Lecture 4 – Censoring and truncation

Introduction In this lecture, we consider the case where the dependent variable has limited range, in the sense that it is partially observable. Consider a very special situation where a variable of interest, denoted as Y, is partially observable from Y^*, in the following sense:

$$Y = \begin{cases} Y^* & if\ Y^* < b \\ b & if\ Y^* \geq b \end{cases},$$

where $b > 0$. This is a possible outcome of what is called selective sampling. More specifically, this is what is called top-coding.

Question: Can you name situations for which top-coding is likely to occur?

If one has data on Y, how can we estimate the population mean of Y^* knowing that there is top-coding? A possible suggestion would be to use a sample average which includes all observations, including the top-coded ones. Is this a good idea? Let us set up the problem.

Top-coding and its practical implications Consider an i.i.d. sample of (Y_i, D_i) for $i = 1, \ldots, n$, where D_i is an indicator which is equal to 1 when $Y_i^* < b$. We know from the law of large numbers that the sample average of the Y_i's converges to its expectation, denoted by $\mathbb{E}(Y_i)$. But by the law of total expectation,

$$\begin{aligned} \mathbb{E}(Y_i) &= \mathbb{E}(Y_i|D_i = 1)\Pr(D_i = 1) + \mathbb{E}(Y_i|D_i = 0)\Pr(D_i = 0) \qquad (1) \\ &= \frac{\left[\int_{-\infty}^{b} y f(y)\, dy\right]}{\Pr(D_i = 1)}\Pr(D_i = 1) + b(1 - F(b)) \\ &= \left[\int_{-\infty}^{b} y f(y)\, dy\right] + b(1 - F(b)) \\ &= \int_{-\infty}^{b} y f(y)\, dy + \int_{b}^{+\infty} y f(y)\, dy - \int_{b}^{+\infty} y f(y)\, dy + b(1 - F(b)) \\ &= \int_{-\infty}^{+\infty} y f(y)\, dy - \int_{b}^{+\infty} y f(y)\, dy + b(1 - F(b)) \\ &= \mathbb{E}\left(Y_i^*\right) + \left[+b(1 - F(b)) - \int_{b}^{+\infty} y f(y)\, dy\right]. \end{aligned}$$

To make more progress with respect to the term in square brackets, we use integration by parts on the following integral

$$\begin{aligned} \int_{b}^{+\infty} (1 - F(y))\, dy &= y(1 - F(y))\Big|_{b}^{+\infty} + \int_{b}^{+\infty} y f(y)\, dy \\ &= \lim_{y \to +\infty} y(1 - F(y)) - b(1 - F(b)) + \int_{b}^{+\infty} y f(y)\, dy \\ &= -b(1 - F(b)) + \int_{b}^{+\infty} y f(y)\, dy. \end{aligned}$$

The last equality used a result where

$$\lim_{y \to +\infty} y(1 - F(y)) = 0.$$

Unfortunately, this result does not follow from L'Hôpital's rule. It uses some other trick along with the squeeze theorem.[1] Putting all of this together gives us

$$\mathbb{E}(Y_i) = \mathbb{E}\left(Y_i^*\right) - \int_b^{+\infty} (1 - F(y))\, dy.$$

Question: What does the last line imply about using the sample average of the Y_i's to estimate the mean of Y^*?

Question: Reconsider the situation of hours worked in the light of the preceding discussion. Repeat the arguments above for this case.

Censoring versus truncation The situation considered in the top-coding example is called fixed censoring at point b. We can rewrite the model as $Y = \min(Y^*, b)$ and Y is called a right-censored version of Y^*. One can also think of Y^* as a latent variable, just as in the binary choice situation.

It is possible to have censoring both to the left and to the right. It is also possible to allow for random censoring and even allow b to vary across i.

Question: Think about the last possibilities in the context of the hours worked example.

If it happens that we only get to observe only the values of Y_i for which $D_i = 1$, then we have what is called truncation. So the crucial difference between censoring and truncation is whether or not we get to observe only the values that pass the condition $D_i = 1$.

