Econometrics II

Lecture 9: Sample Selection Bias

Måns Söderbom*

5 May 2011

[*] Department of Economics, University of Gothenburg. Email: mans.soderbom@economics.gu.se. Web: www.economics.gu.se/soderbom, www.soderbom.net.

1. Introduction

In this lecture we discuss how to estimate regressions if your sample is not random, in which case there may be sample selection bias.

References sample selection:

- Greene (2006) 24.5.1-3; 24.5.7.
- Wooldridge (2002) Chapter 17.1-17.2; 17.4 (optional)
- Vella, Francis (1998), "Estimating Models with Sample Selection Bias: A Survey," Journal of Human Resources, 33, pp. 127-169 (optional)
- François Bourguignon, Martin Fournier, Marc Gurgand "Selection Bias Corrections Based on the Multinomial Logit Model: Monte-Carlo Comparisons" DELTA working paper 2004-20, downloadable at http://www.delta.ens.fr/abstracts/wp200420.pdf (optional, but useful background reading for the computer exercise)

2. Sample Selection Bias

- Up to this point we have assumed the availability of a random sample from the underlying population. In practice, however, samples may not be random. In particular, samples are sometimes **truncated** by economic variables.
- We write our equation of interest (sometimes referred to as the 'structural equation' or the 'primary equation') as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{2.1}$$

where $\boldsymbol{x}_i$ is a vector of explanatory variables, all of which are *exogenous in the population*, and ε_i is an error term.

- Suppose selection is determined by the equation

$$z_i^* = \boldsymbol{w}_i'\boldsymbol{\gamma} + u_i \tag{2.2}$$

$$z_i = \left\{ \begin{array}{ll} 1 & \text{if } z_i^* \geq 0 \\ 0 & \text{otherwise} \end{array} \right\}, \tag{2.3}$$

  where $z_i = 1$ if we observe y_i and zero otherwise; the vector $\boldsymbol{w}_i$ is assumed to contain all variables in the vector $\boldsymbol{x}_i$ plus some more variables (unless otherwise stated); and u_i is an error term. We assume we always observe $\boldsymbol{w}_i$ (and thus $\boldsymbol{x}_i$), regardless of whether we observe y_i.

- **Example**: Suppose you want to study how education impacts on the wage an individual *could potentially* earn in the labour market - i.e. the wage offer. Your plan is to run a regression in which log wage is the dependent variable and education is (let's say) the only explanatory variable. You are primarily interested in the coefficient β on education. Suppose in the population, education is uncorrelated with the residual ε_i - i.e. it is exogenous (this can be relaxed, but the model would get more complicated as a result). Thus, with access to a random sample, OLS would be the best estimator.

- Suppose your sample contains a non-negligible proportion of unemployed individuals. For these individuals, there is no information on earnings, and so the corresponding observations cannot be used when estimating the wage equation (missing values for the dependent variable). Thus you're looking at having to estimate the earnings equation based on a non-random sample - what we shall refer to as a **selected sample**. Can the parameters of the wage offer equation - most importantly β - be estimated without bias based on the selected sample?

- The general answer to that question is: It depends! Whenever we have a selected (non-random) sample, it is important to be clear on two things:

  - Circumstances under which OLS estimates, based on the selected sample, will suffer from bias - specifically **selectivity bias** - and circumstances when it won't; and

- If there is selectivity bias in the OLS estimates: how to obtain estimates that are not biased by sample selection.

2.1. When will there be selection bias, and what can be done about it?

I will now discuss estimation of the model above under:

- (i) the assumptions made above (observability of $\boldsymbol{w}_i, \boldsymbol{x}_i, z_i$; exogeneity of $\boldsymbol{w}_i, \boldsymbol{x}_i$);
- and (ii) $u_i, \varepsilon_i, \sim$ Bivariate normal $[0, 0, 1, \sigma_\varepsilon, \rho]$.

Note that part (ii) implies

$$E(\varepsilon_i | u_i) = \beta_\lambda u_i$$

whereβ_λ measures the covariance between ε_i and u_i (prove this).

The fundamental issue to consider when worrying about sample selection bias is **why** some individuals will not be included in the sample. As we shall see, sample selection bias can be viewed as a special case of **endogeneity bias**, arising when the selection process **generates** endogeneity in the selected sub-sample.

In our model sample selection bias arises when the residual in the selection equation (i.e. u_i) is correlated with the residual in the primary equation (i.e. ε_i), i.e. whenever $\beta_\lambda \neq 0$. To see this, we will derive the expression for $E(y_i | \boldsymbol{w}_i, z_i = 1)$, i.e. the expectation of the outcome variable conditional on observable $\boldsymbol{w}_i$ and selection into the sample.

