Truncamiento y Censuramiento

The Author

26 de noviembre de 2016

1. Introduction

In this chapter we turn to several missing data problems. It is critical—even more so than in the previous chapters—to distinguish between assumptions placed on the population model and assumptions about how the data were generated. Under random sampling, the interesting issues concern the population model and assumptions we make about distributional features in the population. With nonrandom sampling, we must take particular care in stating assumptions about the population and separately stating assumptions on the sampling scheme.

We first study the general problem of **data censoring**. With data censoring we can still randomly sample units from the relevant population, but we face the problem that one or more of the variables is censored: we only observe the response over a certain range. The case of data **truncation** is a situation where we do not randomly draw units from the population; rather, we randomly sample from a subpopulation defined in terms of one or more of the observed variables. Naturally, the population parameters cannot always be identified with such sampling schemes, but they can be under suitable assumptions.

Truncation effects arise when one attempts to make inferences about a larger population from a sample that is drawn from a distinct subpopulation. For example, studies of income based on incomes above or below some poverty line may be of limited usefulness for inference about the whole population. Truncation is essentially a characteristic of the distribution from which the sample data are drawn. To continue the example, suppose that instead of being unobserved, all incomes below the poverty line are reported as if they were at the poverty line. The **censoring** of a range of values of the variable of interest introduces a distortion into conventional statistical results that is similar to that of truncation. Unlike truncation, however, censoring is essentially a defect in the sample data. Presumably, if they were not censored, the data would be a representative sample from the population of interest.

2. Useful Math

In both censoring and truncation, we need to work with truncated random variables. That is, a random variable that we can only observe in some specific range. There are

some useful properties:
$f(x|x > a) = \frac{x}{\mathbb{P}[x>a]}$. Which in the case of a random variable $Y \sim N(\mu, \sigma^2)$:

$$f(x|x > a) = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{1 - \Phi(a)}$$

Sea $c = \frac{a-\mu}{\sigma}$

Demostración.

$$f(x|X > c) = \frac{d}{dx}F(x|X > c) = \frac{d}{dx}\mathbb{P}[X < x|X > c]$$

Por regla de Bayes:

$$f(x|X > c) = \frac{d}{dx}\frac{\mathbb{P}[X < x, X > c]}{\mathbb{P}[X > c]}$$
$$f(x|X > c) = \frac{\frac{d}{dx}\int_c^x f(m)dm}{\mathbb{P}[X > c]}$$

Por Th. Fundamental del cálculo:

$$f(x|X > c) = \frac{f(x)}{1 - \Phi(c)} = \frac{\phi(x)}{1 - \Phi(c)}$$

□

Por otro lado, la esperanza condicional tiene la siguiente expresión :

$$\begin{aligned}\mathbb{E}[Z \mid Z > c] &= \int_c^\infty z f(z|Z > c)dz \\ &= \int_c^\infty z\frac{\phi(z)}{1 - \Phi(c)}dz \\ &= \frac{1}{1 - \Phi(c)}\int_c^\infty z\phi(z) = \frac{\phi(c)}{1 - \Phi(c)}\end{aligned}$$

3. Data Truncation

We do not sample from the full distribution of a variable; instead from just a subpopulation. Were we to estimate the effects for the subpopulation we have with OLS, our estimates would be biased as the expected conditional is not linear, even if it's linear for the not-truncated variable. To see that:

$$y = x\beta + \epsilon,\ \mathbb{E}[\epsilon \mid x] = \mathbb{E}[\epsilon]$$

$$\mathbb{E}[y \mid y > a] = \mathbb{E}[x\beta + \epsilon \mid \epsilon > a - x\beta]$$

With $\epsilon \sim (0, \sigma^2)$. Thus:

$$\mathbb{E}[y \mid y > a] = x\beta + \sigma \cdot \mathbb{E}[Z \mid Z > \frac{a - x\beta}{\sigma}] = x\beta + \sigma \cdot \frac{\phi(\frac{a-x\beta}{\sigma})}{1 - \Phi(\frac{a-x\beta}{\sigma})}$$

The conditional mean is therefore a nonlinear function of a, σ, x, and β. Regardless of the direction of the truncation, were we to magically have the average marginal effect for the truncated variable, we would underestimate the effect for the untruncated. That is:

$$\frac{\partial \mathbb{E}[y \mid x]}{\partial x} = \beta\delta$$

With $\delta \in [0, 1]$— Remember β is the average marginal effect for the truncated variable. However, we cannot consistently estimate the effect for the subpopulation with OLS. Instead, use MLE with the distribution of the truncated variable:

$$f(y|y > c) = f(y|x\beta + \epsilon > c) = f(y = \frac{f(x)}{1 - \Phi(c)}$$

Of course, correct estimation of β will depend on the strong distributional assumptions of MLE.

