TOOLS FOR INTUITION ABOUT SAMPLE SELECTION BIAS AND ITS CORRECTION *

Ross M. Stolzenberg
University of Chicago

Daniel A. Relles
RAND Corporation

We provide mathematical tools to assist intuition about selection bias in concrete empirical analyses. These new tools do not offer a general solution to the selection bias problem; no method now does that. Rather, the techniques we present offer a new decomposition of selection bias. This decomposition permits an analyst to develop intuition and make reasoned judgments about the sources, severity, and direction of sample selection bias in a particular analysis. When combined with simulation results, also presented in this paper, our decomposition of bias also permits a reasoned, empirically-informed judgment of when the well-known two-step estimator of Heckman (1976, 1979) is likely to increase or decrease the accuracy of regression coefficient estimates. We also use simulations to confirm mathematical derivations.

Sample selection is said to occur when data for a variable are missing for some cases and present for others. Heckman (1976) made sociologists acutely aware that sample selection bias on a dependent variable in a regression can cause severe bias in estimates of regression coefficients. A variety of sample selection bias corrections is now available, as is extensive information about the deficiencies, difficulties, and limitations of all of these correction techniques (Winship and Mare 1992).[1] At present, no technique or combination of techniques appears to offer universal or even predictable rescue from the sometimes severe problems of selection bias. Further, simulation studies and analyses of survey data provide both anecdotal and systematic evidence that these corrections can and do go awry under ordinary circumstances, sometimes grossly worsening estimates rather than improving them, without providing any indication that a problem has occurred. Thus, in spite of the elegant models and inventive methods brought to bear on sample selection bias, researchers often have little more than their intuition to guide them.

For the simplified situations usually found in introductory regression texts, Berk (1983) showed that it is fairly easy to examine selection bias pictorially and to develop intuition about the cause, size, and direction of selectivity effects. However, Berk also showed that this intuition is difficult to obtain from pictures when examples are only slightly more complicated. As we show in this paper, the usual mathematics of Heckman's (1976) selection model does not lend itself to intuitive thinking either, in part because it relies on recalculating regression coefficients after adding a new independent variable, the inverse Mills Ratio. This ratio is a nonlinear and otherwise generally unfamiliar function based on the results of a probit analysis of selection in the sample over which the regression is estimated. As a consequence, it is often difficult to develop intuition about selection bias in any particular empirical analysis.

Our main purpose in this paper is to provide mathematical tools to assist intuition about sample bias in concrete empirical

* Direct Correspondence to Ross M. Stolzenberg (r-stolzenberg@uchicago.edu). This research was supported by the National Institute for Child Health and Human Development (Grant No. P50 HD12639), The RAND Corporation, and the Graduate Management Admission Council. The comments and criticisms of reviewer George Farkas and of Deputy Editor Trond Petersen materially improved this work.

[1] See Duan et al. (1984), Goldberger (1980), Lillard, Smith, and Welch (1986), Little (1985), Little and Rubin (1987), Nelson (1984), Paarsch (1984), Stolzenberg and Relles (1990). Wainer (1986) records sharp debate between Heckman and Robb, Glynn, Laird and Rubin, and Tukey.

analyses. These new tools are not a general solution to the selection bias problem; no method provides that at this time. But they do offer a new approximation and decomposition of the bias. Most important, this decomposition gives a simplified but still rigorous explanation of how selection bias occurs and how the usual two-step selection correction can worsen estimates. Mathematically, the new approximation we offer is the product of a few familiar or easily computed quantities. Most of these quantities can be calculated from the data under analysis, while ranges of reasonable values must be selected for others. We use simulation studies to test the approximation methods we present here. These simulations also provide additional useful information about the circumstances under which selection bias corrections are likely to improve estimation.

We begin with an anatomy of the two-step selection correction. Then we present derivations that simplify its mathematics, simulation studies that confirm the accuracy of the mathematics, and conclusions about appropriate strategies for dealing with selection bias.

A MODEL OF CENSORING BIAS

Heckman's Model

For convenience, our selection model follows Heckman (1976:476, 1979:154). We consider the case with only one independent variable (described in Appendix A). Appendix B considers the multiple regression case, for which the derivations are tedious, although results are nearly as simple.

Let equation 1 be a *regression equation* of substantive interest:

$$Y_1 = \beta_0 + \beta_1 X + \sigma\varepsilon. \qquad (1)$$

X is the regression independent variable, Y_1 is the regression dependent variable, and $\sigma\varepsilon$ is the regression error term, where σ is a scalar and ε is normally distributed with a mean of 0 and a variance of 1 ($N_{(0,1)}$). β_0 and β_1 are regression coefficients.

For the same data for which equation 1 is defined, we also define equation 2, which is called the *selection equation:*

$$Y_2 = \alpha Z + \delta. \qquad (2)$$

δ is also normally distributed $N_{(0,1)}$. Z is the selection equation independent variable, and α is the coefficient of Z. Z may be identical to X . T is a scalar called the selection threshold. The value of Y_1 is observed for some cases (selected cases) and is missing for other cases (censored cases). A data case is selected if $Y_2 > T$ for that case; the case is censored if $Y_2 \leq T$. We do not need a constant term in the selection equation because that term is absorbed into the selection threshold, T.

In a variation on the classic example of sample selection, we might wish to estimate a regression of earnings (Y_1) of married women on their years of schooling (X). However, for purposes of this example only, assume that many married women can choose between paid work in the labor market and unpaid work at home. Also for purposes of this example, assume that women make this choice on the basis of the occupational socioeconomic status (Y_2) provided by the job they would obtain *if they chose market work.* If Y_2 exceeds some threshold (T), then women choose market work and their earnings are observed; otherwise, the women lack observable earnings data, and they fall from the sample when the regression of Y_1 is computed.

For a given value of Z, the probability of selection is determined by T, α, and the random error term δ. The larger the value of α, the more that sample selection depends on the value of Z. The larger the value of T, the lower the probability that *any* observation will be selected, regardless of the value of Z. If α equals 0, then selection is random and merely reduces sample size. If T equals $-\infty$, then all cases are selected no matter how large the value of α. If T equals $+\infty$, then no cases are selected, no matter how small the value of α.

Heckman (1976) observed that there is a potential bias in using only selected cases to estimate equation 1. He computed the conditional expectation of Y_1 *given that* Y_1 *is observed*, as:

$$E(Y_1 \mid Y_2 > T) = \beta_0 + \beta_1 X + \sigma\, \rho_{\varepsilon\delta}\, \lambda(T - \alpha Z), \qquad (3)$$

where $\rho_{\varepsilon\delta}$ is the correlation between ε and δ, and λ is the reciprocal of the Mills Ratio function as defined in equation 4.

$$\lambda(T-\alpha Z) = \phi(T-\alpha Z)/[1-\Phi(T-\alpha Z)], \quad (4)$$

where $\phi(T-\alpha Z)$ is the normal density function (the height of the normal curve) evaluated at $T-\alpha Z$, and $\Phi(T-aZ)$ is the normal cumulative distribution function (the area under the normal curve) evaluated at $T-\alpha Z$. Thus, in the presence of selection, the original regression equation is not appropriate because it omits an independent variable that belongs in the equation. This omitted variable is $\lambda(T-\alpha Z)$, and its coefficient is an estimate of $\sigma\ \rho_{\varepsilon\delta}$. If $\lambda(T-\alpha Z)$ is substantially correlated with X and Y_1, then, instead of estimating

$$Y_1 = \beta_0 + \beta_1 X + \sigma\varepsilon, \quad (5)$$

we should estimate

$$Y_1 = \beta_0 + \beta_1 X + \sigma\,\rho_{\varepsilon\delta}\,\lambda(T-\alpha Z) + \sigma'\varepsilon', \quad (6)$$

where ε' has a mean of 0 but is not normally distributed, and σ' is a constant not necessarily equal to σ.

