Censored Models

Li and Racine (2007, Chapter 11)

Yasuyuki Matsumura (yasu0704xx [at] gmail.com)

January 20, 2025

Graduate School of Economics, Kyoto University

Parametric Censored Models

Type-1 Tobit Model

- Consider the following latent variable model:

$$Y_i^\star = X_i^T\beta + \epsilon_i, \quad i = 1, \cdots n,$$

  where $X_i \in \mathbb{R}^q$ is an explanatory vector, β is a $q \times 1$ vector of coefficients, and ϵ_i is a mean zero disturbance term.
- $Y_i^\star$ is a latent variable, which we cannot observe. Instead, we observe

$$\begin{aligned} Y_i &= Y_i^\star 1(Y_i^\star > 0) \\ &= \max\{X_i^T\beta + \epsilon_i, 0\}. \end{aligned}$$

- Note that the "cutoff" is set equal to 0 without loss of generality. That is, we expect that Y_i (ϵ_i) is censored at 0 (resp. $-X_i^T\beta$).

Parametric Approach

- Popular parametric approaches include MLE and Heckit. [1]
- These approaches demand the following distributional assumption:

$$\epsilon_i | X_i \sim \text{Normal}(0, \sigma^2).$$

  Since $Y_i^\star$ is censored, for example, by top coding, the distribution of $Y_i^\star$ cannot be identified without this assumption.
- In other words, these parametric approaches do not allow for the heteroscedasticity of ϵ_i (Arabmazar and Schmidt 1981).

[1] Amemiya (1984)：Tobit モデルのサーベイ論文；Amemiya (1985)：教科書.

Semiparametric Type-1 Tobit Models

Semiparametric Type-1 Tobit Models

- We introduce the following semiparametric type-1 Tobit model:

$$\begin{aligned} Y_i^\star &= X_i^T \beta + \epsilon_i, \\ Y_i &= Y_i^\star 1(Y_i^\star > 0). \end{aligned}$$

- For identifying the moments of $Y_i^\star$, we need additional assumptions.
- Powell (1984) proposes to assume that med$(\epsilon_i | X_i) = 0$.
- Chen and Khan (2000) proposes a estimation procedure which requires weaker assumptions for identification than Powell (1984).

Semiparametric Censored Regression Models

Powell (1984): CLAD

- Consider the semiparametric type-1 Tobit model:

$$Y_i^\star = X_i^T \beta + \epsilon_i,$$
$$Y_i = Y_i^\star 1(Y_i^\star > 0) = \max\{Y_i^\star, 0\}.$$

- Assume that $\text{med}(\epsilon_i | X_i) = 0$. Noting that the "monotonicity" of median [2], we obtain $\text{med}(Y_i | X_i) = \max\{\text{med}(Y_i^\star | X_i), 0\} = \max\{X_i^T \beta, 0\}$, which implies that the above model can be rewritten as

$$Y_i = \max\{X_i^T \beta, 0\} + \epsilon_i,$$
$$\text{med}(\epsilon_i | X_i) = 0.$$

[2] max と med の順番を入れ替えても大丈夫ということ．max でなくても，単調変換なら入れ替え可.

- Powell (1984) proposes the following censored least absolute deviations estimator:

$$\hat{\beta}_{clad} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \max\{X_i^T \beta, 0\}|$$
$$= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} 1(X_i^T \beta > 0)|Y_i - X_i^T \beta|.$$

- Computation is sometimes complex [3]. See Buchinsky (1994); Khan and Powell (2001).
- Powell (1984) establishes the $\sqrt{n}$-consistency and asy. normality:

$$\sqrt{n}(\hat{\beta}_{clad} - \beta) \xrightarrow{d} \text{Normal}(0, V_{clad}^{-1}),$$

where

$$V_{clad} = 4f^2(0)\mathbb{E}[1(X_i^T \beta > 0)X_i X_i^T]$$

and $f(0)$ is the density of ϵ_i at the origin.

[3] β が 2 つの役割をもつことに起因する：データの選択，係数の値の決定.

