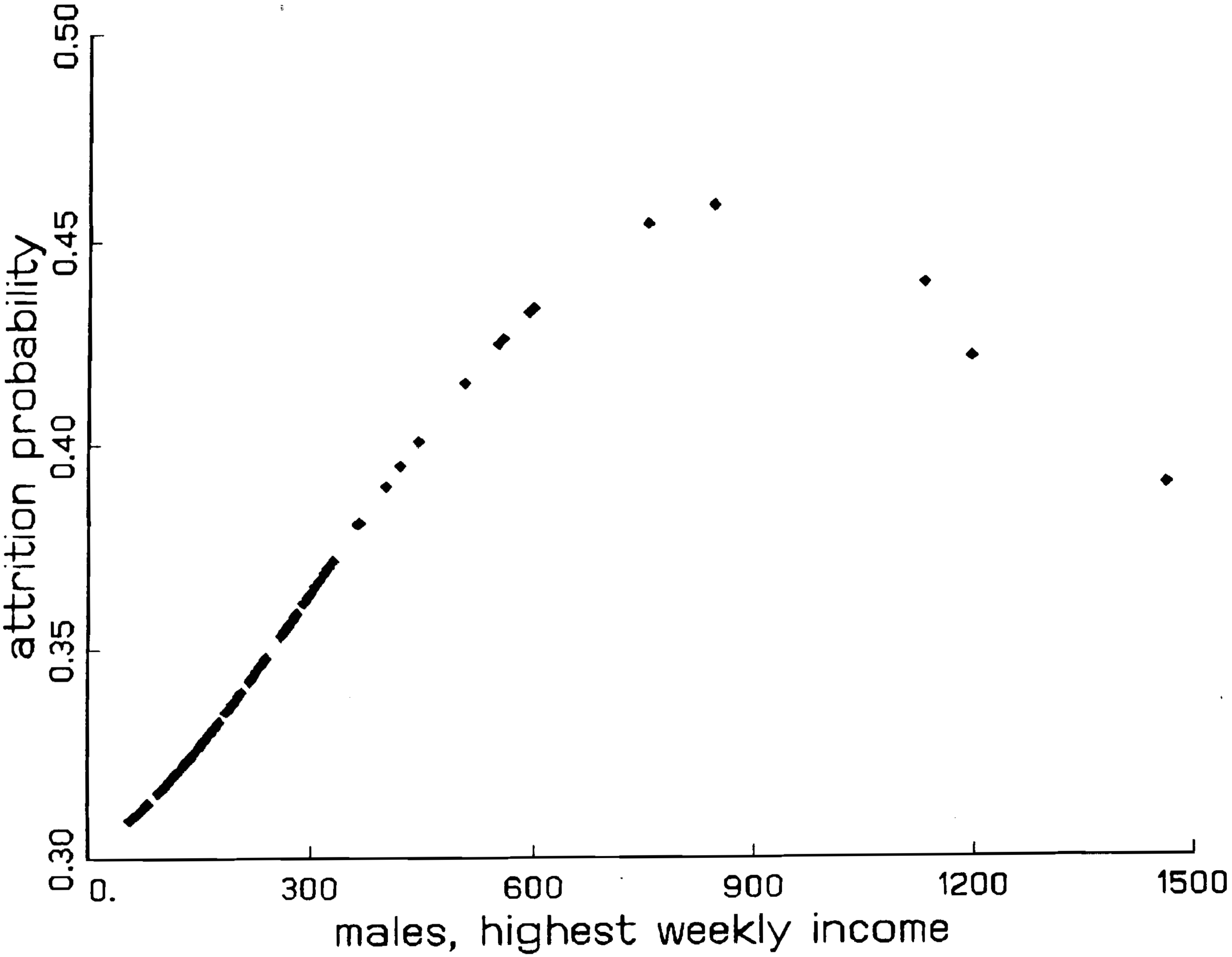
