SELECTIVITY BIAS: NEW DEVELOPMENTS†

Varieties of Selection Bias

By JAMES HECKMAN*

This paper considers a prototypical problem in the econometrics of selection bias: estimating the impact of unionism on wage differentials. The statistical structure of this problem is the same as that of many other self-selection problems in economics. The union wage differential question is of interest in its own right, and has been the subject of numerous papers evaluated in the landmark study of H. Gregg Lewis (1986).

I consider the following questions. 1) What are the parameters of economic interest? 2) Under what conditions can they be identified? 3) How can alternative nonparametric procedures set forth in the literature aid in securing the parameters of interest? 4) What is the status of the evidence on appropriate methods for estimating union-nonunion wage differentials? The answer to the third question is particularly relevant in view of Lewis' pessimistic and influential summary of existing studies that attempt to correct for self-selection bias. He complains about the wide variation in the estimates produced in different studies. Such variability may be due to the imposition of false distributional assumptions—an issue that motivates the recent literature on nonparametric estimation.

I. What Are the Parameters of Economic Interest?

Which parameters are we seeking to estimate without self-selection bias? Suppressing individual subscripts, a switching regressions model of unionism writes union wages Y_1 as $Y_1 = X_1\beta_1 + U_1$, $E(U_1) = 0$; nonunion wages Y_0 as $Y_0 = X_0\beta_0 + U_0$, $E(U_0) = 0$; and the sectoral choice equation as $I = Z\gamma + V$, $E(V) = 0$; where $I \geq 0$ implies that union sector one is chosen $(D = 1)$. Otherwise $D = 0$.

The observed wage Y is

$$(1)\quad Y = 6Y_1D + Y_0(1-D)$$
$$= (X_1\beta_1)D + (X_0\beta_0)(1-D)$$
$$+ DU_1 + (1-D)U_0.$$

The case $I = Y_1 - Y_0$ is the Roy model, put in linear regression form. Lung Fei Lee (1978) considers more general self-selection rules. Following much of the literature on this problem, assume that (U_1, U_0, V) are independent of the regressors (X_1, X_0, Z). In the original models, (U_1, U_0, V) were assumed to be joint normal.

The most commonly used specification of this model assumes that $\beta_1 = \beta_0$ except for the intercepts α_1, α_0, and $X_1 = X_0$. Letting β be the common slope coefficient vector and X the vector of regressors excluding the intercept, we obtain

$$(2)\quad Y = \alpha_0 + D(\alpha_1 - \alpha_0) + X\beta + (1-D)U_0 + DU_1.$$

The empirical literature on the simultaneity problem in estimating the impact of union status focuses on the stochastic dependence between D and the disturbance $\{U_0 + D(U_1 - U_0)\}$. But the parameters of interest are often not well defined, and are also defined differently in various studies—sometimes inconsistently in the same study.

†*Discussants*: Paul Rudd, University of California-Berkeley; Whitney Newey, Princeton University; Roger Klein, Bell Communications Research.

*Department of Economics, Yale University, 37 Hillhouse, New Haven, CT 06520. This research was supported by NSF grant no. 87-11845. Steve Cameron, Kyung Soo Choi, Bo Honoré, Whitney Newey, and James Walker gave helpful comments.

Most economists have been content with estimating an average effect. But which average? Two have been suggested in the literature. Sometimes they have been confused. Under certain conditions, they are the same. When they are not, they differ in the assumptions that must be invoked to identify them in the presence of various selection rules.

The first average is the experimental treatment average: what is the effect of randomly (i.e., as in an ideal laboratory experiment) moving a worker from the nonunion sector to the union sector? In terms of equation (2), this parameter is

$$(3) \qquad (\alpha_1 - \alpha_0).$$

In the more general model, it is often defined as

$$(4) \qquad \tilde{X}_1\beta_1 - \tilde{X}_0\beta_0,$$

where $\tilde{X}_1$, $\tilde{X}_0$ are the endowments of the skills in each sector assumed available to the "average" worker in the unionized sector ($\tilde{X}_1 = E(X_1|D=1)$, $\tilde{X}_0 = E(X_0|D=1)$).

