A General Approach to Incorporating Selectivity in a Model

William Greene*
Department of Economics, Stern School of Business,
New York University,
June, 2006

1. Introduction

Based on the wisdom in Heckman's (1979) treatment of the linear model, there seems to be a widespread tendency (temptation) to extend his approach to other frameworks by mimicking his two step approach. Thus, for example, Wynand and van Praag (1981), in an early application, proposed to fit a probit model with sample selection with the following two steps:

Step 1. Fit the probit model for the sample selection equation.
Step 2. Using the selected sample, fit the second step probit model merely by adding the inverse Mills ratio from the first step to the main probit equation as an additional independent variable.

Another, similar application to the Poisson regression model is Greene (1994). This approach is inappropriate for several reasons

- The impact on the conditional mean of the model of interest will not take the form of an inverse Mills ratio. That is specific to the linear model. (See Terza (1995) for a development in the context of the Poisson regression.)
- The bivariate normality assumption needed to justify the inclusion of the inverse Mills ratio in the second model generally does not appear anywhere in the model.
- The dependent variable, conditioned on the sample selection, is unlikely to have the distribution described by the model in the absence of selection. That would be needed to use this approach. Note that this even appears in the canonical linear case. The normally distributed disturbance in the absence of sample selection has a nonnormal distribution in the presence of selection. That is the salient feature of Heckman's development.

Counterparts to these three problems will show up in any nonlinear model. One cannot generally 'correct for selectivity' by dropping the inverse Mills ratio into the model at a convenient point.

We describe an internally consistent method of incorporating 'sample selection' in a model. The method is based on the premise that motivated Heckman's canonical studies on the subject, that the force of 'sample selectivity' is exerted through the behavior of the unobservables in the model. As such, the key to modeling the effect is to introduce the unobservables that might be affected into the model in a reasonable way that maintains the internal consistency of the model itself. For example, in the Poisson model, the standard approach to introducing unobserved heterogeneity is through the conditional mean, specifically,

$$(1) \qquad \lambda_i(\varepsilon_i) = \exp(\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i)$$

* 44 West 4th St., Rm. 7-78, New York, NY 10012, USA, Telephone: 001-212-998-0876; e-mail: wgreene@stern.nyu.edu, URL www.stern.nyu.edu/~wgreene.

The negative binomial model arises if it is assumed that the unobserved heterogeneity, ε_i, has a log gamma distribution. 'Selectivity' would arise if the unobserved heterogeneity in this conditional mean is correlated with the unobservables in the sample selection mechanism.

This note describes an appropriate method of incorporating sample selection in a nonlinear model. This approach to selection in the linear model has only been extended to a relatively small range of other models, such as the Poisson regression Terza (1995, 1998)), (Greene (1997), the multinomial logit model (Terza (2002)) and the probit model (Wynand and van Praag (1981), Boyes et al. (1989), Greene (1992)). Wooldridge (2002) states "Heckman's approach is known to only work for specialized nonlinear models, such as an exponential regression model... Unfortunately, [it] does not extend easily to general nonlinear models." The method described here of handling many types of nonlinear models is, in fact, quite straightforward with modern software.

2. A Generic Model for Sample Selection

The generic model will take the form

(2)

$$
\begin{aligned}
z_i^* &= \boldsymbol{\alpha}'\mathbf{w}_i + u_i \;\; u_i \sim N[0,1], \\
z_i &= \mathbf{1}(z_i^* > 0) && \text{(probit selection equation)} \\
\lambda_i|\,\varepsilon_i &= \boldsymbol{\beta}'\mathbf{x}_i + \sigma\varepsilon_i,\ \varepsilon_i \sim N[0,1] && \text{(index function with heterogeneity)[1]} \\
y_i\,|\mathbf{x}_i,\, \varepsilon_i &\sim f(y_i \,|\, \mathbf{x}_i,\varepsilon_i) && \text{(index function model for outcome)} \\
[u_i,\varepsilon_i] &\sim \mathrm{N}[(0,1),(1,\rho,1)] \\
y_i,\mathbf{x}_i &\text{ are observed only when } z_i = 1.
\end{aligned}
$$

'Selectivity' is transmitted through the parameter ρ. At first blush, the framework might seem to impose an additional random element, ε, and parameter, it's variance, σ^2, on the main model. However, if one is to argue that sample selection has an impact on the model, then the implication is that there exists the unobserved ε through which it operates. As such, this is a natural formulation of the problem, and if the model is to be internally consistent, the additional elements are unavoidable.

