

Notes on Sample Selection Models

© Bronwyn H. Hall 1999,2000,2002

February 1999 (revised Nov. 2000; Feb. 2002)

1 Introduction

We observe data (X, Z) on N individuals or firms. For a subset $N_1 = N - N_0$ of the observations, we also observe a dependent variable of interest, y_1 but this variable is unobserved for the remaining N_0 observations. The following model describes our estimation problem:

$$y_{1i} = X_i\beta + \nu_{1i} \quad \text{if } y_{2i} > 0 \quad (1)$$

$$y_{1i} = \text{not observed} \quad \text{if } y_{2i} \leq 0$$

$$y_{2i} = Z_i\delta + \nu_{2i} \quad (2)$$

$$D_{2i} = 1 \quad \text{if } y_{2i} > 0$$

$$D_{2i} = 0 \quad \text{if } y_{2i} \leq 0$$

The equation for y_{1i} is an ordinary regression equation. However, under some conditions we do not observe the dependent variable for this equation; we denote whether or not we observe its value by a dummy variable D_{2i} . Observation of the dependent variable y_{1i} is a function of the value of another regression equation (the selection equation, which relates a latent variable y_{2i} to some observed characteristics Z_i). The variables in X_i and Z_i may overlap; if they are identical this will create problems for identification in some cases (see the discussion below).

Examples are married women's labor supply (where the first equation is the hours equation and the second equation is an equation for the difference between the market and the unobserved reservation wage) and the firm size and growth relationship (where the first equation is the relation between growth and size and the second equation describes the probability of exit between the first and second periods).

2 Bias Analysis

Suppose that we estimate the regression given in equation (1) by ordinary least squares, using only the observed data. We regress y_{1i} on X_i , using $i = N_0 + 1, \dots, N$ observations. When are

the estimates of β obtained in this way likely to be biased? We can analyze this question without assuming a specific distribution for the ν s. Compute the conditional expectation of y_1 given X and the probability that y_1 is observed:

$$E[y_1|X, y_2 > 0] = X\beta + E[\nu_1|\nu_2 > -Z\delta]$$

From this expression we can see immediately that the estimated β will be unbiased when ν_1 is independent of ν_2 (that is, $E[\nu_1|\nu_2] = 0$), so that the data are missing "randomly," or the selection process is "ignorable." That is the simplest (but least interesting) case.

Now assume that ν_1 and ν_2 are jointly distributed with distribution function $f(\nu_1, \nu_2; \theta)$ where θ is a finite set of parameters (for example, the mean, variance, and correlation of the random variables). Then we can write (by Bayes rule)

$$E[\nu_1|\nu_2 > -Z_i\delta] = \frac{\int_{-\infty}^{\infty} \int_{-Z_i\delta}^{\infty} \nu_1 f(\nu_1, \nu_2; \theta) d\nu_2 d\nu_1}{\int_{-\infty}^{\infty} \int_{-Z_i\delta}^{\infty} f(\nu_1, \nu_2; \theta) d\nu_2 d\nu_1} = \lambda(Z\delta; \theta) \quad (3)$$

$\lambda(Z\delta; \theta)$ is a (possibly) nonlinear function of $Z\delta$ and the parameters θ . That is, in general the conditional expectation of y_1 given X and the probability that y_1 is observed will be equal to the usual regression function $X\beta$ plus a nonlinear function of the selection equation regressors Z that has a non-zero mean.¹ This has two implications for the estimated β s:

1. The estimated intercept will be biased because the mean of the disturbance is not zero. (In fact, it is equal to $E_i[\lambda(Z_i\delta; \theta)]$).
2. If the X s and the Z s are not completely independently distributed (i.e., they have variables in common, or they are correlated), the estimated slope coefficients will be biased because there is an omitted variable in the regression, namely the $\lambda(Z_i\delta; \theta)$, that is correlated with the included variables X .

Note that even if the X s and the Z s are independent, the fact that the data is nonrandomly missing will introduce heteroskedasticity into the error term, so ordinary least squares is not fully efficient (*Why?*²).

This framework suggests a semi-parametric estimator of the sample selection model, although few researchers have implemented it (see Powell, *Handbook of Econometrics*, Volume IV, for more discussion of this approach). Briefly, the method would have the following steps:

¹Although I will not supply a proof, only for very special cases will this term be mean zero. For example, in the case of bivariate distributions with unconditional means equal to zero, it is easy to show that $\lambda(\cdot)$ has a nonzero mean unless the two random variables are independent.