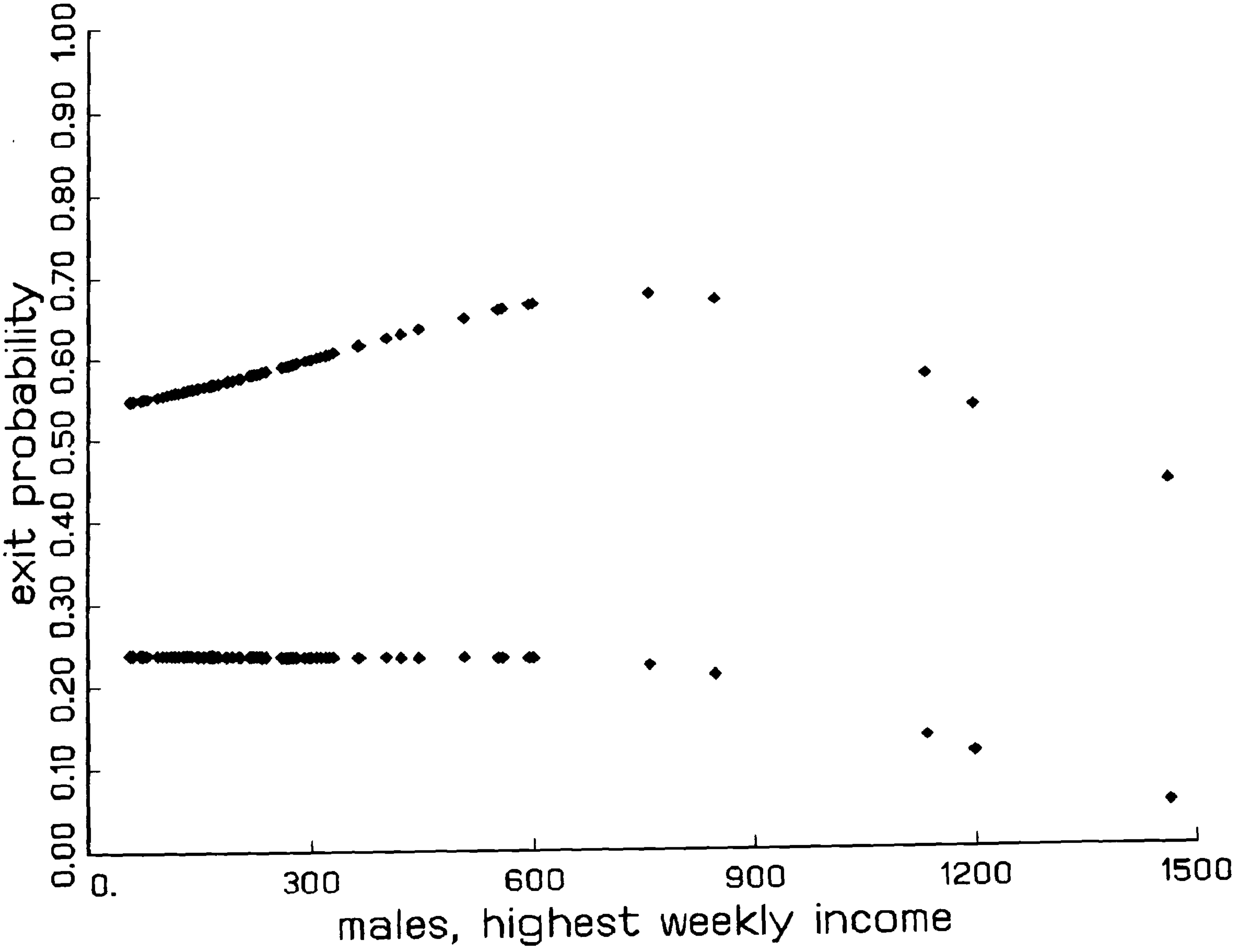


