Lecture 3: Corner solutions, censored data and selected samples

Luc Behaghel (luc.behaghel@ens.fr)
Paris School of Economics, Master PPD

January 2009

1 Introduction

We continue our exploration of non standard dependent variables. The "mechanics" is similar to models we saw in the previous chapters (CMLE, OLS), and we will not insist on it. Instead, our focus will be on two things: (i) when to use which model; (ii) under which conditions the models are identified. Question (ii) is very important for selectivity correction models.

The different cases we are going to consider are:

1. **"Corner solution" models.** Cases where y is (roughly) continuous over strictly positive values but is 0 for a non trivial fraction of the population. This is a case where y is perfectly observed. But the spike at 0 arises as a corner solution to an optimization problem: the agent maximizes his utility by choosing the amount of y to buy under the constraint that his demand is positive ($y \geq 0$). *Examples: amount spent on DVD purchase; hours worked during a week.*

2. **Censored regression models.** y is censored when the value it takes are unknown when it is above (or below) a threshold c. This a missing data problem. Observations for which $y \geq c$ are in the sample (in particular, we observe covariates x), but the exact value of y is not observed. *Example: top coded income.*

3. **Models with selectivity correction.** Selection models apply when the sample available for estimation is a *non random* sample of the population of interest. *Examples: wages observed only among employed people; non response in a survey.*

2 "Corner solution" models

We call "corner solution" models the models that apply in cases where y is always positive but is 0 for a nontrivial fraction of the population. This typically arises for consumption: some individuals do not want the good ($y = 0$) while others demand different quantities ($y > 0$); or for labor supply: some people do not participate ($y = 0$) while others offer strictly positive hours ($y > 0$).

This setting is not very different from the Poisson model we saw in the previous lecture. The main difference is that y is now supposed to be continuous (rather than discrete) over positive values. Just like in the Poisson case, a linear model is inappropriate (it could yield *negative* predicted outcomes) and taking the log of y is not feasible due to the values where $y = 0$.

The standard model in that case is the **Tobit** model, defined as a latent variable model:

$$\begin{aligned} y^* &= x\beta + \varepsilon \\ y &= \max(0, y^*) \\ \varepsilon|x &\sim N(0, \sigma^2) \end{aligned}$$

Given these assumptions, the contribution to the likelihood for observations with $y_i = 0$ is

$$\Pr(y_i = 0|x_i) = \Pr(\varepsilon_i < -x_i\beta|x_i) = 1 - \Phi\left(\frac{x_i\beta}{\sigma}\right)$$

and

$$f(y_i|x_i) = f\left(\frac{\varepsilon}{\sigma} = \frac{y_i - x_i\beta}{\sigma}\right) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)$$

for observations with $y_i > 0$.

Therefore, the log-likelihood is

$$\log L(y|x;\beta,\sigma) = \sum_{i=1}^{N}\left\{1(y_i = 0)\log\left[1 - \Phi\left(\frac{x_i\beta}{\sigma}\right)\right] + 1(y_i > 0)\log\left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right]\right\}.$$

The Tobit model is easy to use in practice, as its likelihood is already coded in econometric softwares. The main difficulty is to compute marginal effects. In most cases, we are not interested in the index variable (y^*) itself. We might be interested in two versions of the CEF of y: $E(y|x)$ or $E(y|y > 0, x)$. To see what these two quantities mean, take the labor supply example. Assuming that x is age, $\frac{\partial E(y|x)}{\partial x}$ tells us by how much the average number of hours supplied increases when age increases; this cumulates the impact on the extensive margin (new entrants in the labor force) and the impact on the intensive margin (changes in hours for those already participating). By contrast, $\frac{\partial E(y|y>0,x)}{\partial x}$ only describes the intensive margin.

Of course, the two CEF are connected:

$$E(y|x) = \Pr(y > 0|x)E(y|x, y > 0) = \Phi\left(\frac{x\beta}{\sigma}\right)E(y|x, y > 0).$$

Now, $E(y|y > 0, x)$ is the mean of a truncated normal. Here is a useful result: if $z \sim N(0,1)$, then $E(z|z > c) = \phi(c)/(1 - \Phi(c))$. Therefore,

$$\begin{aligned} E(y|x, y > 0) &= x\beta + E(\varepsilon|\varepsilon > -x\beta) \\ &= x\beta + \sigma E\left(\tfrac{\varepsilon}{\sigma}|\tfrac{\varepsilon}{\sigma} > -\tfrac{x\beta}{\sigma}\right) \\ &= x\beta + \sigma\tfrac{\phi(-x\beta/\sigma)}{1-\Phi(-x\beta/\sigma)} \\ &= x\beta + \sigma\tfrac{\phi(x\beta/\sigma)}{\Phi(x\beta/\sigma)} \\ &= x\beta + \sigma\lambda\left(\tfrac{x\beta}{\sigma}\right) \end{aligned}$$

where $\lambda = \phi(.)/\Phi(.)$ is called the **inverse Mills ratio**.

Finally,

$$E(y|x, y > 0) = x\beta + \sigma\lambda\left(\frac{x\beta}{\sigma}\right) \tag{1}$$

and

$$E(y|x) = \Phi\left(\frac{x\beta}{\sigma}\right)x\beta + \sigma\phi\left(\frac{x\beta}{\sigma}\right).$$

From these expressions, you can see that taking the marginal effects will yield complicated expressions depending on β and on the value x_0 at which the effects are evaluated.[1]

Important note: if we are only interested in $E(y|y > 0, x)$, why can't we simply ignore the observations for which $y = 0$ and run OLS on observations with $y > 0$? We will go back to this in section 4, but equation (1) gives you a first (formal) answer: $E(y|y > 0, x)$ is **not** equal to $x\beta$. There is an additional term $\sigma\lambda(\frac{x\beta}{\sigma})$. OLS of y on x in the sample with $y > 0$, because they omit this term, would yield biased estimates (omitted variable bias).

