NBER TECHNICAL PAPER SERIES

CORRECTING FOR TRUNCATION BIAS CAUSED BY A LATENT TRUNCATION VARIABLE

David E. Bloom

Mark R. Killingsworth

Technical Working Paper No. 38

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 1984

In addition to three anonymous referees, we thank the following individuals for helpful comments and suggestions, without implying that they are responsible for any of the opinions in this paper: John Abowd, Orley Ashenfelter, Henry Braun, Charles Brown, Stephen Cosslett, Henry Farber, William Greene, Gilles Greinier, John Ham, Jerry Hausman, James Heckman, Nicholas Kiefer, Steven Klepper, Robert Moffitt, Randall Olsen, Cordelia Reimers, Kenneth Small, David Stapleton, James Trussell and Clifford Winston. The research reported here is part of the NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #38
June 1984

Correcting for Truncation Bias Caused by a Latent Truncation Variable

ABSTRACT

We discuss estimation of the model

$$Y_i = X_i b_Y + e_{Yi}$$

$$T_i = X_i b_T + e_{Ti}$$

when data on the continuous dependent variable Y and on the independent variables X are observed iff the "truncation variable" $T > 0$ and when T is latent. This case is distinct from both (i) the "censored sample" case, in which Y data are available iff $T > 0$, T is latent and X data are available for all observations, and (ii) the "observed truncation variable" case, in which both Y and X are observed iff $T > 0$ and in which the actual value of T is observed whenever $T > 0$. We derive a maximum-likelihood procedure for estimating this model and discuss identification and estimation.

David E. Bloom
Department of Economics
Harvard University

Mark R. Killingsworth
Department of Economics
Rutgers, The State University

Consider estimation of the model

$$(1.1) \qquad Y_i = X_i b_Y + e_{Yi}$$

$$(1.2) \qquad T_i = X_i b_T + e_{Ti}$$

when data on the continuous dependent variable Y and on the independent variables X are observed iff the "truncation variable" $T > 0$ and when T is latent (i.e., not observed). This case is distinct from both (i) the "censored sample" case, in which Y data are available iff $T > 0$, T is latent and X data are available for all observations, and (ii) the "observed truncation variable" case, in which both Y and X are observed iff $T > 0$ and in which the actual value of T is observed whenever $T > 0$.[1]

In these alternative cases, one can estimate the parameters of (1.2) directly, and thus can make a selection bias correction to estimate (1.1) consistently using only observations with $T > 0$. When T is latent, however, there are no X (or T) data for "zeroes" (observations with $T \leq 0$), and the data for "ones" (observations with $T > 0$) exhibit no observed variation in T. Hence, in this case one cannot estimate directly the probability that an observation will have positive T or use existing methods for selection bias correction to estimate (1.1) consistently. However, as we now show, one can correct for selection bias and obtain consistent estimates of (1.1) even when T is latent and even when X is available only for observations in the truncated sample.[2]

First, the probability density of a Y value of Y_i for observation i in the population, conditional on its characteristics X_i and on the fact that it is in the truncated sample (has $T_i > 0$), is

$$(2) \qquad c_Y(Y_i \mid T_i > 0, X_i) = m_Y(Y_i \mid X_i) \Pr\{T_i > 0 \mid Y_i, X_i\} / \Pr\{T_i > 0 \mid X_i\}$$

where m_z and c_z are the marginal and conditional probability density functions for any variable z. We follow most previous work on censoring and truncation by assuming that e_Y and e_T are bivariate normal mean-zero random variables, uncorrelated with the X, with variances σ_{YY} and σ_{TT}, respectively, and with covariance σ_{TY}. Together with (2), this implies that

(3) $$m_Y(Y_i \mid X_i) = c_{e_Y}(e_{Yi} \mid X_i) = m_{e_Y}(e_{Yi}) = f(e_{Yi}/\sigma^{1/2}_{YY})/\sigma^{1/2}_{YY}$$

where f is the standard normal probability density function. Also, by (1.2),

(4) $$\Pr\{T_i > 0 \mid X_i\} = \Pr\{e_{Ti} > -X_i b_T \mid X_i\} = \int_{-J_i/\sigma^{1/2}_{TT}}^{\infty} f(t)dt = 1 - F(-J_i/\sigma^{1/2}_{TT})$$