The log likelihood function for the full model is the joint density for the observed data. When z_i equals one, $(y_i,\mathbf{x}_i,z_i,\mathbf{w}_i)$ are all observed. We seek $f(y_i,\, z_i{=}1|\mathbf{x}_i,\mathbf{w}_i)$. To obtain it, proceed as follows:

$$f(y_i,\, z_i{=}1|\mathbf{x}_i,\mathbf{w}_i) \;=\; \int_{-\infty}^{\infty} f(y_i,z_i{=}1|\mathbf{x}_i,\mathbf{w}_i,\varepsilon_i)\, f(\varepsilon_i)\mathrm{d}\varepsilon_i. \tag{3}$$

Conditioned on ε_i, z_i and y_i are independent. Therefore,

$$f(y_i,z_i{=}1|\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\varepsilon}_i) \;=\; f(y_i|\mathbf{x}_i,\varepsilon_i)\mathrm{Prob}(z_i{=}1|\mathbf{w}_i,\varepsilon_i). \tag{4}$$

The first part, $f(y_i\,|\mathbf{x}_i,\varepsilon_i)$ is the conditional index function model, however specified. By joint normality, $f(u_i|\varepsilon_i) = N[\rho\varepsilon_i\,,\,(1\text{-}\rho^2)]$, so $u_i = \rho\varepsilon_i + (\sqrt{1-\rho^2}\,)v_i$ where $v_i \sim N[0,1]$. Therefore, $\mathrm{Prob}(z_i{=}1|\mathbf{w}_i,\varepsilon_i)$ is

[1] The use of the linear index form is a convenience. The random component, ε, could enter the model in some other form, with no change in the general approach.

(5) $$\text{Prob}(z_i=1|\mathbf{w}_i,\varepsilon_i) \;=\; \Phi\left([\boldsymbol{\alpha}'\mathbf{w}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right).$$

Combining terms and using the earlier approach, the unconditional joint density is obtained by integrating ε out of the conditional density. Recall $\varepsilon_i \sim N[0,1]$, so $f(\varepsilon_i)$ is simply $\phi(\varepsilon_i)$. Thus,

(6) $$f(y_i,z_i=1|\mathbf{x}_i,\mathbf{w}_i) \;=\int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i,\varepsilon_i)\; \Phi\left([\boldsymbol{\alpha}'\mathbf{w}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right) \phi(\varepsilon_i)d\varepsilon_i\,.$$

By exploiting the symmetry of the normal cdf

(7) $$\text{Prob}(z_i=0|\mathbf{w}_i,\varepsilon_i) \;=\; \Phi\left(-[\boldsymbol{\alpha}'\mathbf{w}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right)$$

(8) $$\text{Prob}(z_i=0|\mathbf{w}_i) \;=\int_{-\infty}^{\infty} \Phi\left(-[\boldsymbol{\alpha}'\mathbf{w}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right) \phi(\varepsilon_i)d\varepsilon_i\,.$$

Expressions (6) and (8) can be combined by using the symmetry of the normal cdf,

(9) $$f(y_i,z_i|\mathbf{x}_i,\mathbf{w}_i) = \int_{-\infty}^{\infty} [(1-\,z_i) + z_i\, f(y_i|\mathbf{x}_i,\varepsilon_i)]\, \Phi\left((2z_i-1)[\boldsymbol{\alpha}'\mathbf{w}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right) \phi(\varepsilon_i)d\varepsilon_i\,,$$

where for z_i equal to zero, $f(y_i,z_i \mid \mathbf{x}_i,\mathbf{w}_i)$ is just $\text{Prob}(z_i=0|\mathbf{w}_i)$. Maximum likelihood estimates of the model parameters are obtained by maximizing the full log likelihood function,

(10) $$\log L \;=\; \sum_{i=1}^{N} \log f(y_i,z_i \mid \mathbf{x}_i,\mathbf{w}_i)$$

with respect to the model parameters $[\boldsymbol{\beta},\sigma,\boldsymbol{\alpha},\rho]$.

This formulation is not completely novel. Most of the elements appear in Terza (1995, 1998) in applications to the Poisson regression and some related variants, e.g., in Greene (1997).[2] This note adds to those results by extending them to the generic framework suggested here, and by proposing a convenient and surprisingly straightforward method of doing the estimation

3. Maximizing the Log Likelihood

Let $v \;=\; \varepsilon/\sqrt{2}$, $\theta = \sigma\sqrt{2}$, $\tau = \sqrt{2}\,[\rho/\sqrt{1-\rho^2}\,]$, and $\boldsymbol{\gamma} = [1/\sqrt{1-\rho^2}\,]\boldsymbol{\alpha}$. After making the change of variable and reparameterizing the probability as before, we obtain

(11) $$f(y_i,z_i=1|\mathbf{x}_i,\mathbf{w}_i) \;=\; \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \exp(-v^2)\, f(y_i|\mathbf{x}_i,v_i)\Phi(\boldsymbol{\gamma}'\mathbf{w}_i + \tau v_i)\, \mathrm{d}v_i$$

[2] Both Terza (1995, 1998) and Greene (1997) focused on the Poisson regression model. Greene employed the quadrature method described here. Terza also suggests the possibility of maximizing the log likelihood with quadrature, but then derives the actual conditional mean post selection, and also proposes nonlinear least squares as the preferred method. It will generally not be possible to derive the explicit form of the conditional mean; the Poisson (and a few other models) are unusual in the simple form, $E[y|\mathbf{x},\varepsilon] = \exp(\boldsymbol{\beta}'\mathbf{x} + \sigma\varepsilon)$. The simulation based approach is new with this survey.