We begin by deriving $E(y_i | \boldsymbol{w}_i, u_i)$:

$$\begin{aligned} E(y_i | \boldsymbol{w}_i, u_i) &= \boldsymbol{x}_i' \boldsymbol{\beta} + E(\varepsilon_i | \boldsymbol{w}_i, u_i) \\ &= \boldsymbol{x}_i' \boldsymbol{\beta} + E(\varepsilon_i | u_i) \\ E(y_i | \boldsymbol{w}_i, u_i) &= \boldsymbol{x}_i' \boldsymbol{\beta} + \beta_\lambda u_i. \end{aligned} \tag{2.4}$$

Note that the exogeneity assumption for $\boldsymbol{w}_i$ enables us to go from the first to the second line (let's be strict on ourselves: by 'exogeneity' we mean that $\boldsymbol{w}_i$ is independent of ε_i, which is stronger than assuming these terms are uncorrelated); assuming bivariate normality enables us to go from the second to the third line.

Since u_i is not observable, eq (2.4) is not directly usable in applied work (since we can't condition on unobservables when running a regression). To obtain an expression for the expected value of y_i conditional on observables $\boldsymbol{w}_i$ and the actual selection outcome z_i, we make use of the law of iterated expectations, enabling us to write:

$$E\left(y_i|\boldsymbol{w}_i, z_i\right) = E\left[E\left(y_i|\boldsymbol{w}_i, u_i\right)|\boldsymbol{w}_i, z_i\right].$$

Hence, using (2.4) we obtain

$$\begin{aligned} E\left(y_i|\boldsymbol{w}_i, z_i\right) &= E\left[\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda u_i\right)|\boldsymbol{w}_i, z_i\right], \\ E\left(y_i|\boldsymbol{w}_i, z_i\right) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda E\left(u_i|\boldsymbol{w}_i, z_i\right), \\ E\left(y_i|\boldsymbol{w}_i, z_i\right) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda h\left(\boldsymbol{w}_i, z_i\right), \end{aligned}$$

where $h\left(\boldsymbol{w}_i, z_i\right) = E\left(u_i|\boldsymbol{w}_i, z_i\right)$ is some function.

Because the selected sample has $z_i = 1$, we only need to find $h\left(\boldsymbol{w}_i, z_i = 1\right)$. Our model and assumptions imply

$$E\left(u_i|\boldsymbol{w}_i, z_i = 1\right) = E\left(u_i|u_i \geq -\boldsymbol{w}_i'\boldsymbol{\gamma}\right),$$

and so we can use our 'useful result' appealed to in the previous lecture:

$$E\left(e|e > c\right) = \frac{\phi\left(c\right)}{1 - \Phi\left(c\right)}, \tag{2.5}$$

where e follows a standard normal distribution, c is a constant, ϕ denotes the standard normal probability density function, and Φ is the standard normal cumulative density function. Thus

$$\begin{aligned} E\left(u_i | u_i \geq -\boldsymbol{w}_i'\boldsymbol{\gamma}\right) &= \frac{\phi\left(-\boldsymbol{w}_i'\boldsymbol{\gamma}\right)}{1-\Phi\left(-\boldsymbol{w}_i'\boldsymbol{\gamma}\right)} \\ E\left(u_i | u_i \geq -\boldsymbol{w}_i'\boldsymbol{\gamma}\right) &= \frac{\phi\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right)}{\Phi\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right)} \equiv \lambda\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right), \end{aligned}$$

where $\lambda\left(\cdot\right)$ is the inverse Mills ratio (see Section 1 in the appendix for a derivation of the inverse Mills ratio).

We now have a *fully parametric expression* for the expected value of y_i, conditional on observable variables $\boldsymbol{w}_i$, and selection into the sample ($z_i = 1$):

$$E\left(y_i | \boldsymbol{w}_i, z_i = 1\right) = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right).$$

2.1.1. Exogenous sample selection: $E\left(\varepsilon_i |\ u_i\right) = 0$

- Assume that the unobservables determining selection are independent of the unobservables determining the outcome variable of interest:

  $$E\left(\varepsilon_i |\ u_i\right) = 0.$$

  In this case, we say that sample selection is **exogenous**, and - here's the good news - we can estimate the main equation of interest by means of OLS, since

  $$E\left(y_i | \boldsymbol{w}_i, z_i = 1\right) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

  hence

  $$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varsigma_i,$$

where ς_i is a mean-zero residual that is uncorrelated with $\boldsymbol{x}_i'$ in the selected sample (recall we assume exogeneity in the population). Examples:

- Suppose sample selection is randomized (or as good as randomized). Imagine an urn containing a lots of balls, where 20% of the balls are red and 80% are black, and imagine participation in the sample depends on the draw from this urn: black ball, and you're in; red ball and you're not. In this case sample selection is independent of **all** other (observable and unobservable) factors (indeed $\boldsymbol{\gamma} = 0$). Sample selection is thus exogenous.
- Suppose the variables in the $\boldsymbol{w}$-vector affect the likelihood of selection (i.e. $\boldsymbol{\gamma} \neq 0$). Hence individuals with certain observable characteristics are more likely to be included in the sample than others. Still, we've assumed $\boldsymbol{w}$ to be *independent* of the residual in the main equation, ε_i, and so sample selection remains **exogenous**. In this case also - no problem.

