Estimating Models with Sample Selection Bias: A Survey

Francis Vella

ABSTRACT

This paper surveys the available methods for estimating models with sample selection bias. I initially examine the fully parameterized model proposed by Heckman (1979) before investigating departures in two directions. First, I consider the relaxation of distributional assumptions. In doing so I present the available semi-parametric procedures. Second, I investigate the ability to tackle different selection rules generating the selection bias. Finally, I discuss how the estimation procedures applied in the cross-sectional case can be extended to panel data.

I. Introduction

The ability to estimate and test econometric models over nonrandomly chosen subsamples is unquestionably one of the more significant innovations in microeconometrics. Since James Heckman’s seminal work on sample selection bias, the economics literature has abounded with empirical applications employing his proposed methodology. Although Heckman’s ideas initially had a more significant impact on empirical studies, the recent interest in semi- and nonparametric estimation of econometric models has revitalized the theoretical interest in the sample selection model. Despite its wide applicability the model initially considered by Heckman had a rather limited structure and was highly parameterized. Subsequent papers, however, have extended it in two important directions. First, although the original model accounted for a selection process captured by a dichotomous outcome subsequent approaches have incorporated different censoring rules in the selection equation. Second, the popularity of semi- and nonparametric econometrics has seen the relaxation of many of the model’s assumptions. Given the prominence of the
sample selection model in microeconometrics, it is useful to survey the literature motivated by Heckman's initial investigations.

The primary objective of this paper is to provide an intuitive discussion of the ideas underlying the various estimators available for models contaminated with selection bias. I attempt to cover a wide range of estimators and provide insight into how they eliminate selection bias. I do not, however, provide a detailed discussion of their properties nor do I provide efficiency comparisons for different estimators suitable for the same model. This decision does not reflect any opinion regarding the importance of these issues but is based on the length of the discussion that would be required.

Some of the following discussion can be found in existing surveys on subsections of the literature I cover (see, for example, Maddala 1983, Verbeek and Nijman 1992a, Powell 1994, Pagan and Ullah 1997). Accordingly I do not cover some details to the same extent as these related studies; these associated works should be treated as complements to this paper. These studies are particularly important for the semiparametric procedures that are generally associated with substantially more technical details.

The following section provides a brief intuitive discussion of sample selection bias. Section III provides a formal version of the model discussed in Section II. Sections IV and V are devoted to estimation with cross-sectional data with a dichotomous selection rule. First, I examine maximum likelihood estimation before focusing on two-step estimation. In both instances I examine how the underlying parametric and distributional assumptions can be relaxed. Section V also discusses the computation of the conditional expectations from these models and presents the methodology of Manski (1989) for computing bounds for the conditional expectations. Section VI focuses on models with alternative forms of selection bias. Section VII is devoted to the estimation of panel data models and Section VIII concludes.

II. Sample Selection Bias

Before proceeding to a discussion of the estimators available for the correction of sample selection bias, it is useful to provide an intuitive discussion of the primary issue. To do so, consider the framework in which the issue of selectivity bias first arose, namely the determinants of wages and labor supply behavior of females (see Gronau 1974, Heckman 1974).

Consider a population of women where only a subsample is engaged in market employment and report wages. Suppose I am interested in identifying the determinants of wages and labor supply behavior of females (see Gronau 1974, Heckman 1974). A population of women where only a subsample is engaged in market employment and report wages. Suppose I am interested in identifying the determinants of wages and labor supply behavior of females (see Gronau 1974, Heckman 1974).
be similar to the average characteristics of the population. Now consider where the
decision to work is no longer random and consequently the working and nonworking
samples potentially have different characteristics. Sample selection bias arises when
some component of the work decision is relevant to the wage determining process.
That is, when some of the determinants of the work decision are also influencing
the wage. When the relationship between the work decision and the wage is purely
through the observables, however, one can control for this by including the appro-
priate conditioning variables in the wage equation. Thus, sample selection bias will
not arise purely because of differences in observable characteristics.

If I now assume the unobservable characteristics affecting the work decision are
correlated with the unobservable characteristics affecting the wage, however, I gener-
ate a relationship between the work decision and the process determining wages.
Controlling for the observable characteristics when explaining wages is insufficient,
as some additional process is influencing the wage, namely, the process determining
whether an individual works. If these unobservable characteristics are correlated with
the observables then the failure to include an estimate of the unobservables will lead
to incorrect inference regarding the impact of the observables on wages. Thus, a
bias will be induced due to the sample selection.

This discussion highlights that sample selectivity operates through unobservable
elements and their correlation with observed variables, although often one can be
alerted to its possible presence through differences in observables across the two
samples. However, this latter condition is by no means necessary, or even indicative,
of selection bias. Although this example is only illustrative, it highlights the gener-
ality of the issues and their relevance to many economic examples. The possibility of
sample selection bias arises whenever one examines a subsample and the unobserv-
able factors determining inclusion in the subsample are correlated with the unobserv-
able variables influencing the variable of primary interest. This possibility arises in many
economic applications and, accordingly, the methodology for controlling for selec-
tion bias has become commonly employed in microeconometrics.

III. The Model

The conventional sample selection model has the form:

\[
\begin{align*}
(1) \quad y_i^* &= x_i'\beta + \varepsilon_i; \quad i = 1, \ldots, N \\
(2) \quad d_i^* &= z_i'\gamma + \nu_i; \quad i = 1, \ldots, N \\
(3) \quad d_i = 1 &\text{ if } d_i^* > 0; \quad d_i = 0 \text{ otherwise} \\
(4) \quad y_i &= y_i^* \cdot d_i
\end{align*}
\]

where \( y_i^* \) is a latent endogenous variable with observed counterpart \( y_i \), \( d_i^* \) is a latent
variable with associated indicator function \( d_i \), reflecting whether the primary depend-
ent variable is observed and where the relationships, between \( d_i \) and \( d_i^* \) and \( y_i \) and
\( y_i^* \) respectively, are shown in (3) and (4). (1) is the equation of primary interest and
(2) is the reduced form for the latent variable capturing sample selection; \( x_i \) and \( z_i \)
are vectors of exogenous variables; \( \beta \) and \( \gamma \) are vectors of unknown parameters; and
\( \varepsilon_i \) and \( \nu_i \) are zero mean error terms with \( E[\varepsilon_i \mid \nu_i] \neq 0 \). I let \( N \) denote the entire sample size and use \( n \) to denote the subsample for which \( d_i = 1 \). At this point I allow \( z_i \) to contain at least one variable which does not appear in \( x_i \), although this is sometimes seen to be a controversial assumption. I return to a discussion of it later in this paper. Although this exclusion restriction is typically not necessary for parametric estimation, it is generally crucial for semi-parametric procedures. For the remainder of the paper I assume that \( x_i \) is contained in \( z_i \). The primary aim is to consistently estimate \( \beta \).

It is well known that ordinary least squares (OLS) estimation of \( \beta \) over the subsample corresponding to \( d_i = 1 \) will generally lead to inconsistent estimates due to the correlation between \( x_i \) and \( \varepsilon_i \) operating through the relationship between \( \varepsilon_i \) and \( \nu_i \). A number of remedies, however, exist. The first is maximum likelihood estimation and relies heavily on distributional assumptions regarding \( \varepsilon_i \) and \( \nu_i \). A second approach is characterized by two-step procedures which approximate or eliminate the nonzero expectation of \( \varepsilon_i \) conditional on \( \nu_i \).

IV. Maximum Likelihood Estimation

A. Parametric Methods

The first solution to sample selection bias was suggested by Heckman (1974) who proposed a maximum likelihood estimator. This requires distributional assumptions regarding the disturbances and Heckman made the following assumption:

Assumption 1.

\[ \varepsilon_i \text{ and } \nu_i, i = 1 \ldots N, \text{ are independently and identically distributed } N(0, \Sigma) \]  

where

\[ \Sigma = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\nu} \\ \sigma_{\varepsilon\nu} & \sigma_\nu^2 \end{pmatrix} \]

and \( (\varepsilon_i, \nu_i) \) are independent of \( z_i \).

Using Assumption 1 it is straightforward to estimate the parameters for the model in Section III by maximizing the following average log likelihood function:

\[
L = \frac{1}{N} \sum_{i=1}^{N} \left\{ d_i \ast \ln \left[ \int_{-\infty}^{\infty} \phi_{\nu}(y_i - x'_i \beta, \nu) \, dv \right] + (1 - d_i) \ast \left[ \ln \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_{\nu}(\varepsilon, \nu) \, d\varepsilon \, dv \right] \right\}
\]

where \( \phi_{\nu} \) denotes the probability density function for the bivariate normal distribution. This is closely related to the Tobit estimator although it is less restrictive in that the parameters explaining the censoring are not constrained to equal those explaining the variation in the observed dependent variable. For this reason it is also
known as Tobit type two (see, for example, Amemiya 1984). As estimation heavily relies on the normality assumption the estimates are inconsistent if normality fails. This is an unattractive feature although it is straightforward to test the normality assumption using tests such as those proposed by Gourieroux, Monfort, Renault, and Trognon (1987) and Chesher and Irish (1987). It is also possible to employ the conditional moment framework of Newey (1985) and Tauchen (1985) as discussed in Pagan and Vella (1989).

It is clear that estimation would be simplified if \( \phi_{ev} = 0 \) as (5) would then reduce to the product of the two marginal likelihoods. That is, a product of the likelihood function explaining whether \( d_i \) was equal to 1 or 0, over the entire \( N \) sample, and the likelihood function explaining the variation in \( y_i \) for the \( n \) subsample satisfying \( d_i = 1 \). Alternatively, when \( \sigma_{ev} \neq 0 \), it is necessary to evaluate double integrals. Moreover, because there is no selection bias when \( \sigma_{ev} = 0 \) a test of this hypothesis is a test of whether the correlation coefficient \( \rho_{ev} \) is equal to zero as this parameter captures the dependence between \( \varepsilon \) and \( v \). Alternatively, one could estimate under the null hypothesis of no selection bias and test \( \sigma_{ev} = 0 \) via the Lagrange multiplier or the conditional moment approaches.

Throughout the paper I apply several estimators to a common data set. The empirical example chosen is based on a sample of female youth from the National Longitudinal Survey (NLS). Using these data, I focus on the estimation of wage equation parameters while accounting for the possibility that the unobservables determining the work decision are correlated with those determining wages. Initially, I examine a sample of 2,300 females taken from the 1987 wave of the NLS. Column 1 of Table 1 reports the coefficients from estimating a log wage equation over the subsample of 1,569 workers. The estimated standard errors are reported in parentheses below the estimated coefficients. The coefficients are largely consistent with expectations although the implied rate of return to schooling is high. In Column 2, I report the corresponding estimates from estimating the model by maximum likelihood. I do not report the estimates for the participation decision. The estimates are quite different from those in Column 1. Most notably, with the exception of the experience effects, the coefficients on the remaining terms are all smaller in absolute size. This suggests that their effect is partially captured through the correlation between the errors. This is supported by the size of the estimated correlation coefficient, which strongly suggests the presence of selectivity. In fact, the size of the estimate, and its associated standard error, suggest model misspecification. I return to this below.

When the model is estimated by maximum likelihood the parameter estimates are fully efficient. This is an important characteristic as several alternative estimators do not require the same parametric assumptions for consistency. The relaxation of

1. Note, however, that it is possible to obtain the maximum likelihood estimates from this model using readily available packages such as *Limdep* (Greene 1995).
2. The model is estimated using the maximum likelihood estimation program for the selection model in *Limdep*.
3. The variables in the participation equation include those in the wage regression plus three regional dummy variables and two dummy variables denoting whether the individual was married and the status of their health.
parametric assumptions is, however, accompanied by an efficiency loss. Accordingly, the maximum likelihood estimates are a benchmark to examine the efficiency loss of these procedures under normality (see, for example, Nelson 1984). Furthermore, much of the development of this literature is devoted to the tradeoff between efficiency and robustness. Although maximum likelihood is the best when the model is correctly specified there exists a willingness to trade off some efficiency for estimators that are more robust to distributional assumptions. The ease of implementation is also an important issue.

One way to relax normality, while remaining in the maximum likelihood framework, was suggested by Lee (1982, 1983). Lee proposes transforming the stochastic components of the model into random variables that can be characterized by the bivariate normal distribution. For example, suppose the errors \( \varepsilon \) and \( v \) are drawn respectively from the nonnormal but known distributions \( F(\varepsilon) \) and \( G(v) \). It is possible to transform \( \varepsilon \) and \( v \) into normal disturbances via the functions \( J_1 \) and \( J_2 \), which involve the inverse standard normal distribution function, such that:

\[
\begin{align*}
(6) \quad \varepsilon^* &= J_1(\varepsilon) = \Phi^{-1}[F(\varepsilon)] \\
(7) \quad v^* &= J_2(v) = \Phi^{-1}[G(v)]
\end{align*}
\]

where the transformed errors \( \varepsilon^* \) and \( v^* \) now have standard normal distributions. The joint distribution of the transformed errors is now fully characterized through the bivariate normal distribution. Furthermore, it is possible to construct a likelihood function for the transformed errors as is done in Equation (5) noting, however, that an additional set of parameters, characterizing \( F(\varepsilon) \) and \( G(v) \), must be estimated.

I noted above that the model is similar to the Tobit model. Another closely related alternative is proposed by Cragg (1971). Cragg assumes, in the empirical example I provide, that although the workers satisfy the condition that \( z'y + v_i > 0 \), some of the nonworkers may also satisfy this condition but may not contribute positive hours due to some additional form of censoring. That is, each individual may have to satisfy two conditions, or clear two hurdles, to be observed working. For obvious reasons this is known as the “double hurdle” model. Cragg illustrates how the likelihood function has to be adapted to account for this additional possibility of censoring. A similar idea is adopted in Blundell and Meghir (1987) who illustrate how the censored regression model can be extended to allow for additional censoring rules of interest. This approach is particularly useful in the labor supply context (see, for example, Blundell, Ham, and Meghir 1987) as it allows for nonparticipation not only to reflect supply decisions but also to incorporate the possibility that willing participants face market constraints. A special case, in which the Tobit model accounts for the possibility of an additional form of censoring, is known as p-Tobit (see Deaton and Irish 1982).