OLS under truncation Assume a linear regression model for the latent variable Y_i^*, i.e., $Y_i^* = X_i'\beta + \varepsilon_i$. Assume zero conditional mean for ε_i. Consider an i.i.d. sample of (Y_i, X_i, D_i) for $i = 1, \ldots, n$, where D_i is an indicator which is equal to 1 when $Y_i^* > 0$. Under truncation, we only get to observe units for which $D_i = 1$. The OLS estimator is given by

$$\hat{\beta} = \left(\sum_i X_i' X_i\right)^{-1} \sum_i X_i' Y_i.$$

Calculating the expectation of $\hat{\beta}$ gives us

$$\mathbb{E}\left(\hat{\beta}|X\right) = \left(\sum_i X_i' X_i\right)^{-1} \sum_i X_i' \mathbb{E}(Y_i | D_i = 1, X_i).$$

Let us now calculate $\mathbb{E}(Y_i|D_i = 1, X_i)$:

$$\mathbb{E}(Y_i|D_i = 1, X_i) = \mathbb{E}\left(Y_i^*|Y_i^* > 0, X_i\right) = \mathbb{E}\left(X_i'\beta + \varepsilon_i | \varepsilon_i > -X_i'\beta, X_i\right) = X_i'\beta + \mathbb{E}\left(\varepsilon_i | \varepsilon_i > -X_i'\beta, X_i\right). \quad (2)$$

As a result,

$$\mathbb{E}\left(\hat{\beta}|X\right) = \beta + \left(\sum_i X_i' X_i\right)^{-1} \sum_i X_i' \mathbb{E}\left(\varepsilon_i | \varepsilon_i > -X_i'\beta, X_i\right).$$

Question: What does the preceding result tell you about the bias of the OLS estimator under truncation? Can you say something about the direction of the bias? Finally, try sketching a picture for the simple regression case.

[1] Apparently, actuaries call this result the Darth Vader rule. Don't ask.

There are specific results that are obtained if we make additional assumptions about the distribution of ε_i, say normality. A particularly useful result for truncated normal random variables is as follows: Let $Y^* \sim N\left(\mu, \sigma^2\right)$ and suppose that we truncate Y^* to lie on the interval (a, b). Let this new variable be denoted by Y. Then, we have

$$\mathbb{E}(Y) = \mu - \sigma \frac{\phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}.$$

The preceding result can be used to find an expression for $\mathbb{E}\left(\varepsilon_i | \varepsilon_i > -X_i'\beta, X_i\right)$. If we assume normality for the ε_i, i.e., $\varepsilon_i | X_i \sim N\left(0, \sigma^2\right)$, we have

$$\mathbb{E}\left(\varepsilon_i | \varepsilon_i > -X_i'\beta, X_i\right) = \frac{\sigma\phi\left(\frac{-X_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{-X_i'\beta}{\sigma}\right)} = \frac{\sigma\phi\left(\frac{X_i'\beta}{\sigma}\right)}{\Phi\left(\frac{X_i'\beta}{\sigma}\right)}.$$

It turns out that $\left|\mathbb{E}\left(\hat{\beta}|X\right)\right| < |\beta|$ under normality. This is called attenuation bias. Clearly, the attenuation bias is larger when the truncation is much more severe.

Question: Is there a way to repair OLS under normality and restore unbiasedness?

OLS under censoring Let us rework the details of the previous section for the OLS estimator under fixed censoring at zero. We have two options:

1. Apply OLS to the entire sample. In this case, we have

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}|X\right) &= \left(\sum_i X_i'X_i\right)^{-1} \sum_i X_i' \mathbb{E}(Y_i|X_i) \\
&\overset{(1)}{=} \left(\sum_i X_i'X_i\right)^{-1} \sum_i X_i' \left[\mathbb{E}(Y_i|D_i = 1, X_i)\Pr(D_i = 1|X_i) + \mathbb{E}(Y_i|D_i = 0, X_i)\Pr(D_i = 0|X_i)\right] \\
&= \left(\sum_i X_i'X_i\right)^{-1} \sum_i X_i' \mathbb{E}(Y_i|D_i = 1, X_i)\Pr(D_i = 1|X_i) \\
&\overset{(2)}{=} \left(\sum_i X_i'X_i\right)^{-1} \sum_i X_i' \left[X_i'\beta + \mathbb{E}\left(\varepsilon_i|\varepsilon_i > -X_i'\beta, X_i\right)\right]\Pr(D_i = 1|X_i).
\end{aligned}$$

   Question: The expression is a bit unwieldy but it shows you that $\mathbb{E}\left(\hat{\beta}|X\right)$ is nonlinear in β. Is the OLS estimator unbiased?

2. Apply OLS only to the units for which $D_i = 1$. The steps are exactly the same as the case for OLS under truncation, i.e.

$$\mathbb{E}\left(\hat{\beta}|X\right) = \beta + \left(\sum_i X_i'X_i\right)^{-1} \sum_i X_i' \mathbb{E}\left(\varepsilon_i|\varepsilon_i > -X_i'\beta, X_i\right).$$

   Question: Is there a way to repair OLS under normality and restore unbiasedness?