²Questions in italics throughout these notes are exercises for the interested reader.

1. Estimate the probability of observing the data (equation (??)) using a semi-parametric estimator for the binary choice model (*why are these estimates consistent even though they are single equation?*).
2. Compute a fitted value of the index function $\hat{y}_{2i} = Z_i\hat{\delta}$.
3. Include powers of \hat{y}_{2i} in a regression of y_{1i} on X_i to proxy for $\lambda(Z_i\delta; \theta)$. It is not clear how many to include.

Note that it is very important that there be variables in Z_i that are distinct from the variables in X_i for this approach to work, otherwise the regression will be highly collinear. Note also that the propensity score approach of Rubin and others is related to this method: it uses intervals of \hat{y}_{2i} to proxy for $\lambda(Z_i\delta; \theta)$, interacting them with the X variable of interest (the treatment).

3 Heckman Estimator

Semi-parametric estimation can be difficult to do and has very substantial data requirements for identification and for the validity of finite sample results. Therefore most applied researchers continue to estimate sample selection models using a parametric model. The easiest to apply in the case of sample selection is the bivariate normal model, in which case the selection equation becomes the usual Probit model. There are two approaches to estimating the sample selection model under the bivariate normality assumption: the famous two-step procedure of Heckman (1979) and full maximum likelihood. I will discuss each of these in turn. Although ML estimation is generally to be preferred for reasons discussed below, the Heckman approach provides a useful way to explore the problem.

The Heckman method starts with equation (3) and assumes the following joint distribution for the ν s:

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right] \quad (4)$$

where N denotes the normal distribution. Recall that the variance of the distribution in a Probit equation can be normalized to equal one without loss of generality because the scale of the dependent variable is not observed. Using the assumption of normality and the results in the Appendix on the truncated bivariate normal, we can now calculate $E[y_1|y_2 > 0]$.

$$\begin{aligned} E[y_1|y_2 > 0] &= X\beta + E[v_1|v_2 > -Z\delta] = X\beta + \rho\sigma_1\lambda\left(\frac{-Z\delta}{1}\right) \\ &= X\beta + \rho\sigma_1\frac{\phi(-Z\delta)}{1 - \Phi(-Z\delta)} = X\beta + \rho\sigma_1\frac{\phi(Z\delta)}{\Phi(Z\delta)} \end{aligned} \quad (5)$$

Let's interpret this equation. It says that the regression line for y on X will be biased upward when ρ is positive and downward when ρ is negative, since the inverse Mills ratio is always positive (see the Appendix). The size of the bias depends on the magnitude of the correlation, the relative variance of the disturbance (σ_1), and the severity of the truncation (the inverse Mills ratio is larger when the cutoff value $Z\delta$ is smaller – see the figure in the Appendix). Note that when ρ is zero there is no bias, as before.³

Also note that the simple Tobit model, where y_1 and y_2 coincide and ρ is therefore one, can be analyzed in the same way, yielding

$$E[y_1|y_1 > 0] = X\beta + \sigma_1 \frac{\phi(X\beta)}{\Phi(X\beta)}$$

In this case, because the second term is a monotonic declining function of $X\beta$, it is easy to see that the regression slope will be biased downward (*Why?*).

3.1 Estimation using Heckman's Method

Equation (5) suggests a way to estimate the sample selection model using regression methods. As in the semi-parametric case outlined above, we can estimate β consistently by including a measure of $\phi(Z\delta)/\Phi(Z\delta)$ in the equation. Heckman (1979, 1974?) suggests the following method:

1. Estimate δ consistently using a Probit model of the probability of observing the data as a function of the regressors Z .
2. Compute a fitted value of the index function or latent variable $\hat{y}_{2i} = Z_i\hat{\delta}$; then compute the inverse Mills ratio $\hat{\lambda}_i$ as a function of \hat{y}_{2i} .
3. Include $\hat{\lambda}_i$ in a regression of y_{1i} on X_i to proxy for $\lambda(Z_i\delta)$. The coefficient of $\hat{\lambda}_i$ will be a measure of $\rho\sigma_1$ and the estimated ρ and σ_1 can be derived from this coefficient and the estimated variance of the disturbance (which is a function of both due to the sample selection; see Heckman for details).

The resultant estimates of β , ρ , and σ_1 are consistent but not asymptotically efficient under the normality assumption. This method has been widely applied in empirical work because of its relative ease of use, as it requires only a Probit estimation followed by least squares, something which is available in many statistical packages. However, it has at least three (related) drawbacks:

³In the normal case, $\rho = 0$ is equivalent to the independence result for the general distribution function.