### B. Semi-nonparametric Methods

Although the method proposed by Lee avoids the imposition of joint normality, it requires that the marginal distribution of the primary equation’s disturbances is known. To avoid distributional assumptions it is possible to employ the general esti-
mation strategy of Gallant and Nychka (1987) who approximate the underlying true joint density with:

\[
\begin{align*}
    b_{cv} &= \left( \sum_{k=0}^{K} \sum_{j=0}^{J} \pi_{kj} \varepsilon^{k} v^{j} \right) \phi_{c} \phi_{v}
\end{align*}
\]

where \( b_{cv} \) denotes the true joint density; \( \phi_{c} \) and \( \phi_{v} \) denote the normal densities for \( \varepsilon \) and \( v \) respectively; and \( \pi_{kj} \) denotes unknown parameters. The basic idea is to multiply the product of these two marginal normal densities by some suitably chosen polynomial such that it is capable of approximating the true joint density. The estimate of \( b_{cv} \) must represent a density and this imposes some restrictions on the chosen expansion and the values of \( \pi_{kj} \). Gallant and Nychka show that the estimates of \( \beta \) and \( \gamma \) are consistent providing the number of approximating terms tends to infinity as the sample size increases. Although Gallant and Nychka provide consistency results for their procedure, they do not provide distributional theory. When \( K \) and \( J \) are treated as known, however, inference can be conducted as though the model was estimated parametrically.

Although the sample selection model is explicitly considered by Gallant and Nychka (1987) empirical applications remain scarce. One exception is Melenberg and van Soest (1993) who examine the determinants of the wages of married women, while accounting for the market work decision, using data for the Netherlands. They also extend the model to allow for heteroskedasticity of a known form. Melenberg and van Soest conclude that the Gallant and Nychka approach was effective in accounting for the apparent departures from normality revealed by the data.

V. Two-Step Estimation

Although the semi-nonparametric procedures can be computationally challenging, the maximum likelihood procedures of Heckman and Lee are relatively straightforward. However, their use in empirical work is relatively uncommon. The more frequently employed methods for sample selection models are two-step estimators. In considering the two-step procedures, it is useful to categorize them into three "generations." The first fully exploits the parametric assumptions. The second relaxes the distributional assumptions in at least one stage of estimation. The third is semi-parametric in that it relaxes the distributional assumptions.

A. Parametric Two-Step Estimation

To examine the two-step version of the fully parameterized model I retain Assumption 1. The primary equation of interest over the \( n \) subsample corresponding to \( d_{i} = 1 \) can be written:

\[
    y_{i} = x_{i}' \beta + \varepsilon_{i}; \quad i = 1, \ldots, n
\]

recalling OLS estimation leads to biased estimates of \( \beta \) because \( E[\varepsilon_{i}|z_{i}, d_{i} = 1] \neq 0 \) (that is, the conditional mean of \( y \) is misspecified). The general strategy proposed by Heckman (1976, 1979) is to overcome this misspecification through the inclusion
of a correction term that accounts for \( E[\varepsilon_i|z_i, d_i = 1] \). To employ this approach, take the conditional expectation of (9) to get:

\[
E[y_i|z_i, d_i = 1] = x'_i\beta + E[\varepsilon_i|z_i, d_i = 1]; \quad i = 1, \ldots, n.
\]

Using Assumption 1 and the formula for the conditional expectation of a truncated random variable note that

\[
E[\varepsilon_i|z_i, d_i = 1] = \frac{\sigma_v}{\sigma_z^2} \left\{ \phi\left(\frac{z_i'\gamma}{\sigma_z}\right) \right\}
\]

where \( \phi(\cdot) \) and \( \Phi(\cdot) \) denote the probability density and cumulative distribution functions of the standard normal distribution. The term in curly brackets is known as the inverse Mills ratio.\(^4\)

To obtain an estimate of the inverse Mills ratio I require the unknown parameters \( \gamma \) and \( \sigma_z \). By exploiting the latent structure of the underlying variable capturing the selectivity process, and the distributional assumptions in Assumption 1, I can estimate \( \gamma/\sigma_z \) by Probit.\(^5\) Thus the two-step procedure suggested by Heckman (1976, 1979) is to first estimate \( \gamma \) over the entire \( N \) observations by maximum likelihood Probit and then construct an estimate of the inverse Mills ratio. One can then consistently estimate the parameters by OLS over the \( n \) observations reporting values for \( y_i \) by including an estimate of the inverse Mills ratio, denoted \( \lambda_i \), as an additional regressor in (9). More precisely, estimate:

\[
y_i = x'_i\beta + \mu\lambda_i + \eta_i
\]

by OLS to obtain consistent estimates of \( \beta \) and \( \mu \), where \( \eta_i \) is the term I use throughout the paper to denote a generic zero mean error uncorrelated with the regressors and noting \( \mu = \sigma_v/\sigma_z^2 \). This procedure is also known as a “control function” estimator (see, for example, Heckman and Robb 1985a,b).

The t-test on the null hypothesis \( \beta = 0 \) is a test of \( \sigma_v = 0 \) and represents a test of sample selectivity bias. Melino (1982) shows this represents the optimal test of selectivity bias, under the maintained distributional assumptions, as it is based on the same moment as the Lagrange multiplier test. That is, both the Lagrange multiplier test and the t-test for the coefficient on \( \lambda_i \) are based on the correlation between the errors in the primary equation and the errors from the selection equation. Note that the inverse Mills ratio is the error from the Probit equation explaining selection. I return to this interpretation of the inverse Mills ratio below.

The Heckman two-step estimator is straightforward to implement and the second step is only complicated by the standard errors having to be adjusted to account for the first step estimation (see, for example, Heckman 1976, 1979, Greene 1981, Maddala 1983).\(^6\) However, one concern is related to identification. Although the inverse

\(^4\) It is useful to note that the inverse Mills ratio is also the generalized residual for the Probit model (see Gourieroux et al. 1987, Vella 1993).

\(^5\) It is frequently assumed in Probit models that \( \sigma_v = 1 \) in order to identify \( \gamma \).

\(^6\) There exists a substantial literature on the estimation of the covariance matrix for the two-step sample selection estimator and related models (see, for example, Lee, Maddala and Trost 1980 for discussion of other models). However, note that the covariance matrix can be estimated employing the general strategy of Newey (1984).
Mills ratio is nonlinear in the single index \((z'/\gamma)\) the function mapping this index into the inverse Mills ratio is linear for certain ranges of the index. Accordingly the inclusion of additional variables in \(z_i\) in the first step can be important for identification of the second step estimates. However, there are frequently few candidates for simultaneous exclusion from \(x_i\) and inclusion in \(z_i\). In fact, many theoretical models impose that no such variable exists. For example, empirical models based on the Roy (1951) model often employ the estimated covariances to infer the nature of the sorting. The underlying economic model often imposes the same variables to appear in both steps of estimation. Thus many applications constrain \(x_i = z_i\) and identify \(\beta\) through the nonlinearity in the inverse Mills ratio. As the inverse Mills ratio is often linear, however, the degree of identification is often "weak" and this results in inflated second step standard errors and unreliable estimates of \(\beta\). This has proven to be a major concern (see, for example, Little 1985) and remains a serious point of contention.

Given this is a relatively important issue for empirical work, it has been the object of several Monte Carlo investigations (see, for a recent example, Leung and Yu 1996). Although most studies find that the two-step approach can be unreliable in the absence of exclusion restrictions Leung and Yu (1996) conclude that these results are due to the experimental designs. They find that the Heckman two-step estimator is effective providing at least one of the \(x_i's\) displays sufficient variation to induce tail behavior in the inverse Mills ratio. An examination of the inverse Mills ratio reveals that although it is linear over the body of permissible values the single index can take, it becomes nonlinear at the extreme values of the index. Accordingly, if the \(x_i's\) display a relatively large range of values, even in the absence of exclusion restrictions it is likely that the data will possess values of the single index which induce the nonlinearity and this assists in model identification. Despite this finding, however, these two-step procedures should be treated cautiously when the models are not identified through exclusion restrictions.

Column 3 of Table 1 reports the parametric two-step estimates for the wage equation discussed in the previous section. The inclusion of the correction term has a substantial impact on the coefficients. With the exception of the union coefficient the estimates are estimated with notably less precision than the OLS and maximum likelihood estimates. Based on conventional wisdom, the coefficient reflecting the return to schooling seems low. As the model is identified through several exclusion restrictions, one would suspect that the collinearity should not be a concern. It is unclear what is generating these results although the failure of normality is one possibility. To examine this I tested for nonnormality in the Probit equation explaining participation through an examination of the third and fourth moments using conditional moment tests. The results suggested these higher moments were inconsistent with those from normal distribution although neither of the rejections occurred at the 5 percent level.\(^7\) Below I focus on the presence of nonnormality in the wage equation.

\(^7\) These tests were conducted via artificial regressions in which the sample moments were regressed against an intercept and the scores from the Probit model. Under the null the coefficients on the intercepts should each equal zero. For this application the \(t\)-statistics for the intercepts for the third and fourth moments are 1.815 and 1.916 respectively.
It is worth reformulating the Heckman two-step estimator from a different and more restrictive perspective as this provides some insight into models I examine below. By imposing the restrictive assumption that the parameters are the same for each subsample, it is possible to view the sample selection model as a model with a censored endogenous regressor. That is, rewrite the model as:

\begin{align}
  y_i &= x_i^\prime \beta + \theta d_i + \varepsilon_i; \quad i = 1, \ldots, N \\
  d_i^* &= z_i^\prime \gamma + v_i; \quad i = 1, \ldots, N \\
  d_i &= 1 \text{ if } d_i^* > 0; \quad d_i = 0 \text{ otherwise}
\end{align}

where rather than sample selection, I have an endogenous dummy variable. Estimating \( \beta \) and \( \theta \) over the whole, or any chosen, subsample results in inconsistent estimates due to the correlation between \( d_i \) and \( \varepsilon_i \) operating through the nonzero covariance \( \sigma_{\varepsilon \varepsilon} \). This is known as an "endogenous treatment model" and is closely related to the sample selection model (see, for example, Heckman 1978).8

It is well known, (see, for example, Hausman 1978 and Heckman 1978), that the inconsistency in (11) can be overcome by; i) projecting \( d_i \) onto \( z_i \) to obtain \( \hat{d}_i \) and then replacing \( d_i \) with \( \hat{d}_i \); or ii) obtaining the residuals from this projection, \( \hat{v}_i \) and including both \( \hat{v}_i \) and \( d_i \) in (11). A similar approach, which exploits the distributional assumptions and the dichotomous nature of the \( d_i \), involves estimating \( \gamma \) by Probit and then computing the corresponding Probit residual. That is, by using our distributional assumptions one can rewrite (11) as:

\begin{align}
  y_i &= x_i^\prime \beta + \theta d_i + \mu v_i + \eta_i.
\end{align}

The Probit residual is known as a generalized residual (see Gourieroux et al. 1987) and has the form:

\[ d_i = \frac{\sigma_{\varepsilon \varepsilon}}{\sigma_\varepsilon^2} \left[ \frac{\phi(z_i^\prime \hat{\gamma})}{\Phi(z_i^\prime \hat{\gamma})} \right] + (1 - d_i) \left[ -\frac{\Phi(z_i^\prime \hat{\gamma})}{1 - \Phi(z_i^\prime \hat{\gamma})} \right]. \]

This can be identified as the inverse Mills ratio for the entire sample. This term possesses two important characteristics of a residual. First, it has mean zero over the whole sample. Second, it is uncorrelated with the variables that appear as explanatory variables in the first step Probit model.

As the inclusion of the generalized residual accounts for the correlation between \( \varepsilon_i \) and \( d_i \), it is possible to estimate \( \beta \) over either subsample corresponding to \( d = 0 \) or \( d = 1 \) after including the generalized residual. This model is identified without exclusion restrictions due to the nonlinearity of the residual. Also note that the generalized residual is uncorrelated with the \( z_i^\prime \)'s, over the whole sample, by construction. Thus the consequences of a high degree of collinearity between the generalized residual and the \( z_i^\prime \)'s, which is a concern in the sample selection model, does not arise.9

8. Examples of endogenous treatments include the impact of union status on wages or the effect of some government intervention on labor market outcomes.
9. The zero correlation between the generalized residual and the \( z_i^\prime \)'s is due to the derivation of the generalized residual as the score for the intercept from the Probit model evaluated at each data point. Given the definition of the score, and the inclusion of the \( z_i^\prime \)'s in the Probit model, it follows that for each of the \( k \)
An advantage of this interpretation is that it generalizes to alternative forms of censoring. Moreover, if I assume $E[e_i|v_i]$ is a linear function, I can also relax the distributional assumptions for $v_i$. I return to this below.

The parametric procedures are based on the exploiting the relationship between $e_i$ and $v_i$ operating through the distributional assumptions. Bivariate normality dictates that the relationship between the disturbances is linear. Accordingly, one may test, or even correct, for departures from normality by including terms that capture systematic deviations from linearity. Lee (1984) suggests approximating the true density by the product of normal density and a series of Hermite polynomials. Although the test that Lee motivates is based on the Lagrange multiplier framework, he also presents a variable addition type test in which (10) is augmented with the additional terms. Pagan and Vella (1989) adopt a similar approach and, following Gallant and Nychka (1987), approximate the bivariate density of $e_i$ and $v_i$ as:

$$b_{ev} = \left( \sum_{k=0}^{K} \sum_{j=0}^{J} \pi_{kj} e_i v_j \right) \phi_{ev}$$

recalling $\phi_{ev}$ is the bivariate normal density; the $\pi$’s denote unknown parameters; and $\pi_{00} = 1$. If we set $K = 0$, let $\phi_{ev}^*$ denote the conditional normal density of $e_i$ given $v_i$, and $p = b_{i}/\phi_{ev}$, then:

$$E[\varepsilon_i|v_i] = \sum_{j=0}^{J} p^{-1} \pi_{0j} (\varepsilon \phi_{\varepsilon_i}, d\varepsilon) v_j^j$$

$$= \sum_{j=0}^{J} p^{-1} \pi_{0j} \rho v_j^{j+1}.$$

Thus under the null hypothesis of joint normality:

$$E[\varepsilon_i|d_i = 1] = \pi_{00} E[v_i|d_i = 1] + \sum_{j=1}^{J} \pi_{0j} E[v_i^{j+1}|d_i = 1]$$

because $p = 1$ under the null hypothesis. A test of normality is to add on the higher order terms and test whether they are jointly zero. To compute these terms one can use the recursive formula provided in Bera, Jarque, and Lee (1984). They are proportional to the inverse Mills ratio and take the form $E[v_i^{j+1}|z_i, d_i = 1] = (z_i' \gamma)^j / \Phi(z_i' \gamma)$. Thus one computes these higher order terms and inserts them in (10) and tests whether they are jointly significant. Given the nature of the Hermitian expansion, the additional terms employed by Pagan and Vella are similar to those suggested by Lee.

