Selectivity Models

Li and Racine (2007, Chapter 10)

Yasuyuki Matsumura

January 15, 2025

Graduate School of Economics, Kyoto University

Sample Selection Issues

Sample Selection

- Sample selection issues frequently arise in empirical studies.
- We concern that the treatment effect for those "selected as treated" will differ from that for persons randomly selected from the general population.
- Pioneering parametric approaches to deal with sample selection can be found in Heckman (1976, 1979).

Semiparametric Type-2 Tobit Models

Type-2 Tobit Models

- The Type-2 Tobit model is the four equation system:

$$
\begin{aligned}
Y_{1i}^{\star} &= X_{1i}^{T}\beta_1 + u_{1i}, \\
Y_{2i}^{\star} &= X_{2i}^{T}\beta_2 + u_{2i}, \\
Y_{1i} &= 1(Y_{1i}^{\star} > 0), \\
Y_{2i} &= Y_{2i}^{\star} \times 1(Y_{1i} = 1).
\end{aligned}
$$

- The variables $(Y_{1i}^{\star}, Y_{2i}^{\star})$ are latent (unobserved).
- The observed variables are $(Y_{1i}, Y_{2i}, X_{1i}, X_{2i})$.
- Effectively, $Y_{2i}^{\star}$ is observable only when $Y_{1i} = 1$, equivalently when $Y_{1i}^{\star} > 0$.
- Typically, the second equation is of interest, e.g. the coefficient β_2.

Type-2 Tobit Models: Estimation

- The Type-2 Tobit model is a classical selection model introduced by Heckman (1976).
- It is conventional to assume that the error terms (u_{1i}, u_{2i}) are independent of $X_i = (X_{1i}, X_{2i})$.
- For details of Heckman's estimation procedure, see the attached pdf file.
- Heckman's estimation is one of the parametric approaches, as he imposes the following parametric distributional assumptions on the joint distribution of the errors:

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

Semiparametric Type-2 Tobit Models

- Here we do not impose parametric distributional assumptions on the joint distribution of the errors.
- Assume that (u_{1i}, u_{2i}) are independent of $X_i = (X_{1i}, X_{2i})$.
- Then, we obtain

$$\begin{aligned}\mathbb{E}(Y_{2i} \mid X_i, Y_{1i} = 1) &= X_{2i}^T \beta_2 + \mathbb{E}(u_{2i} \mid X_i, Y_{1i} = 1) \\ &\equiv X_{2i}^T \beta_2 + g(X_{1i}^T \beta_1),\end{aligned}$$

  where

$$g(z) = \mathbb{E}(u_{2i} \mid u_{1i} > -z) = 1 - F_{u_2|u_1}(-z),$$

  and $F_{u_2|u_1}(\cdot)$ is the conditional CDF of u_{2i} given u_{1i}.
- The functional form of $g(\cdot)$ is unknown.

- The simple regression of Y_{2i} on X_{2i} using the available data yields

$$Y_{2i} = X_{2i}^T \beta_2 + g(X_{1i}^T \beta_1) + \epsilon_{2i},$$
$$\mathbb{E}(\epsilon_{2i} \mid X_i, Y_{1i} = 1) = 0,$$

which is a partially linear single index model.

- Here we review the following estimation methods:
  - Powell (1987),
  - Ichimura and Lee (1991),
  - Gallant and Nychka (1987),
  - Heckman (1990)[1].

[1] He proposes an estimation of intercept.

Powell (1987)

- Powell (1987) proposes a two-step estimation procedure.
- Infeasible Estimation: Define $Z_i = X_{1i}^T \beta_1$. If Z_i were observed, the regression is

$$Y_{2i} = X_{2i}^T \beta_2 + g(Z_i) + \epsilon_{2i},$$

which is a partially linear model.

- This can be estimated using Robinson's approach.
- Note that the intercept is absorbed by $g(\cdot)$, and that it must be excluded from X_{2i} [2] .

[2] Recall the identification conditions of semiparametric partially linear models.

