

# Lecture 2: Models with Self-Selection

Wanchuan Lin

09/21/2007

## 1 Definition and Examples of Selection Bias

Sample selection bias refers to problems where the dependent variable is only observed for a restricted, nonrandom sample.

Endogeneity refers to the that an independent variable included in the model is potentially a choice variable, correlated with unobservables relegated to the error term. The dependent variable, however, is observed for all observations in the data.

1. In the Truman vs. Dewey election, it was predicted that Dewey would win because the sample was gathered using a phone sample. Most poorer people did not have a phone at the time, and were therefore not included in the sample. Wealth could have been used as a selection variable in examining the likelihood of being used in the sample.

2. You give a mail survey to see what factors affect time spent watching TV. An unmeasured variable, laziness, may affect the amount of TV watched and whether or not you return your survey. The sample is thus not representative.

3. Suppose attendance affects student score on exam and propose to divide your students into two groups: those with high attendance, and those with low attendance. An unmeasured variable, diligence, may affect both attendance and exam score. Higher values of diligence lead to higher attendance, and so students in the high attendance category are likely to have higher values of diligence than students in the low attendance group. As a consequence of this, regressing exam score on attendance overestimates the influence of attendance because it gives attendance credit for the influence of diligence. The problem is the students have chosen their attendance levels, rather than having this level forced upon them as would be the case in a controlled experiment.

## 2 Selection Biased Estimation

### 2.1 Mathematic Supplement

A **truncated distribution** is the part of an untruncated distribution that is above or below some specified value.

**Theorem 1**<sup>1</sup> *Density of a Truncated Random Variable*

If a continuous random variable  $x$  has pdf  $f(x)$  and  $a$  is a constant, then

$$f(x|x > a) = \frac{f(x)}{\Pr(x > a)}$$

The proof follows from the definition of condition probability and amounts merely to scaling the density so that it integrates to one over the range above  $a$ . NOTE that the truncated distribution is a conditional distribution.

Most recent application based on continuous random variables use the **truncated normal distribution**. If  $x$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then

$$\Pr(x > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha) \quad (1)$$

where  $\alpha = (a - \mu)/\sigma$  and  $\Phi(\cdot)$  is the standard normal cdf.

We are usually interested in the mean and variance of the truncated random variable. They would be obtained by the general formula:

$$E[x|x > a] = \int_a^{\infty} x f(x|x > a) dx$$

for the mean and likewise for the variance.

For the truncated normal distribution, we have the following theorem:<sup>2</sup>

**Theorem 2**<sup>3</sup> *Moments of the Truncated Normal Distribution*

If  $x \sim N[\mu, \sigma^2]$  and  $a$  is a constant, then

$$E[x|truncation] = \mu + \sigma\lambda(\alpha) \quad (2)$$

$$Var[x|truncation] = \sigma^2 [1 - \delta(\alpha)] \quad (3)$$

where  $\alpha = (a - \mu)/\sigma$ ,  $\phi(\alpha)$  is the standard normal density and

$$\lambda(\alpha) = \phi(\alpha) / [1 - \Phi(\alpha)] \quad \text{if truncation is } x > a \quad (4)$$

$$\lambda(\alpha) = -\phi(\alpha) / \Phi(\alpha) \quad \text{if truncation is } x < a \quad (5)$$

and

$$\delta(\alpha) = \lambda(\alpha) [\lambda(\alpha) - \alpha]$$

An important result is

$$0 < \delta(\alpha) < 1 \quad \text{for all values of } \alpha$$

A result that we will use at several points below is  $d\phi(\alpha)/d\alpha = -\alpha\phi(\alpha)$ . The function in (4) is also called the **hazard function** for the standard normal distribution.

<sup>1</sup>Theorem 22.1 of *Econometric Analysis* in fifth edition by William H. Greene

<sup>2</sup>Detail may be found in Johnson, Kotz, and Balakrishnan.

<sup>3</sup>Theorem 22.2 of *Econometric Analysis* in fifth edition by William H. Greene.

## 2.2 Biased Estimation

For showing the biased estimation from a regression based on a selection sample, what we need to compute is the expectation  $E(y_i|y_i > 0, x_i)$  where  $y_i$  is a typical value of the dependent variable in function  $y_i = x_i'\beta + \varepsilon_i$  where  $\varepsilon_i \sim N[0, \sigma_\varepsilon^2]$ . To do this we need to know the conditional density function  $f(y_i|y_i > 0)$ .