In the empirical example I included these additional terms to test for nonnormality. The coefficient for the term for $j$ equal to 1 has a $t$-statistic of 2.630 suggesting the results in Columns 2 and 3 are contaminated with nonnormality. There was no evidence that any of the higher order terms were statistically important.

Although the approach of Lee (1984) and Pagan and Vella (1989) is primarily explanatory variables the following condition holds, $\Sigma_{i=1}^{n} z_i v_i = 0$, as these are the first order conditions, defining $\gamma$, for the Probit model.
motivated for testing for departures from normality, it is also possible that this framework can be employed in estimation. Inasmuch as these higher order terms are capable of capturing nonnormality for the sake of testing it is likely that they are able effectively to adjust for nonnormality in estimation. This strategy is suggested in Lee (1982). The extent to which these procedures are successful in capturing nonnormality for estimation purposes is unexplored. In the empirical example the inclusion of these additional terms appeared to have little impact on the coefficients in the wage equation.

It was quickly recognized that the heavy reliance of the two-step procedures on normality could be partially relaxed by replacing Assumption 1 with the relatively weaker Assumption 2:

**Assumption 2**

The distribution of \( v_i \) is known and \( e_i \) is a linear function of \( v_i \).

This presents no advantage over the Heckman two-step procedure if I assume that \( v_i \) is normal as Assumption 2 implies joint normality. It does however allow us to replace the normality of \( v_i \) with alternative distributional assumptions thereby allowing consistent first-step estimation by methods other than Probit. One procedure is suggested by Olsen (1980) who assumes that \( v_i \) is uniformly distributed. One can now replace the inverse Mills ratio with a simple transformation of the least squares residuals derived from the linear probability model (that is, the residuals from regressing \( d_i \) on \( z_i \)). Olsen shows that when the disturbances in the selection equation are uniformly distributed this two-step estimator is consistent. More formally, Olsen shows that:

\[
E[\epsilon_i | z_i, d_i = 1] = \rho \sigma_i(3)^{1/2}(z_i'\gamma - 1).
\]

This procedure generally produces results similar to Heckman two-step procedure. This follows from the high degree of collinearity between the respective corrections. The Olsen estimator requires the exclusion from the primary equation of at least one variable that appears in the reduced form, because the model can no longer be identified through the nonlinear mapping of the index \( z_i'\gamma \) to the correction term. A test of selectivity bias is captured through a test of statistical significance of the coefficient of the correction term as this parameter captures the linear relationship between the two disturbances.

A more general approach, to relax joint normality while remaining within the parametric framework, is proposed by Lee (1982, 1983) and is related to the maximum likelihood estimator discussed above. A useful case is where the marginal distribution of \( e_i \) is normal, and the marginal of \( v_i \) is known but nonnormal. Thus, the distribution of \( e \) and the transformed disturbance \( v^* \) is bivariate normal and their dependence is captured by their correlation coefficient. More important, the relationship between the disturbances is linear. To implement the two-step version of the Lee maximum likelihood estimator, note that \( d_i = 1 \) when \( v_i < z_i'\gamma \). This implies, from (7) that \( J_2(v_i) < J_2(z_i'\gamma) \). It follows that \( Pr[d_i = 1] = \Phi[J_2(z_i'\gamma)] = G(z_i'\gamma) \). Thus I can now write the conditional expectation of (9) as:
Thus, first estimate the $y$ from the discrete choice model by maximum likelihood where one employs $G(v_i)$ as the distribution function for $v_i$. Then substitute the estimate of $y$ into (14) and estimate by least squares.

Lee (1982) generalizes this approach such that $J_2$ is a specified strictly increasing transformation such that $v_i < z_i' \gamma \Leftrightarrow J_2(v_i) < J_2(z_i' \gamma)$. Let $\mu_j = E[J_2(v)]$ denote the expected value of $v$ and let $\sigma_j^2$ denote the variance of $[J_2(v)]$. Furthermore assume that $\varepsilon_i$ can be written as:

$$
\varepsilon_i = \tau[J_2(v) - \mu_j] + \eta_i,
$$

where $\eta_i$ and $J_2(v_i)$ are independent and where $\tau = 0$ if the disturbances are uncorrelated. If I write the conditional mean of the truncated disturbance as:

$$
\mu(J_2(z_i' \gamma)) = E[J_2(v_i) / J_2(v_i) < J_2(z_i' \gamma)]
$$

then:

$$
E[y_i | z_i, d_i = 1] = x_i' \beta + \rho \sigma_z / \sigma_j \left( \frac{\mu(J_2(z_i' \gamma))}{G(z_i' \gamma)} - \mu_j \right).
$$

Although this methodology provides some flexibility, it crucially depends on the assumption in (15). This approach, in the conventional sample selection model, is typically associated with the use of Logit. It is particularly attractive when there are multiple unordered outcomes as maximum likelihood estimation of the first step can be computationally difficult when the errors are normally distributed as the first step would need to be estimated by multinomial Probit. I examine this case below.

**B. Semi-Parametric Two-Step Estimation**

An early criticism of the parametric sample selection estimators was their reliance on distributional assumptions (see, for example, Goldberger 1983). Although this can be relaxed, through the use of different distributional assumptions, it is appealing to consider alternatives that have a limited reliance on parametric assumptions. To consider the available procedures replace Assumption 2 with a weaker statement about the disturbances.

**Assumption 3**

$$
E[\varepsilon_i | z_i, d_i = 1] = g(z_i' \gamma) \text{ where } g \text{ is an unknown function.}
$$

Assumption 3 is known as an index restriction. The parametric two-step approaches implicitly define the function $g(\cdot)$ through the distributional assumptions, or assume it explicitly, but the semi-parametric procedures seek to avoid the imposition of such information.

Estimation under Assumption 3 rather than Assumptions 1 and 2 raises two difficulties. First, it is no longer possible to invoke distributional assumptions regarding $v_i$ to estimate $\gamma$. Second, one cannot use distributional relationships to estimate $E[\varepsilon_i | z_i, d_i = 1]$. The first problem is overcome through nonparametric or semi-
parametric estimation of the binary choice model. For example, it is possible to estimate $\gamma$ by the procedures of Cosslett (1983); Gallant and Nychka (1987); Powell, Stock, and Stoker (1987); Klein and Spady (1993); and Ichimura (1993); without imposing distributional assumptions on $v_i$. With these estimates it is straightforward to compute an estimate of the single index $z_i'\gamma$ and the second difficulty can then be overcome in a number of ways.

Using the index restriction write the conditional expectation of the primary equation as:

$$E[y_i|z_i, d_i = 1] = x_i'\beta + g(z_i'\gamma); \quad i = 1, \ldots, n,$$

noting that it is not possible to distinguish an intercept term in $x_i$ from an intercept in $g(\cdot)$. Accordingly, the intercept term is not identified in the following procedures. I discuss below, however, some ways to infer the value of the intercept. Given consistent estimates of the single index, the issue is how the $g(\cdot)$ function is approximated.

The first suggestion to estimate the model semi-parametrically is found in Heckman and Robb (1985a). They suggest a two-step estimator in which the first step is the nonparametric estimation of $Pr[d_i = 1|z_i]$, which is also known as the propensity score (see, for example, Rosenbaum and Rubin 1983). The second step is to approximate the $g(z_i'\gamma)$, which is equal to $E[\epsilon_i|z_i, d_i = 1]$, through a Fourier expansion in terms of $Pr[d_i = 1|z_i]$.

Cosslett (1991) proposes a two-step procedure in which he first estimates $\gamma$ via the nonparametric maximum likelihood estimator outlined in Cosslett (1983). The first step approximates the marginal distribution function of the selection error, $F(\cdot)$, as a step function constant on a finite number $J$ of intervals $\{I_j = [c_{j-1}, c_j], j = 1, \ldots, J \text{ and } c_0 = -\infty, c_J = \infty\}$. In the second step Cosslett estimates the primary equation while approximating the selection correction, $g(\cdot)$ by $J$ indicator variables $\{1(z_i'\hat{\gamma} \in I_j)\}$. Consistency requires that $J$ increase with the sample size.

Newey (1988) suggests estimating the single index by some semi-parametric procedure. He then approximates $g(z_i'\gamma)$ by $\hat{g}(z_i'\gamma) = \sum_{k=1}^{K} \alpha^k(z_i'\gamma)^{k-1}$ where $\gamma$ is some first step estimate and $K$, denoting the number of terms in the approximating series, is allowed to grow with the sample size. The second step is then estimated by OLS while setting $K$ equal to some fixed number. An advantage of the Newey approach is that the estimates are $\sqrt{n}$ consistent and it is straightforward to compute the second step covariance matrix.

The above estimator employs the orthogonality conditions $E[\epsilon_i - g(z_i'\gamma)|d_i = 1, z_i] = 0$ to define the estimator of $\beta$. Newey argues efficiency gains can be obtained if the additional orthogonality conditions implied by the independence of $\{\epsilon_i - g(z_i'\gamma)\}$ and $z_i$ are exploited. Newey notes that $\{\epsilon_i - g(z_i'\gamma)\}$ is uncorrelated with any function of $z_i$. To employ the additional orthogonality conditions implied by the independence define $\xi_j(\epsilon_i)(j = 1, \ldots, J)$ as some function of $\epsilon_i$, and $\xi_j(z_i'\gamma) = E(\xi_j(\epsilon_i)|z_i, z_i'\gamma)$. Newey then defines a generalized method of moments estimators based on the orthogonality conditions $E[k(\xi_j(\epsilon_i) - \xi_j(z_i'\gamma))$ where $k$ is some function of $z_i$ and $z_i'\gamma$. Pagan and Ullah (1997) pursue the issue of what are the optimal functions to obtain efficiency.

The above estimators avoid the difficulties associated with misspecifying the distribution of the selection equation error by estimating the first step semi-parametrically. Newey (1997), however, establishes the conditions for the consis-
tency of two-step estimators in the presence of misspecification of the distribution error. Newey employs the results of Ruud (1993) which notes the index can be estimated consistently, in the presence of misspecification of the error distribution, up to scale provided \( E[z_i | z' \gamma] = A + B(z' \gamma) \) for fixed vectors \( A \) and \( B \). He then employs the index restriction above to show a consistent estimate of \( \beta \) can be obtained in the following way. First estimate the index assuming some distribution or through OLS. Then include the index in the second step as an additional regressor. Thus, it is similar to the first of Newey’s two procedures outlined above. The key result in Newey (1997) is that rather than estimate \( \gamma \) semi-parametrically it is sufficient to estimate it by some quasi maximum likelihood procedure. This is an important result as it allows the estimation of the first step to be conducted under some maintained distributional assumption and the second step estimates remain consistent even if this assumption is violated. Moreover, under the assumptions of the model the second step only requires a single correction term. This avoids the choice of the number of correction terms.

An alternative approach to the elimination of selection bias is based on an estimation strategy suggested by Robinson (1988) in which the endogeneity is purged from the model through a differencing process. Powell (1987) exploits the index restriction in estimation by identifying observations by their value of this single index. The underlying intuition is that if two observations \( i \) and \( j \) have similar values for the single index generating the selection bias, then it is likely that subtracting the \( j \)th observation from the \( i \)th observation will eliminate the selection bias. Powell (1987) suggests an instrumental variable estimator based on pairwise comparisons of all observations in the sample where the contribution of each comparison is weighted by the difference in the values of the single index. The estimator of \( \beta \), denoted \( \beta_p \), has the form:

\[
\beta_p = \left\{ \left[ \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} m_{ij}(w_{ij}x_{ij})' \right]^{-1} \left[ \left( \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} m_{ij}(w_{ij}y_{ij})' \right) \right] \right\}
\]

where \( w_{ij}, x_{ij}, y_{ij} \) denote \( (w_i - w_j), (x_i - x_j), \) and \( (y_i - y_j) \) respectively and \( m_{ij} \) captures a weight depending on the distance between the values of the single indices for the \( i \)th and \( j \)th observations; and the \( w_{ij} 's \) denote some chosen instruments. The weight is constructed such that observations that are nearby, in terms of the single index, have a greater contribution than those far apart. As the weights are unobserved, the first step is semi-parametric estimation of the single indices \( z_i' \gamma \) and \( z_j' \gamma \).

A similar approach, based on differencing out the selectivity bias, first uses the index restriction to rewrite the primary equation as:

\[
y_i = x_i' \beta + g(z_i' \gamma) + \eta_i; \quad i = 1, \ldots, n.
\]

With an estimate of \( z_i' \gamma \) I condition (17) on \( z_i' \gamma \) to get:

\[
E[y_i | z_i' \gamma] = E[x_i | z_i' \gamma]' \beta + g(z_i' \gamma); \quad i = 1, \ldots, n.
\]

Subtract this conditional expectation from (17) to get:

\[
y_i - E[y_i | z_i' \gamma] = \{x_i - E[x_i | z_i' \gamma]\}' \beta + \eta_i; \quad i = 1, \ldots, n
\]
which can be estimated by OLS because the component reflecting the selection bias has been eliminated. This is even closer to the approach suggested by Robinson (1988). Newey (1990) notes that the Robinson estimator can be implemented as an instrumental variables estimator. In this context (9) would be estimated by using \(\{x_i - E[x_i|z_i']\gamma]\) as instruments for \(x_i\). As \(z_i'\gamma\) and \(E[.|z_i']\gamma]\) are unobserved, it is necessary to estimate them prior to estimation of (18).