Despite the focus of most of the attention on these parameters in the recent empirical literature, they do not always answer economically interesting questions. It is also interesting to know what is the effect of unionism on the unionized. In the context of equation (2), the parameter is

$$(5) \quad E(Y_1 - Y_0|D=1, x, z) = (\alpha_1 - \alpha_0) + E(U_1 - U_0|D=1, z).$$

This is the gain from moving a unionized person with attributes $X = x$ and $Z = z$ from the nonunionized to the unionized sector. This is what Lewis calls the wage gap (p. 11), although he does not consistently define it this way throughout his survey. If $E(U_1 - U_0|D=1, z) = 0$, the two parameters are equivalent and instrumental variables estimators of $(\alpha_1 - \alpha_0)$ are consistent. (See my paper with Richard Robb, 1985; Gregg Duncan and Duane Leigh, 1985; and Christopher Robinson, 1989).

The distinction between (3) or (4) and (5) is crucial in assessing alternative selection bias estimators, and in evaluating union wage impacts. Yet these parameters are often confused. The wide variability in estimates that causes Lewis to dismiss econometric methods for solving self-selection problems arises in part because parameters (3) and (5) are confused in his comparison.

II. What Can Be Identified?

Recent advances in econometrics have focused attention on nonparametric and semiparametric estimation of economic models. Identification theorems are a necessary first step in this process, and provide useful discipline in separating out parameters of interest that can only be identified by invoking arbitrary functional form assumptions from those that are nonparametrically identified. In this section, I prove that, under conventional assumptions about sectoral wage and selection equations, a version of the Roy-Lee model of unionism is nonparametrically identified. Actually a more general model can be identified. To prove this, I adopt a slightly more general notation. Separate out the variables in common with X_1, X_0, and Z and call these X_c. The variables left over are relabeled X_1, X_0 and Z. Let

$$Y_1 = g_1(X_1, X_c) + U_1,$$

$$Y_0 = g_0(X_0, X_c) + U_0,$$

$$I = (Z, X_c)\gamma + V$$

where $I \geq 0$ implies $D = 1$, and $D = 0$ otherwise.

The following theorem can be proved. (It is based on a modification of a theorem in my paper with Bo Honoré, 1990.)

THEOREM: *Let* (U_1, U_0, V) *be median (or mean) zero i.i.d. random variables with density* $f(U_1, U_0, V)$ *distributed independently of* $X = (X_1, X_0, Z, X_c)$. $\mathrm{Var}(V) = 1$. *If*

(A) (U_1, U_0, V) *are continuously distributed with distribution function F and support equal to* R^3;

(B) *Support* $((Z, x_c)\gamma) = R^1$ *for all* x_c *in the support of* X_c. *There is no proper linear subspace of the space generating* (Z, X_c) *hav-*

ing probability 1 *with respect to the distribution of* (Z, X_c);

(C) *The marginal distributions of* (U_1, U_0, V) *have medians (means) equal to* 0. *Then* $F(u_1, v)$, $F(u_0, v)$, $g_1(x_1, x_c)$, $g_0(x_0, x_c)$ *and* γ *are identified from the distributions* $F(y_1|D=1, x_1, x_c, z)$, $F(y_0|D=0, x_0, x_c, z)$ *and the distribution of* D, $Pr(D=1|z, x_c)$.

PROOF:

Using $Pr(D=1|z, x_c)$ it is possible to mimic Charles Manski's (1988) proof of the identifiability of γ and $F(v)$ in the binary choice model. Using the distribution of sampled Y_1 and D, $Pr(Y_1 \leq y_1|D=1,\ X_1=x_1,\ X_c=x_c,\ Z=z) = Pr(g_1(x_1, x_c)+U_1 \leq y_1,\ V \geq -(z,\ x_c)\gamma)/Pr(V \geq -(z,\ x_c)\gamma)$. Fix $\underline{X}_c = x_c$. Define an isoprobability set for Z as $\bar{Z}_p = \{z\colon Pr(D=1|z, x_c) = p\}$. For each element in this set, define another set

$$\left(\bar{\bar{Y}}_1,\ \bar{\bar{X}}_1\right)_{q,p} = \{(y_1, x_1): Pr(Y_1 \leq y_1|D=1,\ X_1=x_1,\ X_c=x_c,\ Z=z) = q = Pr(g_1(x_1, x_c)+U_1 \leq y_1,\ V \geq -(z,\ x_c)\gamma)/Pr(V \geq -(z, x_c)\gamma)\}.$$