Heckman (1976) noted that $T-\alpha Z$ could be estimated as the predicted values in a probit analysis in which the independent variable is Z and the dependent variable is a dummy that equals 0 if Y_1 has a missing value and 1 if the value of Y_1 is not missing. Heckman's "two-step" correction consists of using probit analysis to estimate the value of $T-\alpha Z$ for each data case, calculating the inverse Mills Ratio $\lambda(T-\alpha Z)$ from those estimates for each data case, and then using $\lambda(T-\alpha Z)$ as an additional regressor in equation 1. Thus, in more familiar notation, we estimate

$$Y_1 = \beta_0 + \beta_1 X + \beta_2 M + \sigma'\varepsilon', \quad (7)$$

where M is $\lambda(T-\alpha Z)$ and $\sigma'\varepsilon'$ is the error term for the regression.

Equation 7 is a multiple regression equation in which one independent variable is a ratio of two nonlinear transformations of the predicted values of a probit analysis computed from an overlapping but different data set than that used to calculate the regression equation of interest. *We think it is very difficult to have much intuition about the coefficients of an equation such as this.* In the absence of intuition, one can only use the two-step correction blindly in the hope that it works properly. When all goes well this is a good way to proceed. But when the two-step estimator produces substantive results that seem to contradict reason, then intuition is essential to learn if the problem is caused by substantive or methodological errors.

Next we show some reasons why the two-step estimator is particularly prone to methodological problems. Then we develop an approximation to the two-step selection correction that is mathematically sound and straightforward enough to support intuitive thinking about selectivity bias in regression.

How Things Go Wrong

Under seemingly ordinary circumstances, even when its assumptions and formal requirements are satisfied, the two-step selection bias correction is known to sometimes produce estimates that are farther from true parameter values than estimates obtained by uncorrected ordinary least squares (Hartman 1991).[2] It is not possible to know how often such problems occur, but they do not seem to be rare (Lillard et al. 1986), and sometimes they are catastrophic (Stolzenberg and Relles 1985).

We now consider specific mechanisms that cause these difficulties. According to Equation 7, ignoring sample selection amounts to failing to include M as an independent variable. Elementary properties of regression indicate that if X and M are not correlated, then this omission—and therefore sample selection—does not bias the estimate of β_1. Those same properties of regression indicate that

[2] Technically this is not a failing of the estimator, since its aim is to reduce bias rather than to increase efficiency. But in most situations in which substantive arguments are evaluated by analyzing a single sample of data, bias cannot be distinguished from other sources of error, and the estimator with the smallest total error is preferred. So, as a practical matter, we say that a selection bias correction method "goes wrong" *in a particular analysis* if it produces an estimate of β_1 that is farther from the population value of β_1 than the biased OLS estimate it is intended to correct. In a distribution of the analyses of a number of different samples, bias can be distinguished from random error, and bias and random error can be attacked separately. Hartman (1991) offers a simulation-based comparison of various estimators, including the least-squares and maximum-likelihood versions of Heckman's (1976) model.

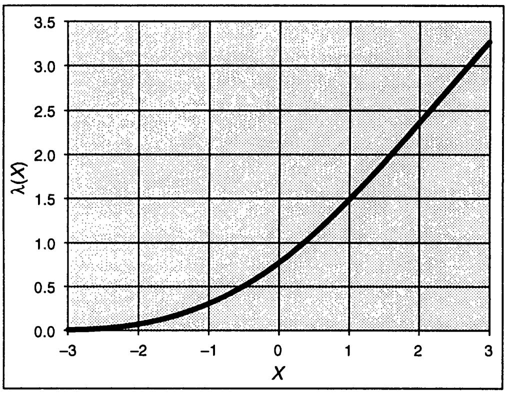

Figure 1. Mills Ratio of X versus X

two things happen as the correlation between X and M increases: First, the estimate of β_1 is increasingly affected by including M in the regression (i.e., bias occurs if M is omitted). And, second, random error in the estimate of β_1 increases when M is included in the regression. As the random error component of the estimate of β_1 grows large, the bias-corrected estimate can become unstable enough to have a substantial chance of being farther from the true population value of β_1 than the OLS estimate it is intended to correct. At extremely high correlations between X and M, muticollinearity or near multicollinearity among regressors occurs, and coefficient estimates become volatile, often taking on substantively ridiculous values characteristic of regressions afflicted by multicollinearity. These problems are not peculiar to the two-step estimator; they tend to occur in any regression in which independent variables are highly correlated. For example, these problems occur frequently in polynomial regressions in which X, X^2, and X^3 are regressors. Confounding the coefficients for X, X^2, and X^3 is acceptable when it is sufficient to know the *combined* effects of all powers of X. However, in correcting for selection bias, the essential purpose is to distinguish the coefficient of X from the coefficient of M.

Why is X so likely to be highly correlated with M? Recall that $M = \lambda(T-\alpha Z)$. X is highly correlated with $\lambda(T-\alpha Z)$ because X and Z are highly correlated and because λ is reasonably linear over the fixed range of values taken by $T-\alpha Z$ in virtually any data set. X and Z are usually highly correlated, or even identical, because variables that cause Y_1 are often the same variables that cause selection. For example, in the hypothetical analysis described above in which Y_1 is earnings and Y_2 is occupational SES, the causes of Y_1 and the causes of Y_2 might be similar, or even the same. And although λ is a nonlinear function and its nonlinearity identifies the selection parameter, its variation in any particular sample of data is approximately linear. Figure 1 plots the function $\lambda(X)$, the inverse of the Mills Ratio function, for values of X, showing its approximate linearity.

Table 1 reports some simple simulation-based estimates of the correlation between $\alpha Z-T$ and $\lambda(T-\alpha Z)$ over different ranges of $\alpha Z-T$. In each simulation, we define a range for $\alpha Z-T$ and then create a data set consisting of (1) 1,000 data points that are equally spaced over this range, and (2) the corresponding value of $\lambda(T-\alpha Z)$ for each data point. We compute the correlation between the data points and the corresponding values of $\lambda(T-\alpha Z)$. Notice that the absolute values of all these correlations exceed .96, and all but two are larger than .99.

Because of the high correlations between Z and $\lambda(T-\alpha Z)$, the correlation between X and Z *that is necessary for substantial selection bias to occur* also tends to increase estimation error substantially, or even introduce collinearity problems when the two-step selection correction is applied. The two-step estimator is a delicate balance of selection bias against errors introduced by adding a regressor that is highly correlated with the regressor of substantive interest.