- Variance estimation can be implemented as follows.
- Assume that ϵ_i is independent of X_i.
- Note that

$$
\begin{aligned}
f(0) &= \lim_{h \to 0} \mathbb{P}(0 \leq \epsilon_i < h) \\
&= \lim_{h \to 0} \mathbb{P}(0 \leq \epsilon_i < h | X_i^T \beta > 0).
\end{aligned}
$$

- Powell suggests to estimate $f(0)$ by

$$
\hat{f}(0) = \frac{1(X_i^T \hat{\beta}_{clad} > 0) 1(0 \leq \hat{\epsilon}_i < h)}{h \sum_{i=1}^n 1(X_i^T \hat{\beta}_{clad} > 0)}.
$$

Extension 1: Estimation of $f(0)$

- Horowitz and Neumann (1987) propose an alternative estimator of $f(0)$.
- To estimate $f(0)$, they use data with $X_i^T \hat{\beta}_{clad} \in \left[-\frac{h}{2}, \frac{h}{2}\right]$.
- Their estimator is given by

$$\hat{f}(0) = \frac{\sum_{i=1}^{n} 1\left(-\frac{h}{2} \leq \hat{\epsilon}_i \leq \frac{h}{2}\right) 1\left(Y_i > 0\right)}{h\left[\sum_{i=1}^{n} 1(X_i^T \hat{\beta}_{clad} > \frac{h}{2}) + \frac{1}{2}\left(1 + \frac{X_i^T \hat{\beta}_{clad}}{\frac{h}{2}}\right) 1\left(-\frac{h}{2} < X_i^T \hat{\beta}_{clad} \leq \frac{h}{2}\right)\right]}.$$

- Hall and Horowitz (1990) suggest to replace the indicator function by a kernel function.

Extension 2: Newey and Powell (1990)

- Newey and Powell (1990) modify the objective function above:

$$\hat{\beta}_{np} = \arg\min_{\beta} \sum_{i=1}^{n} w_i |Y_i - \max\{X_i^T \beta, 0\}|.$$

- They show that the optimal weight is $w_i = 2f(0|X_i)$. The asy. variance is $\{4\mathbb{E}[1(X_i^T\beta > 0) f^2(0|X_i) X_i X_i^T]\}^{-1}$.
- Their estimator achieves the semiparametric efficiency bound for the censored regression model under med$(\epsilon_i|X_i) = 0$.
- If ϵ_i is independent of X_i, then $f(0|X_i) = f(0)$, which implies that $\hat{beta}_{np} = \hat{\beta}_{clad}$.

Extension 3: Other Approaches

- Powell (1986): Additionally assume the symmetry assumption.
- Newey (1991): GMM-based estimation. Assume the symmetry assumption for efficiency.
- Honore and Powell (1994): Identically CLAD; Identically censored least squares.

Nonparametric Heteroscedasticity

Problems Arising with Powell's CLAD

- Recall that Powell's CLAD requires med$(\epsilon_i | X_i) = 0$, which can be interpreted as restrictive [4].
- Avar$(\hat{\beta}_{clad})$ is represented using $\mathbb{E}[1(X_i^T \beta > 0) X_i X_i^T]^{-1}$, which cannot be defined if $\mathbb{E}[1(X_i^T \beta > 0) X_i X_i^T]$ is not of full rank. This problem often arises under heavy censoring (i.e., when $X_i^T \beta$ is negative with high probability).

[4] とはいえ，中央値の識別は，期待値の識別よりもはるかに緩い条件で済むので，CLAD やそれを拡張した打ち切りデータに対する分位点回帰をやろうという話になる．

Chen and Khan (2000)

- Chen and Khan (2000) consider estimation procedures for heteroscedastic censored linear regression models.
- Their approach requires weaker identification conditions than Powell's CLAD.
- They also allow for various degrees of censoring.
- Their main idea is that they model the error term as the product of a homoscedastic error and a scale function of X_i that can be estimated using kernel methods.