Note that

$$f(y_0|y_i > 0) = \frac{d}{dy} F(y_0|y_i > 0) = \frac{d}{dy} \Pr(y_i < y_0|y_i > 0)$$

Also, by Bayes rule,

$$\begin{aligned} \Pr(y_i \leq y_0|y_i > 0) &= \begin{cases} \frac{\Pr(y_i \leq y_0 \cap y_i > 0)}{\Pr(y_i > 0)} & \text{if } y_0 > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{\int_0^{y_0} f(y) dy}{\Pr(y_i > 0)} \end{aligned}$$

Hence,

$$f(y_0|y_i > 0) = \frac{d}{dy} \left[ \frac{\int_0^{y_0} f(y) dy}{\Pr(y_i > 0)} \right] \quad (6)$$

$$= \frac{f(y_0)}{\Pr(y_i > 0)} \quad (7)$$

For a normal distribution  $v \sim N[\mu, \sigma^2]$ , from (7), then

$$f(v|v > a) = \frac{f(v)}{1 - \Phi(a)} = \frac{(2\pi\sigma)^{-1/2} e^{-(v-\mu)^2/(2\sigma^2)}}{1 - \Phi(a)} = \frac{\frac{1}{\sigma} \phi\left(\frac{v-\mu}{\sigma}\right)}{1 - \Phi(a)} \quad (8)$$

where  $\phi(\cdot)$  is the standard normal pdf. Also note that

$$\frac{d}{dz} e^{-z^2/2} = -ze^{-z^2/2} \quad (9)$$

Using the result in (6) and (8) we have now

$$\begin{aligned} E(y_i|y_i > 0, x_i) &= E(x_i'\beta + \varepsilon_i|x_i'\beta + \varepsilon_i > 0, x_i) \\ &= x_i'\beta + E(\varepsilon_i|x_i'\beta + \varepsilon_i > 0, x_i) \\ &= x_i'\beta + \sigma_\varepsilon E\left(\frac{\varepsilon_i}{\sigma_\varepsilon} \mid \frac{\varepsilon_i}{\sigma_\varepsilon} > -\frac{x_i'\beta}{\sigma_\varepsilon}\right) \\ &= x_i'\beta + \sigma_\varepsilon \frac{\int_{-x_i'\beta/\sigma_\varepsilon}^{\infty} \frac{1}{2\pi} ze^{-z^2/2} dz}{1 - \Phi\left(-\frac{x_i'\beta}{\sigma_\varepsilon}\right)} \end{aligned}$$

$$\begin{aligned}
&= x'_i\beta + \sigma_\varepsilon \frac{-\frac{1}{2\pi} z e^{-z^2/2} \Big|_{-\frac{x'_i\beta}{\sigma_\varepsilon}}^{\infty} dz}{1 - \Phi\left(-\frac{x'_i\beta}{\sigma_\varepsilon}\right)}, \text{ from 9} \\
&= x'_i\beta + \sigma_\varepsilon \frac{\phi\left(-\frac{x'_i\beta}{\sigma_\varepsilon}\right)}{1 - \Phi\left(-\frac{x'_i\beta}{\sigma_\varepsilon}\right)} = x'_i\beta + \sigma_\varepsilon \frac{\phi\left(\frac{x'_i\beta}{\sigma_\varepsilon}\right)}{\Phi\left(\frac{x'_i\beta}{\sigma_\varepsilon}\right)} \\
&= x'_i\beta + \sigma_\varepsilon \lambda\left(-\frac{x'_i\beta}{\sigma_\varepsilon}\right) = x'_i\beta - \sigma_\varepsilon \lambda\left(\frac{x'_i\beta}{\sigma_\varepsilon}\right) \tag{10}
\end{aligned}$$

where the  $\lambda(a)$  is called the **inverse Mills ratio** like the function in (4) and (5).

Equation (10) states that if we run a regression of the positive  $y$ 's on  $x$ , then we should also include in the regression the term  $\lambda\left(-\frac{x'_i\beta}{\sigma_\varepsilon}\right)$ . A failure to do so will result in a bias estimate of  $\beta$  due to omitted variable bias.