Table 1. Illustrative Simulated Correlations between $\alpha Z - T$ and $\lambda(T-\alpha Z)$ over Various Ranges of $\alpha Z - T$

Range of $\alpha Z-T$	Correlation between $\alpha Z-T$ and $\lambda(T-\alpha Z)$
(0,1)	–.9996
(–1,1)	–.9960
(0,2)	–.9992
(–1,2)	–.9957
(–2,2)	–.9836
(0,3)	–.9991
(–1,3)	–.9961
(–2,3)	–.9865
(–3,3)	–.9665

Insights from Approximating λ in Simple Regressions

We clarify and simplify the effect of selection bias by using a Taylor series to approximate the function λ. Once the Taylor series approximation and some algebraic manipulation is complete, the individual contributions of each parameter in equations 1 and 2 to the selection bias in β_1 can be seen more clearly, as shown in equation 8. In equation 8, we use $\aleph_{\beta_1}$ (aleph-beta-one) to denote the bias in estimation of β_1 that would occur if no correction for sample selection bias were applied. Thus,

$$\aleph_{\beta_1} \equiv -\lambda'(T-\alpha\bar{Z})\sigma\rho_{\varepsilon\delta}\,\rho_{xz}\,\alpha(s_Z/s_X), \quad (8)$$

where ρ_{XZ} is the correlation between the independent variables in the regression equation (X) and selection equation (Z), s_X and s_Z are the standard deviations of X and Z in the selected part of the sample, $\bar{Z}$ is the mean of Z, and other variables are as defined before. λ' is the first derivative of the reciprocal of the Mills Ratio function. Appendix B derives the multiple regression form of equation 8, (see equation B-6).

Notice that equations 8 and B-6 do *not* require calculation of the inverse Mills Ratio or its first derivative for every case. In the simple regression case, the argument λ' can be calculated by substituting the mean of Z into the estimated probit equation to obtain $\hat{Z} = \alpha\bar{Z} - T$ and multiplying the result by −1. In the multiple regression case, the argument $\lambda'(T - \hat{\alpha}_1\bar{z}_1 - \ldots - \hat{\alpha}_q\bar{z}_q)$ can be calculated by substituting the means of $z_1, z_2, \ldots, z_q$ into the estimated probit equation and multiplying the result by −1. An easier method, depending on the computer program used to calculate the probit analysis, is to simply retain the predicted values from the probit equation, take their mean, and then multiply by −1 (because α and T are constants, the mean of $\alpha Z - T$ equals $\alpha\bar{Z} - T$).[3] Once the argument of λ' is obtained, $\lambda'(T-\alpha\bar{Z})$ or $\lambda'(T - \hat{\alpha}_1\bar{z}_1 - \ldots - \hat{\alpha}_q\bar{z}_q)$ is easily calculated with a spreadsheet program, statistical analysis program, or a table of normal density and normal cumulative distribution functions.[4]

Equation 8 shows the conditions under which sample selection bias is small enough to disregard or too large to ignore. If variables are standardized to a variance of 1, then all components except α are constrained to absolute values between 0 and 1 so that small values for one (or, especially, two) of them are likely to drive selection bias to small values. We now consider each of these components.

(1) *$\lambda'(T-\alpha\bar{Z})$, the first derivative of the function λ.* To obtain the value at which λ' is evaluated, a probit analysis of sample selection is performed, and the mean of predicted values from the probit equation is calculated and multiplied by −1.

(2) *σ, the square root of the regression error variance.* This determines the magnitude of the regression R^2. Bias varies inversely with the regression R^2. If the regression R^2 is large, then selectivity bias is small, other things equal. A regression with a high R^2 can tolerate a lot of sample selection without showing much bias. However, in sociological analyses of individual-level data, the regression R is often between .2 and .5, corresponding to values of σ between .98 and .87 in standardized regressions.

(3) *$\rho_{\varepsilon\delta}$, the correlation between the regression error and selection error terms.* $\rho_{\varepsilon\delta}$ is not directly observable, but like all correlations it is between −1 and +1. Values in this range can be used in sensitivity analyses to determine the likely range of selection bias.

[3] Statistical programs frequently offer the option of retaining predicted probits and/or predicted probabilities. For present purposes, the predicted probit is the appropriate quantity to retain.

[4] The first derivative of the reciprocal of the Mills Ratio function is defined as follows. Where $\phi(x)$ is the normal density function and $\Phi(x)$ is the normal cumulative distribution function, then

$$\lambda'(x)=\{[1-\Phi(x)]\,[-x]\,\phi[x] + [\phi(x)]^2\} \,/\, [1-\Phi(x)]^2.$$

Cell formulas for a two-line Excel 7.0 spreadsheet, which makes these and other calculations, follows at the bottom of this page. Formulas assume that the upper left corner of the sheet is cell *a*1.

x	Normal density (x)	Normal cumulative distribution function (x)	λ(x)	λ′(x)
	=EXP(-.05*A2^2)/SQRT(2*PI())	=NORMSDIST(A2)	=B2/(1-C2)	=((1-C2)*(-A2)*B2+B2^2)/((1-C2)^2)

A correlation of 1 would occur if selection and regression were accomplished by identical processes. A correlation closer to 0 would occur when there is a lot of randomly-missing data or when sample selection is created by a process unrelated to the process described by the regression equation.

(4) ρ_{XZ}, *the correlation between independent variables from the regression equation and the selection equation.* Selection equations, and regression equations often have identical independent variables, in which case the correlation between regression and selection independent variables is 1. Absolute values of ρ_{XZ} between .9 and 1 seem likely in many sociological analyses.

(5) α, *the coefficient of Z in the probit equation.* If α is small, then Z does not explain selection very well, the selection mechanism will have little correlation with X, and the bias will be small. This means that if the probit equation used to estimate $\lambda(T-\alpha Z)$ fits the data poorly, then selection bias is small or the process of selection is misunderstood. In the case of misunderstanding, no statistical method can help. *Otherwise, poor probit fit of the selection equation is evidence that selection bias is likely to be small, other things equal.* In practice, probit fit is the easiest indicator of selection bias to compute and interpret. In most sociological analyses of individual-level data, probit fit is weak, and we therefore expect probit fit in selection models of individual data to be weak as well. Absolute values of α between .1 and .3 are likely in individual-level analyses.

(6) s_Z/s_X , *the standard deviation of the independent variable in the selection equation compared to the standard deviation of X in the selected data.* If the standard deviation of Z is small compared to the standard deviation of X, then bias is reduced. If X and Z are the same variable, as is often the case, then this ratio equals 1 and it neither increases nor decreases the selection bias.