where the index function model now involves $\lambda_i|v_i = \boldsymbol{\beta}'\mathbf{x}_i + \theta v_i$. Maximum likelihood estimates of $[\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, \tau]$ are obtained by maximizing the reparameterized log likelihood. No closed form exists for the function, but in this form, it can be approximated with Hermite quadrature. (See Butler and Moffitt (1982).) The approximation is

$$\text{(12)} \qquad \log L_Q = \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{\pi}}\sum_{h=1}^{H}\omega_h\left[(1-z_i)+z_i f(y_i\,|\,\mathbf{x}_i,v_h)\right]\Phi\left[(2z_i-1)\left(\boldsymbol{\gamma}'\mathbf{w}_i+\tau v_h\right)\right]\right]$$

where v_h and ω_h are the nodes and weights for the quadrature. The BHHH estimator of the asymptotic covariance matrix for the parameter estimates is a natural choice given the complexity of the function. The first derivatives must be approximated as well. For convenience, let

$$\text{(13)} \qquad \begin{aligned} P_{ih} &= f(y_i|\,\mathbf{x}_i, v_h) \\ \Phi_{ih} &= \Phi[(2z_i - 1)(\boldsymbol{\gamma}'\mathbf{w}_i + \tau v_h\,)] \;\; \text{(normal CDF)} \\ \phi_{ih} &= \phi[(2z_i - 1)\,(\boldsymbol{\gamma}'\mathbf{w}_i + \tau v_h\,)] \;\; \text{(normal density)} \end{aligned}$$

and to save some notation, denote the individual terms summed in the log likelihood as $\log L_i$.
We also use the result that $\partial P(.,.)/\partial z = P\times\partial\log P(.,.)/\partial z$ for any argument z which appears in the function. Then,

$$\text{(14)} \qquad \begin{aligned} \frac{\partial\log L_Q}{\partial\begin{pmatrix}\boldsymbol{\beta}\\ \theta\end{pmatrix}} &= \sum_{z_i=1}\frac{1}{L_i}\frac{1}{\sqrt{\pi}}\sum_{h=1}^{H}\omega_h z_i\Phi_{ih}P_{ih}\left[\frac{\partial\log f(y_i\,|\,\mathbf{x}_i,v_h)}{\partial\lambda_i}\right]\begin{pmatrix}\mathbf{x}_i\\ v_h\end{pmatrix} \\ \frac{\partial\log L_Q}{\partial\begin{pmatrix}\boldsymbol{\gamma}\\ \tau\end{pmatrix}} &= \sum_{z_i=1}\frac{1}{L_i}\frac{1}{\sqrt{\pi}}\sum_{h=1}^{H}\omega_h\phi_{ih}[(1-z_i)+z_iP_{ih}]\begin{pmatrix}\mathbf{w}_i\\ v_h\end{pmatrix} \end{aligned}$$

Estimates of the structural parameters, (α,ρ,σ) and their standard errors are computed using the delta method.

Simulation is another effective approach to maximizing the log likelihood function. (See Train (2003) and Greene (2003).) The log likelihood function is

$$\text{(15)} \qquad \log L = \sum_{i=1}^{N}\log\int_{-\infty}^{\infty}\;[(1\text{-}z_i) + z_i\, f(y_i|\mathbf{x}_i,\sigma\varepsilon_i)]\;\;\Phi[(2z_i-1)\left(\boldsymbol{\gamma}'\mathbf{w}_i+\tau\varepsilon_i\right)]\;\;\phi(\varepsilon_i)d\varepsilon_i\,.$$

The simulated log likelihood would be

$$\text{(16)} \qquad \log L_S = \sum_{i=1}^{N}\log\frac{1}{R}\sum_{r=1}^{R}\;[(1\text{-}z_i) + z_i\, f(y_i|\mathbf{x}_i,\sigma\varepsilon_{ir})]\;\;\Phi[(2z_i-1)\left(\boldsymbol{\gamma}'\mathbf{w}_i+\tau\varepsilon_{ir}\right)]$$

where ε_{ir} is a set of R random draws from the standard normal population. (We would propose to improve this part of the estimation by using Halton draws instead. See Train (2003, pp. 224-238).) Derivatives of the simulated log likelihood for the ith observation are

$$\frac{\partial\log L_{S,i}}{\partial\begin{pmatrix}\boldsymbol{\beta}\\ \sigma\end{pmatrix}} = \frac{1}{\log L_{S,i}}\frac{1}{R}\sum_{r=1}^{R}\;z_i\Phi_{ir}P_{ir}\left[\frac{\partial\log f(y_i\,|\,\mathbf{x}_i,\sigma\varepsilon_{ir})}{\partial\lambda_i}\right]\begin{pmatrix}\mathbf{x}_i\\ \varepsilon_{ir}\end{pmatrix}$$

(17)

$$\frac{\partial \log L_{S,i}}{\partial \begin{pmatrix} \gamma \\ \tau \end{pmatrix}} = \frac{1}{\log L_{S,i}} \frac{1}{R} \sum\nolimits_{r=1}^{R} \phi_{ir} [(1 - z_i) + z_i P_{ir}] \begin{pmatrix} \mathbf{w}_i \\ \varepsilon_{ir} \end{pmatrix}$$

where $\Phi_{\boldsymbol{ir}}$, ϕ_{ir} and P_{ir} are defined as in (13) using ε_{ir} in place of v_h.