[1] Actually, one of the marginal effects simplifies to

$$\frac{\partial E(y|x)}{\partial x_k} = \beta_k\Phi\left(\frac{x\beta}{\sigma}\right).$$

3 Censored regression models

We call "censored regression models" the models that apply when the outcome is roughly continuous, but its values are unknown when they are above some threshold c (c itself is known). The typical example is top-coded earnings in a Mincer equation.

This is very close to the interval-coded data problem we saw in lecture 2: y is "interval-coded" only in the last interval. We use a very similar approach, based on a latent model. Specifically, the **censored regression model** is defined by

$$\begin{aligned} y^* &= x\beta + \varepsilon \\ y &= \min(y^*, c) \\ \varepsilon|x &\sim N(0, \sigma^2) \end{aligned}$$

It turns out that this also looks very similar to the Tobit model. If we had assumed that the exact value y was unobserved for $y^* < 0$ (instead of $y^* > c$), the two models would be formally equivalent. However, their interpretations would still be different.

The log-likelihod of the censored regression model is

$$\log L(y|x;\beta,\sigma) = \sum_{i=1}^{N} \left\{ 1(y_i = c_i) \log \left[1 - \Phi\left(\frac{c_i - x_i\beta}{\sigma} \right) \right] + 1(y_i < c_i) \log \left[\frac{1}{\sigma} \phi\left(\frac{y_i - x_i\beta}{\sigma} \right) \right] \right\}.$$

The main difference with the Tobit model is that y^*, not y, is the outcome of interest. The partial effects are therefore very easy to compute:

$$\frac{\partial E(y^*|x)}{\partial x_k} = \beta_k.$$

In the same way as for the interval-coded data case, it turns out that we are able to estimate the same parameters (β, σ) as if there had been no censoring: no information seems to be lost through censoring. But this comes at a cost: we need to assume a specific distribution for $\varepsilon|x$. The censored regression model is thus much less robust than the usual regression model that we could have estimated by OLS without making any distributional assumption.

4 Models with selectivity correction

Let us go back to our problem of estimating a CEF. Here, we assume that we are satisfied with a linear specification (y is not binary nor discrete):

$$\begin{aligned} E(y|x) &= x\beta, \text{ or, equivalently:} \\ y &= x\beta + \varepsilon \text{ with } E(\varepsilon|x) = 0. \end{aligned} \tag{2}$$

Remember the two steps in econometrics: identification and inference. If there was no sample selection problem, we would say that:

1. The model identifies β (for instance, if there is only one explanatory variable, $\beta_1 = \frac{Cov(x,y)}{Var(x)}$)
2. *If we have a random sample of* (x, y), we can infer an estimate of $\frac{Cov(x,y)}{Var(x)}$ using the corresponding empirical moments.

Sample selection problems arise when *we don't have a random sample of* (x, y). Rather, let us note s the indicator variable for being selected into our sample. s splits the population into two subpopulations: those who are observable ($s = 1$), and those who are not observable ($s = 0$). We only have a random sample of those with $s = 1$. Therefore, we cannot identify β from quantities such as $Cov(x, y)$, $Var(x)$, $E(y|x)$ that we have used so far: these quantities are not useful for identification as their empirical counterparts are not

observed. What we observe are empirical analogs of $Cov(x, y|s = 1)$, $Var(x|s = 1)$, $E(y|x, s = 1)$. Therefore, the fundamental identification issue of selection models is: **can we identify the parameter β in model (2) based on quantities that are observable only conditionally on $s = 1$**[2]?

4.1 A simple case of sample selection: truncated regression models

Truncation is a specific case of sample selection where we observe only individuals whose outcome y is below (or above) a known threshold, c. This is close to but different from censored data. In **censored** data, we observe the explanatory variables x for all observations (we observe (x, y) for uncensored observations, and only x for censored observations). In a **truncated** sample, this is not necessary: we observe (x, y) in the selected sample, and we might not observe x outside of the selected sample. So our **identification question** is: can we identify β if we only observe (x, y) when $s \equiv 1(y < c) = 1$?

For instance, imagine a study focusing on poor households. Imagine that our sample was selected so that it only retains workers below the poverty line $y < y_{poverty}$ (where $y_{poverty}$ denotes the poverty line; in France, it is 60% of the median income). We are interested in the following relationship:

$$\ln y = \beta_0 + \beta_1 edu + \varepsilon \tag{3}$$

where edu is the number of years of schooling of the household head. We assume that

$$E(\varepsilon|edu) = 0$$
$$(y, edu) \text{ is observed iff } y < y_{poverty}$$

(In reality, $E(\varepsilon|edu) = 0$ might not be true: there is probably an omitted variable problem here. But we want to abstract from that and to focus on the sample selection, so we maintain this assumption).

To see the consequences of sample selection, let us first represent the relationship *without* sample selection (figure 1), then, *with* sample selection (figure 2). Clearly, the regression slope appears smaller on the selected sample. In other words, by truncating the sample, we have changed the observed relationship between schooling and income. Specifically, this has led us to **underestimate** β_1. What happened?

From the graph, note that wealthier households tend to have more educated household heads. Therefore, if we restrict the analysis to households below the poverty line, we strongly underestimate the average income of more educated households. Hence, we underestimate the income difference between more educated households and less educated households. That is, we systematically underestimate β_1. The use of OLS on the truncated sample leads to a downward bias.