Equation 8 can be used to develop intuition about the amount of selection bias occurring under hypothetical conditions or in a particular data set. With actual data, one would first construct a probit model of selection in the data, calculate and retain predicted values from the probit equation, take the mean of the predicted values, multiply the mean by -1, and evaluate λ' at that value. Most other quantities appearing on the right side of equation 8 can be estimated from sample data. However, for intuitive appeal, in Table 2 we have generated normally distributed random data with censoring of 5, 10, and 20 percent; the values of λ' shown below are based on those data. For other terms on the right side of equation 8 Table 2 displays values we think are commonplace for sociological analysis. The regression R in these examples takes values of .2 and .5, which correspond to σ values of .9798 and .8660; ρ_{XZ} and s_Z/s_X are both set to equal 1, reflecting the common situation in which the same independent variables are used as regressors in both selection and regression equations; the probit coefficient α is equal to .3; the correlation between regression and probit equation residuals is .71 ($.71^2 = .5$; ε explains half the variance in δ). Estimated bias equals the product of the column entries that are *not* shaded. Notice that bias is relatively small in all examples in Table 2.

The simulated results shown in Table 2 suggest that selection bias in standardized regression coefficients is likely to be small under many easily imagined circumstances, even if 10 or 20 percent of data cases are censored. As the fit of the selection model improves, however, selection bias increases. For example, if $\hat{\alpha}$ increased from .3 to .6 in these examples, then the selection bias would double. Poor model fit is seldom viewed as a virtue, but poor fit of the selection model can indicate that substantial selectivity coexists with small selection bias.

Notice that this exercise gives an idea of the direction of bias as well as its size under various conditions. If one hypothesizes and finds a positive coefficient for X in the regression equation, and if simulations like those that follow suggest that bias is likely to be downward, then one might reasonably conclude that one has found a lower-bound estimate of β_1. A downward-biased estimate could be quite useful for testing a substantive hypothesis of a positive coefficient for X, as long as that test found the hypothesized positive coefficient.

An Empirical Example

For a brief empirical example, we consider the regression of the years of schooling com-

Table 2. Application of Equation 8 to Six Hypothetical Examples

	Hypothetical Values					
Component of Bias	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Censoring rate	5%	10%	20%	5%	10%	20%
Mean predicted probit	2.33260	1.82421	1.19641	2.33260	1.82421	1.19641
λ'	.0626	.1488	.3125	.0626	.1488	.3125
Regression R	.2	.2	.2	.5	.5	.5
σ	.9798	.9798	.9798	.8660	.8660	.8660
$\rho_{\varepsilon\delta}$	.71	.71	.71	.71	.71	.71
ρ_{XZ}	1	1	1	1	1	1
α	.3	.3	.3	.3	.3	.3
s_Z / s_X	1	1	1	1	1	1
Estimated bias in standardized regression coefficient $\aleph_{\beta_1}$	–.0131	–.0311	–.0652	–.0115	–.0275	–.0576

Note: The estimated bias equals the product of other column entries in the table that are *not* shaded

pleted by a respondent (EDUC) on the respondent's father's years of schooling (PAEDUC). First we describe how to generate the data used in this analysis. We use data from the public use files of the 1985 NORC General Social Survey (GSS). To simplify calculations, we delete cases with missing values on either variable (leaving a sample of 1,153 cases), and we standardize both variables to a mean of 0 and a standard deviation of 1. Since this is an illustrative example, it is useful to know the impact of sample censoring. So we censor the data ourselves by creating a variable Y_2 equal to the sum of EDUC and a random variable with a mean of 0 and a standard deviation of 1. For the censored sample estimation, we "observe" cases for which Y_2 exceeds –1, and we censor all other cases. Our choice of –1 is arbitrary. After censoring, 893 cases remain (77.45 percent of the sample).

We use STATA (release 4) for statistical calculations. Using the uncensored sample, ordinary least–squares regression yields an estimate of .4898 for the coefficient of PAEDUC ($R^2 = .2392$; $t = 19.060$). Using the censored sample, the coefficient for PAEDUC is .3597 ($R^2 = .1644$; $t = 13.242$). So the difference between the censored and uncensored estimates is about –.13, which is more than five times the standard error of the regression coefficient in the censored sample.

Next, we follow the procedures we recommend for approximating sample selection bias in the coefficient of PAEDUC. We create a dummy variable equal to 1 for observed cases, and equal to 0 for censored cases. We perform a probit regression analysis of this dummy variable, with PAEDUC as the independent variable. The probit analysis yields a coefficient of .4131 for PAEDUC and a constant term of .8160. We retain the predicted values of the probit analysis and find that their mean is .8160. We calculate $\lambda'(-.8160)$ to be .4245. Following our hypothetical examples, we guess a range of values for $\rho_{\varepsilon\delta}$. Assuming that these variables explain about half of the variance in each other, we try .71 (= $\sqrt{.5}$) for the value of $\rho_{\varepsilon\delta}$ (In practice, one should try a range of possible values for $\rho_{\varepsilon\delta}$.) Filling all this information into equation 8 and multiplying yields an approximate bias of –.12, which is about equal to the difference between censored and uncensored estimates. On a purely subjective basis, this seems to be sufficient bias to warrant serious efforts to correct for selection effects. If those efforts failed to produce credible results, we might use our approximation results to suggest interpretation of the censored sample regression estimate as a low estimate of the coefficient of PAEDUC.

How Small Is Small?

A value of $\aleph_{\beta_1}$ that is "big" in one situation may be unimportant in another situation. Subjective judgments of this sort often cause haggling, but most of this subjectivity can be avoided by asking if bias is large compared to sampling error in the estimate of β_1. If bias is small compared to sampling error, then bias can be said to lack practical importance, much like a statistically insignificant regression coefficient.

A standard result in regression is

$$s_{\beta_1} \equiv \sigma/\left(s_X\sqrt{n}\right). \tag{9}$$

Dividing equation 8 by the right hand side of equation 9 and taking absolute values gives equation 10, the formula for comparing bias in the estimate of β_1 to the sampling error in that estimate:

$$\left|\aleph_{\beta_1}/s_{\beta_1}\right| = \left|\rho_{\varepsilon\delta}\rho_{XZ}\alpha\lambda'(T-\alpha\bar{Z})s_Z\sqrt{n}\right|, \tag{10}$$

where n is the sample size. As n grows large, the ratio of bias to sampling error also grows large. Equation B-7 in Appendix B extends equation 10 to multiple regression models. But we think that substantial insight can be gained by applying equations 8 and 10 to simple regression models, even when one's ultimate interest is in multiple regression analysis.

SIMULATIONS

We use simulations to evaluate the Taylor series approximations of λ. These simulations differ from earlier studies, which tested the extent to which two-step estimation corrects selection bias (Nelson 1984; Paarsch 1984; Stolzenberg and Relles 1990). Our simulations are used to check of the accuracy of a mathematically derived method for approximating bias.

To evaluate selection bias, we simulate several data sets, censor them, and fit regressions to the censored data. In all simulated data sets, the true value of the regression coefficient β_0 is 0 and the coefficient of X_1 equals 1. Following Equation 8, our experimental design is based upon six factors: σ, α, T, $\rho_{\varepsilon\delta}$, ρ_{XZ}, and N (the sample size before selection). The values of these factors are as follows: σ, the regression standard error, takes values 2, 1, and .5 (corresponding to regression R^2 of .2, .5, and .8.); α, the selection equation coefficient, takes values 1 and .333333 (corresponding to selection equation R^2 of .5 and .1); T, the selection threshold, equals $\zeta(1+\alpha^2)^{.5}$,where ζ takes values –.674, 0, and .674 (corresponding to selection rates of 25 percent, 50 percent, and 75 percent, respectively; this parameterization is required because the variance of Y_2 varies with α); $(\rho_{\varepsilon\delta})^2$, the squared correlation between ε and δ, takes values of 0, .25, .50, and .75; $(\rho_{X_1Z})^2$, the squared correlation between X_1 and Z, takes values of 0, .25, .50, and .75; N, the sample size before selection, takes values of 200, 500, 1,000, 2,000, and 5,000.

