ECONOMETRICS II, Fall 2022.

Bent E. Sørensen

1 Truncation, Censoring, and Selectivity

1.1 Truncation

Consider the case of a regression model with a truncated sample. We assume

$$y_i = X_i\beta + u_i\,,$$

where u_i is normally distributed with variance σ and the "OLS-assumptions" are satisfied. Data with $y_i > K$ are discarded for some number K (which is often normalized to 0 in textbooks), and this is called *truncation*. You have to memorize this.

If the data are truncated, OLS is biased. $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'u$ and $E(\hat{\beta} - \beta) = (X'X)^{-1}X'Eu$. With no truncation, $Eu = 0$ but with truncation, we have $E(u_i|y_i < K)$=$E(u_i|u_i < K - X_i\beta) = -\sigma\frac{\phi((K-X_i\beta)/\sigma)}{\Phi((K-X_i\beta)/\sigma)}$. (Note: Let us take the results about the mean of a truncated normal as given at this moment, but it is not hard to show and we may do it at some stage.) You can

see that this term is different from zero. You can also see that if $X_i\beta$ is negative and numerically very large for all observations, the bias is almost 0. Why?

The probability that an observation from a distribution with density f is in a small interval of length Δu around u_i is $f(u_i)\Delta u$. (Strictly speaking it would be $\int_{u_i-\frac{\Delta u}{2}}^{u_i+\frac{\Delta u}{2}} f(z)dz$). While densities are not probabilities, it is much easier to use the shorthand of talking about the probability of y_i or u_i. So, because we only have a truncated sample, the probability of observing y_i in the truncated sample is the unconditional probability divided by the probability that $y < K$ as an application of $P(A|B) = P(A \bigcap B)/P(B)$. Here, A is $([y_i - \Delta u/2, y_i + \Delta u/2]$ and the probability of A is $f(u_i)\Delta u = f(y_i - X_i\beta)\Delta u$ when $y_i < K$ and B here is the set $y_i < K$ and the density is zero outside the set B. For f being the normal density (with ϕ denoting the standard normal) we have the probability in the numerator being $\frac{1}{\sigma}\phi(\frac{y_i - X_i\beta}{\sigma})$. We have that the probability in the denominator is $P(y_i < K) = P(X_i\beta + u_i < K) = \Phi(\frac{K - X_i\beta}{\sigma})$. In total we have the truncated density (the limit of Δy going to zero) for observation i:

$$\phi(\frac{y_i - X_i\beta}{\sigma})/(\sigma\Phi(\frac{K - X_i\beta}{\sigma}))\,.$$

As you can convince yourself, this is a density (positive and integrating to unity).

The log likelihood function (ignoring the π term) is

$$\Sigma_{i=1}^{N} - 0.5 \log \sigma^2 - 0.5 \frac{(y_i - X_i\beta)^2}{\sigma^2} - \log \Phi(\frac{K - X_i\beta}{\sigma}) \; .$$

You can think of this as a normal likelihood function with a correction term. Notice that if $X_i\beta$ is very small (large negative number) then the last term is about 0 (the Φ term becomes unity and then the log makes it 0). The logic is that if $X_i\beta$ is small, then y_i is likely small and the chance of observation i being truncated is almost nil so no there is no need to adjust. This is of course what we said above about the bias except when we talk about the likelihood function, this is observation is valid only if we are looking at β values near the true one. If all the $X_i\beta$ terms are very small the last term is always tiny and can be ignored.

1.2 Censoring

Consider the case of a regression model with censored observations. We assume

$$y_i^0 = X_i\beta + u_i \, ,$$

where u_i is normally distributed with variance σ and the "OLS-assumptions" are satisfied. Data with $y_i^0 > K$ are transformed to $y_i = K$. This is called *censoring.* Spend a few minutes on memorizing the this and the difference between truncation

and censoring, it is not smart to lose points on the exam by not doing that.

The probability that an observation is in a small interval of length $\Delta y = \Delta u$ around y_i is $f(y_i)\Delta y$ and is NOT conditional because we observe if the data are censured. $f(u_i)\Delta u = f(y_i - X_i\beta)\Delta u$ is the probability of being in the Δy interval when $y < K$. The only other value y can take is $y_i = K$ and the probability of this is $P(y_i^0 > K) = 1 - \Phi(\frac{K - X_i\beta}{\sigma})$.