4.1.1 OLS bias in truncated samples

Let us make this point slightly more formally. β_1 is defined by

$$E(\ln y|x) = \beta_0 + \beta_1 edu$$

That relationship is displayed by figure 1. By contrast, figure 2 represents the relationship:

$$E(\ln y|x, y < y_{poverty}) = \beta_{0,trunc} + \beta_{1,trunc} edu$$

We have seen graphically that $\beta_1 > \beta_{1,trunc}$. Can we derive this formally? From equation (3), we have:

$$\begin{aligned} E(\ln y|x, y &< y_{poverty}) = E(\beta_0 + \beta_1 edu + \varepsilon|x, \ln y < \ln y_{poverty}) \\ &= \beta_0 + \beta_1 edu + E(\varepsilon|x, \beta_0 + \beta_1 edu + \varepsilon < \ln y_{poverty}) \\ &= \beta_0 + \beta_1 edu + E(\varepsilon|\varepsilon < \ln y_{poverty} - (\beta_0 + \beta_1 edu)) \\ &\equiv \beta_0 + \beta_1 edu + g(\ln y_{poverty} - (\beta_0 + \beta_1 edu)). \end{aligned} \tag{4}$$

[2] Examples of such quantities would be $Cov(x, y|s = 1)$, $Var(x|s = 1)$, $E(y|x, s = 1)$. Unfortunately, as we will see, these quantities are not immediately related to β.