Powell (1987): Feasible Two-Step Estimation

- In practice, Z_i is not observed.
- We implement the following two-step approach.
- 1st Step: Estimate β_1 by a semiparametric binary choice estimator, or by Powell's (1984) CLAD estimator [3].
- 2nd Step: Let $\hat{\beta}_1$ denote the estimator of β_1.
- Replace Z_i with $\hat{Z}_i = X_{1i}^T \hat{\beta}_1$: $Y_{2i} = X_{2i}^T \beta_2 + g(\hat{Z}_i) + \epsilon_{2i}$.
- Estiamte β_2 and $g(\cdot)$ by Robinson's estimator.
- Note that $\hat{Z}_i = X_{1i}^T \hat{\beta}_1$ is a generated regressor, and that the asymptotic distribution will differ from the result presented in Chapter 7.

[3] CLAD 推定量を使うのは，後ほど触れる Type-3 Tobit の 2 段階推定の方が適切なのでは？

Ichimura and Lee (1991)

- Ichimura and Lee (1991) propose a joint estimator for $\theta = (\beta_1^T, \beta_2^T)^T$ based on the nonlinear regression:

$$Y_{2i} = X_{2i} + g(X_{1i}\beta_1) + \epsilon_{2i}$$

  for observations i such that Y_{2i} is observed.
- Their objective function is given by

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in \mathcal{X}) \left[Y_i - X_{2i}^T\beta_2 - \hat{g}(X_{1i}^T\beta_1)\right]^2,$$

  where

$$\hat{g}(X_{1i}^T\beta_1) = \frac{\sum_{j\neq i}(Y_{2j} - X_{2j}^T\beta_2)K_h\left(\frac{(X_{1i}-X_{1j})^T\beta_1}{h}\right)}{\sum_{j\neq i} K_h\left(\frac{(X_{1i}-X_{1j})^T\beta_1}{h}\right)}$$

  is a leave-one-out NW estimator of $\mathbb{E}(Y_{2i} - X_{2i}^T\beta_2 \mid X_{1i}^T\beta_1)$ [4].

[4] 教科書は leave-one-out になっていない.

- Their estimator is an extension of a NLLS Heckit estimator, which is based on the equation

$$Y_{2i} = X_{2i}^T \beta_2 + \sigma_{12} \lambda(X_{1i}^T \beta_1) + \epsilon_{2i}.$$

- Such estimators ignore the first equation in the system.
- This is convenient as it simplifies the estimation.
- However, ignoring relevant information reduces efficiency.
- Ichimura and Lee derive the asymptotic normality:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \text{Normal}(0, A^{-1} \Sigma A^{-1}).$$

- They also derive consistent estimators $\hat{A}$ and $\hat{\Sigma}$.

Gallant and Nychka (1987): Semi-Nonparametric MLE

- Gallant and Nychka suggest approximating the joint density of the error terms $f(u_{1i}, u_{2i})$ by a series expansion:

$$\tilde{f}(u_{1i}, u_{2i}) = \exp\left[-\frac{u_{1i}^2}{2\sigma_1^2} - \frac{u_{2i}^2}{2\sigma_2^2}\right]\left[\sum_{j=0}^{K}\sum_{k=0}^{K}\gamma_{jk}u_{1i}^j u_{2i}^k\right].$$

- $\tilde{f}(u_{1i}, u_{2i})$ is a baseline distribution of a joint normal expression.
- This is accompanied by a power series expansion allowing for a general form of the CDF.
- Using the above joint density formula, we can compute $f(u_{1i}, u_{2i})$ and then construct a log-likelihood function.
- Maximizing the log-likelihood function, we obtain estimators of β_1 and other parameters.
- The estimator has consistency under $K \to \infty$, and $\frac{K}{n} \to 0$ as $n \to \infty$.

- Coppejans and Gallant (2002) show that one can use data-driven methods to select the power series expansion terms when estimating $f(u_{1i}, u_{2i})$.
- Newey (1999) proposes a two-step series-based estimation method.
    - First, estimate β_1 efficiently.
    - Second, select β_2 solving an efficient score equation.
- For details of nonparametric series methods, see Li and Racine (2007, Chapter 15).
- 教科書にタイポが多いのと，元論文にアクセスできないのとで，不正確な内容が含まれているかもしれないため，何かおかしい点があれば指摘していただきたいです.