FIGURE 2: BOUND ON P[y=1|x]



3. SEPARABILITY RESTRICTIONS DERIVED FROM LATENT VARIABLE MODELS

Prevailing practice in the econometric literature on selection is to identify $E(y|x)$ by assuming that $E(y|x,z=1)$ is the sum of $E(y|x)$ and another function that is distinguishable from $E(y|x)$. Suppose it is known that $E(y|x)$ and $E(y|x,z=1)$ have the forms

(22a) $E(y|x) = g_1(x)$

(22b) $E(y|x,z=1) = g_1(x) + g_2(x)$

for some $g_1 \epsilon G_1$ and $g_2 \epsilon G_2$, where G_1 and G_2 are specified families of functions mapping x into the real line. The sampling process identifies $E(y|x,z=1)$; hence $g_1(*) + g_2(*)$ is identified. The functions g_1 and g_2 can be separately identified if knowledge of $g_1(*) + g_2(*)$ is combined with prior restrictions on G_1 and G_2.

The literature provides various specifications for (G_1,G_2) that suffice. These specifications have been motivated by reference to the latent variable model

(23) $y = f_1(x) + u_1$

(24) $E(u_1|x) = 0$

(25) $z = 1[f_2(x) + u_2 > 0]$,

where $[f_1(*),f_2(*)]$ are real functions of x and (u_1,u_2) are unobserved real random variables. Condition (24) normalizes location if $f_1(*)$ is unrestricted but is an assumption otherwise.

The latent variable model implies that

(26a) $E(y|x) = f_1(x)$

(26b) $E(y|x,z=1) = f_1(x) + E[u_1|x,f_2(x)+u_2>0]$.

So (22) holds with

(27a) $g_1(x) = f_1(x)$

(27b) $g_2(x) = E[u_1|x,f_2(x)+u_2>0]$.

Restrictions imposed on $f_1(*)$ translate directly into a specification of G_1. Restrictions imposed on $f_2(*)$ and on the distribution of (u_1,u_2) conditional on x induce a specification of G_2.

Sections 3.1 through 3.3 consider three restrictions that have received considerable attention. In each case I give the resulting specification of (G_1,G_2). The restrictions to be discussed are neither nested nor mutually exclusive. A latent variable model may impose any combination of the three.

3.1. The Model with Conditionally Independent Disturbances

The early literature assumed that u_1 and u_2 are statistically independent conditional on x. This and (24) imply that

$$(28)\quad E(u_1|x,f_2(x)+u_2>0) \quad = \quad E(u_1|x) \quad = \quad 0.$$

So G_2 contains only the function $g_2(x) = 0$. In other words, the conditional mean independence restriction (2) holds.

The model with conditionally independent disturbances imposes no restrictions on $f_1(*)$. Hence this model has no implications beyond (2), which just identifies $E(y|x)$. In practice, researchers have typically imposed supplementary restrictions on $f_1(*)$; most of the applied literature makes $f_1(*)$ linear.

3.2. Parametric Models

A second type of restriction became prominent in the middle 1970s. Suppose that $f_1(*)$ is known up to a finite dimensional parameter β_1, $f_2(*)$ up to a finite dimensional parameter β_2, and the distribution of (u_1,u_2) conditional on x up to a finite dimensional parameter γ. Then (27) implies that G_1 is a finite dimensional family of functions parametrized by β_1 and G_2 is a finite dimensional family parametrized by (β_2,γ). So we may write

$$(29a)\quad E(y|x) \quad = \quad g_1(x,\beta_1)$$

$$(29b)\quad E(y|x,z=1) \quad = \quad g_1(x,\beta_1) \; + \; g_2[x,(\beta_2,\gamma)].$$

Sufficiently strong parametric restrictions identify β_1, hence $E(y|x)$. One widely applied model makes $f_1(*)$ and $f_2(*)$ linear functions, (u_1,u_2) statistically independent of x, and the distribution of (u_1,u_2) normal with mean zero. See Heckman(1976). In this case,

$$(30a)\quad E(y|x) \quad = \quad x'\beta_1$$

$$(30b)\quad E(y|x,z=1) \quad = \quad x'\beta_1 + \gamma\phi(x'\beta_2)/\Phi(x'\beta_2),$$

where $\phi(*)$ and $\Phi(*)$ are the standard normal density and distribution functions and where $\gamma = E(u_1u_2)$. Identification of β_1 hinges on the fact that the linear function $x'\beta_1$ and the nonlinear $\gamma\phi(x'\beta_2)/\Phi(x'\beta_2)$ affect $E(y|x,z=1)$ in different ways.