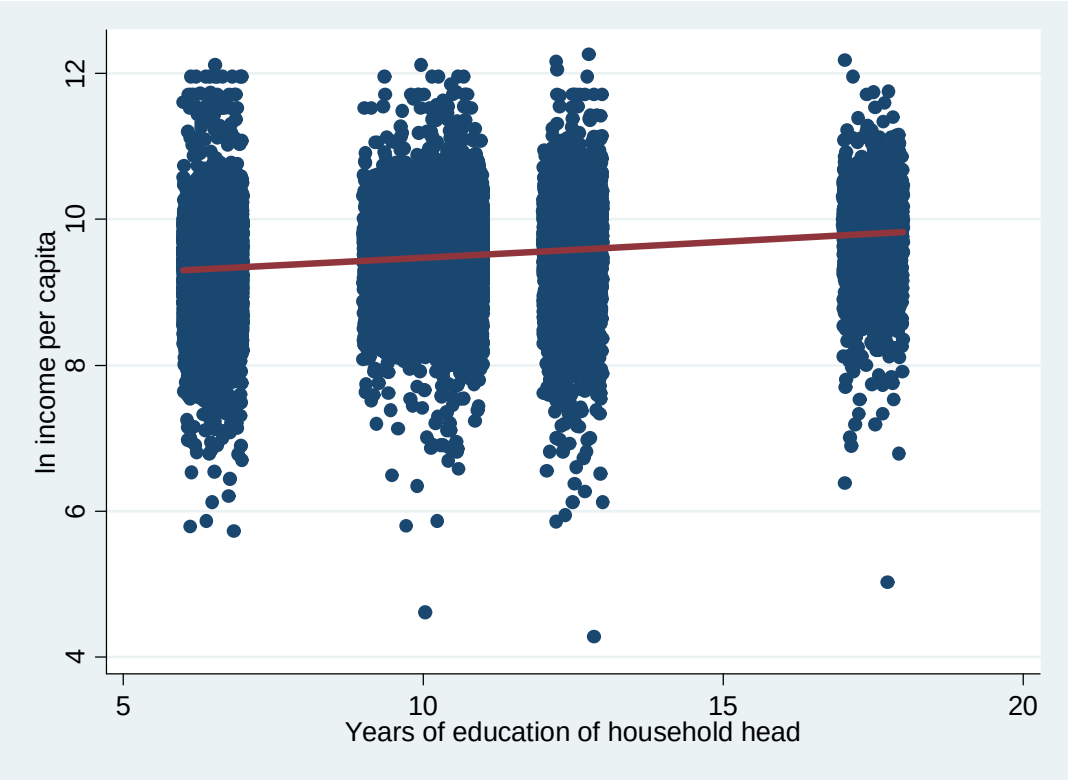

Figure 1: Full sample

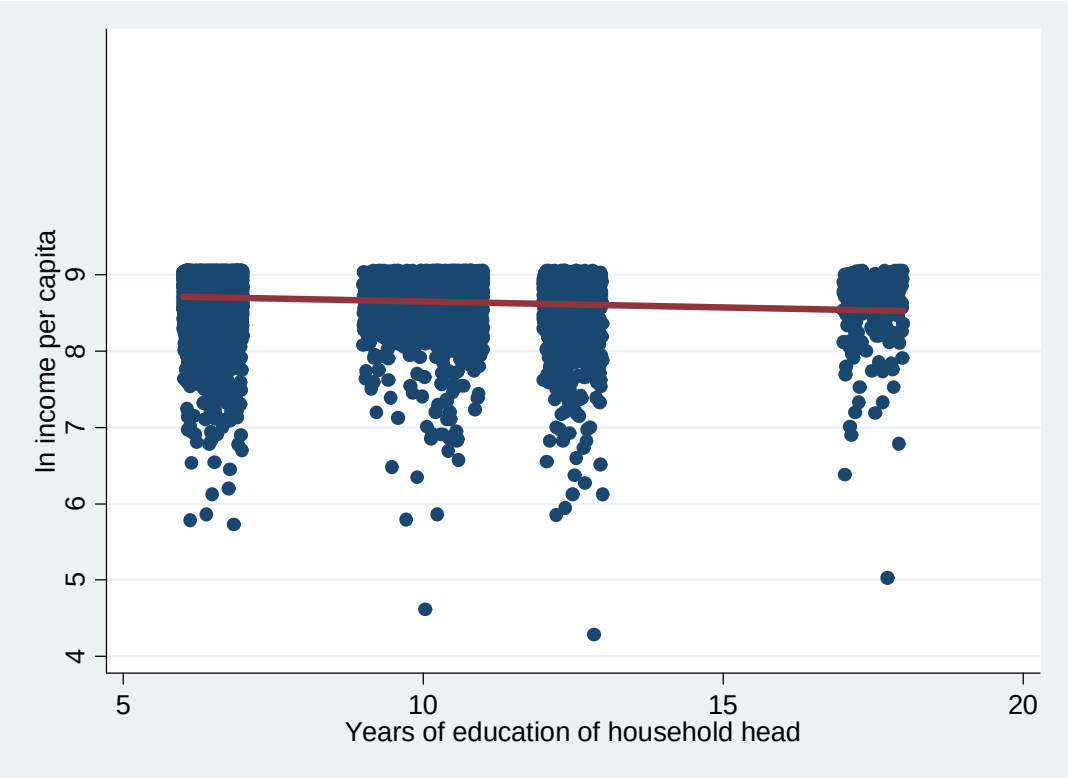

Figure 2: Truncated sample (below poverty line)

$g(.)$ denotes the conditional expectation function such that $g(t) = E(\varepsilon|\varepsilon < t)$. The form of this function is unknown. The last line shows that if we run OLS of y on x in the selected sample, there is an omitted variable, $g(\ln y_{poverty} - (\beta_0 + \beta_1 edu))$. This variable is a function of edu. It is therefore clearly correlated with edu and its omission generates an omitted variable bias.

To know the direction of the bias, we need to assess whether $g(\ln y_{poverty} - (\beta_0 + \beta_1 edu))$ is positively or negatively correlated with edu. Although we don't know the exact form of g, we know that it is an increasing function (as you can see from $g(x) = E(\varepsilon|\varepsilon < x)$: the higher x, the higher the ε's included in the expectation, and the higher the expectation); therefore, as long as β_1 is positive, the omitted term is decreasing in edu. This implies that $Cov(edu, g(\ln y_{poverty} - (\beta_0 + \beta_1 edu))) < 0$: there is a downward bias. This confirms what we saw in the graphs and understood intuitively.

4.1.2 ML estimation

A parametric solution to our problem is to write the likelihood for the *selected* sample. For that, we need to fully specify the distribution of ε. We assume:

$$\begin{gathered} y = x\beta + \varepsilon \\ y \text{ observed iff } y < c \\ \varepsilon|x \sim N(0, \sigma^2) \end{gathered}$$

Then, the probability of observing (y_i, x_i) given that $y_i < c$ is:

$$\begin{aligned} \Pr(y = y_i|x = x_i, y < c) &= \frac{\Pr(y=y_i, y<c|x=x_i)}{\Pr(y<c|x=x_i)} \\ &= \frac{\Pr(y=y_i|x=x_i)}{\Pr(y<c|x=x_i)} \\ &= \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)}{\Phi\left(\frac{c-x_i\beta}{\sigma}\right)} \end{aligned}$$

We can derive the log-likelihood, maximize it with regard to β and σ. Again, in the same way as for the interval-coded data and the censored data cases, we are able to estimate (β, σ): we have identified all the parameters of interest. But, compared to what we would have done if there had been no missing data problem, we had to make a distributional assumption. Moreover, our estimation technique is more complicated than for OLS: we are not using simple moments like $Cov(x, y)$ and $Var(x)$: we need to use maximum of likelihood.

4.2 A general model of sample selection

In general, the reason why we observe an individual or not is more complicated than in the truncated model. For example, in the case of non response to a survey, we observe an individual if he decided to respond to the interviewer; this decision can depend in a complex way of many observed and unobserved determinants. The general model of sample selection uses an auxiliary equation similar to the binary choice model:

$$s = 1 \text{ iff } s^* = z\gamma + v > 0. \tag{5}$$

The equation of interest is unchanged:

$$\begin{gathered} y = x\beta + \varepsilon \\ y \text{ observed iff } s = 1 \end{gathered} \tag{6}$$

The selection equation (5) introduces a new random component, v. We need to make statistical assumptions on v. The key assumption we make is that:

$$(x, z) \text{ is independent from } (\varepsilon, v).$$

How should we interpret this assumption?

- The assumption that x is independent from ε and that z is independent from v are the standard "exogeneity"assumptions. They imply that, if there was no sample selection issue, the selection equation and the equation of interest would not suffer from an endogeneity problem (in particular, there would be no omitted variable). We could estimate the selection equation by Probit, and the equation of interest by OLS.

- But our assumption also means that x is independent from v and that z is independent from ε. In fact, this is not restrictive. Imagine that there is one z_k such that z_k and ε are correlated. This means that we should also control for z_k in the equation of interest. We can easily do so: we need to include z_k in the x vector.

To summarize, the independence assumption implies that no relevant explanatory variable has been included in one equation and omitted from the other, and that the explanatory variables are not correlated with the error terms. These may be strong assumptions; but they allow us to focus on the sample selection issue. What we do **not** impose is that ε and v are independent. They can be correlated (and we will see that this is what creates a selectivity bias). And we have **not** imposed anything on how x and z relate. x and z can be identical (the same factors explain the selection and the outcome of interest). They can partially overlap. They can be fully distinct. We will see that how x and z overlap matters for the existence of a selectivity bias and for the identification of the general selection model.