4. Applications

The following will present four applications of the technique in widely diverse models. The first is based on the quadrature method while the second and third use simulation. Differences between these two methods are not expected to have any implications for computational or statistical efficiency. The Poisson and logit applications are template uses of the method described above. An extension of the method is required for the stochastic frontier model. Finally, the multinomial logit model is an application that is quite straightforward to place in the model framework developed here.

4.1 A Poisson Model with Sample Selection

This application has been well developed, for example in Terza (1995, 1998) who proposed FIML and nonlinear least squares approaches. (See, as well, Greene (1997, 2003) who applied the FIML approach developed here. Setting up the model for maximum likelihood estimation, we use (9),

(18) $$f(y_i,z_i|\mathbf{x}_i,\mathbf{w}_i) = \int_{-\infty}^{\infty} [(1 - z_i) + z_i\, f(y_i|\mathbf{x}_i,\varepsilon_i)]\, \Phi\left((2z_i - 1)[\boldsymbol{\alpha}'\mathbf{w}_i + \rho\varepsilon_i] / \sqrt{1-\rho^2}\right) \phi(\varepsilon_i) d\varepsilon_i\,,$$

with

(19) $$f(y_i \mid \mathbf{x}_i,\varepsilon_i) = \frac{\exp(-\lambda_i \mid \mathbf{x}_i,\varepsilon_i)(\lambda_i \mid \mathbf{x}_i,\varepsilon_i)^{y_i}}{\Gamma(y_i + 1)},\ \lambda_i \mid \mathbf{x}_i, \varepsilon_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i + \sigma\varepsilon_i).$$

As noted earlier, both Greene and Terza used Hermite quadrature to maximize the log likelihood function.

Greene (1992, 1994, 1997) analyzed credit card account default and credit behavior (derogatory reports) using a sample of applications and first year histories from a major credit card vendor. The data used in those studies are described in Table 1. The variable *cardhldr* is a binary variable which indicates whether the individual's application for the major credit card was accepted. The selected sample is the 10,499 individuals in the sample whose credit card application was accepted. (The sample is actually 'choice based.' In the true population at the time, the acceptance rate for this brand of credit card was considerably lower than this. The point is addressed in Greene (1992).)

In the count model application, the number of derogatory reports is analyzed. A major derogatory report is a report to the credit reporting agency of an account that becomes 60 days delinquent. The large majority of individuals in the study (and the population) have zero reports. But, the values in the sample ranged from zero to 22 for the full sample and zero to 14 in the selected sample. The force of the selection in this application should be substantial, since the number of major derogatory reports is a major, indeed, the dominant criteria for whether an application for a credit card is accepted or rejected. Table 2 shows the estimated acceptance (selection) equations. The 'uncorrected' equation is estimated as the first step in the estimation to provide starting values for the FIML estimator. The 'corrected' equation is estimated jointly with the count model. As anticipated, the effect of the sample selection is substantial. Table 3 gives the estimates of the count models. Once again, as expected, the estimates change considerably when the nonrandom sampling is accounted for. The mean predicted number of derogatory reports in the selected sample is 0.09634. The counterpart of the uncorrected model is 0.15325.

There is a minor extension of this model that might be interesting for this application. The major derogatory reports variable, as well as all the covariates in that equation are, in fact, observed for all observations. Thus, to use the full sample of data, the appropriate log likelihood would be

$$(20)\qquad f(y_i,z_i|\mathbf{x}_i,\mathbf{w}_i) = \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i,\varepsilon_i)]\ \Phi\left((2z_i-1)[\boldsymbol{\alpha}'\mathbf{w}_i+\rho\varepsilon_i]/\sqrt{1-\rho^2}\right)\ \phi(\varepsilon_i)d\varepsilon_i\,,$$

We leave this for further investigation.

4.2 A Binary Logit Model with Sample Selection

To illustrate the technique using the simulation based estimator, we construct a binary logit model subject to sample selection. The immediate obstacle to direct FIML estimation is the lack of a functional form for the joint distribution of a normally distributed ε and the logistically distributed variable that underlies the logit model. We use the template described above, instead. The main equation of interest is

$$(18)\qquad \text{Prob}(y_i{=}1|\mathbf{x}_i.\varepsilon_i)\ =\ \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i+\sigma\varepsilon_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i+\sigma\varepsilon_i)},\ \varepsilon_i \sim \text{N}[0,1]$$

The simulated log likelihood function is

$$(19)\qquad \log L_S = \sum\nolimits_{i=1}^{N} \log \frac{1}{R}\sum\nolimits_{r=1}^{R} \left[(1-z_i)+z_i\left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i+\sigma\varepsilon_{ir})}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i+\sigma\varepsilon_{ir})}\right)\right]\Phi[(2z_i-1)\left(\boldsymbol{\gamma}'\mathbf{w}_i+\tau\varepsilon_{ir}\right)]$$

The binary logit model with normal heterogeneity would be a special case of the multinomial logit model in Terza (2002).

