Econometrics II TA Session *

Wang Xin

Nov 13, 2023

1 Review of Truncated Regression Model

1.1 Truncated normal variable

Let's start from a random variable $Z \sim N(0,1)$, and the distribution function is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right),$$

$$\Phi(z) = \int_{-\infty}^{z} \phi(z) dz.$$

Accordingly, the distribution function of $X \sim N(\mu, \sigma^2)$ can be derived as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

Pay attention to the cumulative distribution function of x, it should be

$$F_x(a) = \int_{-\infty}^{a} f(x)dx = \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Now we assume an interval (K_1, K_2), where K_1 and K_2 are known. The conditional density $X|K_1 < X < K_2$ (using the definition of conditional density) is given by

$$f(x|K_1 < x < K_2) = \frac{f(x)}{\int_{K_1}^{K_2} f(x)dx} = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{K_2-\mu}{\sigma}\right) - \Phi\left(\frac{K_1-\mu}{\sigma}\right)}$$

*The codes are cited from documents by Hiroki Kato.

Then

$$
\begin{aligned}
E[x|K_1 < x < K_2] &= \int_{K_1}^{K_2} x f(x|K_1 < x < K_2) dx \\
&= \sigma \frac{\phi\left(\dfrac{K_1-\mu}{\sigma}\right) - \phi\left(\dfrac{K_2-\mu}{\sigma}\right)}{\Phi\left(\dfrac{K_2-\mu}{\sigma}\right) - \Phi\left(\dfrac{K_1-\mu}{\sigma}\right)} + \mu
\end{aligned}
$$

1.2 Truncated regression model

We assume that y_i, $\mathbf{x}_i$ are continuous random variables and the selection rule is

$$
y_i = \mathbf{x}_i \boldsymbol{\beta} + u_i \quad \text{if } a_1 < y < a_2
$$

where $u_i \sim N(0, \sigma^2)$. with a_1, a_2 known constant values. If y_i falls into (a_1, a_2), then we observe both y_i, $\mathbf{x}_i$. If not, we do not observe y_i or $\mathbf{x}_i$. Therefore, groups of data, which exists in reality but does not fit the selection rule, is not included or observed.

The probability density function of y_i conditional on $\mathbf{x}_i$ and $a_1 < y < a_2$ should be

$$
p_\theta(y_i|\mathbf{x}_i, a_1 < y < a_2) = \frac{f(y_i|\mathbf{x}_i)}{P(a_1 < y < a_2|\mathbf{x}_i)}
$$

where $\theta = (\boldsymbol{\beta}, \sigma^2)'$. Considering the distributional assumption and taking variable transformation, the conditional distribution of y_i is given by

$$
p(y_i|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)^2\right) = \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right),
$$

where $\phi(\cdot)$ is the standard normal density function. Moreover, the probability of observation $y_i > 0$ is given by

$$
\begin{aligned}
P(a_1 < y < a_2|\mathbf{x}_i) &= P(a_1 < \mathbf{x}_i\boldsymbol{\beta} + u_i < a_2|\mathbf{x}_i) \\
&= P\left(\frac{a_1 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} < \frac{u_i}{\sigma} < \frac{a_2\mathbf{x}_i\boldsymbol{\beta}}{\sigma}|\mathbf{x}_i\right) \\
&= \Phi\left(\frac{a_2 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)
\end{aligned}
$$

where $\Phi(\cdot)$ is the standard normal cumulative density function.

Thus, the log-likelihood function is supposed to be

$$
M_n(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma}\frac{\phi\left(\dfrac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\dfrac{a_2 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\dfrac{a_1 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}\right)
$$