What does the full model imply for the estimation of β? To answer this question, we first recognize that, given the sample we have, it is not useful to model $E(y|x)$ (as this is not observed), but $E(y|x, s = 1)$. We have:

$$\begin{aligned} E(y|x, z, s = 1) &= E(x\beta + \varepsilon|x, z, v > -z\gamma) \\ &= x\beta + E(\varepsilon|x, z, v > -z\gamma) \\ &= x\beta + E(\varepsilon|v > -z\gamma) \end{aligned}$$

(To go to the last line, we have used the assumption that $(\varepsilon, v) \perp (x, z)$.)

4.2.1 Selectivity bias

The second term can be viewed as the omitted variable if we were to run OLS of y on x in the observed sample. Does this create a bias? Remember that omitting a variable only creates a bias if the omitted variable is correlated with the explanatory variables. Therefore, the question is: is x likely to be correlated with $E(\varepsilon|v > -z\gamma)$?

The answer is of course specific to each application. So let us take an example.

Example 1: Sample selection due to non response

In a randomized trial, job search assistance has been randomly allocated to members of the treatment group ($T = 1$) while the control group ($T = 0$) has followed the usual track. We have already seen that the average treatment effect, β, can be estimated by writing the simple model:

$$\begin{gathered} y = \alpha + \beta T + \varepsilon \\ E(\varepsilon|T) = 0 \end{gathered}$$

If we could observe the outcome (the employment status, y) for everyone, we could estimate β by OLS. This is equivalent to computing the difference between the sample means in the two groups: $\widehat{\beta} = \overline{y}^{T=1} - \overline{y}^{T=0}$. This simplicity is due to the initial randomization: the two groups are initially comparable, so any difference in outcome can be attributed to the treatment.

Instead, assume that we need to run a phone survey to know y. A pollster calls the two groups after six months, and asks them whether they are employed or not. Response rates to such a survey are low, typically 60%. Note $R = 1$ if the person responds. We can compute the average employment rates only among the respondents of the two groups. We note them $\overline{y}^{T=1,R=1}$ and $\overline{y}^{T=0,R=1}$. We are now worried that $\widehat{\beta}_R = \overline{y}^{T=1,R=1} - \overline{y}^{T=1,R=1}$ may no longer be an unbiased estimate of β. Indeed, the initial treatment and

control groups were comparable; but what we now observe are selected subgroups within each group. Are these selected subgroups still comparable?

Again, let us write the model for which we have an empirical analog:

$$E(y|T, R = 1) = \alpha + \beta T + E(\varepsilon|R = 1).$$

We also need to model R, the response behavior. We focus on three possible determinants of response to the survey:

- T may be the first one. For instance, jobseekers who have benefited from job search assistance may be particularly happy (or unhappy) with the public employment service; that may change their response behavior.
- The mood at survey time, v_1, may be another reason for responding or not.
- The intrinsic motivation of the jobseeker to find a job, v_2, can also impact his willingness to respond.

Note that v_1 and v_2 are unobserved. You can think of other determinants; I have chosen these three determinants because they illustrate the different cases.

Given these determinants, the selection equation is:

$$R = 1 \text{ iff } R^* = \delta_0 + \delta_1 T + \gamma_1 v_1 + \gamma_2 v_2 > 0.$$

Therefore, applying the computations above, the "omitted variable" is $E(\varepsilon|\gamma_1 v_1 + \gamma_2 v_2 > -\delta_0 - \delta_1 T)$. We can now concretely ask: is this correlated with T? There are different cases:

1. **The intrinsic motivation of the jobseeker has no impact on his response behavior:** $\gamma_2 = 0$. By normalization, let us assume $\gamma_1 = 1$. The question becomes: is $E(\varepsilon|v_1 > -\delta_0 - \delta_1 T)$ correlated with T? It seems reasonable to assume that it is not. ε is the idiosyncratic ability to find a job, and v_1 is the mood at the time of the survey. We can assume that they are uncorrelated (at least once we have introduced some other controls like age,...). Therefore, $E(\varepsilon|v_1) = E(\varepsilon) = 0$. This implies that $E(\varepsilon|\gamma_1 v_1 > -\delta_0 - \delta_1 T) = 0$. 0 is clearly not correlated with T. We conclude that, in this first case, there is no omitted variable bias if we simply ignore selection and run OLS on the observed sample. The intuition is quite simple: we only observe part of the treatment and of the control groups, i.e. those in good mood at the time of the survey. But, *in terms of employment status,* these two subgroups are representative of the initial treatment and control groups. Therefore, the benefits of randomization are not lost. The *observed* control and treatment groups are comparable, and their comparison allows us to identify the impact of the treatment. More formally, lets us define the "average treatment effects among respondants" (ADR) and let us compare it to β, the average treatment effect in the population. We have:

$$\begin{aligned} ADR &\equiv E(y|T = 1, R = 1) - E(y|T = 0, R = 1) \\ &= [\alpha + \beta + E(\varepsilon|v_1 > -\delta_0 - \delta_1)] - [\alpha + E(\varepsilon|v_1 > -\delta_0)] \\ &= [\alpha + \beta] - \alpha \\ &= \beta. \end{aligned}$$