There is a common perception that the normal-linear model generalizes the model with conditionally independent disturbances. Barros(1988) observes that the two models are, in fact, not nested. The normal-linear model permits u_1 and u_2 to be dependent but assumes linearity of $[f_1(*),f_2(*)]$, normality of (u_1,u_2), and independence of (u_1,u_2) from x. The model with conditionally independent disturbances assumes u_1 and u_2 to be independent conditional on x but restricts neither the form of $[f_1(*),f_2(*)]$, the distribution of u_1 conditional on x, nor the distribution of u_2 conditional on x. Given this, Barros argues that the model with conditionally independent disturbances warrants renewed attention.

3.3. Index Models

Parametric models have increasingly been criticized for their fragility; seemingly small misspecifications may generate large biases in estimates of $E(y|x)$. Several articles have reported that estimates obtained under the normal-linear model are sensitive to misspecification. Hurd(1979) has shown the consequences of heteroskedasticity. Arabmazar and Schmidt(1982) and Goldberger(1983) have described the effect of non-normality.

The lack of robustness of parametric models is particularly severe when the x components that enter g_2 are the same as those that determine g_1. In this case, identification of g_1 hinges entirely on the imposed functional form restrictions. Recognition of this has led to the recent development of a third class of models, one which weakens functional form restrictions at the cost of imposing exclusion restrictions.

Let $h_1(x)$ and $h_2(x)$ be "indices" of x; that is, many-to-one functions of x. Suppose that $f_1(x)$ is known to vary with x only through $h_1(x)$. Suppose that $f_2(x)$ and the distribution of (u_1,u_2) conditional on x are known to vary with x only through $h_2(x)$. Then G_1 is a family of functions that depend on x only through $h_1(x)$ and G_2 is a family of functions that depend on x only through $h_2(x)$. So we may write

$$(31a)\quad E(y|x) = g_1[h_1(x)]$$

$$(31b)\quad E(y|x,z=1) = g_1[h_1(x)] + g_2[h_2(x)].$$

An example is the model in which $f_1(x) = f_1[h_1(x)]$, $f_2(x) = f_2[h_2(x)]$, and (u_1,u_2) is statistically independent of x. This model weakens the assumptions of the normal-linear model in some respects but strengthens them in others. The index model does not force f_1 and f_2 to be linear nor the distribution of (u_1,u_2) to be normal. On the other hand, it assumes that f_1 and f_2 are determined by distinct indices, a condition not imposed by the normal-linear model.

When combined with restrictions on the family G_1 of feasible regression functions, index restrictions can identify g_1. Powell(1987) expresses the basic idea, which is to difference-out the function g_2 as in fixed effects analyses of panel data.

Let (ξ,ρ) denote a pair of points in the support of x such that $h_2(\xi) = h_2(\rho)$ but $h_1(\xi) \neq h_1(\rho)$. For each such pair, (31b) implies that

$$(32) \quad E(y|x=\xi,z=1) - E(y|x=\rho,z=1) = g_1[h_1(\xi)] - g_1[h_1(\rho)].$$

The left-hand side of (32) is identified by the sampling process. The right-hand side is determined by the function of interest g_1 and not by the "nuisance" function g_2. Hence (32) restricts g_1. Identification hinges on whether the support of x contains enough pairs (ξ,ρ) for (32) to pin g_1 down to a single function within the family of feasible functions G_1.

The statistics literature on "projection pursuit" regression offers approaches to the estimation of g_1 when the family G_1 is restricted only qualitatitively. See Huber(1985). Econometricians studying index models have typically assumed that g_1 is linear. See Ichimura and Lee(1988), Powell(1987), and Robinson(1988) for alternative estimation approaches. The first two papers are concerned with an extension of the index model in which the form of the index function h_2 is not known but is estimable.