Thus, our experimental design has 1,440 cells (3×2×3×4×4×5). For each cell, we perform 50 simulations, yielding 72,000 total simulations. For each simulation, we perform an ordinary least–squares regression, a probit analysis of selection bias, and a two-step regression for correcting selection bias, for a total of 216,000 analyses.

The Pearsonian correlation between $\aleph_{\beta_1}$ and the actual selection bias is .99. The Taylor series estimate explains 97.68 percent of variance in observed bias.[5] Thus, the Taylor approximation of λ is sufficiently accurate to give confidence in equations 8 and 10.

OLS versus the Two-Step Estimator

Our simulations also permit comparison of the accuracy of Heckman's (1976) two-step estimates to uncorrected OLS estimates of selected data. Knowing the correct value of β_1, we estimate it using both techniques and compare the results. First, we simply tabulate the proportion of simulations in which OLS is more accurate than the two-step estimator. In the 1,440 cells in the experimental design, that proportion ranged from 0 percent to 98 percent, with a mean of 42.6 per-

[5] The equation is:

$$\text{Actual Bias} = -.0017 + .8505\ \aleph_{\beta_1}.$$

Similar results were obtained in regressions without the constant term.

Table 3. Two Nonlinear Probit Models of the Probability that OLS Gives a More Accurate Estimate of β_1 than the Two-Step Selection Correction Estimator, by Sample Size before Selection

	Sample Size before Selection				
Independent Variable	200	500	1,000	2,000	5,000
A. Using Bias in Standardized Regression Coefficient as Predictor					
Constant	.006 (.333)	.284 (12.313)	.231 (8.768)	.300 (7.685)	.273 (6.225)
$\aleph_{\beta_1}\left(s_{X_1}/s_{Y_1}\right)$	.619 (.549)	-17.026 (-11.288)	-20.822 (-12.477)	-39.664 (-15.797)	-52.953 (-18.877)
$\left[\aleph_{\beta_1}\left(s_{X_1}/s_{Y_1}\right)\right]^2$	-44.211 (-3.105)	87.253 (4.518)	67.638 (3.252)	184.499 (5.824)	277.735 (7.851)
$\left[\aleph_{\beta_1}\left(s_{X_1}/s_{Y_1}\right)\right]^3$	127.037 (3.051)	-180.220 (-3.141)	-89.499 (-1.492)	-308.940 (-3.332)	-472.978 (-4.566)
R^2	.249	.661	.767	.787	.812
Number of cases	288	288	288	288	288
B. Using Ratio of Bias to Standard Error of β_1 as Predictor					
Constant	-.010 (-.619)	.3030 (15.542)	.251 (13.331)	.305 (10.563)	.300 (8.410)
$\aleph_{\beta_1}/s_{\beta_1}$	.341 (3.280)	-.727 (-9.766)	-.641 (-13.394)	-.643 (-11.548)	-.667 (-17.180)
$\left[\aleph_{\beta_1}/s_{\beta_1}\right]^2$	-.599 (-4.527)	.172 (2.995)	.080 (3.168)	.009 (.421)	.032 (3.672)
$\left[\aleph_{\beta_1}/s_{\beta_1}\right]^3$	.141 (3.524)	-.025 (-2.404)	-.009 (-2.890)	.002 (1.069)	-.000 (-0.274)
R^2	.358	.773	.888	.891	.881
Number of cases	288	288	288	288	288

Note: Numbers in parentheses are *t*-statistics. Regressions are based on 1,440 data cases, each case corresponding to one cell of the experimental design described in the text. Each cell of the design contains 50 simulations. In each panel of this table, each of the five columns of the table reports an analysis of 288 cases. The dependent variable in each regression is the probit of the proportion of simulations in which the OLS estimate of β_1 is more accurate than the estimate obtained with the two-step estimator.

cent and a standard deviation of 21.2 percent. Large sample size alone is not sufficient to make the two-step estimator better than OLS: In simulations based on samples of 5,000 cases, OLS outperforms the two-step correction in 34.5 percent of the simulations.

We use these simulations to model the probability that OLS is more accurate than the two-step correction. To fit these models, we define a data set in which each cell of the experimental design represents one case. The dependent variable in these analyses is the probit of the proportion of simulations in which the OLS estimate of β_1 is more accurate than the estimate obtained with the two-step estimator.[6]

We calculate two sets of regression analyses in which the dependent variable described above (Y_1) is regressed on a measure

[6] The variable we wish to explain with these regressions is a proportion, and therefore it is limited to values between 0 and 1. Probit regression constrains predicted values from this regression to the interval (0,1). Probit regression of grouped data is accomplished by taking the inverse normal cumulative distribution function of that proportion, then regressing the transformed proportion on independent variables of interest. See Hanushek and Jackson (1977).

of bias severity. In the first set, the measure of bias severity is the expected bias measured in units of the standardized regression coefficient in the regression of Y_1 on X (this is $\left|\aleph_{\beta_1}\left(\sigma_X \sigma_{Y_1}\right)\right|$). In the second set, the measure of bias severity is the expected bias divided by the standard error of the regression coefficient for X($\left|\aleph_{\beta_1}/s_{\beta_1}\right|$) We fit cubic polynomial regressions to allow the effects to level off as the severity of selection bias gets very large or very small. Analyses are stratified by simulation sample size and are reported in Table 3. R^2 statistics indicate that these models fit rather well for sample sizes of 1,000 or more.

Regressions reported in Table 3 (or graphs drawn from them) can be used to estimate how severe bias must be before the two-step estimator becomes more accurate than OLS some percentage of the time. For simulations of 1,000 cases before sample selection, we cannot be 95 percent sure that the two-step correction outperforms OLS unless selection bias in the regression coefficient estimate is *at least 4 times the standard error of* β_1 . We cannot be 80 percent sure the correction is better unless bias is 2.14 times the standard error of β_1. When bias severity is measured in standardized units, the results are much the same: Unless the bias changes the standardized regression coefficient by least .146, we cannot be 95 percent sure that the two-step estimator performs better than OLS. Unless bias changes the standardized coefficient by at least .064, we cannot be even 80 percent sure that the two-step estimator performs better than OLS. *In short, the two-stage estimator worsens estimation unless selection bias is severe.*

Our third approach to comparing the accuracy of the two-step estimator to OLS is based on the difference in root-mean-squared errors (RMSE) of the two estimates. Our simulations show that RMSE tends to be lower for OLS than for the two-step estimator when $\aleph_{\beta_1}/s_{\beta_1}$ is less than .4. For values of $\aleph_{\beta_1}/s_{\beta_1}$ between .4 and 1.0, the correction performs only marginally better than OLS.