Using all x_1 and y_1 pairs within this set, it is possible to recover g_1 up to an additive constant. From knowledge of g_1 up to a constant, it is possible to trace out the distribution of (U_1, V) up to a constant for U_1 for each value of p, using all values of q. Using (C), it is possible to identify the constant. Since x_c was arbitrary this completes the proof for $g_1, \gamma, F(u_1, v)$. By a parallel argument, g_0 and $F(u_0, v)$ can be identified.

To identify β_1, β_0, augmented to be the coefficients of (X_1, X_c) and (X_0, X_c), respectively, it is necessary to assume that there is no proper linear subspace of the space generating (X_1, X_c) and (X_0, X_c) with probability one with respect to the distributions of these random variables. Under the conditions of the theorem, it is thus possible to nonparametrically identify $E(U_1 - U_0|D=1, z, x_c)$, and hence it is possible to nonparametrically identify the effect of unionism on the unionized

$$E(g_1 - g_0 + U_1 - U_0|D=1, z, x).$$

The crucial feature of (B) is the presence of a regressor in Z that traces out the probability of selection. The g_1 and g_0 functions can be constants as in the classical Roy model. The argument fails if we permit the distribution of U to depend on X in a general way. However, it is obviously possible to permit the distribution to depend on X_c in a general way. The theorem can be generalized in many ways. Normalizations other than $\mathrm{Var}(V)=1$, such as $\gamma'\gamma=1$, are clearly possible. The support of either U_1 or U_0 need not be R^1. Different variables may be in common in the Y_1, I equations than in the Y_0, I equations.

The intuition underlying this theorem is very simple. From Manski, the parameters of the sectoral choice probability $(Pr(D=1|z, x_c))$ can be identified. For each value of the sectoral choice probability (p), stratify the data on y_1 and (x_1, z, x_c). Fixing (z, x_c) so that $Pr(D_1=1|z, x_c)=p$, we may trace out a set of (x_1, y_1) values such that $Pr(Y_1 \leq y_1,\ \ V \geq -(z, x_c)\gamma) = q$. For each (x_1, x_c) in the set, we find the associated y_1. This identifies g_1 up to a constant c. Transforming the data by $y_1 - (g_1 + c)$, we can trace out the distribution $F(u_1, v)$. Doing this for all x_c, p, and q, using the mean (or median) zero assumption to identify c, and repeating the exercise for Y_0, we can identify the stated parameters of the theorem.

Note that, under the conditions of the theorem, it is not possible to identify the full joint distribution $F(u_1, u_0, v)$. This is intuitively reasonable since, by hypothesis, we never observe Y_1 and Y_0 for the same person. Placing a restriction on the decision rule such as the hypothesis that sectoral choices are made on the basis of income maximization, I becomes

$$I = g_1(X_1, X_c) - g_0(X_0, X_c) + U_1 - U_0.$$

Honoré and I establish under the conditions of the theorem that the joint distribution $F(u_1, u_0)$ can be nonparametrically identified so that it is possible to identify dependence between Y_1 and Y_0.

Identification results of the sort reported here are very fragile. If the Z regressors are finite valued, identification fails. The continuity of the underlying distribution and the assumption that $(z, x_c)\gamma$ can be varied over the real line play a crucial role in securing identification.