4. Data Censoring

Let y^* be the dependent variable of interest.

$$y^* = x\beta + u$$
$$\mathbb{E}[u \mid x] = \mathbb{E}[u]$$

The problem we study in this chapter occurs when we observe only a censored version of y. For example, if we only observe wealth up to some point, then we are sampling from the random variable: $w = min(y, r_i)$ which would indicate right censoring. A very common problem in microeconomic data is censoring of the dependent variable. When the dependent variable is censored, values in a certain range are all transformed to (or reported as) a single value. Conventional regression methods fail to account for the qualitative difference between limit (zero) observations and nonlimit (continuous) observations.
When data are censored, the distribution that applies to the sample data is a mixture of discrete and continuous distributions. That is, even if we have correctly specified a distribution for the uncensored variable, we cannot estimate that model; we have to estimate the model for the censored variable. If the censoring is from below, that is:

$$y = \begin{cases} a & \text{si } y^* \leq a \\ y^* & \text{si } y^* > a \end{cases}$$

Then, the expected value that applies to the observed data:

$$\mathbb{E}[y] = \Phi(\frac{a-\mu}{\sigma}) \cdot a + (1 - \Phi(\frac{a-\mu}{\sigma})) \cdot [\mu + \sigma \cdot \frac{\phi(\frac{a-\mu}{\sigma})}{1 - \Phi(\frac{a-\mu}{\sigma})}]$$

Whereas when the censoring is from above:

$$y = \begin{cases} a & \text{si } y^* \geq a \\ y^* & \text{si } y^* < a \end{cases}$$

$$\mathbb{E}[y] = (1 - \Phi(\frac{a-\mu}{\sigma})) \cdot a + \Phi(\frac{a-\mu}{\sigma}) \cdot [\mu - \sigma \cdot \frac{\phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{a-\mu}{\sigma})}]$$

Thus, as you can probably see, the average marginal effect for the censored variable is different than the average marginal effect for the original variable. That is:

$$\frac{\partial \mathbb{E}[y^*]}{\partial x} \neq \frac{\partial \mathbb{E}[y]}{\partial x}$$

$$\frac{\partial \mathbb{E}[y]}{\partial x} = \beta \mathbb{P}[a < y^*]$$

That is, were we use the average marginal effects of the censored variable to estimate the average marginal effects for uncensored variable, we would underestimate the effects for the latter: Çensoring leads to an attenuation of the marginal effect of X relative to its effect in the latent regression.". The solution, then, to the problem of how to estimate the average marginal effects for our variable of interest is to work with the observed distribution and estimate the parameters with maximum likelihood. The observed distribution, then, will be a bernoulli like distribution, where the non censored part will be the continuous distribution of the latent variable, and the censored part will be the probability of being censored. That is:

$$f(y|x) = \{f^*(y^*|x)\}^{I(y^*>a)} \{\mathbb{P}[y^* \leq a]\}^{I(y^* \leq a)}$$

Of course, correct estimation of β will depend on the strong distributional assumptions of MLE.

5. Sample Selection

We sample from an incomplete distribution of the dependent variable, conditional on the values for some other independent variable. Suppose that y and z have a bivariate distribution with correlation ρ. We are interested in the distribution of y given that z exceeds a particular value. Intuition suggests that if y and z are positively correlated, then the truncation of z should push the distribution of y to the right.
The truncated joint density is thus:

$$f(y, z|z > a) = \frac{f(y, z)}{\mathbb{P}[z > a]}$$

Then, if both random variables have a bivariate normal distribution with μ_y, σ_y

$$\mathbb{E}[y \mid z > a] = \mu_y + \rho \cdot \sigma_y \cdot \frac{\phi(\frac{a-\mu}{\sigma})}{1 - \Phi(\frac{a-\mu}{\sigma})}$$

Thus, as always, the average marginal effect for our observed variable won't correspond to the average marginal effect of our modeled variable y.