### 3 Heckman two-step Method

#### 3.1 Motivate

The Heckman bivariate normal selection model represents the classic way for dealing with selection on unobservables. Selection on unobservables occurs when the error term in the outcome equation is correlated with the treatment, or with selection into the sample being used for estimation.

We begin with the Heckman model because it is the classic model in the literature. Relative to instrumental variables and panel models, it imposes, arguably, the strongest assumptions.

#### 3.2 Basic wage equation model

##### Setup

The context in which the bivariate normal selection model was developed is that of estimating a population wage equation when only wage information on workers is observed.

Note that interest in the population wage equation, rather than the wage equation for observed workers, is important here. The latter can easily be obtained by simple OLS estimation using data on workers and their wages.

The usual setup is as follows. we have a wage equation (outcome equation)

$$W_i = \beta X_i + \varepsilon_i$$

where  $W$  is the wage,  $X$  are observed variables related to productivity and  $\varepsilon$  includes all unobserved determinants of wages.

We assume that  $W$  is observed only for workers; it does not matter whether  $X$  is observed for just workers or for everyone, as this information will only be used for workers.

A reduced form employment equation (Participation equation) is given by

$$E_i^* = Z_i\gamma + \mu_i$$

where  $E_i^*$  is a latent index that can be thought of as representing the difference between the observed wage and the reservation wage. The latter is the lowest wage at which the individual is willing to accept employment.

We observe only an indicator variable for employment, defined as  $E = 1$  if  $E_i^* > 0$  and  $E = 0$  otherwise.

## Assumptions

In addition to the basic structure, the Heckman model requires the following

- (a)  $(\varepsilon, \mu) \sim N(0, 0, \sigma_\varepsilon^2, \sigma_\mu^2, \rho_{\varepsilon\mu})$ , bivariate normal distribution;
- (b)  $(\varepsilon, \mu)$  is independent of  $X$  and  $Z$ ;
- (c)  $var(\mu) = \sigma_\mu^2 = 1$

The first assumption represents a very strong functional form assumption - namely joint normality of the distribution of the error terms in the participation and outcome equation.

The second assumption is also strong; it assumes that both error terms are independent of both sets of observables.

The final assumption is the standard normalization for the probit selection equation, which is identified only up to scale.

## The selection problem

Now consider taking expectation of the wage equation conditional on working. Doing so yields

$$E(W_i | E_i = 1, X_i) = E(W_i | X_i, Z_i, \mu_i) = \beta X_i + E(\varepsilon_i | X_i, Z_i, \mu_i)$$

The first equality just recognizes the fact that the variables determining employment in this model are  $Z$  and  $\mu$ . The second equality comes from the fact that the expected value of  $X$  given  $X$  is just  $X$ .

The final term can be simplified by noting that selection into employment does not depend on  $X$ , only on  $Z$  and  $\mu$ . Not only that, but it depends on  $Z$  and  $\mu$  in a particular manner. This reasoning leads to

$$E(W_i | E_i = 1, X_i) = \beta X_i + E(\varepsilon_i | E_i = 1) = \beta X_i + E(\varepsilon_i | \mu_i > -Z_i\gamma)$$

Thus, if we estimate the model using only data on workers, we do not get the population wage equation, but rather something else.

As a result of this term, OLS estimation on a sample of workers generally provides inconsistent estimates of the parameters of the population wage equation.

## Key insights

One way to think about this, and this is the first key insight in the Heckman (1979) *Econometrica* paper, is that this is an omitted variables problem.

The second key insight is that an estimate of the omitted variable would solve the omitted variables problem and thereby the selection problem as well.

The third key insight is that under the assumptions listed above, we do have such an estimate, up to an unknown parameter. This estimate follows from standard properties of the bivariate normal distribution.

In formal terms, the final insight implies that

$$E(\varepsilon_i | \mu_i > -Z_i\gamma) = \rho_{\varepsilon\mu} \sigma_{\varepsilon} \lambda_i(-Z_i\gamma) = \theta \lambda_i(-Z_i\gamma)$$

### 3.3 Two step estimation of the wage equation model

#### The estimator

The first method for estimating the bivariate normal selection model is that due to Heckman (1979). It is sometimes called the "Heckman two-step" method.

Indeed, the model itself is sometimes misleadingly referred to as the Heckman two-step model. This is misleading because there are alternative ways of estimating the model in only one step, as will be shown later on. It is clearer, and more descriptive, to call the model the bivariate normal selection model.