The logit equation is used to model *default* on the credit card account. The selection model arises as default is only observed for those with *cardhldr* = 1, and without question, the determination of cardholder status is based on the vendor's attempt to forecast the probability of default. We will demonstrate the technique developed here by using a binary logit model for the default equation. (We concede the issue of probit versus logit as a functional form is a minor point. The purpose here is to demonstrate the use of the proposed method, not to redo the original study with an eye toward obtaining new results.)

We note a small, but potentially substantive difference between this model and the bivariate probit model employed in Greene (1992) and Boyes, Hoffman and Low (1989). The variance of ε is a free parameter in this model. It is assumed to equal one in the bivariate probit model. Our model could be employed with a probit, rather than a logit formulation to investigate the implication. (Of course, for the probit model, the bivariate model has already been established in the literature.) We leave this for further research to analyze. It should also be noted that the presence of the unobserved ε implies that the marginal distribution for the default model is something other than the logit model – it reverts to the logit model when σ equals zero.

Greene (1992, 1997) analyzed usage and default patterns for a sample of individuals applying for and using a major credit card. He used a bivariate probit model to accommodate the sample selection issue. Descriptive statistics for a subset of the variables in the data set of 13,444 observations are as follows. We are interested in the probability of default, conditioned on cardholder status. Estimates of the probit cardholder status equation are given in Table 2. The 'Corrected' estimates are those estimated jointly with the default equation. The 'Uncorrected results are estimated in isolation. Both use the full sample of 13,444 observations. Table 3 presents the estimates for the logit default models. These are based on the subsample of 10,499 observations whose applications were accepted. Notwithstanding the intuitive appeal of the formulation, it appears that the impact of the sampling mechanism here is fairly small. The results are quite similar for both sets of results. For this particular application, we could speculate on why that might be the case. The cardholder status variable is not actually being determined jointly with the default outcome, as the model would suggest. The cardholder status is the outcome of a screening procedure that is outsourced by the credit card vendor, while the default outcome is determined by the individual. Thus, the first step screen arguably represents the attempt by the credit scorer to forecast the default indicator, and, by implication, to predict whether ε is likely to be large or small. By this interpretation, we would not expect ρ to be very large.

4.3 Sample Selection in a Stochastic Frontier Model

We will apply the preceding technique to the stochastic frontier model. The notation and mechanics of the procedure differ slightly in this context, though the overall development is the same as in the generic model given earlier. A slight change in notation is employed to maintain consistency with the familiar standard in this literature. The modified frontier model is

$$d_i^* = \boldsymbol{\alpha}'\mathbf{z}_i + w_i,\ d_i = 1(d_i^* > 0), \qquad \text{(Selection equation)}$$

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \sigma_v v_i - u_i, \qquad \text{(Stochastic frontier model)}$$

$$u_i = |\sigma_u U_i| \text{ with } U_i \sim N[0,1],$$

$$(v_i,w_i) \sim \text{ Bivariate standard normal with } [(0,0),(1,\ \rho,\ 1)],$$

$$(y_i,\mathbf{x}_i) \text{ only observed when } d_i = 1.$$

Thus, the selection operates through the heterogeneity component of the production model, not the inefficiency. The observation status is not viewed as a function of the level of inefficiency. (Note, for convenience later, we have moved the scale parameters on v and u explicitly into the structural model.) For convenience, the observation subscript will be omitted in this part of the derivation. The selection equation may be recomposed as follows: Write the bivariate distribution of v and w in terms of the conditional distribution of w given v and the marginal distribution of v. From the bivariate normality assumption,

$$w|v = \rho v + h \text{ where } h \sim N[0,\ (1 - \rho^2)] \text{ and } h \text{ is independent of } v.$$

Therefore, $\quad d^*|v = \boldsymbol{\alpha}'\mathbf{z} + \rho v + h,\ \ d = 1(d^* > 0|v).$

Then, using familiar results for the probit model,

$$\text{Prob}[d = 1 \text{ or } 0 \mid \mathbf{z},v] = \Phi\left[(2d-1)\left(\frac{\boldsymbol{\alpha}'\mathbf{z} + \rho v}{\sqrt{1-\rho^2}}\right)\right].$$

As before, we consider the sample in two parts. For the selected observations, $d = 1$, conditioned on v, the joint density for y and d is the product of the marginals since conditioned on v, y and d are independent. Thus,

$$f(y,d{=}1|\mathbf{x},\mathbf{z},v) \;=\; f(y|\mathbf{x},v)\ \mathrm{Prob}(d{=}1|\mathbf{z},v).$$