2. **Intrinsic motivation has an impact on response behavior:** $\gamma_2 = 1$. Since we know that we can ignore v_1, let us assume that $\gamma_1 = 0$. The question now is: is $E(\varepsilon|v_2 > -\delta_0 - \delta_1 T)$ correlated with T? Well, this time, we tend to believe that the ability to find a job is positively correlated with the intrinsic motivation: $E(\varepsilon|v_2)$ is not 0 for any v_2. Does it necessarily create a bias? There are two subcases:

   (a) $\delta_1 = 0$. This is the case where the treatment actually has no impact on response behavior. In that case, $E(\varepsilon|v_2 > -\delta_0 - \delta_1 T) = E(\varepsilon|v_2 > -\delta_0)$. Due to randomization, $(\varepsilon, v_2) \perp T$, so this term is *not* correlated with T. Therefore, there is no omitted variable bias: we can run OLS. This may seem strange: The observed control and treatment groups are no longer representative of the initial control and treatment groups. Instead, they are selected subgroups that tend to be highly

motivated and to more quickly find jobs. *But the selection is the same in the control and the treatment group.* As a consequence, we can look at the difference between the observed treatment and control groups: as the two groups are similarly selected, the only reason they differ is the difference of treatment. Formally, the average difference among respondents (ADR) is

$$\begin{aligned} ADR &\equiv E(y|T=1,R=1) - E(y|T=0,R=1) \\ &= [\alpha + \beta + E(\varepsilon|v_2 > -\delta_0)] - [\alpha + E(\varepsilon|v_2 > -\delta_0)] \\ &= \beta. \end{aligned}$$

(b) $\delta_1 > 0$. Let us assume that the treatment group was happy with the program and is therefore more willing to respond to the survey. Then $E(\varepsilon|v_2 > -\delta_0 - \delta_1 T)$ actually varies with T. Specifically, the average intrinsic motivation to find a job is lower in the *observed* treatment group than in the *observed* control group (the former includes everyone with $v_2 > -\delta_0 - \delta_1$, and the latter only those with $v_2 > -\delta_0$). Therefore, the difference in outcome between these two groups is due to two sources: the treatment, but also the impact of intrinsic motivation on finding a job. Formally,

$$\begin{aligned} ADR &\equiv E(y|T=1,R=1) - E(y|T=0,R=1) \\ &= [\alpha + \beta + E(\varepsilon|v_2 > -\delta_0 - \delta_1)] - [\alpha + E(\varepsilon|v_2 > -\delta_0)] \\ &= \beta + [E(\varepsilon|v_2 > -\delta_0 - \delta_1) - E(\varepsilon|v_2 > -\delta_0)]. \end{aligned}$$

$E(\varepsilon|v_2 > -\delta_0 - \delta_1) - E(\varepsilon|v_2 > -\delta_0)$ is the selection bias. It is negative. The impact of the treatment is therefore underestimated.

Extrapolating from this example, you can see the two necessary conditions under which sample selection induces a selectivity bias for OLS:

1. The unobserved determinants of the equation of interest and of the selection equation are correlated. *In cases 2a and 2b, we assumed that intrinsic motivation v_2 and the ability to find a job ε are correlated.*
2. Some of the explanatory variables of the selection equation also enter the equation of interest (z and x overlap). *In case 2b, T entered the two equations.*

Note that the first condition concerns unobservables. It cannot be tested easily. This is a matter of judgement on the case at hand: sometimes, you will have reasons to believe that selection can be neglected, sometimes you won't (this is called *incidental selection).* We therefore now turn to models that are robust to incidental selection. It turns out that they are also useful to test *ex post* whether selection is or not an issue.

4.3 The Heckit model

Heckman (1979) proposes a simple approach to estimating the general sample selection model when we suspect incidental sample selection. The basic insight is to base the estimation on the "complete" CEF

$$E(y|x,z,s=1) = x\beta + E(\varepsilon|v > -z\gamma). \tag{7}$$

This requires to have an estimate of $E(\varepsilon|v > -z\gamma)$. Heckman assumes that ε and v follow a bivariate normal distribution and have a coefficient of correlation ρ:

$$(\varepsilon, v) \sim N(0, 0, \sigma_\varepsilon, 1, \rho).$$

Note that a normalization is imposed: $\sigma_v = 1$. We have already seen this: v is the error term in the binary choice equation (2). We saw in lecture 1 that its variance is not identified.

Consider the auxiliary regression of ε on v:

$$\varepsilon = \beta_0 + \beta_1 v + \eta.$$

From the bivariate normal distribution assumption, it turns out that $E(\eta|v) = 0$ (and not only $E(\eta v) = 0$). From the usual formulas, we also have

$$\begin{aligned}\beta_1 &= \frac{Cov(\varepsilon, v)}{\sigma_v^2} = \frac{\rho\sigma_\varepsilon\sigma_v}{\sigma_v^2} = \rho\sigma_\varepsilon \text{ and} \\ \beta_0 &= E(\varepsilon) - \beta_1 E(v) = 0.\end{aligned}$$

Therefore, the "omitted variable" $E(\varepsilon|v > -z\gamma)$ has a simple expression:

$$\begin{aligned}E(\varepsilon|v > -z\gamma) &= E(\rho\sigma_\varepsilon v|v > -z\gamma) + E(\eta|v > -z\gamma) \\ &= \rho\sigma_\varepsilon E(v|v > -z\gamma).\end{aligned}$$

You should recognize the term we have already encountered in the truncated sample case (equation 4): if $\rho \neq 0$, the selection in (5) creates a truncature problem in equation (6). In the terms of the previous example, selecting only those who responded to the survey, if they are the most motivated to find a job, amounts to truncating the sample in the employment equation – by keeping only those with the highest unobserved propensity to find a job.