2.1.2. Endogenous sample selection: $E(\varepsilon_i|\ u_i) \neq 0$

Sample selection results in bias if the unobservables ε_i and u_i are correlated, i.e. $\beta_\lambda \neq 0$. Recall:

$$E(y_i|\boldsymbol{w}_i, z_i = 1) = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda(\boldsymbol{w}_i'\boldsymbol{\gamma})$$

- This equation tells us that the expected value of y_i, given $\boldsymbol{w}_i$ and observability of y_i (i.e. $z_i = 1$) is equal to $\boldsymbol{x}_i'\boldsymbol{\beta}$, **plus** an additional term which is the product of the covariance of the error terms β_λ and the inverse Mills ratio evaluated at $\boldsymbol{w}_i'\boldsymbol{\gamma}$. Hence in the selected sample, actual y_i is written as the sum of expected y_i (conditional on $\boldsymbol{w}$ and selection) and a mean-zero residual:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda(\boldsymbol{w}_i'\boldsymbol{\gamma}) + \varsigma_i,$$

- It follows that if, based on the selected sample, we use OLS to run a regression in which y_i is the dependent variable and $\boldsymbol{x}_i$ is the set of explanatory variables, then $\lambda(\boldsymbol{w}_i'\boldsymbol{\gamma})$ will go into the residual;

and to the extent that $\lambda\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right)$ is correlated with $\boldsymbol{x}_i$, the resulting estimates will be biased unless $\beta_\lambda = 0$. Omitted variables bias, right?

2.1.3. An example

Based on these insights, let's now think about estimating the following simple wage equation based on a selected sample.

$$\ln w_i = \beta_0 + \beta_1 educ_i + \varepsilon_i,$$

- Always when worrying about endogeneity, you need to be clear on the underlying mechanisms. So begin by asking yourself: What factors are likely to go into the residual ε_i in the wage equation? Clearly individuals with the same levels of education can obtain very different wages in the labour market, and given how we have written the model it follows by definition that the residual ε_i is the source of such wage differences. To keep the example simple, suppose I've convinced myself that the (true) residual ε_i consists of two parts:

  $$\varepsilon_i = \theta_1 m_i + e_i,$$

  where m_i is personal 'motivation', which is unobserved (note!) and assumed uncorrelated with education in the population (clearly a debatable assumption, but let's keep things reasonably simple), θ_1 is a positive parameter, and e_i reflects the remaining source of variation in wages. Suppose for simplicity that e_i is independent of all variables except wages.

- I know from my econometrics textbook that there will be sample selection bias in the OLS estimator if the residual in the earnings equation ε_i is correlated with the residual in the selection equation. Let's now relate this insight to economics, sticking to our example. Since motivation (m_i) is (assumed) the only economically interesting part of ε_i, I thus need to ask myself: Is it reasonable to assume that motivation is uncorrelated with education **in the selected sample**? For now, maintain the assumption that motivation and education are uncorrelated in the population - hence had there been no sample selection, education would have been exogenous and OLS would have been fine.

- Still - and this is the *key point* - I may suspect that selection into the labour market depends on education **and** motivation:

$$z_i = \left\{ \begin{array}{ll} 1 & \text{if } \gamma \cdot educ_i + (\theta_2 m_i + \eta_i) \geq 0 \\ 0 & \text{otherwise} \end{array} \right\},$$

  where θ_2 is a positive parameter and η_i is a residual independent of all factors except selection. Because m_i is unobserved it will go into the residual, which will consist of the two terms inside the parentheses (.).

- The big question now is whether the factors determining selection are correlated with the wage residual $\varepsilon_i = \theta_1 m_i + e_i$. There are only three terms determining selection. Two of these are η_i and $educ_i$, and they have been assumed uncorrelated with ε_i. But what about motivation, m_i? Abstracting from the uninteresting case where θ_1 and/or θ_2 are equal to zero, we see that i) motivation determines selection; and ii) motivation is correlated with the wage residual since $\varepsilon_i = \theta_1 m_i + e_i$. So clearly we have endogenous selection.

- Does this imply that education is correlated with ε_i **in the selected sample**? Yes it does. The intuition as to why this is so is straightforward. Think about the characteristics (education and motivation) of the people that are included in the sample.

  - Someone with a **low** level of education must have a **high** level of motivation, otherwise he or she is likely not to be included in the sample (recall: the selection model implies that individuals with **low** levels of education and **low** levels of motivation are those most unlikely to be included in the sample).

  - In contrast, someone with a **high** level of education is fairly likely to participate in the labour market even if he or she happens to have a relatively low level of motivation.

- The implication is that, **in the sample**, the average level of motivation among those with little education will be higher than the average level of motivation with those with a lot of education. In other words, education and motivation are negatively correlated **in the sample**, even though this is not the case in the population.

- And since motivation goes into the residual (since we have no data on motivation - it's unobserved), it follows that education is (negatively) correlated with the residual in the selected sample. And that's why we get selectivity bias.