III. The Current State of the Art in Semiparametric Estimation of Selection Models

Identification is only a necessary first step toward estimation. Important progress on the consistent nonparametric estimation of selection models has been made by Stephen Cosslett (1990), Ronald Gallant and Douglas Nychka (1987), Hidehiko Ichimura and Lee (1990), Don Andrews (1989), Whitney Newey (1988), and James Powell (1989). It is interesting to compare what is theoretically possible, and economically interesting to know, with what has been produced in the recent literature. The models differ in the assumptions made and results that can be obtained. All authors except Gallant-Nychka assume that g_1 and g_0 are linear functions: $g_1 = (X_1, X_c)\beta_1$, $g_0 = (X_0, X_c)\beta_0$. Gallant-Nychka make smoothness and continuity assumptions about the distribution of the errors stronger than those used in my identification theorem, and produce consistent nonparametric estimators of the densities $f(u_1, v)$ and $f(u_0, v)$ and the parameters β_1, β_0 and γ. They do not produce a distribution theory for their estimator.

The other papers do not make strong independence assumptions and take as their point of departure the conditional mean-index-function representation of the sample selection problem. Focusing on the equation for Y_1,

$$E(Y_1|D=1, x_1, x_c, z) = (x_1, x_c)\beta_1 + E(U_1|D=1, z, x_c), \tag{6}$$

where $E(U_1|D=1, z, x_c) = \phi((z, x_c)\gamma)$ and where $(z, x_c)\gamma$ is an index function (minus the inverse function of $Pr(D=0|z, x_c)$ when V is independent of $(Z, X_c)\gamma$). The dependence of U_1 on Z and X_c comes strictly through the index. The objective of these papers is to eliminate the contaminating effect of $E(U_1|D=1, z, x_c)$ in forming regression estimates of β_1. Ichimura-Lee use the index property to jointly estimate β_1, γ by kernel methods. They assume access only to truncated samples and hence do not assume access to information on γ or $Pr(D=1|z, x_c)$ in forming their estimator. All of the other authors do. Andrews and Newey approximate the conditional mean of U_1 by series expansion methods assuming censored samples. Cosslett exploits the index function property and approximates $E(U_1|D=1, z, x_c)$, using step functions based on his nonparametric estimator of γ and the distribution of V. Powell uses kernel function methods and assumes censored sampling, and uses the information on γ to form approximate equivalence classes within which $Pr(D=1|z, x_c)$ and hence $E(U_1|D=1, z, x_c)$ are approximately constant. Within the approximate equivalence classes, he differences out the conditional mean. As sample size increases and kernel window widths contract, the approximation becomes arbitrarily good. All of these authors except Cosslett produce a large sample distribution theory for their estimators.

These papers present various identification criteria. The most agnostic assume truncated sampling so there is no information available to estimate $Pr(D=1|z, x_c)$. Ichimura-Lee forego identification of components of β_1 that are associated with variables in the index function. "ϕ" may be linear in $(z, x_c)\gamma$ and so such components are not identified. Models that exploit the index property more fully and assume access to censored samples require that there be one regressor in Z and that $\phi((z, x_c)\gamma)$ does not lie in the space spanned by X_1, X_c. Both groups of models absorb the intercept into the definition of $E(U_1|D=1, z, x_c)$. Thus *none* of these papers except for Gallant-Nychka produces consistent estimators of parameters (3), (4), or (5).

With enough variation in Z and continuity in the densities, it is possible to separate $E(U_1|D=1, z, x_c)$ from $(x_1, x_c)\beta_1$. An index function assumption strengthened with sufficient variation in Z will work. Thus for those Z such that $Pr(D=1|z, x_c)$ becomes arbitrarily close to unity, $E(U_1|D=1, z, x_c)$ becomes arbitrarily close to 0. With sufficient variation in the (X_1, X_c) for this set of Z, β_1 can be identified by least squares. (This is an instance of Gary Chamberlain's 1986 "identification at infinity.") Using the set of (z, x_c) that set $Pr(D=1|z, x_c)$ close to one, it is possible to modify Andrews' and Newey's procedure to consistently estimate the intercepts of equation (6). Rather than estimating all parameters on what may be a small probability set, I recommend a two-step procedure. 1) Exploit the results of Andrews and Newey to estimate the parameters that can be identified from the available distribution of the data. 2) Fixing these parameters, use the values of z and x_c that set $P(d=1|z, x_c)$ close to one, and in the limit becomes one, to identify and estimate the remaining parameters. The analogy to this strategy in density estimation is the difference between estimating a mode and a tail of a density. Different points of expansion estimate different parameters. The distribution theory for the second step of this order statistic estimator remains to be developed. With knowledge of $E(U_0|D=0, z, x_c)$ obtained from a parallel analysis of the Y_0 equation, it is possible to use an iterated expectation argument to exploit $E(U_0|z, x_c) = 0$ to find

$$E(U_0|D=0, z, x_c)$$
$$= -E(U_0|D=1, z, x_c)\frac{Pr(D=1|z, x_c)}{Pr(D=0|z, x_c)}.$$