Note that if $\rho = 0$, there is no selectivity issue: $E(\varepsilon|v > -z\gamma) = 0$. This is case 1 in the previous example.

We now need to compute an estimate of the mean of a truncated normal distribution ($E(v|v > -z\gamma)$). We have already given the formula for corner solution models, but let us derive it formally here. The density of a truncated normal distribution truncated above c is $\frac{\phi(v)}{1-\Phi(c)}$. Therefore, its mean is:

$$\begin{aligned}E(v|v > s) &= \int_c^\infty \frac{v\phi(v)}{1-\Phi(c)} dv \\ &= \frac{1}{1-\Phi(c)} \int_c^\infty \frac{v \exp(-v^2/2)}{\sqrt{2\pi}} dv \\ &= -\frac{1}{1-\Phi(c)} \left[\frac{\exp(-v^2)}{\sqrt{2\pi}}\right]_c^\infty \\ &= \frac{\phi(c)}{1-\Phi(c)} \\ &= \frac{\phi(-c)}{\Phi(-c)} \\ &= \lambda(-c)\end{aligned}$$

(where $\lambda(.) = \frac{\phi(.)}{\Phi(.)}$ is the inverse Mills ratio). Replacing c by $-z\gamma$, we get

$$E(v|v > -z\gamma) = \lambda(z\gamma) = \frac{\phi(z\gamma)}{\Phi(z\gamma)}.$$

Finally, we can rewrite the CEF in (7) as

$$E(y|x, z, s = 1) = x\beta + \rho\sigma_\varepsilon \frac{\phi(z\gamma)}{\Phi(z\gamma)}.$$

We can therefore write the following model for the selected sample:

$$\begin{gathered}y = x\beta + \rho\sigma_\varepsilon \frac{\phi(z\gamma)}{\Phi(z\gamma)} + \eta \\ \text{with } E(\eta|x, z) = 0.\end{gathered}$$

This is a perfectly "well-behaved" linear regression model, with explanatory variables x and $\frac{\phi(z\gamma)}{\Phi(z\gamma)}$. We just need values for $\frac{\phi(z\gamma)}{\Phi(z\gamma)}$. Heckman (1979) proposes the following procedure (also known as the **"Heckit" procedure**):

1. Estimate the selection equation (5) by probit.

2. Use the estimates $\widehat{\gamma}$ to form an estimate of the inverse Mills ratio for each observation in the selected sample:

$$\widehat{\lambda}_i = \frac{\phi(z_i\widehat{\gamma})}{\Phi(z_i\widehat{\gamma})},$$

and regress y on x and $\widehat{\lambda}$ to get estimates of β and $\rho\sigma_\varepsilon$.

Note 1: If we apply this method sequentially, we cannot directly use the standard errors obtained in the second step. Indeed, given the assumptions made, one can show that η is necessarily heteroskedastic. Moreover, the usual standard errors do not take into account the fact that $\widehat{\lambda}$ has been estimated (introducing some noise). This is a usual problem with two-step estimators: you have already encountered it with 2SLS. In practice, we therefore do not implement the two steps "manually" but rather use built-in functions in econometric packages that provide correct standard errors.

Note 2: It is also possible to estimate σ_ε and ρ separately. ρ is a particularly interesting parameter, as it says whether selection is incidental or not. Remember that if $\rho = 0$, selection is not incidental. The t-stat of ρ therefore allows us to test the assumption H0 that selection is not incidental.

Note 3: As we have written a fully parametric model (since we have fully specified the distribution of (ε, v)), it is possible to use CMLE. This is in principle more efficient, but less used in practice.

Example 2: Heckit estimation of a wage equation

Estimating a wage equation is a standard application of the Heckit model. Sample selection naturally occurs in that context, as we only observe wages for a selected sample of the population, those who have a job. Simple labor supply theory tells us that those who offer to work are those who can get a wage that is above their reservation wage. Therefore, the selection process seems to be directly connected to the wage. Selection is likely to be incidental.

A standard statistical[3] model for women's labor supply and wage is:

$$s = 1 \text{ iff } s^* = \gamma_0 + \gamma_1 edu + \gamma_2 age + \gamma_3 kids + v > 0. \tag{8}$$

and

$$\begin{gathered} \ln w = \beta_0 + \beta_1 edu + \beta_2 age + \varepsilon \\ w \text{ observed iff } s = 1 \end{gathered} \tag{9}$$

The "augmented" equation of interest is

$$\ln w = \beta_0 + \beta_1 edu + \beta_2 age + \rho\sigma_\varepsilon \widehat{\lambda} + \eta. \tag{10}$$

Stata's "heckman" command provides us with the probit estimates of the selection equation as well as the estimates of the (augmented) equation of interest (10), with the appropriate standard errors (figure 3).

For the sake of comparison, figure 4 displays the output when we simply run OLS on equation (9) without accounting for sample selection. See the impact on the schooling coefficient. Does the difference between the heckit and the OLS coefficients go in the direction you expected? Is the difference statistically significant?

4.4 An important caveat of sample selectivity correction models: the need for instruments

So far, everything went very well: we have been able to correct for sample selectivity in a rather simple way. But the question is: is this robust? Under which conditions have we been able to separate the effects of sample selectivity from the effects we were interested in? You may recognize this as the question of identification. Identification actually is a key concern in models with sample selectivity correction.