- Illustration: Figure 2 in the appendix.

2.2. How correct for sample selection bias?

I will now discuss the two most common ways of correcting for sample selection bias.

2.2.1. Method 1: Inclusion of control variables

The first method by which we can correct for selection bias is simple: include in the regression observed variables that control for sample selection. In the wage example above , if we had data on motivation, we could just augment the wage model with this variable:

$$\ln w_i = \beta_0 + \beta_1 educ_i + \theta_1 m_i + e_i.$$

More generally, recall that

$$E\left(y_i | \boldsymbol{w}_i, u_i\right) = \boldsymbol{x}_i' \boldsymbol{\beta} + \beta_\lambda u_i,$$

and so if you have data on u_i, we could just use include this variable in the model as a control variable for selection and estimate the primary equation using OLS. Such a strategy would completely solve the sample selection problem.

Clearly this approach is only feasible if we have data on the relevant factors (e.g. motivation), which may not always be the case. The second way of correcting for selectivity bias is to use the famous **Heckit method**, developed by James Heckman in the 1970s.

2.2.2. Method 2: The Heckit method

We saw above that

$$E\left(y_i|\boldsymbol{w}_i, z_i = 1\right) = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right).$$

Using the same line of reasoning as for 'Method 1', it must be that if we had data on $\lambda\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right)$, we could simply add this variable to the model and estimate by OLS. Such an approach would be fine. Of course, in practice you would never have direct data on $\lambda\left(\boldsymbol{w}_i'\boldsymbol{\gamma}\right)$. However, the functional form $\lambda\left(\cdot\right)$ is known - or, rather, assumed (at least in most cases) - and $\boldsymbol{w}$ is (it is assumed) observed. If so, the only missing ingredient is the parameter vector $\boldsymbol{\gamma}$, which can be estimated by means of a probit model. The Heckit method thus consists of the following two steps:

1. Using **all** observations - those for which y_i is observed (selected observations) and those for which it is not - and estimate a probit model where z_i is the dependent variable and $\boldsymbol{w}_i$ are the explanatory variables. Based on the parameter estimates $\hat{\boldsymbol{\gamma}}$, calculate the inverse Mills ratio for each observation:

$$\lambda\left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right) = \frac{\phi\left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right)}{\Phi\left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right)}.$$

2. Using the selected sample, i.e. all observations for which y_i is observed, and run an OLS regression in which y_i is the dependent variable and $\boldsymbol{x}_i$ **and** $\lambda\left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right)$ are the explanatory variables:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda\left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right) + \varsigma_i.$$

   This will give consistent estimates of the parameter vector $\boldsymbol{\beta}$. That is, by including the inverse Mills ratio as an additional explanatory variable, we have corrected for sample selection bias.

Important considerations

- The Heckit procedure gives you an estimate of the parameter β_λ, which measures the covariance between the two residuals ε_i and u_i. Under the null hypothesis that there is no selectivity bias, we have $\beta_\lambda = 0$. Hence testing $H_0 : \beta_\lambda = 0$ is of interest, and we can do this by means of a conventional t-test. If you cannot reject $H_0 : \beta_\lambda = 0$ then this indicates that sample selection does not result in significant bias, and so using OLS on the selected sample without including the inverse Mills ratio is fine - all this, provided the model is correctly specified (i.e. all the underlying assumptions hold), of course.

- We assumed above that the vector $\boldsymbol{w}_i$ (the determinants of selection) contains all variables that go into the vector $\boldsymbol{x}_i$ (the explanatory variables in the primary equation), and possibly additional variables. In fact, it is highly desirable to specify the selection equation in such a way that there is *at least one* variable that determines selection, and which has no direct effect on y_i. In other words, it is important to impose at least one *exclusion restriction.* The reason is that if $\boldsymbol{x}_i = \boldsymbol{w}_i$, the second stage of Heckit is likely to suffer from a collinearity problem, with very imprecise estimates as a result. Recall the form of the regression you run in the second stage of Heckit:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda \left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right) + \varsigma_i.$$

  Clearly, if $\boldsymbol{x}_1 = \boldsymbol{x}$, then

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda \left(\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}\right) + \varsigma_i.$$

  Remember that collinearity arises when one explanatory variable can be expressed as a **linear** function of one or several of the other explanatory variables in the model. In the above model $\boldsymbol{x}_i$ enters linearly (the first term) and **non-**linearly (through inverse Mills ratio), which seems to suggest that there will not be perfect collinearity. However, if you look at the graph of the inverse Mills ratio (see Figure 1 in the appendix) you see that it is **virtually linear over a wide range**

of values. Clearly had it been exactly linear there would be no way of estimating

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_\lambda \lambda\left(\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}\right) + \varsigma_i.$$

because $\boldsymbol{x}_i$ would then be perfectly collinear with $\lambda\left(\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}\right)$. The fact that Mills ratio is virtually linear over a wide range of values means that you can run into problems posed by severe (albeit not complete) collinearity. This problem is solved (or at least mitigated) if $\boldsymbol{w}_i$ contains one or several variables that are not included in $\boldsymbol{x}_i$. Similar to identification with instrumental variables, the exclusion restriction has to be justified theoretically in order to be convincing. And that, alas, is not always straightforward.