Heckman (1990): Intercept Estimation

- In the semiparametric Type-2 Tobit model, we cannot identify an intercept term, which cannot be separated from $g(\cdot)$.
- One might be interested in the intercept, for example, when determining "wage gaps" between unionized and non-unionized workers, or when decomposing wage differentials between different socioeconomic groups, etc $\cdots$.

- Letting μ denote the intercept, we write

$$Y_{2i}^{\star} = \mu + \tilde{X}_{2i}^{T}\delta + u_{2i},$$

where $X_{2i} = (1, \tilde{X}_{2i}^{T})^{T}$, and $\beta_2 = (\mu, \delta^{T})^{T}$.

- Then, we obtain

$$\mathbb{E}(Y_{2i} \mid X_i, Y_{1i} = 1) = \mu + \tilde{X}_{2i}^{T}\delta + g(X_{1i}^{T}\beta_1).$$

- Recall the definition of $g(\cdot)$:

$$g(z) = \mathbb{E}(u_{2i} \mid u_{1i} > -z),$$

which leads to

$$\lim_{z\to\infty} g(z) = \mathbb{E}(u_{2i}) = 0, \text{ or}$$
$$\lim_{z\to\infty} \mathbb{E}(Y_{2i} - \tilde{X}_{2i}^{T}\delta \mid Y_{1i} = 1, X_{1i}^{T}\beta_1 > z) = \mu.$$

- Heckman (1990) suggests to use the observations such that $\mathbb{E}(u_{2i} \mid Y_{1i} = 1) = g(X_{1i}^T \beta_1)$, i.e., the observations such that $g(\cdot)$ satisfies $g(-\infty) = 0$.
- Thus, μ can be rewritten as

$$\mu = \mathbb{E}(Y_{2i} - \tilde{X}_{2i}^T \delta \mid Y_{1i} = 1, X_{1i}^T \beta_1 > \gamma_n),$$

where $\gamma \to \infty$ is a bandwidth.

- This can be estimated by

$$\tilde{\mu} = \frac{\sum_{i=1}^n (Y_{2i} - \tilde{X}_{2i} \hat{\delta}) Y_{1i} 1(X_{1i}^T \hat{\beta}_1 > \gamma_n)}{\sum_{i=1}^n Y_{1i} 1(X_{1i}^T \hat{\beta}_1 > \gamma_n)}.$$

Extension: Andrews and Schafgans (1998)

- Since the indicator function $1(\cdot)$ is not differentiable, it is difficult to examine the asymptotic distribution of $\tilde{\mu}$.
- Andrews and Schafgans (1998) suggest to replace the indicator function with a smoothed non-decreasing CDF $s(\cdot)$, which satisfies

$$s(z) = 0 \text{ for } z \leq 0,$$
$$s(z) = 1 \text{ for } z \geq b \text{ for some } 0 < b < \infty, \text{ and}$$
$$s(\cdot) \text{ has third bounded derivatives.}$$

- They estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^{n}(Y_{2i} - \tilde{X}_{2i}\hat{\delta})Y_{1i}s(X_{1i}^T\hat{\beta}_1 > \gamma_n)}{\sum_{i=1}^{n} Y_{1i}s(X_{1i}^T\hat{\beta}_1 > \gamma_n)}.$$

- They find that the asymptotic distribution has a non-standard rate, depending on the distribution of $X_{1i}^T\beta_1$.

Semiparametric Type-3 Tobit Models

Type-3 Tobit Models

- The Type-3 Tobit model is the four equation system:

$$\begin{aligned} Y_{1i}^{\star} &= X_{1i}^T \beta_1 + u_{1i}, \\ Y_{2i}^{\star} &= X_{2i}^T \beta_2 + u_{2i}, \\ Y_{1i} &= \max\{Y_{1i}^{\star}, 0\}, \\ Y_{2i} &= Y_{2i}^{\star} \times 1(Y_{1i} > 0). \end{aligned}$$

- The difference from the Type-2 is that Y_{1i} is censored than binary.
- We observe $Y_{2i}^{\star}$ only when there is no censoring on $Y_{1i}^{\star}$.
- Typically, the second equation is of interest, e.g. the coefficient β_2.