Maximum likelihood estimation The first step is to set up the likelihood function. We need the density of the data. Note that the data depends on whether there is censoring or truncation.

1. When we have a truncated sample, the likelihood function is given by

$$L(\theta) = \prod_{i:D_i=1} f(Y_i|D_i=1,X_i).$$

2. When we have a censored sample, the likelihood function is given by

$$\begin{aligned} L(\theta) &= \prod_{i:D_i=1} f(Y_i|D_i=1,X_i) \prod_{i:D_i=0} \Pr\left(Y_i^* \leq 0|D_i=0,X_i\right) \\ &= \prod_{i:D_i=1} f(Y_i|D_i=1,X_i) \prod_{i:D_i=0} \Pr(D_i=0|X_i). \end{aligned}$$

Question: What do you notice is the difference between the likelihood (log-likelihood) functions under censoring and truncation?

If we make an assumption of normality, $\varepsilon_i|X_i \sim N\left(0,\sigma^2\right)$, just as we did earlier, then it is possible write down the log-likelihood a little bit more explicitly. Here, we will say much more about the censored regression model. The likelihood function is given by:

$$\begin{aligned} \log L\left(\beta,\sigma^2\right) &= \sum_{i:D_i=1} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2\right] + \sum_{i:D_i=0} \log \Pr\left(\varepsilon_i \leq -X_i'\beta|X_i\right) \\ &= \sum_{i:D_i=1} \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2\right] + \sum_{i:D_i=0} \log \Pr\left(\frac{\varepsilon_i}{\sigma} \leq -\frac{X_i'\beta}{\sigma}\middle|X_i\right) \\ &= \sum_{i:D_i=1} \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2\right] + \sum_{i:D_i=0} \log \Phi\left(-\frac{X_i'\beta}{\sigma}\right). \end{aligned}$$

Note that you have access to the whole ML toolkit at this stage.

Question: Try deriving the score and establish a connection with the discussion regarding OLS estimation.The asymptotic variance requires us to derive the second derivative of the log-likelihood. Try deriving this as well.

Question: Revisit the top-coding example. Set up the log-likelihood for the mean of the latent variable Y_i^*, denoted as $\mu = \mathbb{E}\left(Y_i^*\right)$. It may help to assume that $Y_i^* \sim N(\mu,1)$. Notice that this is a more realistic version of Example 1 in Freedman's book.

Unfortunately, since ML relies on pre-specifying the density of the data, all desirable properties of ML may no longer hold once we make a mistake in specifying this density. As mentioned in our discussion of binary choice models, there are alternatives that are semiparametric in nature. These semiparametric alternatives are beyond the scope of the course.

Marginal effects at an "average" and average marginal effects Recall that calculating marginal effects has to start from the conditional expectation of Y_i given X_i. We will work out the censored regression case and leave the truncated regression case as an exercise. Assume once again the case where we have left-censoring at zero. Recall that (compare with (1)):

$$\mathbb{E}(Y_i|X_i) = \mathbb{E}(Y_i|D_i=1,X_i)\Pr(D_i=1|X_i) + \mathbb{E}(Y_i|D_i=0,X_i)\Pr(D_i=0|X_i).$$

Taking derivatives with respect to the kth regressor, we obtain

$$\begin{aligned}\frac{\partial \mathbb{E}(Y_i|X_i)}{\partial X_{ik}} &= \frac{\partial \mathbb{E}(Y_i|D_i=1,X_i)}{\partial X_{ik}}\Pr(D_i=1|X_i)+\frac{\partial \Pr(D_i=1|X_i)}{\partial X_{ik}}\mathbb{E}(Y_i|D_i=1,X_i) \qquad (3)\\ &= \left(\beta_k+\frac{\partial \mathbb{E}\left(\varepsilon_i|\varepsilon_i>-X_i'\beta,X_i\right)}{\partial X_{ik}}\right)\Pr(D_i=1|X_i)\\ &\quad +\frac{\partial \Pr(D_i=1|X_i)}{\partial X_{ik}}\left(X_i'\beta+\mathbb{E}\left(\varepsilon_i|\varepsilon_i>-X_i'\beta,X_i\right)\right)\end{aligned}$$