Lee (1994) suggests estimation of (18) by instrumental variables. Lee’s model also incorporates endogenous regressors in the primary equation. Given the structure of the estimator Lee describes his procedure as semi-parametric two-stage least squares. To ensure his estimator has desirable properties, Lee employs the trimming function \(\tau(x)\). Lee’s estimator is then defined as:

\[
\hat{\beta}_{SLS} = \left[\hat{\hat{x}}'\hat{\hat{x}}^{-1}\hat{\hat{x}}'\hat{\hat{y}}\right]^{-1}\left[\hat{\hat{x}}'\hat{\hat{x}}^{-1}\hat{\hat{x}}'\hat{\hat{y}}\right]
\]

where \(\hat{\hat{x}}, \hat{\hat{y}}\) are matrices with typical elements \(\{\tau(x)(z_i - E[z_i|x_i'])\}, \{\tau(x)(x_i - E[x_i|z_i']\gamma]\) and \(\{\tau(x)y_i - E[y_i|z_i']\gamma]\), respectively.

A closely related estimator to that proposed by Powell (1987) is Ahn and Powell (1993). The innovation in the Ahn and Powell procedure is their use of nonparametric kernel methods to compute the propensity scores \(Pr[d_i = 1|z_i]\) and \(Pr[d_i = 1|z_i]\). They then use these probabilities in place of the estimated single indices \(z_i'\gamma\) and \(z_j'\gamma\) in the computation of the weights \(m_{ij}\). This is an important variation on Powell (1987) as it relaxes the single index assumption. It is accompanied by a substantial increase in computational requirements, however, as it is necessary to estimate the first step nonparametrically. Although the second step is \(\sqrt{n}\) consistent, one would expect some efficiency loss due to the manner in which the first step is estimated. This issue is addressed by Newey and Powell (1993) who examine the efficiency bounds for selection models where the index restriction is relaxed. Another study that relaxes the index restriction is Choi (1990). However, rather than employ kernel methods in the first step to compute the propensity score, \(Pr[d_i = 1|z_i]\), Choi employs series expansions.

A relative unexplored issue in the estimation of the sample selection model is heteroskedasticity. Heteroskedasticity creates problems for consistency but it is relatively difficult to tackle without distributional assumptions. This follows from the inability to distinguish between heteroskedasticity of an unknown form and the unspecified relationship between the single index and the correction factor. Donald (1995) addresses this issue by first assuming \(\varepsilon_i\) and \(v_i\) are bivariate normal with covariance \(\Omega(z_i)\) where the diagonals of \(\Omega(z_i)\) are \(\sigma_{\varepsilon_i}(z_i)\) and \(\sigma_{\sigma_i}(z_i)\) with off-diagonal \(\sigma_{\varepsilon\sigma}(z_i)\). From bivariate normality one can write:

\[
E[y_i|z_i, d_i = 1] = x_i'\beta + \frac{\sigma_{\varepsilon_i}(z_i)}{\sigma_{\sigma_i}(z_i)} \lambda \left(z_i'\gamma\right)
\]

where \(\lambda(\cdot)\) is the inverse Mills ratio function. Donald writes \(q(z_i) = \sigma_{\varepsilon_i}(z_i)/\sigma_{\sigma_i}(z_i)\) and \(h(z_i) = z_i'\gamma/\sigma_{\sigma_i}(z_i)\) and rewrites (19) as:

\[
E[y_i|z_i, d_i = 1] = x_i'\beta + q(z_i)\lambda(h(z_i)).
\]

Dividing the regression form of (20) by \(\lambda_i = \lambda(h(z_i))\) gives:
where given estimates of $\lambda_i$, one can eliminate the term $q(z_i)$ via the methods discussed above although the error in (21) is heteroskedastic. Donald estimates the index equation nonparametrically via series expansions and employs the normality assumption to estimate $\lambda_i$. He then replaces $\lambda_i$ with its estimate in (21) and computes the following expectation:

$$E \left( \frac{y_i}{\lambda_i} \right) | z_i = E \left( \frac{x_i}{\lambda_i} \right) | z_i \beta + q(z_i)$$

and subtracts (22) from (21). He then estimates $\beta$ by OLS over the differenced sample and shows the resulting estimator is $\sqrt{n}$ consistent and asymptotically normal.

The final semi-parametric procedure for the conventional sample selection model that I consider was proposed by Ichimura and Lee (1991) and is based on their estimator for models with multiple indices. Although this procedure is a single equation estimator, it is well motivated in the sequential equation framework. Recall that the model for the subsample has the following form:

$$E[y_i | z_i, d_i = 1, z_i] = x_i' \beta + g(z_i' \gamma); \quad i = 1, \ldots, n.$$  

which implies:

$$E[(y_i - x_i' \hat{\beta}) | z_i' \gamma] = g(z_i' \gamma).$$

Equations (23) and (24) characterize the relationship between $\beta$, $\gamma$, and $g(\cdot)$. The Ichimura and Lee procedure is based on the following iterative nonlinear least squares approach. With estimates of $\beta$ and $\gamma$ employ (24) to estimate $g(\cdot)$ nonparametrically. Then using (23) and our estimate of $g(\cdot)$ one can estimate $\beta$ and $\gamma$. Ichimura and Lee show that this provides consistent and asymptotically normal estimates for $\beta$ and $\gamma$.

It was noted above that the intercept in the primary equation was not identified through the two-step semi-parametric procedures. However, often the intercept may have economic content of interest. Heckman (1990) suggests the following estimator for the intercept:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} (y_i - x_2 \hat{\beta}_2) * d_i * I(z_i' \gamma > w)}{\sum_{i=1}^{n} d_i * I(z_i' \gamma > w)}$$

where the $x$ and $\beta$ vectors are partitioned into $[1 : x_2]$ and $[\beta_1 : \beta_2]$ respectively and $w$ reflects a smoothing parameter. Thus the basic idea is to get the average value of the deviation $y_i - x_2 \hat{\beta}_2$ for observations where the expected value of the error approaches zero as $n$ goes to infinity. Andrews and Schafgans (1996) adopt a similar approach but replace the indicator function with a smooth function.

Although the parametric two-step procedures have proven extremely popular with those doing empirical work, an examination of the literature indicates the semi-
parametric methods have been less frequently employed. This is partially due to the relative difficulty in implementation and the estimation of the associated covariance matrices required for inference. It is also partially due to a growing feeling that the parametric procedures perform well if the conditional mean of the model is correctly specified. Two papers that provide estimates from a range of two-step procedures are Newey, Powell and Walker (1990) and Melenberg and van Soest (1993). Newey, Powell and Walker (1990) examine the data on married women’s hours of work employed in Mroz (1987) and employ a number of two-step semi-parametric procedures that they compare with the two-step parametric models. They find very little difference between the point estimates and conclude, on the basis of their findings in combination with those of Mroz, that specification of the regression function appears to be more important than the error distributions for these data. Lee (1990) investigates the efficiency loss from employing the semi-parametric procedure of Powell and the inefficient procedure proposed by Newey. Lee concludes from Monte Carlo experiments that those estimators have variances ten times the size of the semi-parametric efficiency bounds. This suggests the efficiency losses are nontrivial.

Although I do not apply all the above estimators to our wage equation example, I do employ the simpler of the two Newey (1988) procedures. I adopt an even simpler version in that I estimate the index parametrically by Probit. I employ this index as the rejection of normality in the participation equation is only relatively weak. In Column 4 of Table 1, I provide the resulting wage equation parameters when the number of approximating terms is two.10 This was chosen on the basis of the $t$-statistics on the additional terms. The striking feature of this column is the similarity of the results with the parametric two-step estimates. This suggests that the inclusion of these terms is essentially performing the same task as the inclusion of the inverse Mills ratio. A similar finding was obtained when the index was computed as the predicted values from regressing $d_i$ on $z_i$. However, this suggests that the absence of normality is not the cause of the unexpected coefficients reported in Column 3. Despite the exclusion restrictions I employ it appears the collinearity of the correction term with the regressors is the cause of the movement in the coefficients. Although I do not report the results, I also estimated the model in the manner suggested by Newey (1997). I did this by first using the index from the Probit model and secondly, the index from the linear probability model. In both instances the estimates were almost identical to those reported in Columns 3 and 4. This suggests the assumptions required by Newey (1997) are satisfied in this application.

C. Conditional Expectations and Bounds

Although I have focused on the estimation of the $\beta$, one may be interested in the conditional expectation of $E[y_i|z_i, d_i]$.11 This originally arose in Lee (1978) and Willis and Rosen (1979) where the model comprised two mutually exclusive and

---

10. For the remainder of the paper I do not account for the estimation of $\gamma$ in the calculation of the second step standard errors. I choose to do this in order to avoid the issues which arise in adjusting these two-step covariance matrices.

11. When $d$ is a “treatment” the objective may be to evaluate the impact of the “treatment effect.” This may be defined as $E[y_i|x_i, z_i, d_i = 1] - E[y_i|x_i, z_i, d_i = 0]$. Discussion related to the estimation of treatment effects can be found in, for example, Heckman (1990).
Table 1
Cross-Sectional Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS</th>
<th>MLE</th>
<th>Parametric Two-Step</th>
<th>Semiparametric Two-Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>-.113</td>
<td>-.069</td>
<td>-.046</td>
<td>-.033</td>
</tr>
<tr>
<td></td>
<td>(.039)</td>
<td>(.045)</td>
<td>(.062)</td>
<td>(.040)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.049</td>
<td>-.037</td>
<td>-.025</td>
<td>-.027</td>
</tr>
<tr>
<td></td>
<td>(.046)</td>
<td>(.055)</td>
<td>(.073)</td>
<td>(.046)</td>
</tr>
<tr>
<td>Rural</td>
<td>-.244</td>
<td>-.176</td>
<td>-.139</td>
<td>-.137</td>
</tr>
<tr>
<td></td>
<td>(.039)</td>
<td>(.044)</td>
<td>(.066)</td>
<td>(.042)</td>
</tr>
<tr>
<td>School</td>
<td>.112</td>
<td>.068</td>
<td>.049</td>
<td>.043</td>
</tr>
<tr>
<td></td>
<td>(.010)</td>
<td>(.011)</td>
<td>(.022)</td>
<td>(.014)</td>
</tr>
<tr>
<td>Union</td>
<td>.095</td>
<td>.071</td>
<td>.108</td>
<td>.111</td>
</tr>
<tr>
<td></td>
<td>(.037)</td>
<td>(.042)</td>
<td>(.044)</td>
<td>(.037)</td>
</tr>
<tr>
<td>Exper</td>
<td>.012</td>
<td>.088</td>
<td>.080</td>
<td>.096</td>
</tr>
<tr>
<td></td>
<td>(.034)</td>
<td>(.032)</td>
<td>(.040)</td>
<td>(.035)</td>
</tr>
<tr>
<td>Exper2</td>
<td>-.005</td>
<td>-.003</td>
<td>-.001</td>
<td>-.002</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td></td>
<td></td>
<td>-.121</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td></td>
<td></td>
<td>-.902</td>
<td></td>
</tr>
<tr>
<td>$(z'\gamma)$</td>
<td></td>
<td></td>
<td></td>
<td>.419</td>
</tr>
<tr>
<td>$(z'\gamma)^2$</td>
<td></td>
<td></td>
<td></td>
<td>(.166)</td>
</tr>
<tr>
<td>Constant</td>
<td>-.409</td>
<td>.526</td>
<td>.955</td>
<td>.110</td>
</tr>
<tr>
<td></td>
<td>(.161)</td>
<td>(.164)</td>
<td>(.400)</td>
<td>(.176)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.139</td>
<td>.161</td>
<td>.164</td>
<td></td>
</tr>
</tbody>
</table>

exhaustive sectors. Wages were observed for each individual for the sector in which the individual was located. One objective in those studies was to compute the wage for the sector in which the individual was not observed. I now examine the case where $y_i$ is only observed for a subsample. The generalization to the case of $y_i$ observed for everyone, although individuals are in different sectors, is straightforward.

Suppose the two "sectors" refer to market and nonmarket employment and the variable $y$ reflects the offered market wage. Furthermore, assume the errors are bivariate normally distributed and I estimate the model by the Heckman two-step procedure. Denote the parameter estimates for the subsample of those engaged in market employment, as $\beta_M$. Consider the expectation $CE_{il} = x_i'\beta_M$. This represents the expected market wage for an individual randomly selected from the sample. That is, the conditioning set does not contain any information regarding the sector in which
the individual is actually located. This is the approach adopted in Lee (1978) and Willis and Rosen (1979). However, the expectation $CE_{1i}$ can be “improved” via the inclusion of information relaying the chosen sector. For example, $CE_{1i} = E[y_i|z_i, d_i = 1] = x_i'\beta_M + \mu \lambda_i^M$ represents the expected wages for those already located in the respective sectors, noting that the $\lambda^M$ denotes the inverse Mills ratios for those in the market sector and the $\mu$ is the estimated parameter capturing the covariance between the errors across equations. This latter expectation varies from $CE_{1i}$ in that it includes the respective returns to the unobservables associated with market sector (see, for example, Vella 1988). Accordingly, one may consider the following counterfactual wages as conditional expectations of interest. Namely, $CE_{2i} = E[y_i|z_i, d_i = 0] = x_i'\beta_M + \mu \lambda_i^N$ which represents the expected wages for those in the nonmarket sector if they obtained market employment noting that $\lambda^N$ is the inverse Mills ratio for those in the nonmarket sector. Once again the term $\mu \lambda_i^N$ captures the market return to the unobservables. Lee (1995) extends and generalizes this approach by providing a general strategy for estimating the conditional expectations of the outcomes not chosen. Furthermore, Lee provides the formulae for the conditional expectations of outcomes models with polychotomous outcomes in models with sample selection bias.

A shortcoming of the above approach is its reliance on distributional assumptions to obtain estimates of the $\lambda_i$ and $\mu$. An additional consideration is the use of the residual in the construction of the counterfactual wages. Anderson-Schaffner (1994) shows this is only valid when the distribution, and implicit prices, of the various components comprising the unobservables are the same for the two subsamples $d_i = 0$ and $d_i = 1$. Although this appears to be a relatively innocuous assumption for many applications it is useful to note this requirement.

