Econometrica, Vol. 49, No. 3 (May, 1981)

NOTES AND COMMENTS

SAMPLE SELECTION BIAS AS A SPECIFICATION ERROR: COMMENT

BY WILLIAM H. GREENE[1]

IN A RECENT PAPER, J. J. Heckman [**1**] describes a general model for the treatment of sample selection bias. He presents a consistent two step estimator and formulas for appropriate asymptotic standard errors. The second step consists of ordinary least squares applied to the original model augmented by a constructed regressor which accounts for the bias due to nonrandom sampling. (The procedure is described in detail in [**1**, p. 157].) It is claimed that the conventionally computed standard errors for the second step estimator always underestimate the correct asymptotic standard errors. We show in this note that the claim is incorrect and that the conventional "incorrect" standard errors can either be larger or smaller than their "correct" counterparts. As a consequence, the usual OLS standard errors cannot be used as lower bounds in statistical inferences.

For practical applications, the formulas given in [**1**] represent a straightforward but rather involved and potentially very expensive calculation (if the sample size is large). We show below, however, that Heckman's asymptotic covariance matrix can be considerably simplified. The calculations can be done using a small set of moment matrices, and are not much more burdensome than the usual least squares computations.

To begin, it is necessary to correct two typographical errors on page 159 of [**1**]; in the definitions of $\theta_{ii'}$ and $\Omega_{ii'}$, $X_{2i}\sum X_{2i}'$ should be $X_{2i}\sum X_{2i'}'$. Henceforth, for ease of exposition, the subscript i' will be replaced with j and the apostrophe will be used only to denote a column vector. Now, with these corrections, we may write

$$\psi = \operatorname*{plim}_{I_1, I\to\infty} (\psi_1 + \psi_2), \text{ where}$$

$$\psi_1 = \frac{\sigma_{11}}{I_1}\sum_{i=1}^{I_1}\eta_i\begin{bmatrix} X_{1i}'X_{1i} & X_{1i}'\lambda_i \\ \lambda_i X_{1i} & \lambda_i^2 \end{bmatrix}$$

and

$$\psi_2 = \frac{C^2}{II_1}\sum_{i=1}^{I_1}\sum_{j=1}^{I_1}\theta_{ij}\begin{bmatrix} X_{1i}'X_{1j} & X_{1i}'\lambda_j \\ \lambda_i X_{1j} & \lambda_i\lambda_j \end{bmatrix},$$

where η_i and θ_{ij} are defined on p. 159. (Note, this first simplification obviates π_{ij} and Ω_{ij}.)

[1] The helpful comments of J. J. Heckman and an anonymous referee are gratefully acknowledged.

It is claimed [**1**, p. 159] that because ψ_2 is positive definite when C is nonzero, $B\psi B$ which is the correct asymptotic covariance matrix when λ_i must be estimated, gives larger standard errors (has larger diagonal elements) than the standard covariance matrix, $\sigma_{11}B$. Thus, the usual procedure (least squares) understates the standard errors of the parameter estimates and overstates significance levels. The claim is not true.

Consider, first, the case in which X_{2i} is 1 for all i. That is, the selection is random, with probability of inclusion in the selected sample equal to $\Phi(\beta_2/\sigma_2)$, a constant. (This presumes the case in Heckman's footnote 7, in which X_{1i} does not contain a constant term.) In this case, Z_i, η_i, λ_i, and θ_{ij} all become constants, Z, η, λ, and θ respectively. It is readily verified that in this case, $\text{plim}\,\psi_1$ reduces to $\eta\sigma_{11}B^{-1}$. Also, ψ_2 can now be written

$$(*)\quad \psi_2 = \theta C^2 \frac{I_1}{I}\begin{bmatrix} \bar{X}_1'\bar{X}_1 & \bar{X}_1'\lambda \\ \lambda\bar{X}_1 & \lambda^2 \end{bmatrix}$$

where $\bar{X}_1'$ is the $(K_1 \times 1)$ vector of sample means of the regressors in the selected sample. Now, suppose, in addition, that the data are measured so that $\bar{X}_1 = 0$.[2] Then, with λ a constant, B becomes block diagonal, as does B^{-1}. Moreover, ψ_2 becomes

$$\psi_2 = \begin{bmatrix} 0 & 0 \\ 0 & \theta C^2 k\lambda^2 \end{bmatrix},$$

where k is defined in [**1**, p. 159]. Combining all of these, we obtain

$$B\psi B = \begin{bmatrix} \eta\sigma_{11}\left(\underset{I_1\to\infty}{\text{plim}} \sum_{i=1}^{I_1} \frac{X_{1i}'X_{1i}}{I_1}\right)^{-1} & 0 \\ 0 & \sigma_{11}\left(\frac{\eta + \theta\rho^2 k}{\lambda^2}\right) \end{bmatrix}.$$

The northwest block gives the asymptotic covariance matrix for $\sqrt{I_1}\,(\hat{\beta}_1 - \beta_1)$. Since $C^2/\sigma_{11} = \rho^2 < 1$, and $-1 < Z_i\lambda_i - \lambda_i^2 < 0$ (see [**1**, p. 157]), $0 < \eta < 1$. Thus, for the assumed case, the correct asymptotic covariance for $\sqrt{I_1}\,(\hat{\beta}_1 - \beta_1)$ is $\eta\sigma_{11}B$, each diagonal element of which is smaller than the corresponding value in the standard estimator $\sigma_{11}B$, not larger. For the estimator $\sqrt{I_1}\,(\hat{C} - C)$, the comparison can go either way; the correct standard error is smaller if $\eta + \theta\rho^2 k < 1$, which is not precluded. For this simple case, it can be shown that η and θ are both functions of k, and for $k = 1/2$ the above inequality holds for all ρ^2. Thus, the OLS standard errors may also overestimate the correct standard error for $\hat{C}$.