Under normality, we have

$$\begin{aligned}\frac{\partial \mathbb{E}(Y_i|X_i)}{\partial X_{ik}} &= \left(\beta_k+\frac{\partial \mathbb{E}\left(\varepsilon_i|\varepsilon_i>-X_i'\beta,X_i\right)}{\partial X_{ik}}\right)\Pr(D_i=1|X_i)+\frac{\partial \Pr(D_i=1|X_i)}{\partial X_{ik}}\left(X_i'\beta+\mathbb{E}\left(\varepsilon_i|\varepsilon_i>-X_i'\beta,X_i\right)\right)\\ &= \left[\beta_k+\sigma\frac{\Phi\left(\frac{X_i'\beta}{\sigma}\right)\phi'\left(\frac{X_i'\beta}{\sigma}\right)\frac{\beta_k}{\sigma}-\phi\left(\frac{X_i'\beta}{\sigma}\right)\phi\left(\frac{X_i'\beta}{\sigma}\right)\frac{\beta_k}{\sigma}}{\left[\Phi\left(\frac{X_i'\beta}{\sigma}\right)\right]^2}\right]\Phi\left(\frac{X_i'\beta}{\sigma}\right)\\ &\quad +\phi\left(\frac{X_i'\beta}{\sigma}\right)\frac{\beta_k}{\sigma}\left[X_i'\beta+\frac{\sigma\phi\left(\frac{X_i'\beta}{\sigma}\right)}{\Phi\left(\frac{X_i'\beta}{\sigma}\right)}\right]\\ &= \left[\beta_k\Phi\left(\frac{X_i'\beta}{\sigma}\right)+\beta_k\frac{\Phi\left(\frac{X_i'\beta}{\sigma}\right)\phi'\left(\frac{X_i'\beta}{\sigma}\right)-\phi\left(\frac{X_i'\beta}{\sigma}\right)\phi\left(\frac{X_i'\beta}{\sigma}\right)}{\Phi\left(\frac{X_i'\beta}{\sigma}\right)}\right]to\\ &\quad +\frac{\beta_k}{\sigma}\phi\left(\frac{X_i'\beta}{\sigma}\right)X_i'\beta+\beta_k\phi\left(\frac{X_i'\beta}{\sigma}\right)\frac{\phi\left(\frac{X_i'\beta}{\sigma}\right)}{\Phi\left(\frac{X_i'\beta}{\sigma}\right)}\\ &= \beta_k\Phi\left(\frac{X_i'\beta}{\sigma}\right)+\beta_k\phi'\left(\frac{X_i'\beta}{\sigma}\right)+\frac{\beta_k}{\sigma}\phi\left(\frac{X_i'\beta}{\sigma}\right)X_i'\beta.\end{aligned}$$

To simplify the preceding expression further, note that

$$\phi'(z)=-z\phi(z).$$

This trick exploits the form of the standard normal pdf. To see this, note that

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)\Rightarrow\phi'(z)=\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}z^2\right)\right](-z)=-z\phi(z)$$

Therefore, we have

$$\frac{\partial \mathbb{E}(Y_i|X_i)}{\partial X_{ik}}=\beta_k\Phi\left(\frac{X_i'\beta}{\sigma}\right).$$

An alternative way to rewrite the derivative of $\mathbb{E}(Y_i|D_i=1,X_i)$ with respect to the kth regressor as follows. Let $z_i = X_i'\beta/\sigma$.

$$\begin{aligned}\frac{\partial \mathbb{E}(Y_i|D_i=1,X_i)}{\partial X_{ik}} &= \beta_k + \sigma \frac{\Phi(z_i)\,\phi'(z_i)\,\frac{\beta_k}{\sigma} - \phi(z_i)\,\phi(z_i)\,\frac{\beta_k}{\sigma}}{[\Phi(z_i)]^2} \\ &= \beta_k\left[1 - \frac{z_i\phi(z_i)}{\Phi(z_i)} - \left(\frac{\phi(z_i)}{\Phi(z_i)}\right)^2\right]. \end{aligned} \tag{4}$$

I hope you know how to estimate marginal effects at the "average" and average marginal effects given the preceding expression. Finally, standard errors are available via the delta method or the parametric bootstrap.

A structural model of female labor supply Consider the following Cobb-Douglas utility function summarizing preferences for work and leisure for the ith woman:

$$U_i = [W_i(H_i + e_i) + V_i]^{\alpha}[1-(H_i+e_i)]^{\beta}, \tag{5}$$

where W_i is real wage, V_i is real property income, H_i is proportion of time spent at work, and e_i is an unobservable error representing unobserved heterogeneity across women. Note that, in contrast to our discussion at the very beginning of the course, the total time endowment T available is equal to 1 for all women. So, $L_i = 1 - H_i$ is the proportion of time spent on leisure (or non-market activities). As noted before, it may happen that some women will choose to not participate in market activities. This is the corner solution that we sketched a while back.