Manski (1989) focuses on the estimation of bounds for the conditional expectation (that is, $E[y_i|z_i]$ over the support of $z_i$) when $y_i$ is only observed for either $d_i = 1$ or $d_i = 0$ but not both. Manski considers the case where $z_i$ and $d_i$ are observed over the whole sample and $E[y_i|z_i, d_i = 1]$ is observed. First, it is straightforward to see that:

\begin{equation}
E[y_i|z_i] = E[y_i|z_i, d_i = 1] \Pr[d_i = 1|z_i] + E[y_i|z_i, d_i = 0] \Pr[d_i = 0|z_i].
\end{equation}

Manski assumes that the support of $y_i$ conditional on $d_i = 0$ and $z_i$ is known and lies in the interval $[K_{Lz}, K_{Uz}]$ which implies $K_{Lz} \leq E(y_i|z_i, d_i = 0) \leq K_{Uz}$. This, in turn with (25), implies:

\begin{align*}
E(y_i|z_i, d_i = 1) \Pr(d_i = 1|z_i) + K_{Lz} \Pr(d_i = 0|z_i) & \leq E(y_i|z_i) \\
& \leq E(y_i|z_i, d_i = 1) \Pr(d_i = 1|z_i) + K_{Uz} \Pr(d_i = 0|z_i).
\end{align*}

The components of the bound are readily available in most contexts and Manski discusses the methodology for implementing the bound. The first important feature of (26) is that rather than focusing on a point estimate it provides a bound. A second feature is that it can be implemented nonparametrically as the components of (26) can be estimated from sample data without the imposition of parametric assumptions. A possible criticism is that the estimated bounds may be too wide to be informative. Although this represents information in itself, Manski (1994) shows how the bounds can be tightened through the use of additional information such as functional form.
VI. Sample Selection Models with Alternative Censoring Rules

The estimators discussed thus far have been limited to a dependent variable in the selection equation that takes the value of zero or one. Although the vast majority of empirical studies address this form of selection bias either naturally or through manipulation of the model, nevertheless the selection process cannot be easily put into this framework for a wide family of models. Moreover, the dependent variable in the reduced form may contain additional information that can be exploited in estimation. Some of these commonly encountered forms of selectivity, however, can also be put into the above framework where the only reformulation required is a slight generalization of the original model and the specification of the censoring function determining the selection. The model has the form:

\[ y^* = x_i'\beta + \varepsilon_i; \quad i = 1, \ldots, N \]

\[ d^*_i = z_i'\gamma + v_i; \quad i = 1, \ldots, N \]

\[ d_i = h(d_i) \]

\[ y_i = j(d_i, y^*_i) \]

where I assume that \( \varepsilon_i \) and \( v_i \) are bivariate normally distributed and at this point I restrict \( \beta \) to be constant for all values of \( d_i \). The selection mechanism now has the generic form \( h(\cdot) \) and the process determining the observability of \( y_i \) has the form \( j(\cdot) \).

A. Ordered Censoring Rules

The first case I examine is where \( h(\cdot) \) generates a series of ordered outcomes through the following rule:

\[ d_i = 1 \text{ if } -\infty < d^*_i \leq 0; \quad d_i = 2 \text{ if } 0 < d^*_i \leq \mu_1; \ldots, \]

\[ d_i = 3 \text{ if } \mu_1 < d^*_i \leq \mu_2; \ldots, d_i = J \text{ if } \mu_{J-1} < d^*_i; \]

where the \( \mu \)'s denote separation points satisfying \( \mu_0 < \mu_2 \cdots < \mu_J \) where \( \mu_0 \) and \( \mu_J \) equal \( -\infty \) and \( +\infty \) respectively, and noting I may only observe \( y_i \) for a specified value(s) of \( d_i \). That is, the \( j(\cdot) \) function specifies that \( y_i = y^*_i \ast I(d_i = j) \). It is now necessary to incorporate the ordering of the outcomes when accounting for the selection bias. This model is considered in Vella (1993) and following that general methodology I estimate the first step by ordered Probit to obtain estimates of the \( \mu \)'s and \( \gamma \)'s. I then compute the generalized residuals for each outcome, \( d_i = j \), which take the form:

\[ \frac{\phi(\hat{\mu}_{j-1} - z_i'\hat{\gamma}) - \phi(\hat{\mu}_j - z_i'\hat{\gamma})}{\Phi(\hat{\mu}_j - z_i'\hat{\gamma}) - \Phi(\hat{\mu}_{j-1} - z_i'\hat{\gamma})} \]
which I include as the selection correction rather than the Inverse Mills ratio. I can then estimate over the various subsamples corresponding to different values of \( d_i \). A similar approach has also been suggested in specific contexts by Jimenez and Kugler (1987) and others. Following Lee (1982) it is possible to capture departures from normality in the \( \epsilon_i \) by powering up the generalized residual by the index \( z_i^\gamma \) and its higher powers.

A second model that exploits the ordering in the selection equation is the continuous selection model of Garen (1984). Garen considers the case where \( d_i^* \) is continuously observed (namely, \( d_i = d_i^* \)) and a subset of the \( N \) observations corresponds to each permissible value of \( d_i^* \). Rather than estimate a different set of parameters for each value of \( d_i^* \) Garen suggests estimation of:

\[
y_i = \mathbf{x}_i \beta + d_i \alpha + (d_i^* v_i) \theta_1 + v_i \theta_2 + \eta_i
\]

where I have replaced the \( y_i^* \) and \( d_i^* \) with their observed counterparts. To implement this procedure I obtain an estimate of \( v_i \), denoted \( \hat{v}_i \), by estimating the reduced form by ordinary least squares. I then replace \( v_i \) with \( \hat{v}_i \) and estimate (31) by ordinary least squares. Note that the ordering of the outcome variable in the selection equation is important as it ensures the residual has the appropriate interpretation.

**B. Tobit Type Censoring Rule**

A further type of commonly encountered censoring is one in which the dependent variable in the selection equation is partially observed. This model is also known as Tobit type three. For example, in the labor supply context one often not only observes if the individual works but also observes the number of hours they work. This information can be exploited in estimation. To capture this process specify \( h(\cdot) \) as:

\[
d_i = \begin{cases} 
d_i^* & \text{if } d_i^* > 0 \\
0 & \text{otherwise}
\end{cases}
\]

and specify the \( j(\cdot) \) function as:

\[
y_i = y_i^* \cdot I(d_i > 0).
\]

Thus, the censoring variable is observed whenever it is greater than some threshold and equal to zero otherwise. Furthermore, the dependent variable on the primary equation is only observed when the censoring variable is positive. The appropriate way to estimate the censoring equation is by Tobit. Following Vella (1993) I compute the generalized residuals that take the form:

\[
(1 - I_i) * \left( \frac{-\phi(z_i^\gamma)}{\Phi(z_i^\gamma)} \right) + I_i * (d_i - z_i^\gamma).
\]

Note that when the second step estimation is only over the sample for which \( I_i = 1 \), denoting \( d_i > 0 \), the residuals have a very simple form. Wooldridge (1994) notes that this approach does not require exclusion restrictions as one obtains identification via the variation in \( d_i \) providing it is not included in the conditional mean of \( y_i \). It
is clear that this procedure is closely related to the original Heckman (1976) two-step procedure. Accordingly, I refer to this as a control function procedure. Note that the strong reliance on normality of this control function estimator could be relaxed in a number of ways. First, one could relax normality in the second step by taking a series expansion around \( \hat{v} \), which, for the observations corresponding to \( I_i = 1 \), is equal to \( d_i - z_i'\gamma \). Second, to relax normality in both steps one could estimate \( \gamma \) semi-parametrically by using the procedures in Powell (1984, 1986). Using this semi-parametric estimate of \( \gamma \), the residuals for the uncensored observations could be estimated. Then the primary equation could be estimated by OLS while including the estimated residual, and possibly its higher order terms, as additional regressors.

A number of semi-parametric procedures exist for this model. Lee (1994) generalizes the approach in Lee (1992) which proposes a semi-parametric estimator for the truncated regression model. The basic idea is similar to Powell (1987) and Ahn and Powell (1994) for the conventional form of selection. Lee notes that approach of Powell only exploits the index restriction, although it is possible to employ the underlying assumption that the joint distribution of \( \varepsilon \) and \( v \) is independent of \( z \) in estimating the model. To understand the Lee approach, let us introduce some additional notation. First, the variables \( y, d, x, z, \varepsilon, \) and \( v \) are the random variables appearing in the latent model and the terms \( z_i, \varepsilon_i, \) and \( v_i \) denote sample observations. Finally, the \( z_i \) denotes a parameter. Thus, the expectation \( E[y|d > 0, z] \) is the conditional expectation of \( y \) conditional on the set \( \{(v,z)|v > -z_i'\gamma, z_i'\gamma > z_i'\gamma\} \) where \( z_i \) can be chosen. Now recalling \( x \) is a subset of \( z \), I can take the conditional expectation of \( y \) conditional on \( d > 0 \) and \( z \). This gives:

\[
E[y|d > 0, z] = x'\beta + E[\varepsilon|v > -z_i'\gamma, z_i'\gamma]
\]

remembering \( d > 0 \) implies \( v > -z_i'\gamma \). Now the joint density of \( \varepsilon_i \) and \( v_i \) conditional on \( d > 0 \) and \( z_i'\gamma \) is the same as the density of any \( \varepsilon_j \) and \( v_j \) conditional on \( v_j > -z_i'\gamma, z_i'\gamma \), and \( z_j'\gamma > z_i'\gamma \). Now, note that \( z_i'\gamma > z_i'\gamma \) and \( v > -z_i'\gamma \) implies \( d = z_i'\gamma + v > 0 \). This implies that for each \( z_i \):

\[
E[y|v > -z_i'\gamma, z_i'\gamma > z_i'\gamma] = E[x|z_i'\gamma > z_i'\gamma]'\beta + E[\varepsilon|v > -z_i'\gamma, z_i'\gamma > z_i'\gamma].
\]

Lee then defines an estimator:

\[
\hat{\beta} = \left\{ \sum_i (x_i - E[x|z_i, \gamma])' (x_i - E[x|z_i, \gamma]) \right\}^{-1} \sum_i (x_i - E[x|z_i, \gamma])' (y_i - E[y|z_i, \gamma])
\]

where, following Lee, \( E[(\cdot)|z_i, \gamma] = E[(\cdot)|v > -z_i'\gamma, z_i'\gamma > z_i'\gamma] \). To implement this estimator a first step estimator of \( \gamma \) is necessary. For the simulation results reported in Lee (1994) \( \gamma \) is estimated via the procedure in Lee (1992). Estimates of the expectations that appear in the formula of \( \hat{\beta} \) can be obtained via a bivariate kernel. Finally, note that although I do not show it explicitly in the formula for \( \hat{\beta} \), it is necessary to trim the data on the basis of the values of the single index \( z_i'\gamma \). This is done to satisfy conditions necessary for consistency.

Honore, Kyriazidou, and Udry (1997) provide two semi-parametric procedures of this model that are applicable under different assumptions. The first assumes conditional symmetry of the disturbances (\( \varepsilon, v \)). By assuming that, conditional on \( (x, z) \), the disturbances (\( \varepsilon, v \)) are distributed like \((-\varepsilon, -v)\) the marginal distribution of \( v \) is symmetric and thus the selection equation can be estimated by the procedures
proposed by Powell (1984, 1986). To estimate the primary equation, they restrict
the sample to the observations that satisfy \(-z'y < d < z'y\). For this trimmed sample,
the conditional symmetry assumption implies that the \(e\) are symmetrically distributed
around 0. Accordingly, the selection bias has been purged from the trimmed data.
Honore, Kyriazidou, and Udry (1997) propose least squares or least absolute devia-
tion estimation over the trimmed sample. The second estimator assumes \((\epsilon, v)\) are
independent of \(z\). The first step involves estimation by some appropriate procedure
to obtain \(y\). The second step employs a pairwise trimming that is similar to that in
Honore and Powell (1994). The underlying idea is that the difference of independent
and identically distributed random variables is symmetrically distributed around
zero. Due to the censoring and selection processes, however, this differencing strat-
egeny will not produce differences that are symmetrically distributed around zero. The
authors propose the following trimming strategy: \(v_i > \max(-z_i'\gamma, -z_i'\gamma)\) and \(v_j >
\max(-z_j'\gamma, -z_j'\gamma)\). Conditional on this trimming the \(\epsilon_i\) and \(\epsilon_j\) are independent and
identically distributed and thus the pairwise differences based on this trimmed sam-
ple will be distributed symmetrically around zero. Thus, under this assumption of
independence the second step can be estimated using the pairwise differenced estima-
tors of Honore and Powell (1994). An attractive feature of the procedures suggested
by Honore, Kyriazidou, and Udry (1997) is that their implementation does not re-
quire the use of smoothing parameters.

Chen (1996) also proposes two estimators for this model under the assumption
of independence. The first estimator is based on a trimming scheme such that the
term required to account for the selection bias is equal to a constant. Thus, by estimat-
ating over the trimmed sample the correction term is captured in the constant. For
example, Chen suggest the following conditional expectation \(E[y|z, v > 0, z'\gamma > 0]\), which, under the model’s structure has the regression form \(y_i = x_i'\beta + \tau\) where
\(\tau\) is a constant for the trimmed sample. Thus, the estimation strategy is to obtain an
estimate of \(\gamma\) and \(v_i\) and then estimate \(\beta\) and \(\tau\) by OLS over the trimmed sample
which satisfy \(v_i > 0\) and \(z_i'\gamma > 0\). Chen also suggests breaking the data into multiple
groups and estimating a different \(\tau\) for each group.