- Finally, always remember that in order to use the Heckit approach, you must have data on the explanatory variables for both selected and non-selected observations. This may not always be the case.

Quantities of interest Now consider partial effects. Suppose we are interested in the effects of changing the variable x_k. It is useful to distinguish between two quantities of interest:

- The effect of a change on x_k on expected y_i in the population:

$$\frac{\partial E\left(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}\right)}{\partial x_k} = \beta_k$$

For example, if x_k is education and y_i is wage offer, then β_k measures the marginal effect of education on expected wage offer in the population.

- The effect of a change on x_k on expected y_i for individuals in the population for whom y_i is observed:

$$\frac{\partial E\left(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, z_i = 1\right)}{\partial x_k} = \beta_k + \beta_\lambda \frac{\partial \lambda\left(\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}\right)}{\partial x_{ki}}.$$

Recall that

$$\lambda'(c) = -\lambda(c)\left[c + \lambda(c)\right],$$

hence

$$\frac{\partial E\left(y_i | \boldsymbol{x}_i'\boldsymbol{\beta}, z_i = 1\right)}{\partial x_k} = \beta_k - \beta_\lambda \gamma_k \lambda\left(\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}\right)\left[\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}} + \lambda\left(\boldsymbol{x}_i'\hat{\boldsymbol{\gamma}}\right)\right].$$

It can be shown that $c + \lambda(c) > 0$, hence if β_λ and γ_k have the **same sign**, this partial effect is lower than that on expected y_i in the population. In the context of education and wage offers, what is the intuition of this result? [Hint: increase education and less able individuals will work.]

Estimation of Heckit in Stata In Stata we can use the command **heckman** to obtain Heckit estimates. If the model is

$$y_i = \beta_0 + \beta_1 x1_i + \varepsilon_i,$$

$$z_i = \left\{ \begin{array}{cc} 1 & \text{if } \gamma_0 + \gamma_1 z1_i + \gamma_2 x1_i + u_i \geq 0 \\ 0 & \text{otherwise} \end{array} \right\},$$

the syntax has the following form

```
heckman y x1, select (z1 x1) twostep
```

where the variable y_i is **missing** whenever an observation is not included in the selected sample. If you omit the twostep option you get full information maximum likelihood (FIML) estimates. Asymptotically, these two methods are equivalent, but in small samples the results can differ. Simulations have taught us that FIML is more efficient than the two-stage approach but also more sensitive to mis-specification due to, say, non-normal disturbance terms. In applied work it makes sense to consider both sets of results.

EXAMPLES: See Section 2.1-2.3 in appendix.

2.3. Extensions of the Heckit model

2.3.1. Non-continuous outcome variables

We have focused on the case where y_i, i.e. the outcome variable in the structural equation, is a continuous variable. However, sample selection models can be formulated for many different models - binary response models, censored models, duration models etc. The basic mechanism generating selection bias remains the same: correlation between the unobservables determining selection and the unobservables determining the outcome variable of interest.

Consider the following binary response model with sample selection:

$$\begin{aligned} y_i &= 1\left[\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0\right] \\ z_i &= 1\left[\boldsymbol{w}_i'\boldsymbol{\gamma} + u_i > 0\right], \end{aligned}$$

where y_i is observed only if $z_i = 1$, and $\boldsymbol{w}$ contains $\boldsymbol{x}$ and at least one more variable. In this case, probit estimation of $\boldsymbol{\beta}$ based on the selected sample will generally lead to inconsistent results, unless ε_i and u_i are uncorrelated. Assuming that $\boldsymbol{w}$ is exogenous in the population, we can use a two-stage procedure very similar to that discussed above:

1. Obtain $\hat{\boldsymbol{\gamma}}$ by estimating the participation equation using a probit model. Construct $\hat{\lambda}_{i2} = \lambda\left(\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\right)$.

2. Estimate the structural equation using probit, with $\hat{\lambda}_{i2}$ added to the set of regressors:

$$\Pr\left(y_i = 1 | \boldsymbol{x}_i, z_i = 1\right) = \Phi\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \rho_1\hat{\lambda}_{i2}\right),$$

   where ρ_1 measures the correlation between the residuals ε_i and u_i (note: correlation will be the same as the covariance, due to unity variance for the two residuals)

This is a good procedure for testing the null hypothesis that there is no selection bias (in which case $\rho_1 = 0$). If, based on this test we decide there is endogenous selection, we might choose to estimate the two equations of the model simultaneously (in Stata: **heckprob**). This produces the right standard errors, and recovers the structural parameters $\boldsymbol{\beta}$ rather than a scaled version of this vector.

