Censored and Truncated Regression Models
Economics 583 Fall 2009

Eric Zivot

December 6, 2009
**Truncated and Censored Samples**

Consider the linear regression model

\[ y_i = x'_i \beta + \varepsilon_i, i = 1, \ldots, n \]

*Censoring*

This occurs when we observe \( x_i \) for the entire sample, but for some observations we only have limited information about \( y_i \)

*Truncation*

This occurs when we only observe \( y_i \) and \( x_i \) for those cases in which \( y_i \) is above or below some threshold.
**Example:** Tobin’s study of Household expenditure (Censored Regression)

Consumer maximizes utility by purchasing durable goods under constraint that total expenditures do not exceed income

\[
\text{expenditure of durables} \geq \text{cost of least expensive durable good}
\]

If available income is less than least expensive durable good then no expenditure is observed. Don’t know how much a household would have spent if a durable good could be purchased for less than the least expensive item.

**Example:** Expenditure on Booze

Model how much an individual spends on alcohol in a given month. A significant fraction have zero expenditure.
**Example:** Hausman and Wise’s analysis of the New Jersey negative income tax experiment (Truncated Regression)

Goal: Estimate earnings function for low income individuals

Truncation: Individuals with earnings greater than $1.5 \times$ poverty level were excluded from the sample.

Two types of inference:

1. Inference about entire population in presence of truncation

2. Inference about sub-population observed
Truncated Regression Model

\[ y_i = x_i' \beta + \varepsilon_i, \; i = 1, \ldots, n \]
\[ \varepsilon_i \sim iid \; N(0, \sigma^2) \]

Truncation from below

observe \( y_i \) and \( x_i \) for \( y_i > c \)

Truncation from above

observe \( y_i \) and \( x_i \) for \( y_i < c \)

Consider the case of truncation from below.
Truncated Distributions

Let $X$ be a continuous random variable with pdf $f(x)$, and let $c$ be a constant. Then

$$f(x|x > c) = \frac{f(x)}{\Pr(X > c)} = \frac{f(x)}{\int_{-\infty}^{c} f(x) \, dx}$$

Truncated Normal Distribution

let $X \sim N(\mu, \sigma^2)$ and let $c$ be a constant. Then

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

$$\Pr(X > c) = \Pr \left( \frac{X - \mu}{\sigma} > \frac{c - \mu}{\sigma} \right) = 1 - \Phi \left( \frac{c - \mu}{\sigma} \right)$$
The distribution of $X | X > c$ is called *truncated normal* and the density is given by

$$f(x | x > c, \mu, \sigma^2) = \frac{f(x; \mu, \sigma)}{\Pr(X > c)} = \frac{\frac{1}{\sigma} \phi \left( \frac{x-\mu}{\sigma} \right)}{1 - \Phi \left( \frac{c-\mu}{\sigma} \right)}$$
The mean and variance are given by

\[ E[X|X > c] = \mu + \sigma \cdot \lambda \left( \frac{c - \mu}{\sigma} \right) \]
\[ \text{var}(X|X > c) = \sigma^2 \left( 1 - \delta \left( \frac{c - \mu}{\sigma} \right) \right) \]

where

\[ \lambda(x) = \frac{\varphi(x)}{1 - \Phi(x)} = \text{inverse Mills ratio} \]
\[ = \text{normal hazard function} \]
\[ \delta(x) = \lambda(x) (\lambda(x) - x), \ 0 < \delta(x) < 1 \]

Remarks:

1. \( E[X|X > c] > \mu \)
2. \( \text{var}(X|X > c) < \sigma^2 \)
In the truncated regression model with truncation from below

\[ E[y_i | x_i] = x_i'\beta \]

\[ E[y_i | x_i, y_i > c] = x_i'\beta + \sigma \cdot \lambda \left( \frac{c - x_i'\beta}{\sigma} \right) \]

The appropriate regression function to describe the observed data is

\[ y_i = E[y_i | x_i, y_i > c] + \varepsilon_i = x_i'\beta + \sigma \cdot \lambda \left( \frac{c - x_i'\beta}{\sigma} \right) + \varepsilon_i \]

Hence, the usual linear regression model

\[ y_i = E[y_i | x_i] + \varepsilon_i = x_i'\beta + \varepsilon_i \]

is misspecified because it omits the truncation correction variable \( \lambda \left( \frac{c - x_i'\beta}{\sigma} \right) \)
Result: Bias and Inconsistency in Truncated Regression

Because $x_i$ is correlated with $\lambda \left( \frac{c - x_i' \beta}{\sigma} \right)$ the OLS estimator of $\beta$ in the linear regression model with truncation from below is biased and inconsistent.

Consistent and asymptotically normal estimates can be computed using

1. Non-linear least squares estimation of

$$y_i = x_i' \beta + \sigma \cdot \lambda \left( \frac{c - x_i' \beta}{\sigma} \right) + \varepsilon_i$$

Note: truncation correction term is based on $\varepsilon_i \sim iid \ N(0, \sigma^2)$.