Chen’s second estimator also utilizes the assumption of independence but over-
comes the selection bias induced by the incorrect specification of the conditional
mean by estimating the correction term for each observation. Chen estimates \(E[y_i -
x_i'\beta | v_i > -z_i'\gamma, z_i]\) by

\[
\hat{E}(z_i, \beta) = \frac{1}{n - 1} \sum^n_{j \neq i} D_i(\hat{\gamma})(y_j - x_j'\beta)
\]

where \(\hat{\gamma}\) is some appropriate first step estimator and \(D_{ij} = I(d_i - z_i'\gamma > -z_i'\gamma,
-z_i'\gamma > -z_i'\gamma)\). Chen then estimates the second step by weighted semiparametric
least squares by minimizing the following, suitably weighted function

\[
\frac{1}{n} \sum^n_{i=1} (y_i - x_i'\beta - \hat{E}(z_i, \beta))^2.
\]
Table 2
Tobit Type Censoring Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control Function</th>
<th>HKU Estimator</th>
<th>LV Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>-0.129</td>
<td>-0.118</td>
<td>-0.127</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.039)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.052</td>
<td>-0.052</td>
<td>-0.055</td>
</tr>
<tr>
<td></td>
<td>(0.048)</td>
<td>(0.048)</td>
<td>(0.048)</td>
</tr>
<tr>
<td>Rural</td>
<td>-0.257</td>
<td>-0.241</td>
<td>-0.247</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
<td>(0.038)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>School</td>
<td>0.123</td>
<td>0.111</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.010)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>Union</td>
<td>0.089</td>
<td>0.092</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.034)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Exper</td>
<td>0.118</td>
<td>0.110</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.034)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Exper2</td>
<td>-0.005</td>
<td>-0.004</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>( \hat{\nu}_i )</td>
<td>.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.146</td>
<td>0.135</td>
<td>0.133</td>
</tr>
<tr>
<td>Observations</td>
<td>1569</td>
<td>1569</td>
<td>1569</td>
</tr>
</tbody>
</table>

Lee and Vella (1997) suggest an estimator based on (17). They note that the selection bias operates through the reduced form. That is, the selection bias is generated by the presence of \( \nu_i \) in the primary equation. Accordingly, they suggest purging the model of the component contaminated with \( \nu_i \). To do this they propose an estimator based on:

\[
(32) \quad y_i - E[y_i|\nu_i] = \{x_i - E[x_i|\nu_i]\}'\beta + \eta_i; \quad i = 1, \ldots, n.
\]

To implement this procedure they require a \( \sqrt{n} \) consistent semiparametric estimator of the censored model to obtain an estimate of \( \gamma \) and they suggest the use of the estimators proposed by Powell (1984, 1986). They define the residuals for the subsample corresponding to \( d_i > 0 \) as \( \nu_i = d_i - z'\gamma \) and propose estimation by the methodology in (32). Following the logic of Newey (1990), it is clear that in the case of Tobit type censoring it is possible to employ an instrumental variable approach by using \( \{x_i - E[x_i|d_i - z'\gamma]\} \) as instruments for \( x_i \).

I again consider estimation of the wage equation from the NLS data set examined above. As the data provides information on weekly hours worked, I can estimate the selection process by a procedure that accounts for the observability of positive hours for the workers in the sample. In Table 2, I provide the estimates from the two-step control function procedure in addition to the estimates from the procedure that assumes conditional symmetry of Honoré, Kyriazidou, and Udry (1997), denoted...
A number of features are worth noting from this Table. First, the three sets of estimates are very similar. Second, despite the evidence of selection bias, as revealed by the t-statistic on the residual in Column 1 of this table, the estimates are similar to the OLS estimates reported in Table 1. Third, unlike the previous table where the presence of selectivity appeared drastically to change the point estimates the results in this table appear more similar to the OLS estimates. Note that this cannot be attributed to the lack of normality uncovered above as the control function estimates are fully parametric and require normality. One suspects that the stability of the results is partially due to the variation in the residual. Recall that in the conventional selection model all the variation in the correction term comes from the variation in the index as all the observations in the selected subsample have the same dependent variable. This is not true for the Tobit type censoring as variation in the number of hours also contributes to identification. It appears that in this application this additional form of identification is important.

C. Unordered Censoring Rules

A feature of these two extensions of the selectivity model is that although estimation is somewhat complicated by the presence of multiple outcomes it is greatly simplified by the imposition of ordering on the outcomes. When imposing such ordering is not possible, it is necessary to treat the outcomes in the first step as unordered. One possibility would be to estimate the first step by multinomial Probit and then compute the corresponding generalized residual to include as an additional regressor. Such an approach, however, will be difficult to implement whenever there are more than three outcomes. Two alternative approaches are those outlined by Lee (1983a), and Hay (1980) and Dubin and McFadden (1984). To analyze these approaches, consider the following model:

\[
Y_{si} = X_{si} \beta_s + \varepsilon_{si} \\
I_{si}^* = z_{si}' \gamma_s + \nu_{si}
\]

where the number of outcomes is given by \( s = 1, \ldots, M \). This model is characterized by a different parameter vector for each outcome. First, consider the approach of Lee (1983a). Assume the selection rule, determining the chosen outcome, is based on the following rule:

\[
I_i = s \text{ iff } I_{si}^* \geq \max_{j \neq s} I_{sj}^*; \quad j = 1, \ldots, M; \quad j \neq s.
\]

If I let \( \kappa_{si} = \max_{j \neq s} I_{sj}^* - \nu_{si} \), it follows from the selection rule that:

\[
I_i = s \text{ iff } \kappa_{si} < z_{si}' \gamma_s; \quad j = 1, \ldots, M; \quad j \neq s.
\]

Thus the model can now be characterized by a series of observed \( y_{si} \) that are only

12. I do not report the estimates, when applicable, for the intercepts.
13. In a Monte Carlo examination of these estimators Lee and Vella (1997) found that the estimators effectively accounted for selection bias in situations where the unadjusted OLS estimates were significantly different from the true values.
observed if \( \kappa_i < z_i'Y_i \). When the distribution function of \( \kappa_i \) is known, one can proceed in the same manner as for the binary choice model. That is, estimate:

\[
(34) \quad E[y_i|z_i, d_i = 1] = x_i'\beta + \rho \sigma_i \phi(z_i'\gamma)/G(z_i'\gamma)
\]

where either I assume the marginal distribution of the untransformed \( \varepsilon_i \) is normal, or I assume that the relationship between \( \varepsilon \) and transformed \( v_i \) is normal. I require a first step estimate of \( \gamma \) that accounts for the polychotomous nature of \( I_i \). A popular way to proceed is to assume that \( v_i \) has a type 1 extreme value distribution that then allows estimation of \( \gamma \) by multinomial Logit.

An alternative approach is found in Hay (1980) and Dubin and McFadden (1984). This is an extension of the Heckman (1979) estimator to models with polychotomous outcomes and is based on the use of a truncated conditional expectation function. Note from above that \( I_i = s \) iff \( z_i'Y_i - z_i'Y_i > v_i - v_{si} \). If \( \omega_i = z_i'Y_i - z_i'Y_i \) and \( \xi_{is} = v_i - v_{si} \), note that \( I_i = s \) iff \( \omega_i > \xi_{is} \). Thus, taking expectations of (33), conditional on the \( s \)th outcome being observed, gives:

\[
E[y_i|\xi_i] = x_i'\beta + E[\varepsilon_i|\xi_i]
\]

where \( \xi_i \) denotes the vector \( \xi_{si}, \ldots, \xi_{si} \). I can then employ the conditional expectation of the error term via the truncated conditional expectation function.

If one assumes that the \( v_{si} \) are i.i.d type 1 extreme value random variables the \( \xi_{si} \) have a multivariate logistic distribution. Thus:

\[
(35) \quad E[\varepsilon_i|\xi_i] = \sum_{r=1}^{M} \lambda_i[1 - e^{-\omega_i}F(\omega_i)^{-1} [\log F(\omega_i) - \omega_i e^{-\omega_i}F(\omega_i)]
\]

where \( F(\omega_i) = F(\omega_{i1}, \ldots, \omega_{in}) \); \( F(\cdot) \) denotes the multivariate logistic distribution and \( \lambda_i \) is a parameter vector.

An attraction of the Lee procedure is that it possesses a maximum likelihood form while the alternative approach relies on the linearity of the conditional expectation. An examination of published empirical work based on these types of models reveals the Lee approach, employing the correction shown in (34), is more popular than that in (35). Both procedures impose restrictions on the model. The first requires that one can characterize the relationship between \( \varepsilon_{si} \) and the vector of transformed \( v_{si} \) through a bivariate distribution while the second assumes the relationship \( E[\varepsilon_{si}|\xi_i] \) is linear and can be captured through the parameter vector \( \lambda_i \). Schmertmann (1994) shows that the Lee estimator imposes severe restrictions on the covariances between the errors in the indices equations and those in the equations of primary interest. More explicitly, the Lee estimator imposes that the covariance between the errors is the same sign for all outcomes while the Heckman type estimator allows any combination of signs. Schmertmann (1994) illustrates, through Monte Carlo simulations, that the incorrect imposition of this restriction can produce biased estimates.

D. Censoring Rules Based on Multiple Indices

A feature of many of these models, even in the case of unconventional forms of censoring, is that the selection bias is treated purely as a function of a single index. As I noted for the unordered model in many models, however, the sample selection
may be a function of multiple indices. Now consider the following recharacterization of our original model where the selectivity bias is a function of multiple indices:

\[ y_i^* = x_i' \beta + \varepsilon_i \]

\[ d_{1i} = I(z_i' \gamma_1 > -v_{1i}) \]

\[ d_{2i} = I(z_i' \gamma_2 > -v_{2i}) \]

\[ y_i = y_i^* \cdot d_{1i} \cdot d_{2i} \]

where \( I(\cdot) \) is an indicator function and the additional notation is obvious. The sample selection is now based on multiple indices and multiple criteria. The method of estimation relies crucially on: i) the relationship between \( v_{1i} \) and \( v_{2i} \), and ii) the observability of the two indices \( d_{1i} \) and \( d_{2i} \).

The simplest case is where the disturbances are jointly normally distributed; \( v_{1i} \) and \( v_{2i} \) are uncorrelated and both \( d_{1i} \) and \( d_{2i} \) are observed. Then, using the procedures discussed above, it is relatively straightforward to compute the following correction terms to include as regressors in the primary equation:

\[
E[e_i | z_i, d_{1i} = 1, d_{2i} = 1] = \left( \sigma_{\varepsilon_1} \sigma_{\varepsilon_2} \right)^{1/2} \left[ \phi(z_i' \gamma_1) / \Phi(z_i' \gamma_1) \right] + \left( \sigma_{\varepsilon_1} \sigma_{\varepsilon_2} \right)^{1/2} \left[ \phi(z_i' \gamma_2) / \Phi(z_i' \gamma_2) \right].
\]

To implement this model one first independently estimates (37) and (38) by Probit to obtain \( \gamma_1 \) and \( \gamma_2 \). The corresponding two Inverse Mills ratios can then be computed and included as correction terms in the primary equation.

Although this model is easily estimated, it restricts the error terms in the two censoring equations to be uncorrelated. This is an assumption that most empirical studies would be reluctant to impose. Moreover, one could imagine that the two different selection rules would be related in various ways above the nature of the correlation of their respective disturbances. Perhaps the most commonly encountered of this comes under the heading of partial observability examined by Poirier (1980).

In Poirier’s model neither \( d_{1i} \) nor \( d_{2i} \) is observed but I observe their product \( d_{3i} = d_{1i} \cdot d_{2i} \). Furthermore, I assume one only observes the \( y_i \)'s for the subsample corresponding to \( d_{3i} = 1 \). Poirier examines the conditions for the estimation and identification of \( \gamma_1 \) and \( \gamma_2 \) by maximum likelihood while employing \( d_{3i} \) as the dependent variable. Furthermore, he also shows:

\[
E[e_i | z_i, d_{3i} = 1] = \sigma_{\varepsilon_1} \left[ \phi(z_i' \gamma_1) / \Phi(z_i' \gamma_1) \right] \cdot \phi(z_i' \gamma_2) / \Phi(z_i' \gamma_2)
\]

where \( \Phi^b \) denotes the bivariate normal distribution; \( \rho_{12} \) denotes the correlation coefficient for \( v_1 \) and \( v_2 \); and I normalize \( \sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2 = 1 \). To adjust for sample selection in this model one computes the above two additional terms to include as regressors in the conditional mean function for \( y_i \).
The primary feature of the models in this section is that whether an individual is observed in the second step depends on the value of at least two indices in an earlier step. Accordingly, the models of Cragg (1971), Deaton and Irish (1982), and Blundell and Meghir (1987), discussed above, are also members of this family.

E. Conditional Maximum Likelihood

Thus far I have focused on models where the dependent variable in the selection equation is censored and I have a continuous variable in the primary equation. One often confronts a continuous, or partially continuous, dependent variable in the selection equation and a censored or continuous, or partially continuous, dependent variable in the selection equation and a censored or limited dependent variable in the primary equation. The model has the following structure:

(40) \( y_i^* = x_i'\beta + \theta d_i + \epsilon_i; \quad i = 1, \ldots, N \)

(41) \( d_i = z_i'\gamma + \nu_i; \quad i = 1, \ldots, N \)

(42) \( y_i = l(y_i^*); \)

(43) \( y_i = j(d_i) \)

where I assume that the error terms are jointly normally distributed with nonzero covariance and the \( l(\cdot) \) function maps the latent \( y_i^* \) into the observed \( y_i \) noting that at this stage the \( y_i \) is reported for the whole sample. This model is similar to (11)–(13) except that the censoring occurs in the primary equation and not the reduced form. The model in (40)–(42) is considered by Smith and Blundell (1986), where \( l(\cdot) \) generates Tobit type censoring, and by Rivers and Vuong (1989) for the case where \( l(\cdot) \) generates Probit type censoring. Single equation maximum likelihood estimation of (40), while accounting for the form of \( l(\cdot) \), will not produce consistent estimates of \( \beta \) due to the endogeneity of \( d_i \). One method of estimation is conditional maximum likelihood by which one first employs the bivariate normality assumption to rewrite (40) as:

\[
\begin{align*}
    y_i^* &= x_i'\beta + \theta d_i + \mu \nu_i + e_i; \quad i = 1, \ldots, N \\
    \end{align*}
\]

where \( \mu = \sigma_{\epsilon}/\sigma_{\nu}^2 \) and \( e_i \) is a zero mean and normally distributed error term. As \( \nu_i \) is normally distributed, one can obtain a consistent estimate as \( \nu_i = d_i - z_i'\hat{\gamma} \) where \( \hat{\gamma} \) denotes the OLS estimates from (41). One then estimates:

(44) \( y_i^* = x_i'\beta + \theta d_i + \mu \hat{\nu}_i + e_{ii}; \quad i = 1, \ldots, N \)

by maximum likelihood noting that \( e_{ii} = e_i + \mu (\nu_i - \hat{\nu}_i) \) has zero mean and, most important, is normally distributed. The normality is retained as \( \hat{\nu}_i \) is a linear transformation of normally distributed random variables. This differs from the model in (11)–(13) as the censoring of the endogenous variable in the reduced form results in the generalized residuals being nonlinear functions of normally distributed random variables and the variable \( (\nu_i - \hat{\nu}_i) \) is subsequently nonnormal. As the coefficient \( \mu \) captures the correlation between the errors a \( t \)-test on the null \( \mu = 0 \) is a test of the weak exogeneity of \( d_i \).