In short, our simulation-based comparisons of OLS and the two-step estimator suggest that if bias is very severe and the samples are large one can be reasonably confident that the two-step correction improves estimates. If bias is only moderate, however, or if samples have only a few hundred cases, there is considerable risk that the two-step estimator makes estimates worse, not better, even when sample selection is known to be present and the assumptions of the two-step correction method are satisfied. Compared to the conclusions of some earlier simulation studies, the rules of thumb based on our simulations offer more precision and inspire more confidence in the two-step estimator (Hartman 1991; Stolzenberg and Relles 1990). In particular, our results are inconsistent with Hartman's (1991) recommendation for wholesale abandonment of the two-step procedure.

CONCLUSIONS

What is to be done about sample selection bias? Several bias correction procedures are now available, and more are coming. Winship and Mare (1992) review a number of these techniques and conclude that none of them works well all the time. Some techniques make strong assumptions that require more courage than a particular empirical problem might justify. Other techniques are so imprecise that they often rob empirical analyses of their power to answer meaningful questions (Manski 1995). Yet others require data that are rarely available in sociological research (Little and Rubin 1987:230-34; Rubin 1977).

Instead of seeking a method that always corrects sample selection bias, one might use as many different selection correction methods as possible, giving each estimator a vote at the statistical polls. In this approach, if several of these "experts" offer the same correction, that correction is believed to be right. But the sheer inefficiency of many selectivity correction methods introduces randomness which easily can produce agreement by chance. Getting two of these methods to suggest that OLS estimates are too high, for example, may be no more informative than getting heads on two flips of a coin.

Finally, one can perform a significance test for the presence of selection bias (Heckman 1980). Significance tests are important and useful, but they are conceptually removed from estimation. If confidence intervals are large, a badly biased estimate can fail to differ significantly from an unbiased estimate. Thus, the availability of a significance test for selection bias does not lessen the need for a selection bias correction.

We think the safest approach to sample selection bias problems is first to understand how nonrandom selection occurs in one's data. This understanding focuses attention on selection bias as a missing data problem and sometimes can lead to an artful construction of missing values (Braun and Szatrowski 1982). Thinking about the process of sample selection may help reveal whether selection occurs through a process consistent with Heckman's (1976) model. If no such process exists in a particular data set, then it may be possible to rule out selection bias on purely logical grounds. (That is not to say that the data are necessarily unbiased, but, rather, that they are not biased in the way described in Heckman's [1976] paper and in the literature that subsequently grew from it.) If data do appear to be selected as described by Heckman's model, then it is appropriate to consider Heckman's two-step estimator. When Heckman's model is substantively appropriate, one can use equation 8 to assess the probable direction and likely severity of sample selection bias, even if precise values for all terms on the right side of the equation are not available. In practice, one would begin with univariate analyses and equations 8 and 10, and then move to the multivariate analyses described in Appendix B equations B-6 and B-7. These equations sharply reduce the number of factors one must speculate about to forecast the selection bias in a regression analysis. Equations 8 and A-6 (from Appendix A) support considerable intuition about whether selection bias tends to make analyses stringent or lenient tests of the substantive hypotheses under consideration. For example, if the sign of an estimated coefficient is opposite the sign of selection bias affecting that coefficient, then an uncorrected, biased analysis may be sufficient to support important substantive hypotheses, even if precise measurement of effects is impossible. In addition, equation 10 and the regression equations reported in Table 3 can help indicate whether the two-step estimator is likely to improve on OLS estimation.

The bad news here is no surprise: There appears to be no automatic way to diagnose and correct sample selection bias. Analytical methods cannot make imperfect data perfect. To return to an earlier example, there is no way to know for sure how much nonemployed married women would earn if they became employed. But analytical methods can help bridge some of the gap between the data one can get and the data one would like. Intuition, informed judgment, simulation, experimentation, and statistical methods are necessary to understand and manage inevitable problems in data. Selection bias is just one of these problems. The methods we present here, when combined with other tools for selection bias correction, can help provide the information necessary to cope with selection bias problems. We do not expect that these methods can make data perfect. But we do think that these techniques, in combination with other procedures, can help make the analysis of imperfect data informative and infinitely more useful than speculation about important substantive concerns in the absence of any data at all.

***Ross M. Stolzenberg** is Professor of Sociology at the University of Chicago. His current research interests and some recent publications concern the relationship between attitudes and behavior (*Social Forces, *1995;* American Sociological Review, *1995;* American Journal of Sociology, *1994), and the effects of schooling on employment of Asians and Hispanics (*Social Science Research, *1997). This work continues in his current projects: a study of the structure of young adults' attitudes toward work, family and schooling; and a study of the long-term consequences of schooling.*

***Daniel A. Relles** is Senior Statistician at the RAND Corporation. His main research interest is in developing statistical methods to efficiently manage and analyze large data sets. At RAND, he works on a variety of projects that have complex data requirements; all projects involve interdisciplinary teams. He is currently working on a health-related project to model the supply and demand for orthopaedic surgeons through the year 2010, a military logistics project to reduce delays in the repair of Army vehicle and weapons systems, and a resource management project on the efficient dispatching of electric generating capacity, given uncertain power demands.*

Appendix A. The Taylor Series Approximation of the Mills Ratio Inverse

We expand $\lambda(T-\alpha Z)$ in its Taylor series about the mean of Z in the data for which Y_1 is observed (call it $\bar{Z}$), then decompose Z into its projection on X plus a residual. The Taylor series expansion of $\lambda(T-\alpha Z)$ through the linear term is

$$\lambda(T-\alpha Z) \approx \lambda(T-\alpha\bar{Z}) + \alpha\lambda'(T-\alpha\bar{Z})(Z-\bar{Z}). \qquad \text{(A-1)}$$

The linear projection of Z onto X is $[X\rho_{XZ}(s_Z/s_X)]$, where ρ_{XZ} is the correlation between X and Z, and s_X and s_Z denote the standard deviations of X and Z in the regression sample. Hence, the omitted term $[\sigma\rho_{\varepsilon\delta}\lambda(T-\alpha Z)]$ can be decomposed into a constant, a part of which is a multiple of X, and another part that is orthogonal to X. The X term in this decomposition has the coefficient $\aleph_{\beta_1}$, shown in equation 8 (see page 7), which will be absorbed into the estimate of β if selection bias is not corrected. $\aleph_{\beta_1}$ is therefore the bias induced by failing to correct for selection bias. Note that $\lambda'(T-\alpha\bar{Z})$ is a constant, since all of its arguments are constants. To test the Taylor approximation, we perform 72,000 simulations (described in the body of this paper) in which we observe actual bias and calculate the Taylor series estimate of it. The correlation between actual bias and the Taylor approximation is .9883, which seems sufficient justification for the usefulness of the Taylor approximation.

Appendix B. The Taylor Series Approximation in Multiple Regression

Equations B-6 and B-7 below are the multiple regression versions of equations 8 and 10 respectively. We obtain B-6 and B-7 by restating in multiple regression form the expression for selection bias, applying the Taylor series approximation of λ to this restated form, and then simplifying the result by orthogonalizing the variables. Finally, we replace unfamiliar quantities introduced by the orthogonalization with more familiar equivalents. Multiple and partial correlations appearing in equations B-6 and B-7 can be computed using many statistical analysis programs, including SAS and SPSS. Refer to Appendix Table B-1 for definitions of symbols used in Appendix B.