(5) $$\Pr\{T_i > 0 \mid Y_i, X_i\} = \Pr\{e_{Ti} > -X_i b_T \mid e_{Yi}, X_i\}$$

$$= \int_{-K_i/\sigma^{1/2}_{TT\cdot Y}}^{\infty} f(t)dt = 1 - F(-K_i/\sigma^{1/2}_{TT\cdot Y})$$

where F is the standard normal cumulative density function; $J_i = X_i b_T$; $K_i = X_i b_T + \mu_{T\cdot Y}$; $\mu_{T\cdot Y} = \sigma_{TY} e_{Yi}/\sigma_{YY}$ = the mean of e_{Ti} conditional on e_{Yi}; and $\sigma_{TT\cdot Y} = \sigma_{TT}[1-(\sigma^2_{TY}/\sigma_{TT}\sigma_{YY})]$ = the variance of e_{Ti} conditional on e_{Yi}.

Hence, (1)-(5) imply that the likelihood for a sample of observations' Y conditional on their X and given that they are all in the population subset t that is in the truncated sample[3] is

(6) $$L(Y_i \mid T_i > 0, X_i)$$

$$= \prod_{i \epsilon t} [f(e_{Yi}/\sigma^{1/2}_{YY})/\sigma^{1/2}_{YY}][1-F(-K_i/\sigma^{1/2}_{TT\cdot Y})]/[1-F(-J_i/\sigma^{1/2}_{TT})]$$

Note also that the regression function implicit in (1.1) is

$$(7)\qquad E[Y_i \mid T_i > 0,\ X_i] = X_i b_Y + [\sigma_{YT}/\sigma_{TT}^{1/2}][f(X_i b_T/\sigma_{TT}^{1/2})/F(X_i b_T/\sigma_{TT}^{1/2})]$$

(7) usually arises in the analysis of censored samples (see Heckman, 1979, p. 156). However, it is equally applicable to truncated samples.[4]

As regards identification, note first that both (6) and (7) are homogeneous of degree zero in b_t, σ_{TT} and σ_{YT}: that is, these parameters are identifiable only up to a constant of proportionality. Accordingly, σ_{TT} may be normalized to unity without loss of generality.

Second, b_T is not identified if $\sigma_{TY} = 0$. In this case, $\sigma_{TT\cdot Y} = \sigma_{TT}$ and $J_i = K_i$. If so, the second term in the numerator of (6) and the denominator of (6) are equal for all observations and (6) collapses to

$$(8)\qquad L(Y_i \mid T_i > 0,\ X_i) = \prod_{i\varepsilon t} [f(e_{Y_i}/\sigma_{YY}^{1/2})/\sigma_{YY}^{1/2}]$$

which is exactly the likelihood function implicit in OLS estimation. Similarly, the truncated regression function (7) reduces to

$$(9)\qquad E(Y_i \mid T_i > 0,\ X_i) = X_i b_Y$$

In both cases, however, if $\sigma_{TY} \neq 0$, then use of (8) (or (9)) rather than (6) (or (7)) will result in inconsistent estimates of the parameters of (1.1).

Third, following Fisher (1966, especially Chapter 5), we can approximate the nonlinear terms of (7) (with σ_{TT} normalized) to any desired degree of accuracy by a Maclaurin expansion of arbitrarily high degree. The resulting expression can be written in a form that is nonlinear in X but linear in a new set of parameters (that are nonlinear functions of the original parameters). Except in one special case, X enters this transformed equation in an arbitrarily large number of terms all of which are, however, nonlinearly related because of the nonlinearity of the normal distribution. Thus, subject to the normalization $\sigma_{TT} = 1$, the parameters of the truncated regression model (7) are identified.[5]

The sole exception to this result on identification arises when constants are included among the X's in both (1.1) and (1.2) but $b_T = 0$ for all nonconstant X's (e.g., when the only parameter in $X_i b_T$ is an intercept term). In this case, the ratio f(•)/F(•) in (7) is a constant for all observations in the truncated sample and so the intercept term in (1.1) is not identified. (To see this, examine the expansion of (7) and note the existence of a linearity when $b_T = 0$ for all nonconstant X.)