**The first step** of the two-step approach run a **probit model** of participation ( $E$  on  $Z$ ) using all the observations. The estimates of  $\gamma$  from this probit model are then used to construct consistent estimates of the inverse Mills ratio term

$$\hat{\lambda}_i(-Z_i\hat{\gamma}) = \frac{\phi(Z_i\hat{\gamma})}{\Phi(Z_i\hat{\gamma})}$$

In **the second stage**, the outcome equation is estimated by ordinary least squares where the outcome equation includes both the original  $X$  whose coefficients are the parameters of the population wage equation and the constructed value of the inverse Mills ratio, which is

$$W_i = \beta X_i + \theta \hat{\lambda}_i(-Z_i\hat{\gamma}) + e_i$$

This step is carried out only for the uncensored observations and provides consistent and asymptotically normal estimators for  $\beta$  and  $\theta$ .

## Control function estimator

The inverse Mills ratio is sometimes called a "control function" - literally a function that controls for selection bias.

The bivariate normal model is the best-known member of the general class of control function estimators.

As we will see when we discuss two-stage squares in the context of instrumental variables estimation, there is also a control function version of that estimator.

## Interpretation

With the inverse Mills ratio included, and under the assumptions noted above, the coefficients on the  $X$  represent consistent estimates of the population wage equation.

The coefficient on the inverse Mills ratio term estimates  $\rho_{\varepsilon\mu}\sigma_{\varepsilon}$ . Because  $\sigma_{\varepsilon} > 0$  by definition, the sign of this coefficient is the same as the sign of  $\rho$ . The sign of  $\rho$  is often substantively useful information, as it indicates the correlation between the unobservable in the selection and outcome equations.

The standard  $t$ -test of the null that  $\theta = 0$  is a test of the null that there is no selection bias, conditional on the assumptions of model.

You can also back out an estimate of  $\rho$  and test that directly, although that is more complicated to do.

## Standard errors

The standard errors are tricky here for three reasons.

First, the additional variance that results from the generated regressor - namely the inverse Mills ratio term - must be taken into account.

Second, if there is indeed selection, then there is heteroskedasticity. To see this, note that for values of  $X$  that imply low wages, there will be more truncation and thus a lower variance of the error term in a sample of workers.

Third, spatial dependence is induced by the fact that a common  $\hat{\gamma}$  is used to construct the estimated inverse Mills ratio for all of the observations.

Heckman (1979) includes a consistent variance estimator that deals with all of these problems; you can find it in Greene. Stata produces the correct standard errors automatically.

### 3.4 Exclusion restrictions

The bivariate normal selection model is formally identified even if  $Z = X$ . The identification comes from the non-linearity of the inverse Mills ratio. A model that simply included the predicted probability of participation from a linear probability model into the outcome equation would not be identified.

However, the  $X = Z$  case often results in substantial collinearity between the predicted inverse Mills ratio term and the remaining covariates in the outcome equation. This will be especially strong when there is not much variation in the predicted participation probabilities, because then the non-linearities will not play a major role. This collinearity will, as always, lead to large standard errors.

More generally, a large Monte Carlo literature illustrates the poor performance of the bivariate normal model with no exclusion restriction in finite samples. The "exclusion restriction" here is a variable that "belongs" in the participation but not in the outcome equation. In other words, it is an instrument.

The bottom line is that you will have a tough time getting a paper published using the bivariate normal selection model that does not have a compelling exclusion restriction to help with identification.

### 3.5 Partial ML estimation of the wage equation model

The bivariate normal selection model can be estimated by maximum likelihood. Formally, it is partial maximum likelihood, because the observations of non-workers do not contribute an observed wage to the likelihood function.

Estimating the model using ML methods has both advantages and disadvantages. The key advantages are:

- (a) It is most efficient, assuming the bivariate normal assumption is correct;
- (b) The variances are more easily calculated.

The key disadvantages are:

- (a) The ML version relies more heavily on the functional form assumption and so is less robust than the two-step method;
- (b) The model sometimes has trouble converging (though this is probably telling you something about how much confidence to put in the two-step estimates as well).

To foreshadow: Stata can estimate the model both ways. It uses the two-step estimates as starting values for the partial ML estimation.