2. MLE based on truncated normal distribution. MLE will be efficient if $\varepsilon_i \sim iid \ N(0, \sigma^2)$
Maximum Likelihood Estimation

Assume \( \{(y_1, x_1), \ldots, (y_n, x_n)\} \) is a random sample generated from the truncated regression model with truncation from below. Then

\[
f(y_i | x_i, y_i > c, \beta, \sigma^2) = \frac{1}{\sigma} \varphi \left( \frac{y_i - x^T_i \beta}{\sigma} \right) \frac{1}{1 - \Phi \left( \frac{c - x^T_i \beta}{\sigma} \right)}
\]
The likelihood and log-likelihood functions are

\[
L(\beta, \sigma^2 | y, X) = \prod_{i=1}^{n} \frac{1}{\sigma} \varphi \left( \frac{y_i - x_i' \beta}{\sigma} \right) \frac{1}{1 - \Phi \left( \frac{c - x_i' \beta}{\sigma} \right)}
\]

\[
\ln L(\beta, \sigma^2 | y, X) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i' \beta)^2
\]

\[\quad - \sum_{i=1}^{n} \ln \left( 1 - \Phi \left( \frac{c - x_i' \beta}{\sigma} \right) \right)\]

This is a non-linear function of \(\beta\) and \(\sigma^2\) and no analytic solution for the estimator exists. However, Newton-Raphson iteration is straightforward and analytic derivatives for the score and Hessian are readily available.
Reparameterizing the Likelihood Function

Define

\[ \delta = \beta / \sigma, \quad \gamma = 1 / \sigma \Rightarrow \beta = \delta / \gamma, \quad \sigma^2 = 1 / \gamma^2 \]

Then the log-likelihood simplifies to

\[
\ln L(\delta, \gamma | y, X) = -\frac{n}{2} \ln(2\pi) + n \ln(\gamma) - \frac{1}{2} \sum_{i=1}^{n} \left( \gamma y_i - x'_i \delta \right)^2 \\
- \sum_{i=1}^{n} \ln \left( 1 - \Phi \left( \gamma c - x'_i \delta \right) \right)
\]

which is more numerically stable than the original log-likelihood function. Given \( \hat{\delta}_{mle} \) and \( \hat{\gamma}_{mle} \) the invariance property of MLE gives \( \hat{\beta}_{mle} \) and \( \hat{\sigma}^2_{mle} \) and the delta method gives the appropriate asymptotic variance.
Censored Regression Model

\[ y_i^* = x_i' \beta + \varepsilon_i, \ i = 1, \ldots, n \]
\[ \varepsilon_i \sim iid \ N(0, \sigma^2) \]
\[ y_i = y_i^* \text{ if } y_i^* > c \]
\[ = c \text{ if } y_i^* \leq c \]

Censored Normal Distribution

- \( y_i \) has the normal distribution \( N(x_i' \beta, \sigma^2) \) for the non-censored observations

- \( y_i \) has the discrete pdf \( \Pr(y_i^* \leq c) = \Phi \left( \frac{c-x_i' \beta}{\sigma} \right) \) for the censored observations
Hence,

\[ f(y_i|x_i, \beta, \sigma^2, c) = \begin{cases} \frac{1}{\sigma} \varphi \left( \frac{y_i - x_i'\beta}{\sigma} \right) & \text{for } y_i > c \\ \Phi \left( \frac{c - x_i'\beta}{\sigma} \right) & \text{for } y_i = c \end{cases} \]

Define the dummy variable

\[ D_t = \begin{cases} 1 & \text{if } y_t = c \\ 0 & \text{if } y_t > c \end{cases} \]
The likelihood and log-likelihood functions are

\[ L(\beta, \sigma^2 | y, X) = \prod_{i=1}^{n} \left[ \frac{1}{\sigma} \varphi \left( \frac{y_i - x_i' \beta}{\sigma} \right) \right]^{1-D_t} \left[ \Phi \left( \frac{c - x_i' \beta}{\sigma} \right) \right]^{D_t} \]

\[
\ln L(\beta, \sigma^2 | y, X) = \sum_{i=1}^{n} \left\{ (1 - D_t) \ln \left( \frac{1}{\sigma} \varphi \left( \frac{y_i - x_i' \beta}{\sigma} \right) \right) + D_t \ln \left( \Phi \left( \frac{c - x_i' \beta}{\sigma} \right) \right) \right\} \\
= \sum_{\{y_i > c\}} \left\{ 1 \right\} \left[ \ln(2\pi) + \ln(\sigma^2) + \frac{1}{\sigma^2} (y_i - x_i' \beta)^2 \right]

+ \sum_{\{y_i = c\}} \ln \left( \Phi \left( \frac{c - x_i' \beta}{\sigma} \right) \right) \]
Reparameterizing the Likelihood Function

Define
\[ \delta = \beta / \sigma, \quad \gamma = 1 / \sigma \Rightarrow \beta = \delta / \gamma, \quad \sigma^2 = 1 / \gamma^2 \]

Then the log-likelihood simplifies to
\[
\ln L(\delta, \gamma | y, X) = \sum_{i=1}^{n} \left\{ (1 - D_t) \left\{ -\frac{1}{2} \ln (2\pi) + \ln(\gamma) - \frac{1}{2} (\gamma y_i - x_i' \delta)^2 \right\} + D_t \ln \left( \Phi(\gamma c - x_i' \delta) \right) \right\}
\]

which is more numerically stable than the original log-likelihood function. Given \( \hat{\delta}_{mle} \) and \( \hat{\gamma}_{mle} \) the invariance property of MLE gives \( \hat{\beta}_{mle} \) and \( \hat{\sigma}_{mle}^2 \) and the delta method gives the appropriate asymptotic variance.