As the dates of the foregoing citations indicate, the literature on index models is young. The work so far has been entirely theoretical. Empirical applications have yet to appear.

3.4. Latent Variable Models and the Bound Restriction

It is of interest to juxtapose the restrictions on $E(y|x)$ implied by latent variable models with those implied by a bound on the conditional support of y. For purposes of this discussion, I shall suppose that the bound on y is specified properly. This is an assumption in some applications but is a truism when y is definitionally bounded.

Consider a researcher who has specified a latent variable model. The researcher can check whether the hypothesis $[E(y|x) = f_1(x)]$ is consistent with the bound on $E(y|x)$. I use the informal term "check" rather than the formal one "test" intentionally; sampling theory for these bounds-checks remains to be developed.

For example, suppose that the researcher has specified the normal-linear model. Normality of u_1 implies that y has unbounded support conditional on x. Hence acceptance of the normal-linear model implies that the bound on $E(y|x)$ is ineffective. But the bound on the conditional distribution $P(y\leq\tau|x)$, $\tau \epsilon R^1$ is effective, as was shown in Section 2.1. The normal-linear model implies that

$$(33) \quad P(y\leq\tau|x) = \Phi[(\tau-x'\beta_1)/\sigma_1] \qquad \tau \epsilon R^1,$$

where σ_1 is the standard deviation of u_1. So we may check the validity of the model by estimating (β_1,σ_1), computing the estimate of (33), and comparing the result with an estimate of the bound on the conditional distribution function.

It may be thought that bounds-checks of latent variable models are impractical in those applications where the dimension of x is high. The ostensible reason, stated in Section 2.3, is that nonparametric regression estimation tends to perform poorly in high dimensional settings. Nevertheless, informative checks are practical, as follows.

Consider $E(y|x\epsilon A)$, where A is any region in x-space such that $P(x\epsilon A) > 0$. The bound on $E(y|x\epsilon A)$ is easily estimated by (16). Let a latent variable model be specified. The model implies that $E(y|x) = f_1(x)$. Hence it implies that

$$(34)\quad E(y|x\epsilon A) = E(y*1[x\epsilon A])/P(x\epsilon A)$$
$$= E_x\{E(y|x)*1[x\epsilon A]\}/P(x\epsilon A)$$
$$= E_x\{f_1(x)*1[x\epsilon A]\}/P(x\epsilon A).$$

Let $f_{1N}(x)$ be an estimate of $f_1(x)$. Then the latent variable model implies the following estimate of $E(y|x\epsilon A)$:

$$(35)\quad E_N(y|x\epsilon A) \equiv \frac{\sum_{i=1}^{N} f_{1N}(x_i)\ 1[x_i\epsilon A]}{\sum_{j=1}^{N} 1[x_j\epsilon A]} .$$

One may check the latent variable model by comparing (35) with the estimate of the bound on $E(y|x\epsilon A)$.

4. IDENTIFICATION OF TREATMENT EFFECTS

The selection problem studied in Sections 1 through 3 is sometimes confused with the problem of identifying a treatment effect when persons self-select into treatment. This section seeks to clarify the distinction. To keep the presentation simple I restrict attention to a binary treatment. Moreover, I use only probabilistic terms, making no reference to latent variable models. See Heckman and Robb(1985) for a discussion framed in latent variable terms.

Let y denote the relevant outcome variable. Let t be a binary variable indicating receipt of treatment; t = 1 if a person receives treatment and t = 0 if not. A common labor economics application makes y a person's wage and t his participation in some program meant to enhance his human capital.

Let v denote a set of observable variables characterizing the person. Let r be a binary variable indicating the person's preference for treatment; r = 1 if a person prefers treatment and r = 0 if not. The variables r and t are conceptually distinct. With typical survey data, a researcher can observe realizations of t but not of r.