The more general case can be analyzed by assuming that Σ, the asymptotic

[2] This simplifying normalization was suggested by J. J. Heckman.

covariance matrix of $\sqrt{I}(\hat{\beta}_2^* - \beta_2^*)$ is small enough that ψ is dominated by ψ_1. (The matrix Σ enters the computations only through θ_{ij} which can vary independently of η_i.) Since $0 < \eta_i < 1$, for small values of η_i (which occur when ρ^2 is large but the average probability of inclusion in the selected sample is small), the case of $B\psi B$ having smaller diagonal elements than $\sigma_{11}B$ is the logical possibility. Of course, the comparison can also go the other way if Σ is large and η_i tends to be large. The net result is that the OLS standard errors cannot be used to form bounds for appropriate test statistics.

The actual computation of ψ appears at first sight to be rather formidable. As formulated in [**1**] and above, the computation of ψ_1, requires one pass through the data set; but ψ_2 is the sum of I_1^2 rank 1 matrices, and its computation appears to require I_1 passes through the data. In fact, ψ_2 can be computed in a single pass as follows: For convenience in notation, define the vector

$$W'_{1i} = \left[C\sqrt{k}\,(\partial\lambda_i/\partial Z_i) I_1 \right] \begin{bmatrix} X'_{1i} \\ \lambda_i \end{bmatrix}.$$

Then, we may write ψ_2 as

$$\psi_2 = \sum_{i=1}^{I_1} \sum_{j=1}^{I_1} W'_{1i} W_{1j} (X_{2i} \Sigma X'_{2j}).$$

Since the term $X_{2i}\Sigma X'_{2j}$ is a scalar (and Σ is a constant matrix) we may move it between W'_{1i} and W_{1j} and rewrite ψ_2 as

$$\begin{aligned} \psi_2 &= \sum_{i=1}^{I_1} \sum_{j=1}^{I_1} W'_{1i}(X_{2i}\Sigma X'_{2j}) W_{1j} \\ &= \sum_{i=1}^{I} \sum_{j=1}^{I} (W'_{1i}X_{2i})\Sigma(X'_{2j}W_{1j}) \\ &= \left(\sum_{i=1}^{I_1} W'_{1i}X_{2i} \right) \Sigma \left(\sum_{j=1}^{I_1} X'_{2j}W_{1j} \right) \\ &= F\Sigma F' \end{aligned}$$

where F is simply the matrix of cross moments between W_{1i} and X_{2i}.[3] The entire computation can be done in a single pass through the data. Let $r_i = (X_{1i}\hat{\lambda}_i)$, and $r_i^* = (\sqrt{\hat{\eta}_i \hat{\sigma}_{11}})r_i$. Then $\hat{B}$ is the inverse of the second moment matrix of r_i, $\hat{\psi}_1$ is the second moment matrix for r_i^*, and $\hat{F}$ is the cross products matrix of W_{1i} and X_{2i}. The matrix Σ is produced in the first step of Heckman's procedure. The correct

[3] We note in passing that ψ_2 is not positive definite when $C \neq 0$; it is positive semidefinite. When $K_2 < (K_1 + 1)$, the $(K_1 + 1) \times (K_1 + 1)$ matrix ψ_2 has rank K_2, i.e., it is short ranked. For example, see equation $(*)$, in which ψ_2 has rank 1.

estimate of the asymptotic covariance matrix for

$$\begin{bmatrix} \hat{\beta}_1 \\ C \end{bmatrix}$$

is, then,

$$\hat{\Gamma} = \frac{1}{I_1}\hat{B}[\hat{\psi}_1 + \hat{F}\hat{\Sigma}\hat{F}']\hat{B}.^4$$

We note, finally, that in practice, a rather more serious difficulty can arise. All of the above are limiting results; but for practical use, Γ must be estimated using a finite sample. In the computation of $\eta_i = 1 + \rho^2(Z_i\lambda_i - \lambda_i^2)$, ρ^2 is estimated using $\hat{C}^2/\hat{\sigma}_{11}$. Even if the consistent estimator of σ_{11} on page 157 (in which $\hat{C}$ should be $\hat{C}^2$) is used, $\hat{\rho}^2$ can be greater than one in a finite sample.[5] Now, $\hat{Z}\hat{\lambda} - \hat{\lambda}^2$ is always in $(-1, 0)$, so that $\hat{\eta}_i$ could be negative for some or even all i, even though η_i is always in $(0, 1)$. As a consequence Γ can have negative diagonal elements. Given the earlier discussion, there would appear to be no safe haven to which we can retreat in this instance.

Cornell University

Manuscript received October, 1979; revision received April, 1980.

[4] A program which computes all of the estimators and this asymptotic covariance matrix is available from the author at nominal cost.

[5] Heckman notes this possibility in a footnote on page 15 of an earlier version of this paper, NBER Working Paper No. 172, March, 1977.

REFERENCE

[1] HECKMAN, J. J.: "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1979), 153–161.