2.3.2. Non-binary selection equation

Alternatively, it could be that the selection equation is not a binary response model. In the computer exercise we will study the case where selection is modelled by means of a **multinomial logit**. An excellent survey paper in this context is that by Bourguignon, Fournier and Gurgand.

PhD Programme: Econometrics II
Department of Economics, University of Gothenburg
Appendix
Måns Söderbom

1. Derivation of the Inverse Mills Ratio (IMR)

To show $$E(z \mid z > c) = \frac{\phi(c)}{1-\Phi(c)} = \frac{\phi(-c)}{\Phi(-c)}$$

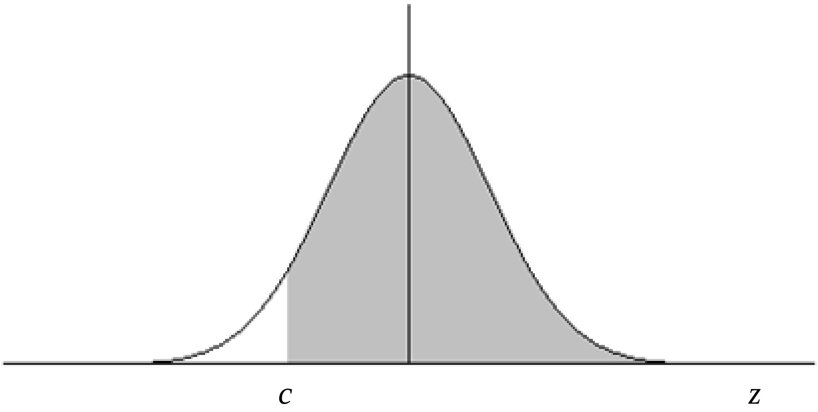
Assume that z is normally distributed:

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(z)dz$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$$

$G(z)$ is the normal cumulative density function (CDF), $\phi(z)$ is the standard normal density function.

We now wish to know the $E(z \mid z > c)$. It is the shaded area in the graph below.



By the characteristics of the normal curve is equal to $[1-\Phi(c)]$. So the density of z is given by

$$\frac{\phi(z)}{[1-\Phi(c)]}, \quad z > c$$

so

$$E(z\,|\,z>c)=\int_c^\infty \frac{z\phi(z)}{[1-\Phi(c)]}dz$$

which can be written using the definitions above as:

$$E(z\,|\,z>c)=\frac{1}{(1-\Phi(c))}\int_c^\infty \frac{z}{\sqrt{2\pi}}.\exp(\frac{-z^2}{2})dz$$

This expression can be written as:

$$E(z\,|\,z>c)=\frac{1}{(1-\Phi(c))}\int_c^\infty -(\frac{d\phi(z)}{dz})dz$$

How do we know that:

$$\frac{d\phi(z)}{dz}=\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})\cdot -z$$

$$\int_c^\infty -(\frac{d\phi(z)}{dz})dz=\int_c^\infty -\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})=0+\frac{1}{\sqrt{2\pi}}\exp(-\frac{c^2}{2})=\phi(c)$$

So:

Lets evaluate $\int_c^\infty \frac{z}{\sqrt{2\pi}}.\exp(\frac{-z^2}{2})dz=$

This can be written as

$$-\frac{1}{[1-\Phi(c)]}\int_c^\infty d\Phi(z)=\frac{\phi(c)}{[1-\Phi(c)]}$$

Recall that for the normal distribution $\phi(c)=\phi(-c)$ and $1-\Phi(c)=\Phi(-c)$

From which it follows that

$$E(z\,|\,z>c)=\int_c^\infty \frac{z\phi(z)}{[1-\Phi(c)]}dz=\frac{\phi(-c)}{\Phi(-c)}$$

It is this last expression which is the inverse Mills ratio.