The "treatment effect" is defined to be

$$(36)\quad E(y|v,r,t=1) - E(y|v,r,t=0),$$

that is, the expected effect on y of receipt of treatment, holding fixed the person's observable characteristics v and his preference for treatment r. Some discussions of the treatment effect suppose that the researcher wants to learn (36). Others suppose that the researcher wants to identify the average treatment effect across persons with different preferences for treatment. The latter quantity is

$$(37)\quad \sum_{r} P(r|v)\ [E(y|v,r,t=1) - E(y|v,r,t=0)].$$

Assume that a researcher observes a random sample of realizations of (y,v,t). Whether the researcher wants to learn (36) or (37), the obvious problem is that the preference indicator r is not observed. How then can the researcher proceed?

The problem of identifying the average treatment effect (37) is easily solved if it is known that, conditioning on v, preference for and receipt of treatment are statistically independent. That is,

$$(38)\quad P(r|v,t) = P(r|v).$$

This restriction holds, for example, in an experiment with randomized assignment to treatment. Given (38), the average treatment effect (37) reduces to

$$(39)\quad E(y|v,t=1) - E(y|v,t=0),$$

an expression which does not involve the unobserved r and is identified by the sampling process. To see this, observe that

$$(40)\quad E(y|v,t=1) - E(y|v,t=0) \quad =$$

$$\sum_r P(r|v,t=1)\ E(y|v,r,t=1) - \sum_r P(r|v,t=0)\ E(y|v,r,t=0).$$

The right-hand side generally differs from the average treatment effect (37) but coincides with it if $P(r|v,t) = P(r|v)$.

Much of the recent literature assumes that persons self-select into treatment. If so (38) does not hold. Rather,

(41) $P(r=t|v,t) = 1.$

Given (41), the preference indicator r is indirectly observable through observation of t. The treatment effect (36) reduces to

(42a) $E(y|v,t=1) - E(y|v,r=1,t=0)$

for persons who are observed to select treatment and

(42b) $E(y|v,r=0,t=1) - E(y|v,t=0)$

for persons who do not select treatment.

Random sampling of (y,v,t) identifies $E(y|v,t)$. The problem is that $E(y|v,r=1,t=0)$ and $E(y|v,r=0,t=1)$ are not identified. In fact, the population contains no persons such that $(r=1,t=0)$ nor any such that $(r=0,t=1)$.

The foregoing makes clear that the problem of identifying a treatment effect when people self-select into treatment is not the same as the problem of selective observation. The selection problem concerns a researcher who selectively observes y and wants to learn the regression $E(y|x)$ on the support of x. The treatment-effect problem concerns a researcher who always observes y and who wants to extrapolate the regression $E(y|v,r,t)$ off the support of $x \equiv (v,r,t)$.

Suppose that treatment is self-selected in the population under observation. It is of interest to ask whether a bound on y implies a bound on the treatment effect (42). The answer is that the magnitude of the effect can be bounded but, in general, not its sign. Suppose it is known that

$$(43a) \quad P\{y \in [L_{0v},L_{1v}]|v,r=1,t=0\} = 1$$

$$(43b) \quad P\{y \in [M_{0v},M_{1v}]|v,r=0,t=1\} = 1.$$

Then the treatment effect must lie in the interval

$$(44a) \quad [E(y|v,t=1)-L_{1v}, \ E(y|v,t=1)-L_{0v}]$$

for persons who select treatment and

$$(44b) \quad [M_{0v}-E(y|v,t=0), \ M_{1v}-E(y|v,t=0)]$$

for those who do not.

5. CONCLUSION

Fifteen years ago few economists paid attention to the fact that selective observation of random sample data has implications for empirical analysis. Then the profession became sensitized to the selection problem. The heretofore maintained assumption, conditional mean independence of y and z, became a standard object of attack. For a while the normal-linear latent variable model became the standard "solution" to the selection problem. But researchers soon became aware that this model does not solve the selection problem. It trades one set of assumptions for another.

Today there is no conventional wisdom. Some applied researchers, such as LaLonde(1986), have leaped from disenchantment with the normal-linear model to the conclusion that econometric analysis is incapable of interpreting observations of natural populations. In rebuttal, Heckman and Hotz(1988) argue that latent variable models are useful empirical tools provided that applied researchers take seriously the task of model specification.