We have the second part above. For the first part,

$$y|\ \mathbf{x},v \;=\; (\boldsymbol{\beta}'\mathbf{x} \;+\; \sigma_v v\,)\ \text{-}\ u$$

where u is the truncation at zero of a normal variable with standard deviation σ_u. The density of u is trivial, since before truncation it ($\sigma_u U$) has zero mean and variance σ_u^2, so $f(u) = 2\phi(u/\sigma_u)$, $u \geq 0$. The Jacobian of the transformation from u to y is $1/\sigma_u$, so by the change of variable, the conditional density is

$$f(y\,|\,\mathbf{x},v) = \frac{2}{\sigma_u}\phi\left(\frac{(\boldsymbol{\beta}'\mathbf{x}+\sigma_v v)-y}{\sigma_u}\right),(\boldsymbol{\beta}'\mathbf{x}+\sigma_v v)-y \geq 0.$$

Therefore, the joint conditional density is

$$f(y,d=1\,|\,\mathbf{x},\mathbf{z},v) = \frac{2}{\sigma_u}\phi\left(\frac{(\boldsymbol{\beta}'\mathbf{x}+\sigma_v v)-y}{\sigma_u}\right)\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z}+\rho v}{\sqrt{1-\rho^2}}\right)$$

To obtain the unconditional density, it is necessary to integrate v out of the conditional density. Thus,

$$f(y,d=1\,|\,\mathbf{x},\mathbf{z}) = \int_v \frac{2}{\sigma_u}\phi\left(\frac{\sigma_v v-(y-\boldsymbol{\beta}'\mathbf{x}))}{\sigma_u}\right)\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z}+\rho v}{\sqrt{1-\rho^2}}\right)f(v)dv$$

The relevant term in the log likelihood is log $f(y,d{=}1|\mathbf{x},\mathbf{z})$. For the nonselected observations, the contribution to the log likelihood is the log of the unconditional probability of nonselection, which is, once again,

$$\mathrm{Prob}(d{=}0|\mathbf{z}) \;=\; \int_v \Phi\left[-\left(\frac{\boldsymbol{\alpha}'\mathbf{z}+\rho v}{\sqrt{1-\rho^2}}\right)\right]f(v)dv$$

The integrals do not exist in closed form, so these terms cannot be evaluated as is. Before proceeding, we note the additional complication, $\boldsymbol{\beta}'\mathbf{x} + \sigma_v v - y = u > 0$, so the density $f(v)$ is not the standard normal that intuition might suggest; it is a truncated normal. The integrals can be computed by simulation. By construction,

$$\int_v \frac{2}{\sigma_u}\phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}+\sigma_v v-y)}{\sigma_u}\right)\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z}+\rho v}{\sqrt{1-\rho^2}}\right)f(v)dv = E_v\left[\frac{2}{\sigma_u}\phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}+\sigma_v v-y)}{\sigma_u}\right)\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z}+\rho v}{\sqrt{1-\rho^2}}\right)\right]$$

so by sampling from the distribution of v, we can compute the function of v and average to obtain the integrals. In order to sample the draws on v, we note the implied truncation,

$$v \geq (y - \boldsymbol{\beta}'\mathbf{x})/\sigma_v \;\text{ or }\; v \geq \varepsilon/\sigma_v.$$

Draws from the truncated normal can be obtained using result (E-1) in Greene (2003). Let A equal a draw from the uniform (0,1) population. The desired draw from the truncated normal distribution will be

$$v_r = \Phi^{-1}\left[\Phi(\varepsilon/\sigma_v) + A_r\Phi(-\varepsilon/\sigma_v)\right]$$

Collecting all terms, then, the simulated log likelihood will be

$$\log L_S = \sum\nolimits_i \log\frac{1}{R}\sum\nolimits_{r=1}^{R} \left\{ d_i\left[\frac{2}{\sigma_u}\phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}+\sigma_v v_{ir} - y)}{\sigma_u}\right)\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z}+\rho v_{ir}}{\sqrt{1-\rho^2}}\right)\right] + (1 - d_i)\left[\Phi\left(\frac{-\boldsymbol{\alpha}'\mathbf{z} - \rho v_{ir}}{\sqrt{1-\rho^2}}\right)\right]\right\}$$

where the draws on v_{ir} are as shown above. Derivatives of this simulated log likelihood are obtained numerically using finite differences.

A final detail concerns the estimated parameters. The preceding shows how to estimate (α, β, σ_u, σ_v). In the stochastic frontier setting, the interesting parameters are $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$. These can be obtained after estimation. Standard errors, if desired, can be computed using the delta method.