From censored samples, it is in principle possible to estimate (3), (4), and (5). It is not necessary to invoke full independence assumptions. Index dependence of the conditional mean suffices. However, it is necessary to exploit more information than is currently used in any paper—except that of Gallant-Nychka. They replace identification at infinity with smoothness and independence conditions that effectively enable them to exploit the benefits of identification at infinity. Full independence allows the analyst to extrapolate out of the sample and to compute the full distribution and not just the mean of union wage impacts.

I am not necessarily advocating the imposition or use of this type of information. However, if it is not imposed, we abandon the pursuit of many economically interesting questions. If it is used, we know the circumstances under which we can address them. This is a fundamental advance in our knowledge about selection bias models.

IV. Simpler Methods May be Robust After All

As previously noted, if $E(U_1-U_0|D=1, z, x_c)=0$, instrumental variables estimators consistently estimate $\alpha_1-\alpha_0$ in equation (2) under standard conditions. Parameters (3) and (5) are the same. In the context of estimating union-nonunion wage gaps, there is accumulating evidence that instrumental variables procedures "work" in the sense of producing approximately the same estimate of parameters (3) and (4) as are obtained from more complicated sample-selection-correction models. There is considerable evidence, however, that D is endogenous. See Duncan-Leigh and Robinson for evidence on these points.

These studies indicate that unobserved (by the econometrician) components of U_1-U_0 contribute negligibly to the endogeneity of D. Several economic models are consistent with this result. Uncertainty about U_1-U_0 at the time union membership decisions are made is one plausible story. Selection on the basis of nonwage components (nepotism, discrimination, etc.) is another source (see Robinson for an excellent discussion of this point). The most general case, that is econometrically the hardest, appears not to be the empirically fruitful one.

REFERENCES

Andrews, Donald, "Asymptotic Normality of Series Estimators For Nonparametric and Semiparametric Regression Models," Discussion Paper No. 874R, Cowles Foundation, Yale University, June 1989.

Chamberlain, Gary, "Asymptotic Efficiency in

Semiparametric Models With Censoring," *Journal of Econometrics*, Spring 1986, *32*, 189–218.

Cosslett, Stephen, "Semiparametric Estimation of a Regression Model with Sample Selectivity," in W. A. Barnett et al., eds., *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press, 1990.

Duncan, Gregory and Leigh, Duane, "The Endogeneity of Union Status: An Empirical Test," *Journal of Labor Economics*, July 1985, *3*, 385–402.

Gallant, Ronald and Nychka, Douglas, "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, March 1987, *55*, 363–90.

Heckman, James and Honoré, Bo, "The Empirical Content of The Roy Model," *Econometrica*, forthcoming 1990.

______ **and Robb, Richard,** "Alternative Methods for Evaluating The Impact of Treatment on Outcomes," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press, 1985.

Ichimura, Hidehiko and Lee, Lung Fei, "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in W. A. Barnett et al., eds., *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press, 1990.

Lee, Lung Fei, "Unionism and Relative Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review*, June 1978, *19*, 415–33.

Lewis, H. Gregg, *Union Relative Wage Effects: A Survey*, Chicago: University of Chicago Press, 1986.

Manski, Charles, "Identification of Binary Response Models," *Journal of the American Statistical Association*, September 1988, *83*, 729–38.

Newey, Whitney, "Two Step Series Estimation of Sample Selection Models," Department of Economics, Princeton University, October 1988.

Powell, James, "Semiparametric Estimation of Censored Selection Models," unpublished manuscript, University of Wisconsin-Madison, July 1989.

Robinson, Christopher, "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy*, June 1989, *97*, 639–67.