Censored data in Social Sciences

Tool: Tobit model.

The Analytics Edge: In some applications, we have only access to censored data. In such case, the value of the observations is only partially known. Examples of censored data include:

- Survival of patients (some patients may leave a medical programme before concluding it);
- Number of extramarital affairs (data collected, for example, by magazine surveys);
- Expenditures on vacations;
- Education testing: if an exam is too long, a lot of people may get a full mark; if it is too hard, a lot of people may get a low mark.

Descriptive analytics—such as the Tobit model—can help us model censored response variables.

Dealing with censored data

Censored-dependent variable

Let's assume that a censored-dependent variable y^* follows a normal distribution with mean μ and standard deviation σ, that is, $y^* \sim N(\mu, \sigma)$. Consider a censored value at C (say, capacity), then, the variable we observe is:

$$y = \begin{cases} y^* & \text{if } y^* \leq C \\ C & \text{otherwise} \end{cases}$$

In this example, illustrated in Figure 0.1, the censored random variable is right-censored and the new variable is a mixture of continuous and discrete points.

Censored regression (Tobit model)

James Tobin, in 1958, proposed a model that deals with censored regression problems. The model is referred to as the Tobit model (from Tobin and probit.) The model is based on the following regression:

$$\underbrace{y_i^*}_{\substack{\text{Latent variable} \\ \text{(unobservable)}}} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \ \ , \forall i = 1, \ldots, n,$$

where $\{\beta_1, \ldots, \beta_p\}$ are the model coefficients, $\{x_1, \ldots, x_p\}$ the predictors, n the number of observations, and ϵ the model error. The Tobit model assumes that $\epsilon_i \sim N(0, \sigma^2)$. The observable variable y_i is left-censored at 0:

$$\underbrace{y_i}_{\substack{\text{observable} \\ \text{variable}}} = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}$$

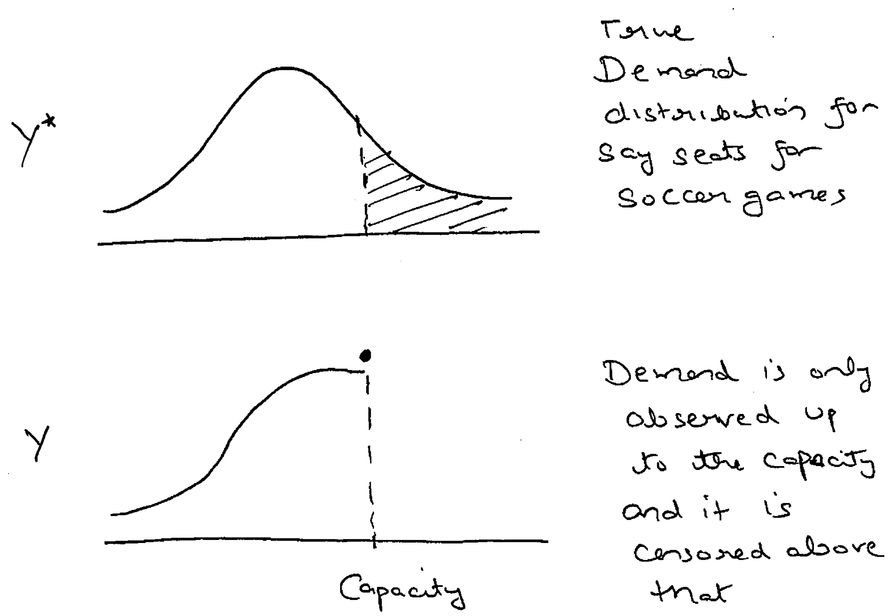

Figure 0.1: Illustration of a right-censored random variable.