The conditional maximum likelihood estimator can be extended to the sample
selection case if the selection rule, captured by the \( j(\cdot) \) function in (43), is within a certain class of functions. For example, if \( y_i \) was only observed when \( d_i > 0 \) then the second step estimation would only involve the subsample satisfying this selection rule and one would still estimate the primary equation by maximum likelihood with \( \hat{v}_i \) included. Moreover, in this subsample case it is necessary to include \( \hat{v}_i \) even if \( \theta = 0 \), whereas this is unnecessary if I observe the whole sample. Finally, despite only observing \( y_i \), for specified values of \( d_i \) I am still able to estimate by maximum likelihood as the error term retains its normality despite the inclusion of \( \hat{v}_i \).

The above discussion illustrates that when the second step estimation of the primary equation is performed by maximum likelihood, it is necessary to impose some restrictions on the mapping from the first step parameter estimates and variables to the residuals operating as correction factors. More explicitly, the inclusion of the correction factor cannot corrupt the normality of the primary equation’s disturbance. This naturally does not apply to models estimated by full information maximum likelihood in which case the selection and primary equation’s dependent variables can take any sensible form and estimation can proceed providing the likelihood function can be constructed. As noted above, however, it is clear that maximum likelihood can be employed in the second step whenever the residuals are a linear function of the variables as this transformation preserves the assumed normality. One particular case of interest is considered by Vella (1992) who examines a model where the primary equation has a binary outcome variable and the selection equation has a dependent variable that is partially observed and has Tobit type censoring. For that model it is possible to perform the reduced form first step estimation over the entire sample by Tobit. One then estimates the Tobit residuals for the subsample corresponding to \( d_i > 0 \) which simply take the form \( \hat{v}_i = d_i - \hat{z}_i^\hat{\gamma} \) where the hats now denote the Tobit estimates. It is then possible to estimate the primary equation by Probit over the subset, satisfying \( d_i > 0 \) while including \( v_i \) as an explanatory variable. Although this is how one would proceed for a model with Probit type censoring, it is possible to estimate a number of models, depending on the form of \( l(\cdot) \), provided they require normality.

VII. Panel Data Estimators

Although sample selection is commonly confronted in cross-sectional studies, it is less frequently considered a concern in panel data estimation. This is partially due to the conception that fixed effects estimation will eliminate most forms of unobserved heterogeneity. Although certain forms of selection bias are eliminated by fixed effects estimators (see Verbeek and Nijman 1992a,b), other forms of selection bias and heterogeneity will not be eliminated. Recent papers have extended the cross-sectional results to the panel data context while exploiting the panel aspect of the data. A feature of these approaches is they provide more economic insight into the processes driving the selection bias and they identify the source of the heterogeneity. These approaches are based on methodologies outlined above, but it is useful to reconsider them in this alternative setting. Consider the following model:
\[(45)\quad y_{it}^* = x_{it}' \beta + \mu_i + \xi_t + e_{it} \]

\[(46)\quad d_{it}^* = z_{it}' \gamma + \alpha_i + \psi_t + v_{it} \]

\[(47)\quad d_{it} = 1 \text{ if } d_{it}^* > 0 \]

\[(48)\quad y_{it} = y_{it}^* \cdot d_{it} \]

where \(i, (i = 1, \ldots, N),\) continues to denote the individual and \(t, (t = 1, \ldots, T),\) denotes the panel. The dependent variable in the primary equation is only observed for the observations satisfying the selection rule (namely, \(d_{it}^* > 0\)). To introduce selection bias, assume the errors for each equation can be decomposed into an individual effect (\(\mu_i\) and \(\alpha_i\)), a time effect (\(\xi_t\) and \(\psi_t\)), and an idiosyncratic effect (\(e_{it}\) and \(v_{it}\)), where each of the error components is assumed to be normally distributed and correlated with the component of the same dimension in other equation. As the treatment of the time effects as random increases the difficulty of estimation, in terms of computational requirements, it is simpler to treat them as fixed time effects and absorbed in \(x_i\) and \(z_i\).

Given the distributional assumptions, it is possible to estimate the parameters by maximum likelihood. This is adopted in Hausman and Wise (1979), which represents the first attempt to account for selectivity in panel data, who examine the impact of endogenous attrition. Keane, Moffitt and Runkle (1988), Nijman and Verbeek (1992), and Verbeek (1990) also examine maximum likelihood estimation of the model under various assumptions for the treatment of the individual effects. However, given the computational demands of estimating by maximum likelihood, induced by the requirement to evaluate multiple integrals, consider the applicability of available simpler or two-step procedures. First consider the conventional fixed effect and random effects estimators. Changing the data for the explanatory variables into deviations from individual means via the following transformations:

\[x_{it}' = x_{it} - \frac{1}{T} \sum_{t=1}^{T} x_{it} d_{it} \quad \text{if } \sum_{t=1}^{T} d_{it} > 0 \]

and doing the same for \(y\), the fixed effects estimators for the unbalanced panel is:

\[(49)\quad \beta_{FE}(U) = \left( \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}' x_{it} d_{it} \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}' y_{it} d_{it} \right) \]

and for the balanced subpanel, \(\beta_{FE}(B)\), I replace \(d_{it}\) with \(c_i = \{(\Pi_{t=1}^{T} d_{it}) = 1\}\). To introduce the random effects estimator I follow Verbeek and Nijman (1992a) and define the following vectors: \(y_i = (y_{i1}, \ldots, y_{iT})', X_i = (x_{i1}, \ldots, x_{iT})', e_i = (e_{i1}, \ldots, e_{iT})\). Also define the number of units for which \(d_{it} = 1\) as \(T_i\) and define a \(T_i \times T\) matrix \(R_i\) transforming \(y_i\) into the \(T_i\) dimensional vector of observed values \(y_{it}^0\). This matrix \(R_i\) is obtained by deleting the rows of the \(T\) dimensional identity matrix corresponding to \(d_{it} = 0\). After defining the \(T\) dimensional unit vector \(t\) the variance of the
error term in (45) can be written $\Omega = \sigma_\\mu^2 I + \sigma_\\nu^2 I$. I can now write the random effects estimators as:

\begin{equation}
\beta_{RE}(U) = (X'X_i^{-1}X_i'X_i^{-1})^{-1}(X'X_i^{-1}X_i'y_i)
\end{equation}
and

\begin{equation}
\beta_{RE}(B) = (Z'X_i^{-1}X_i'Z_i^{-1})^{-1}(Z'X_i^{-1}X_i'y_i)
\end{equation}

where $X_i = R_iX_i$.

Consistency of (49) requires $E[\epsilon_t|x_t, d_t] = 0$ although (50) and (51) require the stronger condition $E[\epsilon_t|x_t, d_t] = 0$. Thus, consistency of the fixed effects estimator requires that $\sigma_\\nu = 0$. That is, the sample selection must operate purely through the individual specific terms. Consistency of the random effects estimator requires the stronger conditions of $\sigma_\\mu = 0$ and $\sigma_\\nu = 0$. Thus it cannot produce consistent estimates if the selection is operating either through the individual and/or the idiosyncratic effects.

An attraction of the fixed effects estimator is that it does not require a model for the selection equation. Nor do they impose any parametric assumptions on the disturbances. An alternative procedure is similar to that proposed in Heckman (1979) for the case where $T = 1$, where one computes the conditional expectation of the random components to include in the conditional mean of $y_i$ to account for the selection bias. This is the procedure adopted in Ridder (1990) and Nijman and Verbeek (1992) and generalized in Vella and Verbeek (1994).

To examine this estimation procedure I follow Vella and Verbeek and consider a general form of censoring. First condition (45) on the vector of all outcomes $d_i$ and the vector of all exogenous variables denoted $z_i$. This gives:

\begin{equation}
E[y_i|z_i, d_i] = x_i\beta + E[\epsilon_t|z_i, d_i] + E[\epsilon_t|z_i, d_i]
\end{equation}

where I replace (47) with $d_t = h(d_t^*)$ where $h$ is some known censoring mechanism. Note that it is necessary to condition on all outcomes simultaneously. Moreover, it is also necessary to assume that the $z_i^*$s are strictly exogenous. Thus, the explanatory variables in the latent model cannot be correlated with the error components. Now condition the error components on $u_t$ where $u_t = \alpha_t + \nu_t$. The expectations of the error components have the following forms:

\begin{equation}
E[\epsilon_t|u_t] = \frac{1}{\sigma_\\nu^2 + T\sigma_\\nu^2} \tilde{u}_t.
\end{equation}

\begin{equation}
E[\epsilon_t|u_t] = \frac{T\sigma_\\nu^2}{\sigma_\\nu^2 + T\sigma_\\nu^2} \tilde{u}_t.
\end{equation}

where $\tilde{u}_t = 1/T \sum_{t=1}^T u_t$. The difficulty in implementing this procedure is the calculation of $E[u_t|z_t, d_t]$. This expectation can be written:

\begin{equation}
E[u_t|z_t, d_t] = \int[\alpha_t + E[\nu_t|z_t, d_t, \alpha_t]]f(\alpha_t|z_t, d_t)d\alpha_t
\end{equation}

where $f(\alpha_t|z_t, d_t)$ represents the conditional density of $\alpha_t$, and $E[\nu_t|z_t, d_t, \alpha_t]$ is the generalized residual from (46). To compute these terms it is useful to note that:
\[
f(\alpha_i | z_i, d_i) = \frac{f(d_i | z_i, \alpha_i) f(\alpha_i | z_i)}{f(d_i | z_i)}
\]

where \(f(d_i | z_i) = \int f(d_i | z_i, \alpha_i) f(\alpha_i | z_i) d\alpha_i\) is the likelihood contribution of individual \(i\) in (45); \(f(\alpha_i | z_i) = f(\alpha_i)\) is a normal density; and \(f(d_i | z_i, \alpha_i) = \Pi_{t=1}^{T_i-1} f(d_i | z_i, \alpha_i)\) where \(f(d_i | z_i, \alpha_i)\) has the form of the likelihood contribution in the cross-sectional case.

The correction terms (52) and (53) require expressions for the likelihood contribution, the generalized residuals and the numerical evaluation of two one-dimensional integrals. Once these terms are computed the primary equation can be estimated over the subsample for which \(y_{it}\) is observed as these terms account for the selection bias. In dichotomous sample selection bias the likelihood function has the Probit form. Thus, one estimates the reduced form by random effects Probit and then employs these reduced form estimates to compute the correction terms. One then estimates the second step by OLS with the additional correction terms. This is the approach suggested in Ridder (1990) and Nijman and Verbeek (1992). Vella and Verbeek (1994) consider the case where the selection bias is generated by a general form of censoring such as those considered in the previous section. Vella and Verbeek also extend the conditional maximum likelihood procedure to panel data models.

An important feature of the Vella and Verbeek (1994) procedure is that it allows the inclusion of \(d_{i,t-1}\) in \(z_{it}\) in the selection equation. Thus, one can disentangle the individual effect from the role of state dependence. This is a useful feature as it ensures that the error components, and subsequent correction terms, do not incorrectly capture the dynamics that should be attributed to lagged dependent variables. This also highlights the use of panel in contrast to cross-sectional data. A complication induced by the introduction of lagged dependent variables is known as the ‘‘initial conditions’’ problem. Whenever it is inappropriate to assume that \(d_{i,0}\) is strictly exogenous, one must account for its endogeneity. An approximate solution is suggested in Heckman (1981) in which the reduced form for \(d_{i,0}\) is approximated using all presample information on the exogenous variables.

A feature of this two-step procedure, as with the parametric procedures for the cross-sectional case, is that one obtains a test of selection bias via the \(t\)-statistics on the correction terms. However, with the panel data procedures estimate two covariances. The first, \(\sigma_{\alpha\alpha}\), captures the correlation between the individual effects while \(\sigma_{\alpha\epsilon}\) captures the covariance between the idiosyncratic shocks.

I now apply this estimation procedure to our data. To construct a panel I obtain the data for the same individuals for the period 1980–87. I then estimate two random effects Tobit models.\(^{14}\) The first model assumes the underlying process is static. The second allows for the possibility of state dependence and thus includes lagged hours as an additional explanatory variable. From these two specifications, I compute the correction terms that I include as additional regressors for the 12,039 working observations. The results are reported in Table 3.