- They assume that

$$\epsilon_i = \sigma(X_i)v_i,$$
$$\mathbb{P}(v_i \leq \lambda | X_i) \equiv \mathbb{P}(v_i \leq \lambda) \text{ for any } \lambda \in \mathbb{R}, X_i\, a.s.,$$
$$\mathbb{E}(v_i) = 0, \mathsf{Var}(v_i) = 1.$$

[5]

- Recalling that $Y_i = \max\{X_i^T\beta + \epsilon_i, 0\}$, we obtain

$$\text{For any } \alpha \in (0,1),$$
$$q_\alpha(X_i) = \max\{X_i^T\beta + c_\alpha\sigma(X_i), 0\},$$

where

$q_\alpha(\cdot)$ denotes the α-th quantile of Y_i given X_i,

c_α denotes the α-th quantile from the (unknown) distribution of v_i.

- Thus, for any $q_{\alpha_j}(X_i) > 0$ for two distinct $\alpha_1 \neq \alpha_2$, we have

$$q_{\alpha_j}(X_i) = X_i^T\beta + c_{\alpha_j}\sigma(X_i) \text{ for } j = 1,\ 2.$$

[5] このように ϵ_i を定めると，正規分布を許容できなくなってしまう（heteroscedasticity を守るならば）．

Chen and Khan (2000): Estimation

- Chen and Khan (2000) propose two estimators of β. One is assuming that v_i has a known parametric distribution. The other does not require such assumptions.
- Here we focus on the latter one.
- Notations:

$$\bar{q}_\alpha(\cdot) = \frac{q_{\alpha_2}(\cdot) + q_{\alpha_1}(\cdot)}{2},$$

$$\Delta q_\alpha(\cdot) = q_{\alpha_2}(\cdot) - q_{\alpha_1}(\cdot),$$

$$\bar{c} = \frac{c_{\alpha_2} + c_{\alpha_1}}{2},$$

$$\Delta c = c_{\alpha_2} - c_{\alpha_1},$$

$$\gamma_1 = \frac{\bar{c}}{\Delta c}\text{: we treat } \gamma_1 \text{ as a nuisance parameter.}$$

- From $q_{\alpha_j}(X_i) = X_i^T\beta + c_{\alpha_j}\sigma(X_i)$, one can show that

$$\bar{q}_\alpha(X_i) = X_i^T\beta + \gamma_1\Delta q_\alpha(X_i) \text{ for } j = 1, 2.$$

- Chen and Khan (2000)'s estimation procedures include the following steps:
- 1st step: Estimate $q_{\alpha_j}(\cdot)$ nonparametrically. Let $\hat{q}_{\alpha_j}(\cdot)$ denote the nonparametric estimator of $q_{\alpha_j}(\cdot)$.
- 2nd step: Regress $\hat{\bar{q}}_{\alpha}(\cdot)$ on X_i and $\Delta\hat{q}_{\alpha}(\cdot)$.
- That is, the estimators of β and γ_1 are given by minimizing (w.r.t. β and γ_1)

$$\frac{1}{n}\sum_{i=1}^{n}\tau(X_i)w\left(\hat{q}_{\alpha_1}(X_i)\right)\left[\hat{\bar{q}}(X_i) - X_i^T\beta - \gamma_1\Delta\hat{q}_{\alpha}(X_i)\right]^2,$$

  where $w(\cdot)$ is a smoothing weight function [6], $\tau(\cdot)$ is a trimming function having compact support.
- Under certain regularity conditions, their estimator $\hat{\beta}$ have the parametric $\sqrt{n}$ rate of convergence, and is distributed asymptotically normally.

[6] $1(\hat{q}_{\alpha_1}(X_i) > 0)$ のかわりのようなもの.

Extension: Cosslett (2004)

- Cosslett (2004) proposes asymptotically efficient likelihood-based semiparametric estimators for censored and truncated regression models.
- See the paper for details.