5.1. Regression in a model of Selection

Thus, we have a model to determine the sample selection:

$$z_i^* = w_i'\gamma + u_i$$

And a model of the variable of interest.

$$y_i = x_i'\beta + \epsilon_i$$

The sample rule is that we sample iff $z_i^* > 0$. Suppose as well that ϵ_i and u_i have a bivariate normal distribution with zero means and correlation ρ.
The problem is that the variable we observe is $y_i|z_i^*$ has the following average marginal effect:

$$\mathbb{E}[y_i \mid z_i^* > 0] = \mathbb{E}[x_i'\beta + \epsilon_i \mid w_i'\gamma + u_i > 0] = x_i'\beta + \mathbb{E}[\epsilon_i \mid u_i > -w_i'\gamma]$$

As we saw before:

$$\mathbb{E}[\epsilon_i \mid u_i > -w_i'\gamma] = 0 + \rho \cdot \sigma_\epsilon \cdot \frac{\phi(\frac{-w_i'\gamma}{\sigma_u})}{1 - \Phi(\frac{-w_i'\gamma}{\sigma_u})}$$

Thus:

$$\mathbb{E}[y_i \mid z_i^* > 0] = x_i'\beta + \rho \cdot \sigma_\epsilon \cdot \frac{\phi(\frac{-w_i'\gamma}{\sigma_u})}{1 - \Phi(\frac{-w_i'\gamma}{\sigma_u})}$$

$$\mathbb{E}[y_i \mid z_i^* > 0] = x_i'\beta + \rho \cdot \sigma_\epsilon \cdot \frac{\phi(\frac{w_i'\gamma}{\sigma_u})}{\Phi(\frac{w_i'\gamma}{\sigma_u})}$$

Thus, the average marginal effect for the truncated variable is not the same as the average marginal effect of our variable of interest. Also, to work with our observed variable, we must recognize the fact that the conditional expectation is non-linear.

5.1.1. 2 step Heckman

To do so, note that although we don't observe z_i^*, we do observe whether it's positive or not:

$$z_i = \begin{cases} 1 & \text{si } z_i^* \geq 0 \\ 0 & \text{si } z_i^* < 0 \end{cases}$$

Thus, $\mathbb{P}[z_i] = \mathbb{P}[z_i^* > 0] = \Phi(\frac{w_i'\gamma}{\sigma_u})$ can be estimated from our observed data. Thus, we can also estimate the mills ratio in the former conditional expectation— after normalizing $\sigma_u = 1$

$$\frac{\phi(\frac{w_i'\hat{\gamma}}{1})}{\Phi(\frac{w_i'\hat{\gamma}}{1})}$$

Then, we can estimate:

$$\mathbb{E}[y_i \mid z_i^* > 0] = x_i'\beta + \rho \cdot \sigma_\epsilon \cdot \frac{\phi(\frac{w_i'\gamma}{\sigma_u})}{\Phi(\frac{w_i'\gamma}{\sigma_u})}$$

With OLS with the following, using both x_i and the mills ratio as variables.

$$y_i = x_i'\beta + \lambda \frac{\phi(\frac{w_i'\hat{\gamma}}{1})}{\Phi(\frac{w_i'\hat{\gamma}}{1})} + \epsilon_i$$

Al utilizar una variable generada por regresión, tendremos que corregir los errores estándar. Although we can recover both ρ, σ_ϵ, the sample selection can be tested with the statistical significance of λ.

5.1.2. Maximum Likelihood

The density of our observed variable, y_i^* is the following:

$$f(y_i^*, z_i | x_i, w_i) = f(y_i, z_i | x_i, w_i, z_i = 1) = f(y_i | x_i, w_i, z_i = 1) \cdot f(z_i | x_i, w_i, z_i = 1)$$

Where $f(z_i | x_i, w_i, z_i = 1)$ is a truncated bernoulli:

$$f(z_i | x_i, w_i, z_i = 1) = \frac{\{\mathbb{P}[z_i = 1 \mid x_i, w_i]\}^{z_i} \{\mathbb{P}[z_i = 0 \mid x_i, w_i]\}^{1-z_i}}{\mathbb{P}[z_i = 1 \mid x_i, w_i]}$$

And:

$$f(y_i | x_i, w_i, z_i = 1) = \frac{f(y_i | x_i, w_i)}{\mathbb{P}[z_i = 1 \mid x_i, w_i]}$$