Figure 1: The Inverse Mills Ratio

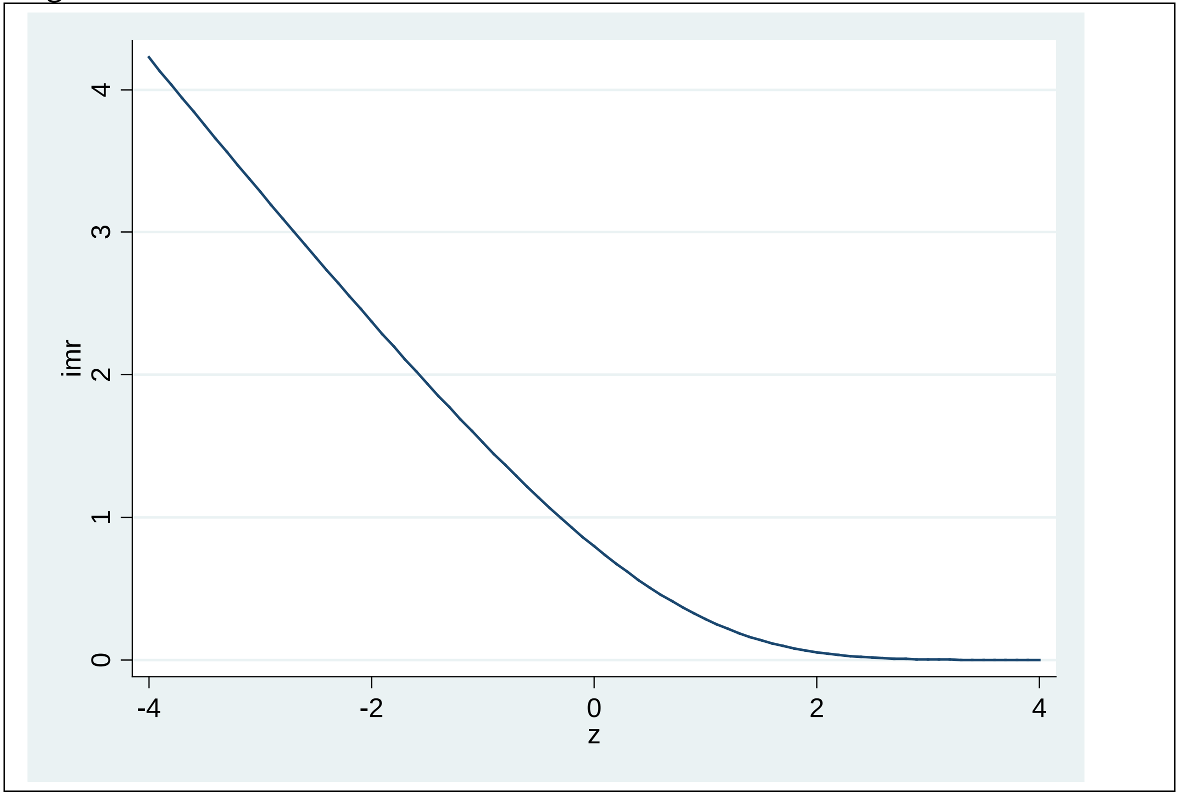
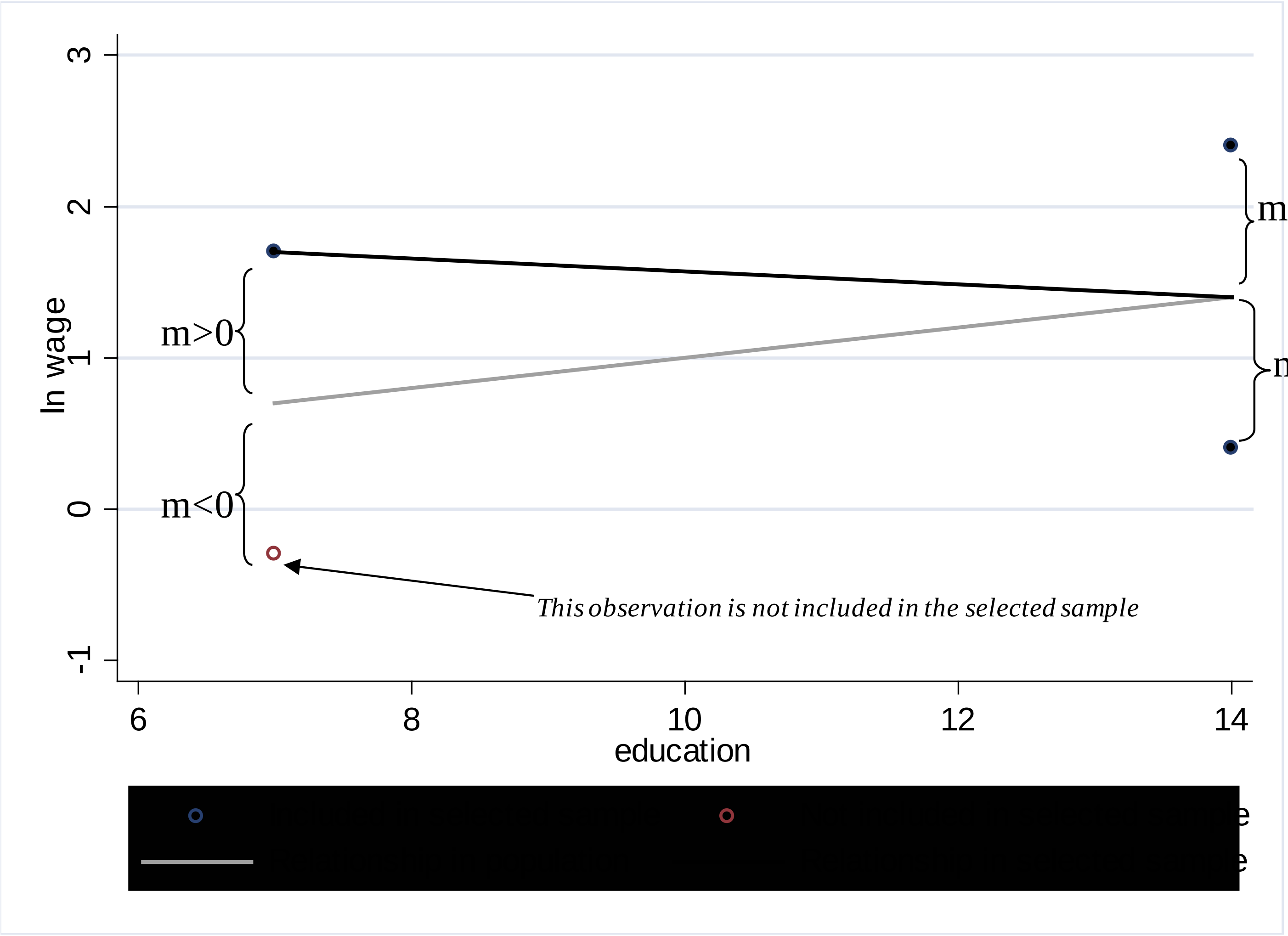