4.4 Sample Selection and a Multinomial Logit Model

Many authors have considered models in which the selection mechanism is extended from the binomial probit model to a multinomial model – the multinomial logit model is the natural choice. Thus, rather than selecting either 'in' or 'out' of the sample (z_i= 0 or 1 in our earlier model), the authors model the case in which z_i may take one of J+1 values, and then the selection is nonrandomly into group j, j = 0,1,...,J. The model to be fit after selection is typically a linear model. Lee's (1983) results for the binary logit model selection mechanism, which (by implication) uses a copula method, have remained the formulation of choice for the past two decades. (For implementation, see, e.g., Econometric Software, Inc. (2003).) Two issues can be raised with respect to this relatively modest extension of the original Heckman (1979) model:

(1) The technique is purely mechanical. It is less than obvious how the force of the sample selection acts to connect the unobservables in the multinomial logit model. Indeed, the presence of the unobservables in the multinomial logit model is, itself, less than obvious. This is fairly straightforward to remedy, as we show below.

(2) Another interesting extension that remains undeveloped in this literature is one that reverses the role of the multinomial outcome model and the sample selection equation.[3] In particular, continuing the analysis of this paper, we are interested in a multinomial outcome, say mode of health treatment, after selection, say, whether an individual had some prior treatment such as visiting a physician.

This model can be easily accommodated in the framework developed here, at the same time, making transparent how the unobservables in the model can play a role in the logit model. We propose the outcome model to be developed around a random utility framework:

$$U_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + u_{ij} + \sigma_j\varepsilon_i, j = 0,1,...,J.$$

$$y_i = j \text{ if } U_{ij} > U_{ik}, \forall \ k \neq j.$$

Conditioned on ε_i, if u_{ij} is distributed as type 1 extreme value, then this produces the multinomial logit model. (See, e.g., Greene (2003) or Hensher, Rose and Greene (2005).) Thus, Terza's (2002) model would be, in this context:

$$f(y_i|\mathbf{x}_i,\varepsilon_i) = \text{Prob}(y_i = j|\text{x}_\text{i},\varepsilon_\text{i}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij} + \sigma_j\varepsilon_i)}{\sum_{j=0}^{J}\exp(\boldsymbol{\beta}'\mathbf{x}_{ij} + \sigma_j\varepsilon_i)}.$$

(Terza considered the case in which the data are characteristics of the individual rather than attributes of the choices. This mandates a renormalization of the coefficients, but is otherwise the same model.)

We assume that the latent heterogeneity enters the utility functions linearly in the same fashion that the observed indicators do. In order to avoid forcing a scaling restriction on the model, there is a separate scale factor in each utility function.[4] Note, finally, for identification, we must have one of the scale parameters normalized at zero, so $\sigma_0 = 0$ is imposed. With this outcome model specified, the rest of the selectivity model is defined by (6) – (9).

5. Conclusion

This note has described a generic approach to modeling sample selection in a nonlinear model. The intuitively appealing approach of inserting an inverse Mills ratio into the model of interest at a convenient point to 'deal with selection bias' is inappropriate for several reasons. On the other hand, in all but the simplest cases, it will not be possible to deduce the appropriate conditional mean or other conditional feature of the model, post selection. The framework gives a straightforward, internally consistent method of introducing selection into a nonlinear model and a method of maximum likelihood estimation that can be easily programmed in any platform that supports either quadrature based or simulation based optimization.

[3] A search of "multinomial logit" + "sample selection" will turn up thousands of references to the first (Lee) model, seemingly most of them related to implementations in *LIMDEP* or *Stata*, but, it appears, none to the second application noted above.

[4] That there is none on the extreme value terms is one of the shortcomings of the multinomial logit model. This particular restriction is easily relaxed. However, after doing so, it would be exceedingly cumbersome to extend the selection model to the expanded form – the less restrictive model is not a minor extension of the MNL model. See Bhat (1995) and Econometric Software (2003).

References

Bhat, C., “A heteroscedastic extreme value model of intercity mode choice,” *Transportation Research B*, 29, 1995, pp. 471–483.

Boyes, W., D. Hoffman and S. Low, “An Econometric Analysis of the Bank Credit Scoring Problem,” *Journal of Econometrics*, 40, 1989, pp. 3-14.

Butler, J. and Moffitt, R., “A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model,” *Econometrica*, 50, 1982, pp. 761-764.

Econometric Software, Inc., *LIMDEP Version 8.0 and NLOGIT Version 3.0*, Plainview, New York, 2003.

Greene, W., “A Statistical Model for Credit Scoring,” Working Paper EC-92-29, Department of Economics, Stern School of Business, New York University, 1992, forthcoming in *Credit Risk: Quantitative Methods and Analysis*, Hensher, D. and S. Jones, eds., Cambridge University Press, 2007.

Greene, W., “FIML Estimation of Sample Selection Models for Count Data,” Working Paper EC-97-02, Department of Economics, Stern School of Business, New York University, 1997.

Greene, W., *Econometric Analysis, 5th Ed.*, Prentice Hall, Englewood Cliffs, 2003.

Heckman, J., “Sample Selection Bias As a Specification Error,” *Econometrica*, 47, 1979, pp. 153-161.

Hensher, D., Rose, J. and W. Greene, *Applied Choice Analysis*, Cambridge University Press, Cambridge, 2005.

Terza, J., “Estimating Count Data Models with Endogenous Switching and Sample Selection,” Department of Economics, Penn State University, Working Paper IPRE-95-14, 1995.