The log-likelihood function is therefore

$$\Sigma_{i=1}^{N} I\{y_i < K\} * [-0.5 \log \sigma^2 - 0.5 \frac{(y_i - X_i\beta)^2}{\sigma^2}] + I\{y_i = K\} * \log\left(1 - \Phi(\frac{K - X_i\beta}{\sigma})\right).$$

Davidson and MacKinnon point out that you can add and subtract $\log \Phi(\frac{K - X_i\beta}{\sigma})$, in which case the likelihood has the form of a sum of a truncated likelihood and a Probit likelihood (with σ identified from the first part). Conceptually this is writing the first part as $P(A) = P(A|B)P(B)$ where A here is the probability of falling in a small interval around y_i and B is the event $y_i < K$. This is not important and may instead be confusing.

1.3 Selection

The general normal selection model is one where y is observed based on some outcome z which we model as a Probit. There are a huge number of applications of this. Say, y is the GPA of a student at, say, Rice, and z is literally the probability of getting selected (admitted) [ignore that students may decline]. U.S. college admission typically depend on a large number of variables such as which state you came from, whether your parents are alumni, and on and on. Assuming you have a sample of students, you might have data for many of these variables but not others. For example, you would likely not observe the quality of the student's essay and this would go into the error term in the admissions equation. If the quality of the students essay also is correlated with the students performance, you would have more efficient inference taking into account that now the errors in the GPA equation is correlated with the error in the selection equation. More importantly, you will get bias if you do not control for this. This section is a little more technical, we will show derivations, but it is more important that you understand intuitively what is going on.

Assume that you are interested in a relation

$$y_i^0 = X_i\beta + u_i \,,$$

which could be wages after participating in a training program. The participation is determined by a probit model, with underlying latent process

$$z_i^0 = W_i\gamma + v_i \,.$$

We assume the error terms are normal and independent across individuals (or whatever the i index stands for). We observe

$$z_i = 1 \;\text{ if }\; z_i^0 > 0; \;\; 0 \text{ otherwise}$$

and

$$y_i = y_i^0 \;\text{ if }\; z_i = 1 \,.$$

If individual i is not selected, we do not observe y_i. The issue here is that if the errors in the selection equation are correlated with the errors in the results from the regression equation are biased. This is sometimes serious bias. (Some people have argued that there is hardly any effect on wages of going the Harvard Business School [I think it was] even though it correlates very strongly with income because the selection criteria correlates so highly with earnings ability.

Denote the variance of u_i by σ^2. The variance of v_i is (as usual for a Probit model) not identified and it is normalized to 1. The covariance of two random variables can always we written as the correlation times the standard deviations of the variables. Here, where one variance is unity, it is convenient to label the correlation ρ and the covariance is then $\rho\,\sigma$.

We want to study the distribution of y_i conditional on $z_i = 1$. We therefore study the conditional distribution of u_i conditional on $z_i = 1$. This is somewhat difficult because $z_i = 1$ is a set of v_i' s. The trick therefore is to use the identity

$$P(A|B) = P(B|A)P(A)/P(B)\,.$$

In our application, we write $f(u_i|z_i = 1)$ as $\frac{P(z_i=1|u_i)f(u_i)}{P(z_i=1)}$ because we can easily find the three terms involved. $f(u_i)$ is just the normal density and $P(z_i = 1)$ is a Probit probability. That conditional term involves z_i which is a function of z_i^0 and we know how to find conditional normals (to remind ourselves, if for any (x, y) we know $Ex, Ey, \sigma_y^2, \sigma_x^2$, and cov_{xy}, then $E(y|x) = Ey + \frac{cov_{xy}}{\sigma_x^2}(x - Ex)$ and

$var(y|x) = \sigma_y^2 - cov_{xy}^2/\sigma_x^2$. Let us take stock

$$\begin{aligned} Ez_i^0 &= W_i\gamma && (1)\\ Ey_i^0 &= X_i\beta && (2)\\ var(z_i^0) &= 1 && (3)\\ var(y_i^0) &= var(u_i) = \sigma^2 && (4)\\ cov(x^0, y^0) &= \rho\sigma && (5) \end{aligned}$$