In principle, estimation of (7) by nonlinear least squares (NLS) or (6) by maximum likelihood (ML) yields consistent estimates of the parameters of (1). An advantage of ML is that the truncated regression function (7) is heteroskedastic (so that NLS will yield incorrect standard errors), whereas correct asymptotic standard errors can be computed directly from the inverted Hessian obtained via ML. Using an earlier version of this paper, Greene (1982) has derived a routine to estimate (6) by ML and has included it in a flexible econometric software package.[6]

We note, in passing, that (6) (or (7)) generalizes the estimator proposed by Heckman (1979) for censored data. Under Heckman's approach, the parameters b_T are estimated in a first-stage probit analysis using data for the entire population. In the second stage of the procedure one estimates (7), using data for the population subset that has T > 0, with the b_T constrained to equal the values obtained in the first stage. However, one could instead estimate (6) (or (7)) directly, using observations with T > 0, but without a first-stage probit analysis. This has the advantage of not constraining the b_T. The disadvantage of this application of (6) to censored data is that it estimates b_T using fewer observations than are used in the Heckman two-stage procedure.

Footnotes

[1] On (i), see Heckman (1979). On (ii), see Amemiya (1973, p. 1015, esp. equations (10)), Hausman and Wise (1976), and Wales and Woodland (1980).

[2] If aggregate data are available on the X characteristics for the general population, one may use the work of Cosslett (1981), who proposes fairly general estimators that make efficient use of such information. Here, however, we are concerned with cases in which such aggregate X data for the population are not available.

[3] Note that (6) is identical in form to the expression in equation (21) in Wales and Woodland (1980). However, Wales and Woodland treat their (21) only as a building-block in the construction of other likelihood functions that require more information than does (21), and do not suggest that their (21) can itself be used to estimate the parameters of their function for Y.

[4] We owe this point to an anonymous referee.

[5] Thus, in the present case, no exclusion restrictions (e.g., that at least one variable that appears in (1.1) may not appear in (1.2) and/or vice versa) are required for identification. What identifies (6) (or (7)) is the nonlinearity of the normal distribution, although the model will be identified under other nonlinear distributions as well. A caveat is in order here: in models like the one developed here, parameter estima-

tes may be sensitive to distributional assumptions. (For example, see Olsen, 1982).

6. After completing this paper, we become aware of the work of Muthen and Joreskog (1983), who also derive the likelihood function (6) and present Monte Carlo results. However, they do not discuss identification of (6) or its non-linear least squares analogue, (7).

REFERENCES

Amemiya, T. (1973), "Regression Analysis When the Dependent Variable is Truncated Normal," Econometrica 41: 997-1016.

Cosslett, S.R. (1981), "Efficient Estimation of Discrete-Choice Models," pp. 51-111 in Manski and McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, Cambridge, MA: MIT Press.

Fisher, F.M. (1966), The Identification Problem in Econometrics, New York: McGraw-Hill.

Greene, W. (1982), LIMDEP Manual, Department of Economics, New York University.

Hausman, J.A. and D.A. Wise (1976), "The Evaluation of Results from Truncated Samples: The New Jersey Income Maintenance Experiment," Annals of Economic and Social Measurement 5: 421-446.

Heckman, J.J. (1979), "Sample Selection Bias as a Specification Error," Econometrica 47: 153-162.

Muthén, B. and K.G. Jöreskog (1983), "Selectivity Problems in Quasi-Experimental Studies," Evaluation Review 7: 139-174.

Olsen, R.J. (1982), "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator," International Economic Review 23: 223-240.

Wales, T.J. and A.D. Woodland (1980), "Sample Selectivity and the Estimation of Labour Supply Functions," International Economic Review 21: 437-468.