Figure 2: Illustration of Sample Selection Bias



The economic model underlying the graph is

$$\ln w = cons + 0.1educ + m,$$

where w is wage, educ is education and m is unobserved motivation.

2. Empirical illustration of the Heckit model

Earnings regressions for females in the US

This section uses the MROZ dataset.[1] This dataset contains information on 753 women. We observe the wage offer for only 428 women, hence the sample is truncated.

```
use C:\teaching_gbg07\applied_econ07\MROZ.dta
```

```
1. OLS on selected sample

reg  lwage educ exper expersq

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   26.29
       Model |  35.0223023     3  11.6741008           Prob > F      =  0.0000
    Residual |  188.305149   424  .444115917           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1509
       Total |  223.327451   427  .523015108           Root MSE      =  .66642

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
       exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
     expersq |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
       _cons |  -.5220407   .1986321    -2.63   0.009    -.9124668   -.1316145
------------------------------------------------------------------------------
```

[1] Source: Mroz, T.A. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," Econometrica 55, 765-799.

2. Two-step Heckit

```
. heckman lwage educ exper expersq, select(nwifeinc educ exper expersq age
kidslt6 kidsge6) twostep

Heckman selection model -- two-step estimates   Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(6)       =    180.10
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1090655    .015523     7.03   0.000     .0786411      .13949
       exper |   .0438873   .0162611     2.70   0.007     .0120163    .0757584
     expersq |  -.0008591   .0004389    -1.96   0.050    -.0017194    1.15e-06
       _cons |  -.5781033   .3050062    -1.90   0.058    -1.175904    .0196979
-------------+----------------------------------------------------------------
select       |
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267472    1.266901
-------------+----------------------------------------------------------------
mills        |
      lambda |   .0322619   .1336246     0.24   0.809    -.2296376    .2941613
-------------+----------------------------------------------------------------
         rho |    0.04861
       sigma |  .66362876
      lambda |  .03226186   .1336246
------------------------------------------------------------------------------
```

3. Simultaneous estimation of selection model

```
. heckman lwage educ exper expersq, select(nwifeinc educ exper expersq age
kidslt6 kidsge6)

Iteration 0:   log likelihood = -832.89777
Iteration 1:   log likelihood =  -832.8851
Iteration 2:   log likelihood = -832.88509

Heckman selection model                         Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(3)       =     59.67
Log likelihood = -832.8851                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1083502   .0148607     7.29   0.000     .0792238    .1374767
       exper |   .0428369   .0148785     2.88   0.004     .0136755    .0719983
     expersq |  -.0008374   .0004175    -2.01   0.045    -.0016556   -.0000192
       _cons |  -.5526974   .2603784    -2.12   0.034     -1.06303   -.0423652
-------------+----------------------------------------------------------------
select       |
    nwifeinc |  -.0121321   .0048767    -2.49   0.013    -.0216903    -.002574
        educ |   .1313415   .0253823     5.17   0.000     .0815931    .1810899
       exper |   .1232818   .0187242     6.58   0.000     .0865831    .1599806
     expersq |  -.0018863   .0006004    -3.14   0.002     -.003063   -.0007095
         age |  -.0528287   .0084792    -6.23   0.000    -.0694476   -.0362098
     kidslt6 |  -.8673988   .1186509    -7.31   0.000     -1.09995   -.6348472
     kidsge6 |   .0358723   .0434753     0.83   0.409    -.0493377    .1210824
       _cons |   .2664491   .5089578     0.52   0.601    -.7310898    1.263988
-------------+----------------------------------------------------------------
     /athrho |    .026614    .147182     0.18   0.857    -.2618573    .3150854
    /lnsigma |  -.4103809   .0342291   -11.99   0.000    -.4774687   -.3432931
-------------+----------------------------------------------------------------
         rho |   .0266078   .1470778                     -.2560319    .3050564
       sigma |   .6633975   .0227075                      .6203517    .7094303
      lambda |   .0176515   .0976057                     -.1736521    .2089552
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.03   Prob > chi2 = 0.8577
------------------------------------------------------------------------------
```