1. The conventional standard error estimates are inconsistent because the regression model in step (3) is intrinsically heteroskedastic due to the selection. (*What is $Var(\nu_1|\nu_2 > 0)$?*) One possible solution to this problem is to compute robust (Eicker-White) standard error estimates, which will at least be consistent.
2. The method does not impose the constraint $|\rho| \leq 1$ that is implied by the underlying model (ρ is a correlation coefficient). In practice, this constraint is often violated.
3. The normality assumption is necessary for consistency, so the estimator is no more robust than full maximum likelihood – it requires the same level of restrictive assumptions but is not as efficient.

For these reasons and because full maximum likelihood methods are now readily available, it is usually better to estimate this model using maximum likelihood if you are willing to make the normal distributional assumption. The alternative more robust estimator that does not require the normal assumption is described briefly in Section 2. The ML estimator is described in the next section.

4 Maximum Likelihood

Assuming that you have access to software that will maximize a likelihood function with respect to a vector of parameters given some data, the biggest challenge in estimating qualitative dependent variable models is setting up the (log) likelihood function. This section gives a suggested outline of how to proceed, using the sample selection model as an example.⁴

Begin by specifying a complete model as we did in equations (1) and (??). Include a complete specification of the distribution of the random variables in the model such as equation (4). Then divide the observations into groups according to the type of data observed. Each group of observations will have a different form for the likelihood. For example, for the sample selection model, there are two types of observation:

1. Those where y_1 is observed and we know that $y_2 > 0$. For these observations, the likelihood function is the probability of the joint event y_1 and $y_2 > 0$. We can write

⁴Several software packages, including TSP, provide the sample selection (generalized Tobit) model as a canned estimation option. However, it is useful to know how to construct this likelihood directly, because often the model you wish to estimate will be different from the simple 2 equation setup of the canned program. Knowing how to construct the likelihood function allows you to specify an arbitrary model the incorporates observed and latent variables.

this probability for the i th observation as the following (using Bayes Rule):

$$\begin{aligned}
\Pr(y_{1i}, y_{2i} > 0 | X, Z) &= f(y_{1i}) \Pr(y_{2i} > 0 | y_{1i}, X, Z) = f(\nu_{1i}) \Pr(\nu_{2i} > -Z_i\delta | \nu_{1i}, X, Z) \\
&= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \int_{-Z_i\delta}^{\infty} f(\nu_{2i} | \nu_{1i}) d\nu_{2i} \\
&= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \int_{-Z_i\delta}^{\infty} \phi\left(\frac{\nu_{2i} - \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1 - \rho^2}}\right) d\nu_{2i} \\
&= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \left[1 - \Phi\left(\frac{-Z_i\delta - \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1 - \rho^2}}\right)\right] \\
&= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \Phi\left(\frac{Z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1 - \rho^2}}\right)
\end{aligned}$$

where we have used the conditional distribution function for the normal distribution given in the appendix to go from the second line to the third line. Thus the probability of an observation for which we see the data is the density function at the point y_1 multiplied by the conditional probability distribution for y_2 given the value of y_1 that was observed.

2. Those where y_1 is not observed and we know that $y_2 \leq 0$. For these observations, the likelihood function is just the marginal probability that $y_2 \leq 0$. We have no independent information on y_1 . This probability is written as

$$\Pr(y_{2i} \leq 0) = \Pr(\nu_{2i} \leq -Z_i\delta) = \Phi(-Z_i\delta) = 1 - \Phi(Z_i\delta)$$

Therefore the log likelihood for the complete sample of observations is the following:

$$\begin{aligned}
\log L(\beta, \delta, \rho, \sigma; \text{thedata}) &= \sum_{i=1}^{N_0} \log [1 - \Phi(Z_i\delta)] \\
&+ \sum_{i=N_0+1}^N \left[-\log \sigma_1 + \log \phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) + \log \Phi\left(\frac{Z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1 - \rho^2}}\right) \right]
\end{aligned}$$

where there are N_0 observations where we don't see y_1 and N_1 observations where we do ($N_0 + N_1 = N$). The parameter estimates for the sample selection model can be obtained by maximizing this likelihood function with respect to its arguments. These estimates will be consistent and asymptotically efficient under the assumption of normality and homoskedasticity of the uncensored disturbances. Unfortunately, they will no longer be even consistent if

these assumptions fail. Specification tests of the model are available to check the assumptions (see Hall (1987) and the references therein).