We begin by re-expressing equation 6 with variables $x_1, x_2, \ldots, x_p$ in place of X and variables $z_1, z_2, \ldots, z_q$ in place of Z:

$$\begin{aligned} Y_1 = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \\ & + \sigma\rho_{\varepsilon\delta}\lambda(T-\alpha_1 z_1 - \ldots - \alpha_q z_q) \\ & + \sigma'\varepsilon', \end{aligned} \qquad \text{(B-1)}$$

where ε' has a mean of 0 but is not normally distributed. Ignoring the selection phenomenon is equivalent to leaving the term $\lambda(T-\alpha_1 z_1 - \ldots - \alpha_q z_q)$ out of the regression equation. Since the order of the independent variables is arbitrary, it is simplest to consider the bias in the estimate of β_1 if $\lambda(T-\alpha_1 z_1 - \ldots - \alpha_q z_q)$ is omitted from the regression.

The two-step estimator requires calculating λ for each data case and reestimating the regression equation with λ added to the list of regressors. We wish to find an expression for the bias that does not require adding a new independent variable, the Mills Ratio inverse, and a subsequent re-estimation of the regression equation. We use a Taylor expansion to estimate the Mills Ratio inverse. In the derivations below, $\hat{\alpha}_1, \ldots, \hat{\alpha}_q$ are the probit estimates of the selection equation independent variable coefficients, $\alpha_1, \ldots, \alpha_q$; $\bar{z}_1, \ldots, \bar{z}_q$ are the mean values of the selection equation independent variables $z_1, \ldots, z_q$; $\hat{Z}$ is the predicted value from the probit equation $(\hat{Z} = \hat{\alpha}_1 z_1 + \ldots + \hat{\alpha}_q z_q - T)$, and K is a constant. The Taylor series expansion of the Mills Ratio inverse $\lambda(T - \hat{\alpha}_1 z_1 - \ldots - \hat{\alpha}_q z_q)$ around $\bar{z}_1, \ldots, \bar{z}_q$ is:

$$\begin{aligned} \lambda(T - \hat{\alpha}_1 z_1 - \ldots - \hat{\alpha}_q z_q) \approx {} & \\ & \lambda(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - z_q \bar{z}_q) \\ & + \lambda'(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q) \\ & \quad (\hat{\alpha}_1 [z_1 - \bar{z}_1) - \ldots - \hat{\alpha}_q [z_q - \bar{z}_q]) \end{aligned}$$

$$\begin{aligned} & = K - \lambda'(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q) \\ & \quad (-\hat{\alpha}_1 z_1 - \ldots - \hat{\alpha}_q z_q) \end{aligned}$$

$$= K - \lambda'(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q)\hat{Z}. \qquad \text{(B-2)}$$

$$\begin{aligned} K = {} & \lambda(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q) \\ & + \lambda'(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q) \\ & \quad (\hat{\alpha}_1 \bar{z}_1 + \ldots + \hat{\alpha}_q \bar{z}_q). \end{aligned} \qquad \text{(B-3)}$$

Substituting the Taylor series approximation for $\lambda(T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q)$ into equation A-1 gives

$$\begin{aligned} Y_1 = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \\ & + \sigma\rho_{\varepsilon\delta}\left(K - \lambda'\left[T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q\right]\hat{Z}\right) \\ & + \sigma'\varepsilon'. \end{aligned} \qquad \text{(B-4)}$$

as the approximate equation to fit that will correct for the sample selection bias. Except for $\hat{Z}$, all terms of $\sigma\rho_{\varepsilon\delta}\left(K - \lambda'\left[T - \hat{\alpha}_1 \bar{z}_1 - \ldots - \hat{\alpha}_q \bar{z}_q\right]\hat{Z}\right)$ are constants. So, $\hat{Z}$, the predicted value from the probit equation, is added as a regressor to the regression equation.

To estimate the bias created by omitting the correction term from the regression, we draw on a standard result from regression theory that is now only rarely cited or taught: If Y_1, x_1, and $\hat{Z}$ are orthogonalized with respect to x_2 through x_p, then the coefficients of x_1 and $\hat{Z}$ in equation B-4 can be estimated by regressing the orthogonalized value of Y_1 on the orthogonalized values of x_1 and $\hat{Z}$ (see Morris and Rolph 1981:78–82), thereby eliminating x_2 through x_p from consideration and making the multiple regression case described in this appendix similar to the simpler case described in the main part of the paper. To orthogonalize Y_1 with respect to x_2 through x_p , regress Y_1 on x_2 through x_p; the residuals of this regression are the orthogonalized values of Y_1 , denoted $\breve{Y}_1$. Orthogonalize x_1 and $\hat{Z}$ similarly, as residuals from their regressions on $x_2, \ldots, x_p$. Denote the orthogonalized values of x_1 and $\hat{Z}$ as $\breve{x}_1$ and $\breve{Z}$, respectively. $\rho_{\breve{x}_1\breve{Z}}$ is the correlation between $\breve{x}_1$ and $\breve{Z}$, $s_{\breve{x}_1}$ and $s_{\breve{Z}}$ are the standard deviations of $\breve{x}_1$ and $\breve{Z}$ (which are the conditional standard deviations of x_1 and $\hat{Z}$). Thus, the multivariate counterpart of equation 8 is equation B-5, where $\aleph_{\beta_1}$ denotes the bias in the estimate of β_1:

$$\aleph_{\beta_1} \approx -\sigma \rho_{\breve{x}_1\breve{Z}} \lambda'(T - \hat{\alpha}_1\bar{z}_1 - \ldots - \hat{\alpha}_q\bar{z}_q)(s_{\breve{Z}}/s_{\breve{x}_1}). \quad \text{(B-5)}$$

Equation B-5 can be written without reference to orthogonalized variables. $\rho_{\breve{x}\breve{Z}}$ is just $\rho_{x_1\hat{Z}.x_2\ldots x_p}$, the partial correlation of x_1 with $\hat{Z}$ net of $x_2, \ldots, x_p$. In addition, $s_{\breve{x}_1}$ can be obtained easily from the standard deviation of x_1 and the multiple correlation of x_1 with x_2 through x_p , $R_{x_1.x_2\ldots x_p}$:

$$s_{\breve{x}_1} = s_{x_1}\sqrt{1-\left(R_{x_1.x_2\ldots x_p}\right)^2}.$$

And $s_{\breve{Z}}$ can be calculated from the standard deviation of $\hat{Z}$ and the multiple correlation of $\hat{Z}$ with x_2 through x_p.

$$s_{\breve{Z}} = s_{\hat{Z}}\sqrt{1-\left(R_{\hat{Z}.x_2\ldots x_p}\right)^2}.$$

So we can write

$$\aleph_{\beta_1} \approx -\sigma \rho_{\varepsilon\delta}\, \rho_{x_1\hat{Z}.x_2\ldots x_p}\, \lambda'(T - \hat{\alpha}_1\bar{z}_1 - \ldots - \hat{\alpha}_q\bar{z}_q)\,(s_{\hat{Z}}/s_{x_1})\frac{\sqrt{1-\left(R_{\hat{Z}.x_2\ldots x_p}\right)^2}}{\sqrt{1-\left(R_{x_1.x_2\ldots x_p}\right)^2}}. \quad \text{(B-6)}$$