The first column of Table 3 presents the OLS estimates. A comparison of these

---

\(^{14}\) In addition to the eight time dummies the explanatory variables are the same as the cross-sectional models.
Table 3
Panel Data Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS without Corrections</th>
<th>OLS with Corrections from Static Model</th>
<th>OLS with Corrections from Dynamic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>-0.138 (0.015)</td>
<td>-0.170 (0.015)</td>
<td>-0.156 (0.015)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.011 (0.017)</td>
<td>-0.010 (0.017)</td>
<td>-0.007 (0.016)</td>
</tr>
<tr>
<td>Rural</td>
<td>-0.137 (0.014)</td>
<td>-0.145 (0.014)</td>
<td>-0.141 (0.014)</td>
</tr>
<tr>
<td>School</td>
<td>0.118 (0.004)</td>
<td>0.124 (0.004)</td>
<td>0.115 (0.004)</td>
</tr>
<tr>
<td>Union</td>
<td>0.170 (0.013)</td>
<td>0.167 (0.013)</td>
<td>0.168 (0.014)</td>
</tr>
<tr>
<td>Exper</td>
<td>0.046 (0.010)</td>
<td>0.017 (0.010)</td>
<td>0.004 (0.010)</td>
</tr>
<tr>
<td>Exper2</td>
<td>-0.0008 (0.0008)</td>
<td>-0.0017 (0.0007)</td>
<td>-0.0007 (0.0008)</td>
</tr>
<tr>
<td>$\sigma_{\mu}$</td>
<td>—</td>
<td>14.908 (0.954)</td>
<td>2.994 (0.194)</td>
</tr>
<tr>
<td>$\sigma_{\nu}$</td>
<td>—</td>
<td>-.577 (0.284)</td>
<td>-2.092 (0.231)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.151</td>
<td>0.191</td>
<td>0.195</td>
</tr>
<tr>
<td>Observations</td>
<td>12,039</td>
<td>12,039</td>
<td>12,039</td>
</tr>
</tbody>
</table>

with the cross-sectional OLS estimates reveal some differences that probably reflect that the individuals in the sample are young. Most notably, the rural effect is smaller, in absolute value, while the union premium is larger for the panel. The larger union effect reflects the larger premium for younger workers. As the panel is extended backwards, the additional observations are for the same workers at an earlier age. Column 2 presents the estimates that use the correction terms from the static model. Recall that as the first step is estimated by Tobit the model is identified through the variation in the residuals. As with the control function estimates reported in Column 1 of Table 2 the wage equation adjusted estimates are similar to the OLS estimates in spite of the strong evidence of selection bias. That is, the estimates in Columns 1 and 2 are generally quite similar despite the large and highly significant coefficients for the selection corrections that reveal that the selection operates both through the individual specific and the idiosyncratic terms. Although the adjusted and unadjusted coefficients are quite similar for most of the variables, they are very different for the experience effects. This suggests the selection bias is operating through these variables. Finally in Column 3 I report the corrected OLS estimates where the correction terms are computed from the dynamic random effects Tobit model. Two points
are worth noting from these estimates. First, for this empirical example the results using either the dynamic and static correction terms are similar. The evidence, however, suggests that the selection bias operates through the experience variables and the correction terms appear to have a different effect on these variables. Second, and perhaps more important, an examination of the Columns 2 and 3 indicates that the individual specific effects appear to be less important in the dynamic model. This suggests that failing to capture state dependence, through the inclusion of a lagged dependent variable, results in assigning too much importance to the individual effects. This highlights the value in exploiting the panel aspect of the data in that one can isolate the individual effects from the state dependence.

A model with such a rich error structure has substantial computational costs to accounting and testing for selectivity via this approach. Accordingly, it would be useful to have some simpler tests available which avoid the computation of the first step random effects model and the subsequent correction terms (52) and (53). Verbeek and Nijman (1992b) propose such a strategy by employing the Hausman (1978) testing framework. Under the null hypothesis of no sample selectivity the probability limits of $\beta_{RE}$ and $\beta_{FE}$ are identical as both procedures are consistent. Under the alternative, however, the two estimators diverge as both are inconsistent, under different forms of selection bias, but there is no reason to suspect that the inconsistency is the same. Verbeek and Nijman (1992b) construct some tests based on these pairwise comparisons.

Verbeek and Nijman (1992b) also suggest some simpler variable addition tests based on the use of variables denoting the number of times a particular observation appears in the sample, and whether or not the observation appears in all periods. Using Monte Carlo methods, Verbeek and Nijman (1992b) concluded these tests have some power although less than that based on the Lagrange multiplier framework that are derived when the model is estimated by maximum likelihood.

Wooldridge (1995) contends that a rejection in the Verbeek and Nijman (1992b) comparison of random versus fixed effects procedures could simply be due to the relative robustness of these procedures to serial dependence and heteroskedasticity rather than their tendency to deviate in the presence of selectivity bias. Wooldridge (1995) outlines a number of estimating procedures, with associated tests, for the above model where the underlying model has a fixed effects structure. Under this assumption fixed effects estimation of the primary equation requires:

$E(e_{it}\mid \mu_i, z_i, d_i) = 0$

and this is the basis of the testing procedure. One possible alternative is:

$E(e_{it}\mid \mu_i, \alpha_i, z_i, v_i, d_i) = \rho v_{it}$.

This implies that $e_{it}$ is mean independent of $(\mu_i, \alpha_i, z_i, v_{i1}, \ldots, v_{i,t-1}, v_{i,t+1}, \ldots, v_{it})$ conditional on $v_{it}$. Under the alternative, with the Wooldridge assumptions, (45), can be written:

$E(y_{it}\mid \mu_i, \alpha_i, z_i, v_i, d_i) = \mu_i + x_{it}\beta + \rho v_{it}$.

15. Recall that when the selection bias operates purely through the individual effects the fixed effects procedure is consistent.
To implement this testing procedure one needs an estimate of $v_t$. This, however, cannot be easily obtained as it depends on $\alpha_i$. Wooldridge suggests employing the Chamberlain (1980) characterization and assuming:

$$\alpha_i = \eta_i + \delta_1 z_{it} + \cdots + \delta_r z_{it} + c_i$$

where it is assumed that $c_i$ and $v_t$ are jointly normally distributed. If I substitute this into (46) I get:

$$d^*_n = z_{it} \gamma + \delta_0 + \delta_1 z_{it} + \cdots + \delta_r z_{it} + h_{it}$$

where $h_{it} = c_i + v_{it}$ and the $h_{it}$ are distributed independently of $z_i$. Wooldridge now writes:

$$E(y_{it}|\alpha_i, z_i, v_i, d_i) = \gamma + x_{it}\beta + \rho(h_{it} - c_i) = (\mu_i - \rho c_i) + x_{it}\beta + \rho h_{it}$$

and thus one can test (55) by testing $\rho = 0$ through the inclusion of $h_{it}$. To implement this testing, and estimating approach, one must implement the following steps. For each period, estimate a cross-sectional Probit model with explanatory variables $z_i$ and dependent variable $d_{it}$ and compute the value of the inverse Mills ratio $h_{it}$. Then estimate:

$$y_{it} = \zeta_i + x_{it}\beta + \rho \hat{h}_{it} + \eta_{it}$$

by fixed effects over the sample corresponding to $d_{it} = 1$, noting $\zeta_i = \mu_i - \rho c_i$. Wooldridge notes that when $\alpha_i$ equals a constant the test is even simpler as one then obtains the inverse Mills ratio from a pooled Probit equation.

Wooldridge also provides several extensions. First, he allows for serial dependence and heterogeneity in the selection equation. Second, he provides the identical estimator for models where the selection mechanism is $d_{it} = \max(0, d^*_{it})$ rather than $d_{it} = I(d^*_{it} > 0)$. For this alternative form of selection one replaces the inverse Mills ratio with the residuals $v_{it} = d_{it} - z_{it} \hat{\gamma}$ where the $\hat{\gamma}$'s are estimated by cross-sectional Tobit models. Furthermore, the second step estimation includes only the observations for which $d_{it} > 0$.

With the exception of simple fixed effects estimators discussed above, the panel data procedures are highly dependent on distributional assumptions. Kyriazidou (1997a) relaxes these assumptions by adapting the methodology of Powell (1987). Kyriazidou examines a model where the time effects, if any, are absorbed into the conditional mean. She also assumes that the individual effects, $\mu_i$, and $\alpha_i$ are fixed. The fixed effects, however, are allowed to be correlated with the explanatory variables and also the error terms. Furthermore, she makes no distributional assumptions regarding $e_{it}$ and $v_{it}$. Kyriazidou explicitly considers the case where $T = 2$, although her approach can be extended to greater values of $T$, and observes that conventional fixed effects estimation over the sample satisfying $d_{it1} = d_{it2} = 1$ eliminates the individual effects. This however does not eliminate the selection bias operating through the correlation between $e_{it}$ and $v_{it}$. In the parametric two-step approach this correlation was accounted for by an additional correction term that relied on the distributional assumptions. Kyriazidou avoids these distributional assumptions by adopting the following approach. Define the first differenced error in the primary equation as $e_{it} = e_{it2} - e_{it1}$. The expectation of this differenced error is a function
of the distribution of \( v_{1i} \) and \( v_{2i} \), in addition to the exogenous variables in the model. This can be written:

\[
(57) \quad E[\epsilon_d | d_{1i} = d_{2i} = 1, \chi] = E[\epsilon_d | v_{1i} < z_{1i}'y + \alpha_i, v_{1i} < z_{2i}'y + \alpha_i, \chi] \\
= \Lambda(v_{1i} < z_{1i}'y + \alpha_i, v_{1i} < z_{2i}'y + \alpha_i)
\]

where \( \Lambda \) is the unknown function determining the value of the “correction term” and \( \chi = [x_{1i}, x_{2i}, z_{1i}, z_{2i}, \alpha_i, \mu_i] \). Now it is possible to rewrite the differenced primary equation over the subsample satisfying \( d_{1i} = d_{2i} = 1 \) as:

\[
y_{id} = x_{id}^\beta + E[\epsilon_d | d_{1i} = d_{2i} = 1, \chi] + \eta_i.
\]

Now, similar to the logic in Powell (1987) and from (57), while assuming that \( \Lambda \) is time invariant, it follows that \( E[\epsilon_d | d_{1i} = d_{2i} = 1] \) usually will not equal zero unless \( z_{1i}'y = z_{2i}'y \). Thus, for any individual for which \( z_{1i}'y = z_{2i}'y \) the first differencing of the primary equation will eliminate both forms of selectivity bias. Furthermore, differencing across observations with “similar” values of \( z_{1i}'y \) and \( z_{2i}'y \) will also approximately eliminate the unobserved conditional expectation. Thus, the procedure suggested by Kyriazidou is the following. First, estimate \( \gamma \) by some procedure that does not impose distributional assumptions on the \( v_{1i} \). One such procedure for panel data is the conditional maximum score estimator of Manski (1987). Using this procedure one obtains consistent estimates of the \( \gamma \), say \( \hat{\gamma} \), and thus one can construct the estimates of the single indices \( z_{1i}'\hat{\gamma} \) and \( z_{2i}'\hat{\gamma} \). The second step then estimates \( \beta \) via the differencing strategy outlined above. More formally:

\[
(59) \quad \hat{\beta} = \hat{S}_{xy}^{-1}\hat{S}_{xy}
\]

where

\[
\hat{S}_{xy} = \frac{1}{n} \sum \hat{\pi}_{in} \Delta x_{i}' \Delta y_{i} \Psi_i
\]

\[
\hat{S}_{xx} = \frac{1}{n} \sum \hat{\pi}_{in} \Delta x_{i}' \Delta x_{i} \Psi_i
\]

where \( \Psi_i \) denotes the first difference operator and \( \hat{\pi}_{in} \) is a weight that declines to zero as \( |z_{1i}'\hat{\gamma} - z_{2i}'\hat{\gamma}| \) increases. As with several earlier semiparametric approaches I have discussed earlier, Kyriazidou suggests estimating this weight via a kernel function.

Kyriazidou (1997b) extends the above model to allow for dynamics in both the primary and selection equations. The model has the form:

\[
y_{it}^* = \delta y_{i,t-1} + x_{it}'\beta + \mu_i + e_{it}
\]

\[
d_{it}^* = \tau d_{i,t-1} + z_{it}'y + \alpha_i + v_{it}
\]

\[
d_{it} = 1 \text{ if } d_{it}^* > 0
\]

\[
y_{it} = y_{it}^* \ast d_{it}
\]

where \( \delta \) and \( \tau \) are unknown parameters capturing the dynamics and thus the model is the same as above except for the added complications introduced through the inclusion of the dynamics. To estimate this model Kyriazidou employs an identifica-
tion strategy similar to that proposed in Kyriazidou (1997a). First, she follows the estimation scheme in the dynamic linear panel data model and identifies the moments conditions implied by the assumptions underlying the model. This is done for the four cases corresponding to i) $\delta \neq 0, \beta = 0, \text{ and } \tau = 0$; ii) $\delta \neq 0, \beta \neq 0, \text{ and } \tau = 0$; iii) $\delta = 0, \beta = 0, \text{ and } \tau \neq 0$; and iv) $\delta \neq 0, \beta \neq 0, \text{ and } \tau \neq 0$. As in the static case, the moments to be satisfied require identifying observations with similar values of the selection equation index at different times. This is done in the same manner as the static case, by weighting the moment conditions for different observations by the difference in their values for the single index across time. Using this weighting scheme and the implied moment conditions for each model Kyriazidou follows Hansen (1982) and provides a generalized method of moments estimator for each model. Note that as the estimation of cases iii) and iv) requires the estimation of a panel data model with a lagged dependent variable it is necessary to employ a suitable first step procedure. For these cases, Kyriazidou employs the distribution free procedure of Honore and Kyriazidou (1997). Accordingly, to implement this procedure it is necessary to ensure the data satisfy the requirements of both estimators.

VII. Conclusions

The estimation of models with sample selectivity bias is unquestionably one of the most commonly encountered problems in micro-econometrics. This is reflected by the popularity of the model among both empirical and theoretical econometricians. Although the original model and estimation procedure proposed by Heckman were relatively simple, the subsequent extension of the model has generated a substantial literature. This paper provides a survey of the available methods for estimating models with sample selection bias. I initially examine the simple fully parameterized model proposed by Heckman before investigating departures in three important directions. First, I examine the relaxation of parametric assumptions. Second, I investigate the ability to tackle different selection rules. Finally, I examine how the methods applied in the cross-sectional case can be extended to panel data.

It should be noted that although I have only focused on estimation issues in this paper a number of important empirical studies have adapted the available sample selection bias estimators to address the problems they confronted. Although I have not attempted to incorporate these papers in this discussion, it is nevertheless valuable to note that many of the important contributions to this literature have been motivated by empirical work where the model under examination could not be estimated by the available procedures.

References

Chen, Songnian. 1996. "Semiparametric Estimation of the Type 3 Tobit Model." Unpublished manuscript, Hong Kong University of Science and Technology.