Type-3 Tobit Models: Parametric Approaches

- Parametric approaches to estimate the Type-3 Tobit models impose parametric distributional assumptions on the joint distribution of the errors.
- Vella (1992, 1998)
- Wooldridge (1994)
- See the attached file for details.

Semiparametric Type-3 Tobit Models

- Here we do not assume that the joint distribution of (u_{1i}, u_{2i}) is known.
- Instead, we have $\mathbb{E}(u_{2i} \mid u_{1i}) = g(u_{1i})$, where $g(\cdot)$ is an unknown function.
- In this case, it is easy to see that $\mathbb{E}(Y_{2i} \mid X_i, u_{1i}) = X_{2i}^T \beta_2 + g(u_{1i})$.
- Thus, we obtain

$$
\begin{aligned}
Y_{2i} &= X_{2i}^T \beta_2 + g(u_{1i}) + v_{2i}, \\
\mathbb{E}(v_{2i} \mid u_{1i}, Y_{1i} > 0) &= 0.
\end{aligned}
$$

Semiparametric Type-3 Tobit Models: Estimation

- If u_{1i} were known, this would be a partially linear model.
- In practice, u_{1i} is unknown.
- We can estimate u_{1i} by Tobit, CLAD, etc $\cdots$.
- Here we review the following estimation methods:
  - Li and Wooldridge (2002),
  - Chen (1997),
  - Honore, Kyriazidou and Udry (1997),
  - Lee (1994),
  - the semiparametric Type-2 Tobit estimator besed on Ichimura (1993); and Ichimura and Lee (1991).

Li and Wooldridge (2002)

- Li and Wooldridge suggest a multistep method to estimate β_2.
- 1st Step: Estimate β_1, for example, by Powell's (1984) CLAD estimator [5] . We assume that for the 1st step there is a $\sqrt{n}$-consistent, and asymptotically normally distributed, estimator for β_1.
- 2nd Step: Replacing u_{1i} with $\hat{u}_{1i}$, we implement Robinson's approach to estimate β_2.
- They establish the $\sqrt{n}$-normality of their estimator $\hat{\beta}_2$:

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \xrightarrow{d} \text{Normal}(0, \Sigma).$$

- Avar$(\hat{\beta}_2)$ can be consistently estimated.

[5] Powell's CLAD (censored least absolute deviation) estimator will be reviewed below in Chapter 11.

- Li and Wooldridge's (2002) estimator does not achieve the semiparametric efficiency bound. Efficient estimation can usually be achieved by a one-step procedure, where β_1 and β_2 are estimated simultaneously as in Ai (1997).
- Powell's CLAD estimator is a parametric approach, i.e., the generated regressor is estimated from a parametric model. Ahn and Powell (1993) suggest to estimate β_1 in the 1st stage using a nonparametric regression model.

Chen (1997)

- Assume that (u_{1i}, u_{2i}) is independent of (X_{1i}, X_{2i}).
- Under the above assumption, it holds that

$$\begin{aligned}
&\mathbb{E}(Y_{2i} \mid X_{1i}, X_{2i}, u_{1i} > 0, X_{1i}^T\beta_1 > 0, Y_{1i} > 0) \\
&\qquad = \mathbb{E}(Y_{2i} \mid u_{1i} > 0, X_i) \\
&\qquad = X_{2i}^T\beta_2 + \alpha_0,
\end{aligned}$$

where α_0 is a constant. Note that α_0 is not the intercept of the original model.