Because y^* is normally distributed, y has a continuous distribution over strictly positive values. In particular, the density of y given $\mathbf{x}$ (the vector of predictors $x_1, \ldots, x_p$) is the same as the density of y^* given $\mathbf{x}$ for positive values. Further,

$$P(y = 0|\mathbf{x}) = P(y^* < 0|\mathbf{x}) = P(\epsilon < -\mathbf{x}\beta|\mathbf{x}) =$$
$$= P(\epsilon/\sigma < -\mathbf{x}\beta/\sigma|\mathbf{x}) = \Phi(-\mathbf{x}\beta/\sigma) = 1 - \Phi(\mathbf{x}\beta/\sigma),$$

where we absorbed the intercept β_0 into a vector of coefficients β, and $\Phi(\cdot)$ is the cumulative distribution function of a normal variable. This expression holds because ϵ/σ has a standard normal distribution and is independent of $\mathbf{x}$. Therefore, if $(\mathbf{x}_i, y_i)$ is a random draw from the population, the density of y_i given $\mathbf{x}_i$ is:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \mathbf{x}_i\beta)^2}{2\sigma^2}} = (1/\sigma)\phi[(y_i - \mathbf{x}_i\beta)/\sigma],\ y_i > 0$$
$$P(y_i = 0|\mathbf{x}_i) = 1 - \Phi(-\mathbf{x}_i\beta/\sigma),$$

where ϕ is the standard normal density function.

From these last two equations, we can obtain the log-likelihood function for each observation i:

$$l_i(\beta, \sigma) = 1(y_i = 0)\log[1 - \Phi(\mathbf{x}_i\beta/\sigma)] + 1(y_i > 0)\log\{(1/\sigma)\phi[y_i - \mathbf{x}_i\beta/\sigma]\}.$$

Notice how this depends on σ, the standard deviation of ϵ, as well as on β. The log-likelihood for a random sample of size n is obtained by summing $l_i(\cdot)$ across all i. The maximum likelihood estimates of β and σ are obtained by maximizing the log-likelihood. This generally requires numerical methods.

Interpreting the model output

From equation $y^* = \mathbf{x}\beta + \epsilon$, we see that the parameter β_j measure the partial effects of x_j on $E(y^*|\mathbf{x})$, where y^* is the latent variable. The variable we want to explain is y, as this is the observed outcome (such as hours worked or amount of charitable contributions).

In Tobit models, two expectations are of particular interest: $E(y|y > 0, \mathbf{x})$, which is sometimes called the *conditional expectation* because it is conditional on $y > 0$, and $E(y|\mathbf{x})$, which is, unfortunately, called the *unconditional expectation.* (Both expectations are conditional on the explanatory variables.) The expectation $E(y|y > 0, \mathbf{x})$ tells us, for given values of $\mathbf{x}$, the expected value of y for the subpopulation where y is positive. Given $E(y|y > 0, \mathbf{x})$, we can easily find $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = P(y > 0|\mathbf{x}) \cdot E(y|y > 0, \mathbf{x}) = \Phi(\mathbf{x}\beta/\sigma) \cdot E(y|y > 0, \mathbf{x}).$$

Skipping some derivations, $E(y|y > 0, \mathbf{x})$ can be expressed as $E(y|y > 0, \mathbf{x}) = \mathbf{x}\beta + \sigma\lambda(\mathbf{x}\beta/\sigma)$, which shows that the expected value of y conditional on $y > 0$ is equal to $\mathbf{x}\beta$ plus a strictly positive term, which is σ times $\lambda(\mathbf{x}\beta/\sigma)$.

Combining the last two expressions, we get:

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\beta/\sigma) \cdot [\mathbf{x}\beta + \sigma\lambda(\mathbf{x}\beta/\sigma)] = \Phi(\mathbf{x}\beta/\sigma)\mathbf{x}\beta + \sigma\phi(\mathbf{x}\beta/\sigma),$$

where the second equality follows because $\Phi(\mathbf{x}\beta/\sigma)\lambda(\mathbf{x}\beta/\sigma) = \phi(\mathbf{x}\beta/\sigma)$. This equation shows that when y follows a Tobit model, $E(y|\mathbf{x})$ is a nonlinear function of $\mathbf{x}$ and β.