One problem with estimation of the sample selection model should be noted: this likelihood is not necessarily globally concave in ρ , although the likelihood can be written in a globally concave manner conditional on ρ . The implication is that a gradient maximization method may not find the global maximum in a finite sample. It is therefore sometimes a good idea to estimate the model by searching over $\rho \subset (-1, 1)$ and choosing the global maximum.⁵

5 Appendix

5.1 Truncated Normal Distribution

Define the standard normal density and cumulative distribution functions ($y \sim N(0, 1)$):

$$\begin{aligned}\phi(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \\ \Phi(y) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp\left(-\frac{1}{2}u^2\right) du\end{aligned}$$

Then if a normal random variable y has mean μ and variance σ^2 , we can write its distribution in terms of the standard normal distribution in the following way:

$$\begin{aligned}\phi(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) = \frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \\ \Phi(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \frac{1}{\sigma} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right) du = \Phi\left(\frac{y-\mu}{\sigma}\right)\end{aligned}$$

The truncated normal distribution of a random variable y with mean zero is defined as

$$E[y|y \geq c] = \frac{\frac{1}{\sigma} \int_c^\infty u \phi(u/\sigma) du}{\frac{1}{\sigma} \int_c^\infty \phi(u/\sigma) du} = \frac{\phi(c)}{1 - \Phi(c)} = \frac{\phi(-c)}{\Phi(-c)}$$

(Can you demonstrate this result?)

⁵TSP 4.5 and later versions perform the estimation of this model by searching on ρ and then choosing the best value.

5.2 Truncated bivariate normal

Now assume that the joint distribution of x and y is bivariate normal:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right]$$

One of the many advantages of the normal distribution is that the conditional distribution is also normal:

$$f(y|x) = N \left(\mu_y + \frac{\rho\sigma_x\sigma_y}{\sigma_x^2}(x - \mu_x), \sigma_y^2(1 - \rho^2) \right) = \phi \left(\frac{y - \mu_y - \frac{\rho\sigma_x\sigma_y}{\sigma_x^2}(x - \mu_x)}{\sigma_y\sqrt{1 - \rho^2}} \right)$$

That is, the conditional distribution of y given x is normal with a higher mean when x and y are positively correlated and x is higher than its mean, and lower mean when x and y are negatively correlated and x is higher than its mean. The reverse holds when x is lower than its mean. In general, y given x has a smaller variance than the unconditional distribution of y , regardless of the correlation of x and y .

Using this result, one can show that the conditional expectation of y , conditioned on x greater than a certain value, takes the following form:

$$E[y|x > a] = \mu_y + \rho\sigma_y\lambda\left(\frac{a - \mu_x}{\sigma_x}\right)$$

where

$$\lambda(u) = \frac{\phi(u)}{1 - \Phi(u)} = \frac{\phi(-u)}{\Phi(-u)}$$

The expression $\lambda(u)$ is sometimes known as the inverse Mills' ratio. It is the hazard rate for x evaluated at a . Here is a plot of the hazard rate as a function of $-u$. It is a monotonic function that begins at zero (when the argument is minus infinity) and asymptotes at infinity (when the argument is plus infinity):

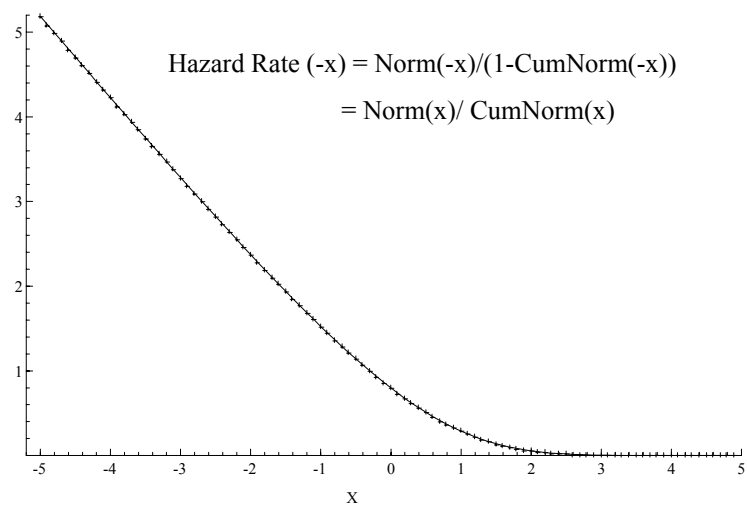


Figure 1: Plot of the Hazard Rate (negative argument)