- 1st Step: Estimate β_1 consistently by Honore and Powell (1984), or by Powell's CLAD. Let $\hat{\beta}_1$ denote the consistent estimator of β_1.
- 2nd Step: Run the following least squares:

$$\min_{\beta_2,\alpha} \frac{1}{n}\sum_{i=1}^{n} 1\{Y_{1i} - X_{1i}^T\hat{\beta}_1 > 0, X_{1i}^T\hat{\beta}_1 > 0\}(Y_{2i} - X_{2i}^T\beta_2 - \alpha)^2.$$

- A problem arising with Chen's (1997) estimator is that it may trim out too many observations, which leads to inefficient estimation.
- Chen (1997) suggests an alternative estimator that trims far fewer data points in finite-sample applications.

Honore, Kyriazidou and Udry (1997)

- Honore, Kyriazidou and Udry (1997) suggest to relax the normality assumption as it can be seen in Heckman (1979).
- Instead, they assume that the distribution of the error terms (u_{1i}, u_{2i}) given regressors X_i is symmetric, with arbitary heteroscedasticity permitted.
- In this case, conditional on the sample selection, the conditional distribution is no longer symmetric.
- Their basic idea is that u_{2i} is symmetrically distributed around 0 if one estimate β_2 using observations for which $-X_{1i}^T\beta_1 < u_{1i} < X_{1i}^T\beta_1$ (i.e., $0 < Y_{1i} < 2X_{1i}^T\beta_1$).

- Honore, Kyriazidou and Udry (1997) suugest the following estimation method.
- 1st Step: Estimate β_1 consistently, for example, by Powell's CLAD. Let $\hat{\beta}_1$ denote the consistent estimator of β_1.
- 2nd Step: Run the following least absolute deviations:

$$\min_{\beta_2} \frac{1}{n} \sum_{i=1}^{n} 1\{0 < Y_{1i} < 2X_{1i}^T \hat{\beta}_1\} \mid Y_{2i} - X_{2i}^T \beta_2 \mid .$$

- They establish the $\sqrt{n}$-normality of their estimator.

Lee (1994)

Under the assumption of independence between the errors and the regressors, Lee (1994, equation 2.12) shows that

$$\begin{aligned} Y_{2i} - \mathrm{E}(Y_2 | u_1 > -X'_{1i}\beta_1, X'_1\beta > X'_{1i}\beta_1) \\ = [X'_{2i} - \mathrm{E}(X'_2 | X'_1\beta_1 > X'_{1i}\beta_1)]\beta_2 + u_{2i}, \end{aligned} \quad (10.25)$$

where u_{2i} satisfies $\mathrm{E}(u_{2i} | u_1 > -X'_{1i}\beta_1, X'_1\beta > X'_{1i}\beta_1) = 0$. Lee suggests first replacing the conditional expectations in (10.25) by kernel estimators (also β_1 needs to be replaced by a first stage estimator) and then applying a least squares procedure to estimate β_2 (which we denote by $\hat{\beta}_2$,*Lee*). Lee establishes the asymptotic normality of $\hat{\beta}_2$,*Lee*.

Comparing the 4 Estimators

	LW	Chen	HKU	Lee
Kernel Methods	Required	-	-	Required
Smoothing Parameter Choice	Insensitive	-	-	Insensitive

- In general, nonparametric kernel methods are sensitive to the choice of smoothing parameters.
- Lee (1994); Min, Sheu and Wang (2003) suggest by Monte-Carlo simulations that the estimators of LW(2002) and Lee (1994) are fairly insensitive to the choice of smoothing parameters.
- The reason is that the semiparametric estimators depend on the average of nonparametric estimators, which are less sensitive to different values of smoothing parameters than a pointwise nonparametric kernel estimator.

Christofides, Li, Liu and Min (2003)

	LW	Chen	HKU	Lee
Dependence Assumption	Yes	Yes	No	Yes
Symmetry Assumption	No	No	Yes	No

- "Dependence Assumption" means that one need to assume that (u_{1i}, u_{2i}) is independent of (X_{1i}, X_{2i}).
- "Symmetry Assumption" means that one need to assume that the distribution of the error terms (u_{1i}, u_{2i}) given regressors X_i is symmetric.
- The symmetry condition is neither weaker nor stronger than the independence condition.