The Univariate Kaplan-Meier CDF Estimator

Kaplan and Meier (1958): Product-Limit Estimator

- There exists a class of semiparametric estimators that employ the so-called Kaplan-Meier estimator of a CDF in the presence of censored data.
- Setup: Consider the following estimands:

$$\text{CDF: } F(\cdot), \text{ or}$$
$$\text{Survival function: } S(\cdot) = 1 - F(\cdot).$$

- Let $\{Y_i\}_{i=1}^n$ be the random sample of interest drawn from $F(\cdot)$.
- Let $\{L_i\}_{i=1}^n$ be random/fixed censoring variables, that are independent of $\{Y_i\}_{i=1}^n$.
- Define $Z_i = \min\{Y_i, L_i\}$, and $\delta_i = 1(Y_i \leq L_i)$. Suppose that we observe only Z_i and δ_i. By construction, we cannot observe the exact value of Y_i if $\delta_i = 0$.

cont'd

- Define the ascending points $c_0, c_1, \cdots, c_m$ at which the CDF $F(\cdot)$ or $S(\cdot)$ is to be evaluated.
- Define $I_j = 1(Y > c_j)$.
- Noting that c's are ascending and so that $I_{j-1} = 1$ if $I_j = 1$, we obtain conditional survival probability:

$$\mathbb{P}(I_j = 1 | I_{j-1} = 1) = \frac{\mathbb{P}(I_j = 1)}{\mathbb{P}(I_{j-1} = 1)} = 1 - \frac{\mathbb{P}(c_{j-1} < Y \leq c_j)}{\mathbb{P}(Y > c_{j-1})}.$$

- By choosing c_0 small enough (say, below the smallest observation in the data), we can always ensure that $\mathbb{P}(I_0 = 1) = 1$. That is, all items survive initially.

Estimation: In the Case of No Censoring

- We can estimate $\mathbb{P}(I_j = 1 | I_{j-1} = 1)$ by the iteration of

$$\tilde{\mathbb{P}}(I_j = 1 | I_{j-1} = 1) = \frac{\tilde{\mathbb{P}}(I_j = 1)}{\tilde{\mathbb{P}}(I_{j-1} = 1)} = \frac{\#\text{ of } Y_i > c_j}{\#\text{ of } Y_i > c_{j-1}}$$
$$= 1 - \frac{\#\text{ of } c_{j-1} < Y_i \le c_j}{\#\text{ of } Y_i > c_{j-1}},$$

which leads to the following estimator of survival probability:

$$\tilde{\mathbb{P}}(I_j = 1) = \prod_{s=1}^{j} \tilde{\mathbb{P}}(I_s = 1 | I_{s-1} = 1)$$
$$= \frac{\#\text{ of } Y_i > c_j}{\#\text{ of } Y_i > c_0} = \frac{\#\text{ of } Y_i > c_j}{n} = 1 - \hat{F}^n(c_j)$$

where $\hat{F}^n(c_j) = \dfrac{\#\text{ of } Y_i \le c_j}{n}$ is the empirical CDF [7].

[7] テキストでは $s = 2$ から計算することになっているが，このスライドでは，$c_0, \cdots$ という点の取り方に consistent な表記に統一した．

Estimation: With the Presence of Censoring

- Similar estimation procedures to above can be implemented: Iteration of

$$\hat{\mathbb{P}}(I_j = 1 | I_{j-1} = 1) = 1 - \frac{\#\text{ of uncensored } c_{j-1} < Y_i \leq c_j}{\#\text{ of } Y_i > c_{j-1}}$$

leads to the following estimator of survival probability:

$$\hat{S}(c_j) = \hat{\mathbb{P}}(I_j = 1) = \prod_{s=1}^{j} \hat{\mathbb{P}}(I_s = 1 | I_{s-1} = 1).$$

- The estimator of CDF is given by $\hat{F}(c_j) = 1 - \hat{S}(c_j)$ [8].

[8] Errata をみると，$s = 2$ から計算することになっているが，このスライドでは，$c_0, \cdots$ という点の取り方に consistent な表記に統一した．

Disclaimer

以降の内容は手書きのノートで替えさせていただきます.

宿題や試験に追われスライドが間に合いませんでした.

余裕があったら春休みの間に Beamer にします...

The Multivariate Kaplan-Meier CDF Estimator

Nonparametric Censored Regression