The mean of z_i^0 conditional on u_i is $W_i\gamma + \frac{\rho\sigma}{\sigma^2}(u_i - 0) = W_i\gamma + \frac{\rho}{\sigma}u_i$, by the usual formula for normal conditionals, and the conditional variance is $1 - \frac{(\rho\sigma)^2}{\sigma^2} = 1 - \rho^2$. So the hard work is to find

$$P(z_i = 1|u_i) = P(z_i^0 > 0|u_i).$$

Now, note that the conditional distribution of z_i^0 can be represented as a random variable ω_i that satisfies $\omega_i \sim N(W_i\gamma + \frac{\rho}{\sigma}u_i, 1 - \rho^2)$. So then we need to find

$$P(\omega_i > 0)$$

As always, we do that by subtracting the mean and dividing by the standard deviation (on both sides) to get to a standard normal:

$$P(\frac{\omega_i - W_i\gamma + \frac{\rho}{\sigma}u_i}{\sqrt{1-\rho^2}} > -\frac{W_i\gamma + \frac{\rho}{\sigma}u_i}{\sqrt{1-\rho^2}}) = \Phi(\frac{W_i\gamma + \frac{\rho}{\sigma}u_i}{\sqrt{1-\rho^2}})$$

or in term of observable variables

$$P(z_i = 1|y_i) = \Phi(\frac{W_i\gamma + \frac{\rho}{\sigma}(y_i - X_i\beta)}{\sqrt{1-\rho^2}}) \,.$$

Now we can write the density for y_i conditional on $z_i = 1$ (and X_i) as

$$f(y_i|z_i = 1) = \Phi(\frac{W_i\gamma + \frac{\rho}{\sigma}(y_i - X_i\beta)}{\sqrt{1-\rho^2}}) * \frac{1}{\sigma}\phi(\frac{y_i - X_i\beta}{\sigma})/\Phi(W_i\gamma) \,.$$

The full likelihood $P(y_i, z_i = 1)$ is the conditional probability $f(u_i|z_i = 1)$ times the probability $P(z_i = 1)$. The contribution of individual i to the likelihood is this joint probability if $Z_i = 1$ plus the probability $z_i = 0$ (where no y is observed); i.e.:

$$I(z_i = 1) * [\Phi(\frac{W_i\gamma + \frac{\rho}{\sigma}(y_i - X_i\beta)}{\sqrt{1-\rho^2}}) * \frac{1}{\sigma}\phi(\frac{y_i - X_i\beta}{\sigma})] + I(z_i = 0) * [1 - \Phi(W_i\gamma)] \,.$$

Finally, we have the the log-likelihood is after re-ordering a bit:

$$\Sigma_{i=1}^{N} I(z_i = 1) * [-0.5 \log \sigma^2 - 0.5\frac{(y_i - X_i\beta)^2}{\sigma^2} + \log \Phi(\frac{W_i\gamma + \frac{\rho}{\sigma}(y_i - X_i\beta)}{\sqrt{1-\rho^2}})] + I(z_i = 0) * \log \Phi(-W_i\gamma) \,.$$

Notice what happens if $\rho = 0$: you have a Probit model and an independent normal which you can estimate by least squares. True, it is strange that y is only observed when $z = 1$, but you do not have to adjust the least squares estimation in the case where the error term in the selection equation is not affecting the error term in the regression.

1.3.1 Heckman correction term for selection

Heckman was the first to consider correction for selection (in his thesis, I think) and this was the basis for his later Nobel prize.

The Heckman correction involves as two-step estimator. Assume you first estimate the Probit equation and then the regression (not recommended—it is always most efficient to estimate the full system, but sometimes we do anyway, at least in a first exploration and earlier it may have been hard numerically to estimate the full system).

Consider the regression

$$y_i = X_i\beta + u_i\,,$$

where you ignore that y_i has been selected based on z. The problem now is that $Eu_i = 0$ if you observed all outcomes (including the ones that were not selected) but if $E(u_i|v_i) = \rho\sigma v_i$, which is easy to see using the standard formula for conditional normals, we are allowed to write u_i as $\rho\sigma v_i + e_i$ where $e_i = u_i - \rho\sigma v_i$ is independent

of v_i. We have

$$y_i = X_i\beta + \rho\sigma v_i + e_i\,,$$

where e_i is independent of X_i but because v_i is instrumental in deciding whether y_i was observed, it is unlikely to have mean zero. In fact, $E(v_i|z_i = 1) = E(v_i|W_i\gamma + v_i > 0) = \frac{\phi(W_i\gamma)}{\Phi(W_i\gamma)}$, where ratio is called the inverse Mill's Ratio.[1] If you have estimated the first state you have an estimate $\hat{\gamma}$ and you run the regression

$$y_i = X_i\beta + \kappa\,\frac{\phi(W_i\hat{\gamma})}{\Phi(W_i\hat{\gamma})} + e_i\,,$$

which is a consistent estimator of β (and approximately unbiased if γ is well estimated). Usually, economists do not attempt to extract the parameters ρ and σ.

[1]To derive the inverse Mill's ratio notice that $\int x\exp(-x^2/2)dx = \int \exp(-x^2/2)(xdx)$ and do a change of variables to $y = \frac{x^2}{2}$ with $dy = xdx$.