[3] This statistical model can actually be derived from a theoretical model (as we did in lecture 1), but, here, we focus on the econometric part.

```
heckman lnW edu age, select(edu age kids)

Heckman selection model                         Number of obs      =     21228
(regression model with sample selection)        Censored obs       =      8871
                                                Uncensored obs     =     12357

                                                Wald chi2(2)       =   1740.09
Log likelihood = -23920.84                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnW          |
         edu |   .0718501   .0018367    39.12   0.000     .0682503    .0754498
         age |   .0193414   .0009323    20.75   0.000     .0175141    .0211686
       _cons |   7.047827   .0481962   146.23   0.000     6.953365     7.14229
-------------+----------------------------------------------------------------
select       |
         edu |   .0518022   .0030606    16.93   0.000     .0458036    .0578008
         age |  -.0020582   .0016668    -1.23   0.217    -.0053251    .0012087
        kids |  -.4695965   .0221245   -21.23   0.000    -.5129596   -.4262333
       _cons |  -.3471823   .0720918    -4.82   0.000    -.4884796    -.205885
-------------+----------------------------------------------------------------
     /athrho |   .0912199   .0457498     1.99   0.046     .0015519    .1808878
    /lnsigma |  -.6187248   .0068088   -90.87   0.000    -.6320697   -.6053798
-------------+----------------------------------------------------------------
         rho |   .0909677   .0453712                      .0015519    .1789404
       sigma |   .5386309   .0036674                      .5314906     .545867
      lambda |    .048998   .0245594                      .0008626    .0971335
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     3.11   Prob > chi2 = 0.0780
------------------------------------------------------------------------------
```

Figure 3: Heckit

```
reg lnW edu age

      Source |       SS       df       MS              Number of obs =   12357
-------------+------------------------------           F(  2, 12354) =  939.57
       Model |   542.76002     2   271.38001           Prob > F      =  0.0000
    Residual |  3568.25184 12354  .288833725           R-squared     =  0.1320
-------------+------------------------------           Adj R-squared =  0.1319
       Total |  4111.01186 12356  .332713812           Root MSE      =  .53743

------------------------------------------------------------------------------
         lnW |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0705477   .0017138    41.16   0.000     .0671883    .0739071
         age |   .0193567   .0009312    20.79   0.000     .0175313     .021182
       _cons |   7.097623   .0411704   172.40   0.000     7.016923    7.178323
------------------------------------------------------------------------------
```

Figure 4: OLS

The problem can be seen from the "augmented" equation

$$y = x\beta + \rho\sigma_\varepsilon \frac{\phi(z\gamma)}{\Phi(z\gamma)} + \eta.$$

Assume that all the variables that have an impact on selection also have an impact on y, that is, $z = x$. In that case, we have

$$y = x\beta + \rho\sigma_\varepsilon \frac{\phi(x\gamma)}{\Phi(x\gamma)} + \eta.$$

Each component of x has two effects on y in the selected sample: one is the direct, causal impact (with coefficient β), and the other is the indirect effect through selection (through coefficient γ, the inverse Mills ratio function, and $\rho\sigma_\varepsilon$). What makes it possible to distinguish the two effects empirically is that we have **assumed** that the direct effect is linear in β and that the selection effect takes place through a very specific, bivariate normal process. But we have made these assumptions for convenience: nothing in the data tells us that they are true. Assume, by contrast, that the direct effect of x on y is non linear, and takes the form $h(x\beta)$. $h(x\beta)$ may very well be collinear to $\frac{\phi(x\gamma)}{\Phi(x\gamma)}$. In that case, there would be no way to empirically distinguish between these two effects.

This is a very important finding: **if all the variables that appear in the selection equation are also included in the equation of interest, then the model is only identified through arbitrary parametric assumptions.** This is bad: we do not want our results to be driven by the arbitrary assumptions we make, but by what the data can say.

A more positive result is the following: **if there is at least one variable (an instrument) that appears in the selection equation but has no direct impact in the equation of interest, then the model is identified without making strong parametric assumption.** The conclusion is that using the heckit model makes only sense if we can reasonably make the assumption that we have such an instrument.

You can get the intuition of this results from writing the conditional expectation in the selected sample as

$$\begin{aligned} E(y|x,z,s=1) &= h(x\beta) + E(\varepsilon|v > -z\gamma) \\ &= h(x\beta) + g(z\gamma) \\ &= h(x\beta) + g(x\gamma_x + \gamma_w w) \end{aligned}$$

where $h(.)$ and $g(.)$ are unknown functions and w is an instrument. From looking at $\frac{\partial E(y|x,z,s=1)}{\partial w} = \gamma_w g'(x\gamma_x + \gamma_w w)$, we can identify the slope of the $g(.)$ function. Indeed, we have estimates of γ_w and γ_x from estimating the selection equation, and we observe the empirical analog to $\frac{\partial E(y|x,z,s=1)}{\partial w}$; we can therefore infer g'. Once we have the form of $g(.)$, we can derive the form of $h(.)$ from the data.

The search for valid instruments is the key in applying selection models. In our application to women's labor supply, we used the number of kids as an instrument. This requires the "exclusion restriction" that the number of kids is not correlated with productivity or with any other determinant of wages. This may not be true if women with kids have to stay home when their kids are sick, etc. In some contexts, there are more natural instruments; for instance, if, due to cost constraints, the phone survey was run only once, each person being called once at a random hour, then, the hour of the call might be a determinant of response and is probably not correlated with the outcome itself. You can use this variable as an instrument.

To summarize, sample selectivity correction requires instruments, that is variables that explain the selection, but have no direct impact on the outcome. The validity of instruments cannot be proven statistically, and needs to be discussed in each situation. What you have learned for IVs also applies for selection models: you need to be creative, to know the setting very well (and to be somewhat lucky!) in order to find credible instruments.