The corner solution arises for some women because the wages that they face does not exceed their own reservation wage. At the corner solution, we have $H_i = 0$ and $L_i = 1$. The marginal rate of substitution at the corner solution is given by

$$MRS_i = \frac{b}{1-b}\cdot\frac{W_i e_i + V_i}{1-e_i},$$

where $b = \beta/(\alpha+\beta)$. Note that the reservation wage is equal to the MRS at the corner solution. This means that ith woman will only work when

$$W_i > \frac{b}{1-b}\cdot\frac{W_i e_i + V_i}{1-e_i} \Rightarrow e_i < 1 - b - \frac{bV_i}{W_i}.$$

This implies that the labor supply function for the ith woman is given by

$$H_i = \begin{cases} 1 - b - \frac{bV_i}{W_i} - e_i & \text{if } e_i \geq 1 - b - \frac{bV_i}{W_i} \\ 0 & \text{if } e_i < 1 - b - \frac{bV_i}{W_i} \end{cases} \tag{6}$$

This is the reason why running a regression of hours worked on the ratio of real property income to real wage, i.e.,

$$H_i = \gamma_0 + \gamma_1\frac{V_i}{W_i} + \varepsilon_i \tag{7}$$

is a bad idea. (Can you recall why?) At first, it may seem that (6) is very different from (7). First, define $\varepsilon_i = -e_i$. Second, under certain restrictions, (7) will reduce to (6). In particular, the restrictions are:

$$\gamma_0 = 1 - b, \gamma_1 = -b \Rightarrow \gamma_0 - \gamma_1 = 1.$$

The last expression is a linear restriction on the coefficients that can be used to test the structural model. This means that we can test from the data whether or not the sample of women have Cobb-Douglas preferences in (5). Of course, we cannot use results from OLS to apply the test.

One can use ML in this situation if we assume that $\xi_i \sim N\left(0, \sigma^2\right)$ and that ξ_i is independent of $X_i = V_i/W_i$ in the following latent variable model:

$$H_i^* = \gamma_0 + \gamma_1 X_i + \xi_i.$$

Note that we have fixed censoring at zero and that H is a left-censored version of H^*. Let $D_i = 1$ if the ith woman is observed to be working. Applying what we have before for the censored regression model, we have the following likelihood function:

$$\begin{aligned} L\left(\gamma_0, \gamma_1, \sigma^2\right) &= \prod_{i:D_i=1} f\left(H_i | D_i = 1, X_i\right) \prod_{i:D_i=0} \Pr\left(H_i^* \leq 0 | D_i = 0, X_i\right) \\ &= \prod_{i:D_i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(H_i - \gamma_0 - \gamma_1 X_i\right)^2\right] \prod_{i:D_i=0} \Pr\left(\xi_i > -\gamma_0 - \gamma_1 X_i | X_i\right) \\ &= \prod_{i:D_i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(H_i - \gamma_0 - \gamma_1 X_i\right)^2\right] \prod_{i:D_i=0} \Phi\left(\frac{\gamma_0 + \gamma_1 X_i}{\sigma}\right) \end{aligned}$$

Once again, the ML toolkit is available at your disposal.

Next, the calculation of marginal effects have a nice interpretation in this context. Recall from (3), along with changes in the notation:

$$\frac{\partial \mathbb{E}\left(H_i | X_i\right)}{\partial X_i} = \frac{\partial \mathbb{E}\left(H_i | D_i = 1, X_i\right)}{\partial X_i} \Pr\left(D_i = 1 | X_i\right) + \frac{\partial \Pr\left(D_i = 1 | X_i\right)}{\partial X_i} \mathbb{E}\left(H_i | D_i = 1, X_i\right).$$

Thus, the marginal effect on hours worked from a change in the property income to wage ratio can be decomposed into two parts that are of policy interest. The first term is the change in the hours worked for women who are already working. The second term is the change in the probability of working for all women (including those that are not working). Notice that these two terms are weighted differently. Of course, the total effect can also be of policy interest.

Question: It is also possible to report the following alternative measure. What exactly do you need to estimate the proportion of the total effect of a change in X_i on labor supply due to the effect from those already working? Is there a computational shortcut? See (4).

Finally, some things to note:

1. The discussion in this section assumes that data is available for X_i is available even for women who are not in the labor market.

2. It is possible to generalize the utility function so that we allow for preferences other than the Cobb-Douglas type.

3. Testing $\gamma_0 - \gamma_1 = 1$ is possible via the trinity.

4. It is also possible to generalize the form of the latent variable model to allow separate influences of real wages and real property income. In this case, it is possible to calculate income and wage effects just as in Mroz (1987).