Econometricians are seeking to widen the menu of separable regression specifications derived from latent variable models. The recent work on index models weakens the parametric assumptions of the normal-linear model at the cost of requiring exclusion assumptions. There is also a revival of interest in the model with conditionally independent disturbances.

I find the current diversity of opinion unsurprising. Moreover, I expect it to persist. Selection creates an identification problem. Identification always depends on the prior knowledge a researcher is willing to assert in the application of interest. As researchers are heterogeneous, so must be their perspectives on the selection problem.

Econometricians can assist empirical researchers by clarifying the nature of the selection problem and by widening the menu of prior restrictions for which estimation methods are available. Work on restrictions derived from latent variable models is welcome. I also believe that researchers should routinely estimate the simple bound developed in Section 2. To bound $E(y|x)$ one need only be able to bound the variable y. One need not accept the latent variable model.

REFERENCES

Arabmazar, A. and Schmidt, P.(1982), "An Investigation of the Robustness of the Tobit Estimator to Non-Normality," Econometrica, 50, 1055-1063.

Barros, R.(1988), "Nonparametric Estimation of Causal Effects in Observational Studies," Economic Research Center, NORC, University of Chicago.

Bierens, H.(1987), "Kernel Estimators of Regression Functions," in T. Bewley(ed.), Advances in Econometrics, Volume I, New York: Cambridge University Press.

Goldberger, A.(1983), "Abnormal Selection Bias," in T. Amemiya and I. Olkin(eds.), Studies in Econometrics, Time Series, and Multivariate Statistics, Orlando: Academic Press.

Gronau, R.(1974), "Wage Comparisons - a Selectivity Bias," Journal of Political Economy, 82, 1119-1143.

Hardle, W.(1988), Applied Nonparametric Regression, Rheinische-Friedrich-Wilhelms Universitat, Bonn, West Germany.

Heckman, J.(1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models, Annals of Economic and Social Measurement, 5, 479-492.

Heckman, J. and Hotz, J.(1988), "Choosing among Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," Department of Economics, Yale University.

Heckman, J. and Robb, R.(1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer(eds.), Longitudinal Analysis of Labor Market Data, New York: Cambridge University Press.

Huber, P.(1981), Robust Statistics, New York: Wiley.

Huber, P.(1985), "Projection Pursuit," Annals of Statistics, 13, 435-475.

Hurd, M.(1979), "Estimation in Truncated Samples When There is Heteroskedasticity," Journal of Econometrics, 11, 247-258.

Ichimura, H. and Lee, L.(1988), "Semiparametric Estimation of Multiple Indices Models: Single Equation Estimation," Department of Economics, University of Minnesota.

Klepper, S. and Leamer, E.(1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables," Econometrica, 52, 163-183.

LaLonde, R.(1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," American Economic Review, 76, 604-620.

Manski, C.(1988a), "Identification of Binary Response Models," Journal of the American Statistical Association, 83, 729-738.

Manski, C.(1988b), Analog Estimation Methods in Econometrics, London: Chapman and Hall.

Manski, C. and Thompson, S.(1987), "MSCORE with NPREG: Documentation for Version 1.4," Department of Economics, University of Wisconsin-Madison.

McFadden, D.(1975), "Tchebyscheff Bounds for the Space of Agent Characteristics," Journal of Mathematical Economics, 2, 225-242.

Piliavin, I. and Sosin, M.(1988), "Exiting Homelessness: Some Recent Empirical Findings," Institute for Research on Poverty, University of Wisconsin-Madison, in preparation.

Powell, J.(1987), "Semiparametric Estimation of Bivariate Latent Variable Models," Social Systems Research Institute Paper 8704, University of Wisconsin-Madison.

Prakasa Rao, B.L.S.(1983), Nonparametric Functional Estimation, Orlando: Academic Press.

Robinson, P.(1988),"Root-N-Consistent Semiparametric Regression," Econometrica, 56, 931-954.

Varian, H.(1985), "Nonparametric Analysis of Optimizing Behavior with Measurement Error," Journal of Econometrics, 30, 445-458.