Table B-1. Symbols Used in Appendix B

Symbol	Definition
$\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_q$	Means of selection equation independent variables $z_1, z_2, \ldots, z_q$.
$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$	Regression equation coefficients.
$\alpha_1, \alpha_2, \ldots, \alpha_q$	Selection equation coefficients.
$\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_q$	Probit estimates of selection equation coefficients.
$\hat{Z}$	Predicted value from probit selection equation, $\hat{Z} = \hat{\alpha}_0 + \hat{\alpha}_1 z_1 + \ldots + \hat{\alpha}_q z_q$.
σ^2	Regression equation error variance.
T	Selection threshold.
ϕ	$\phi(x)$ is the normal density function (height of normal curve) evaluated at x.
Φ	$\Phi(x)$ is the normal curve cumulative distribution function (area under the normal curve) evaluated at x.
λ	Reciprocal of Mills Ratio function, $\lambda(x) = \phi(x)/(1 - \Phi[x])$.
λ'	First derivative of reciprocal of Mills Ratio function, $\lambda'(x) = \dfrac{(1-\Phi[x])\,(-x)\,\phi(x) + (\phi[x])^2}{(1-\Phi[x])^2}$
$R_{\hat{Z}.x_2\ldots x_p}$	Multiple correlation between $\hat{Z}$ and regression independent variables $x_2, \ldots, x_p$.
$R_{x_1.x_2\ldots x_p}$	Multiple correlation between x_1 and regression independent variables $x_2, \ldots, x_p$.
$\aleph_{\beta_1}$	Bias estimate of β_1.
$\rho_{x_1\hat{Z}.x_2\ldots x_p}$	Partial correlation of x_1 with $\hat{Z}$ net of $x_2, \ldots, x_p$.
$\rho_{\varepsilon\delta}$	Correlation between regression and selection equation errors.
$s_{\breve{Z}}$	Standard deviation of $\breve{Z}$.
$s_{\breve{x}_1}$	Standard deviation of $\breve{x}_1$.
	Standard deviation of $\hat{Z}$.
$\rho_{\breve{x}_1\breve{Z}}$	Correlation between $\breve{x}_1$ and $\breve{Z}$ (equal to $\rho_{x_1\hat{Z}.x_2\ldots x_p}$).
$\breve{Z}$	Value of $\hat{Z}$ orthogonalized on x_2 through x_p.
$\breve{Y}_1$	Value of Y_1 orthogonalized on x_2 through x_p.
$\breve{x}_1$	Value of x_1 orthogonalized on x_2 through x_p.

Notice that equation B-6 differs from equation 8 in very straightforward ways, which account for the additional independent variables in the regression and selection equations. $\lambda(T-\hat{\alpha}_1\bar{z}_1-\ldots-\hat{\alpha}_q\bar{z}_q)$ is approximated from the proportion of cases selected using precisely the same procedure in the multiple regression case as is used in the simple regression case; the partial correlation $\rho_{x_1\hat{Z}.x_2\ldots x_p}$ replaces the zero-order correlation ρ_{XZ} and is calculated from the data; and the ratio $\dfrac{\sqrt{1-\left(R_{\hat{Z}.x_2\ldots x_p}\right)^2}}{\sqrt{1-\left(R_{x_1.x_2\ldots x_p}\right)^2}}$ is also calculated from the data.

As in the simple regression case described in the text, it is useful to judge the seriousness of selection bias by comparing $\aleph_{\beta_1}$ to the standard error for β_1. Dividing equation B-6 by the formula for the standard error of β_1, canceling terms and taking the absolute value yields the following result:

$$\left|\aleph_{\beta_1}/s_{\beta_1}\right| \approx -\rho_{\varepsilon\delta}\,\rho_{x_1\hat{Z}.x_2\ldots x_p}\,\lambda'(T-\hat{\alpha}_1\bar{z}_1-\ldots -\hat{\alpha}_q\bar{z}_q)\sqrt{\mathrm{n}}\,(s_{\hat{Z}})\frac{\sqrt{1-\left(R_{\hat{Z}.x_2\ldots x_p}\right)^2}}{\sqrt{1-\left(R_{x_1.x_2\ldots x_p}\right)^2}}. \quad \text{(B-7)}$$

In short, the multiple regression case is a straightforward generalization of the one-variable case.

REFERENCES

Berk, Richard. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Sociological Review* 48:386–98.

Braun, Henry and Theodore Szatrowski. 1982. *The Reconstruction of Ideal Validity Experiments through Criterion-Equating: A New Approach.* Princeton, NJ: Educational Testing Service.

Duan, Naihua, Will Manning, Carl Morris, and Joseph Newhouse. 1984. "Choosing between the Sample Selection Model and the Multi-Part Model." *Journal of Business and Economic Statistic* 2:283–89.

Goldberger, Arthur. 1980 "Abnormal Selection Bias." Department of Economics, University of Wisconsin, Madison, WI. Unpublished manuscript.

Hanushek, Eric and John Jackson. 1977. *Statistical_Methods for Social Scientists.* New York: Academic Press.

Hartman, Raymond. 1991. "A Monte Carlo Analysis of Alternative Estimators in Models Involving Selectivity." *Journal of Business and Economic Statistics* 9:41–49.

Heckman, James. 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* 5:475–92.

———. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 45:153–61.

———. 1980 "Addendum to 'Sample Selection Bias as a Specification Error.'" Pp. 69–74 in *Evaluation Studies Review Annual,* edited by E. Stormsdorfer and G. Farkas. Beverly Hills, CA: Sage.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation." *Journal of Political Economy* 94:489–506.

Little, Roderick. 1985. "A Note about Models for Selectivity Bias." *Econometrica* 53:1469–74.

Little, Roderick and Donald Rubin. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

Manski, Charles. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

Morris, Carl and John Rolph. 1981. *Introduction to Data Analysis and Statistical Inference.* Englewood Cliffs, NJ: Prentice-Hall.

Nelson, Forrest. 1984. "Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection." *Journal of Econometrics* 24:181–96.

Paarsch, Harry. 1984. "A Monte Carlo Comparison of Estimators for Censored Regression Models." *Journal of Econometrics* 24:197–213.

Rubin, Donald. 1977. "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72:538–43.

Stolzenberg, Ross and Daniel Relles. 1985. *Calculation and Practical Application of GMAT Predictive Validity Measures.* Santa Monica, CA: Graduate Management Admission Council.

Stolzenberg, Ross and Daniel Relles. 1990 "Theory Testing in a World of Constrained Research Design: The Significance of Heckman's Censored Sampling Bias Correction for Nonexperimental Research." *Sociological Methods and Research* 18:395–415.

Wainer, Howard, ed. 1986. *Drawing Inferences from Self-Selected Samples.* New York: Springer-Verlag.

Winship, Christopher and Robert Mare. 1992. "Models for Selection Bias." *Annual Review of Sociology* 18:327–50.