Tests for the Exisitence of Sample Selection and Model Specification

Test for the Exisitence of Sample Selection

- Let us test whether sample selection exists or not.
- Consider the following null hypothesis:

$$\mathbb{H}_0^a : \mathbb{E}(u_2|u_1) = 0,$$
$$\mathbb{H}_1^a : \mathbb{E}(u_2|u_1) \neq 0.$$

- A test statistic for $\mathbb{H}_0^a$ is proposed by Li and Wang (1998); and Zheng (1996):

$$\tilde{I}_n^a = \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{j \neq i}^{n_1} \hat{u}_{2i}\hat{u}_{2j} K_h(\hat{u}_{1i}\hat{u}_{1j}),$$

where n_1 denotes the number of observations with $Y_{2i} > 0$, and $\hat{u}_{1i}$ and $\hat{u}_{2i}$ are the OLS residuals.

- Under conditions 10.1, 10.2, and 10.3, we obtain (as $n_1 \to \infty$, $h \to 0$, $n_1 h \to \infty$),

$$n_1 h^{\frac{1}{2}} \frac{\tilde{I}_n^a}{\hat{\sigma}_a} \xrightarrow{d} \text{Normal}(0,1) \text{ under } \mathbb{H}_0^a$$

and

$$\mathbb{P}\left(\left| n_1 h^{\frac{1}{2}} \frac{\tilde{I}_n^a}{\hat{\sigma}_a} \right| > C \right) \to 1 \text{ for any } C > 0 \text{ under } \mathbb{H}_1^a,$$

where $\hat{\sigma}_a^2 = \dfrac{2h}{n_1(n_1-1)} \displaystyle\sum_{i=1}^{n_1} \sum_{j \neq i}^{n_1} \hat{u}_{2i}^2 \hat{u}_{2j}^2 K_h^2(\hat{u}_{1i} - \hat{u}_{1j})$.

Test for Model Specification

- Suppose that we reject $\mathbb{H}_0^a$, that is, we consider there exists sample selection.
- Consider the following null hypothesis:

$$\mathbb{H}_0^b : \mathbb{E}(Y_2|X_2, u_1) = X_2^T\beta_2 + u_1\gamma,$$
$$\mathbb{H}_1^b : \mathbb{E}(Y_2|X_2, u_1) = X_2^T\beta_2 + g(u_1) \text{ with } g(u_1) \neq u_1\gamma.$$

- A test statistic for $\mathbb{H}_0^a$ is proposed by Li and Wang (1998):

$$\tilde{I}_n^b = \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{j\neq i}^{n_1} \tilde{u}_{2i}\tilde{u}_{2j} K_h(\hat{u}_{1i}\hat{u}_{1j}),$$

where n_1 denotes the number of observations with $Y_{2i} > 0$, and $\hat{u}_{1i}$ and $\tilde{u}_{2i}$ are the residuals from OLS β_1 and Li and Wooldridge's (2002) β_2, respectively.

- Under conditions 10.1, 10.2, and 10.3, we obtain (as $n_1 \to \infty$, $h \to 0$, $n_1 h \to \infty$),

$$n_1 h^{\frac{1}{2}} \frac{\tilde{I}_n^b}{\hat{\sigma}_b} \xrightarrow{d} \text{Normal}(0,1) \text{ under } \mathbb{H}_0^b$$

and

$$\mathbb{P}\left(\left| n_1 h^{\frac{1}{2}} \frac{\tilde{I}_n^b}{\hat{\sigma}_b} \right| > C \right) \to 1 \text{ for any } C > 0 \text{ under } \mathbb{H}_1^b,$$

where $\hat{\sigma}_b^2 = \frac{2h}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{j \neq i}^{n_1} \tilde{u}_{2i}^2 \tilde{u}_{2j}^2 K_h^2(\hat{u}_{1i} - \hat{u}_{1j})$.

Nonparametric Sample Selection Model

Das, Newey, and Vella (2003)

- See Das, Newey, and Vella (2003); and Li and Racine (2007, Section 10.4).