Terza, J. “Estimating Count Data Models with Endogenous Switching, Sample Selection and Endogenous Treatment Effects,” *Journal of Econometrics*, 84, 1, 1998, pp. 129-154.

Terza, J., “Alcohol Abuse and Employment: A Second Look,” *Journal of Applied Econometrics*, 17, 2002, pp. 393-404.

Train, K., *Discrete Choice Models with Simulation*, Cambridge University Press, Cambridge, 2003.

Wooldridge, J., “Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification,” *Portuguese Economic Journal*, 1, 2002, pp. 117-139.

Wynand, P. and B. van Praag, “The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection,” *Journal of Econometrics*, 17, 1981, pp. 229-252.

Table 1. Data Used in Analysis of Major Derogatory Reports and Default.

Variable	Symbol	Mean	Standard Deviation
In full sample of 13,444 observations			
Cardholder status (binary)	CARDHLDR	.780943	.413623
Age	AGE	33.6031	10.0141
Yearly income in $	INCOME	34240.5	17774.5
Owns or rents home (binary)	OWNRENT	.455965	.498076
Self employed (binary)	SELFEMPL	.057944	.233646
Months at current residence	CURNTADD	55.3189	63.0897
Yearly income in $	INCOME	34240.5	17774.5
Number of dependents	DEPDNT	1.01726	1.27910
Income per dependent	INCPER	21719.7	13591.2
Holds another major credit card (binary)	CREDMAJR	.813076	.389865
Number of active credit accounts	TRADACCT	6.42205	6.10691
Major derogatory reports	MAJORDRG	0.46281	1.43272
In selected sample of 10,499 observations with CARDHLDR = 1			
Defaulted (binary)	DEFAULT	.094866	0.29304
Major derogatory reports	MAJORDRG	0.15325	0.46157
Yearly income in $	INCOME	35662.5	18395.0
Average yearly card expenditure	AVGYREXP	3012.37	3987.27
Number of dependents	DEPDNT	0.99038	1.27389
Income per dependent	INCPER	22581.4	13755.0
Holds another major credit card (binary)	CREDMAJR	0.84332	0.36352
Number of active credit accounts	TRADACCT	7.11887	6.19853

Table 2. Estimated Probit Models for Cardholder Status

	Corrected	Uncorrected
Variable	Estimate (Standard error)	Estimate (Standard error)
Constant	.3179* (.04276)	.4428* (.04560)
Age	.0004522* (.001255)	-.003439* (.001426)
Income	.1201* (.007565)	.1209* (.008788)
Ownrent	.7535* (.02449)	.1633* (.02767)
SelfEempl	-.2684* (.04176)	-.3355* (.05107)
Curntadd	.0006350* (.0001851)	.00003522 (.0002146)
ρ	-.9562* (.1092)	.0000
Log likelihood		−6899.529

Table 3. Estimated Poisson Models for Major Derogatory Reports

	Corrected	Uncorrected
Variable	Estimate (Standard error)	Estimate (Standard error)
Constant	-1.0160* (.2155)	-2.9379* (.08973)
Income	-.2170* (.04695)	.001189 (.020295)
AvgYrExp	-.2581* (.1065)	.02789* (.003048)
Depdnt	.07922 (.05692)	.1968* (.03234)
IncPer	-.04537 (.06119)	-.04551 (.06547)
Credmajr	-.2033* (.08529)	-.07497 (.07428)
Tradacct	-.03341* (.006382)	.04140* (.003587)
σ	2.2562 (.1092)	.0000
ρ	-.9562* (.1092)	.0000
Log likelihood		-4831.23822
Combined log likelihood	−11212.18	−11730.77

Table 4. Estimated Probit Models for Cardholder Status

	Corrected	Uncorrected
Variable	Estimate (Standard error)	Estimate (Standard error)
Constant	.3453* (.08122)	.4428* (.04560)
Age	-.003423* (.001423)	-.003439* (.001426)
Income	.1211* (.007859)	.1209* (.008788)
Ownrent	.1633* (.02731)	.1633* (.02767)
SelfEempl	-.3359* (.05090)	-.3355* (.05107)
Curntadd	.00003581 (.0002171)	.00003522 (.0002146)
ρ	.1917 (.1275)	.0000

.

Table 5. Estimated Logit Models for Default

	Corrected	Uncorrected
Variable	Estimate (Standard error)	Estimate (Standard error)
Constant	-1.0160* (.2155)	-1.07508* (.1241)
Income	-.2170* (.04695)	-.2168* (.04635)
AvgYrExp	-.2581* (.1065)	-.2582* (.1171)
Depdnt	.07922 (.05692)	.07903 (.05575)
IncPer	-.04537 (.06119)	-.04551 (.06547)
Credmajr	-.2033* (.08529)	-.2034* (.08519)
Tradacct	-.03341* (.006382)	-.03339* (.006411)
σ	.1171 (.3646)	.0000
ρ	.1917 (.1275)	.0000