9 Selectivity Models

9.1 Semiparametric Selection Models

The Type-2 Tobit model is the four equation system

$$\begin{aligned} y_{1i}^* &= X_{1i}'\beta_1 + u_{1i} \\ y_{2i}^* &= X_{2i}'\beta_2 + u_{2i} \\ y_{1i} &= 1\left(y_{1i}^* > 0\right) \\ y_{2i} &= y_{2i}^* 1\left(y_{1i} = 1\right) \end{aligned}$$

The variables (y_{1i}^*, y_{2i}^*) are latent (unobserved). The observed variables are $(y_{1i}, y_{2i}, X_{1i}, X_{2i})$. Effectively, y_{2i}^* is observed only when $y_{1i} = 1$, equivalently when $y_{1i}^* > 0$.

The Type-3 Tobit model is the four equation system

$$\begin{aligned} y_{1i}^* &= X_{1i}'\beta_1 + u_{1i} \\ y_{2i}^* &= X_{2i}'\beta_2 + u_{2i} \\ y_{1i} &= \max\{y_{1i}^*, 0\} \\ y_{2i} &= y_{2i}^* 1\left(y_{1i} > 0\right)) \end{aligned}$$

The difference is that y_{1i} is censored rather than binary. We observe y_{2i}^* only when there is no censoring.

Typically the second equation is of interest, e.g. the coefficient β_2.

The type-2 model is the classic selection model introduced by Heckman.

It is conventional to assume that the errors (u_{1i}, u_{2i}) are independent of $X_i = (X_{1i}, X_{2i})$.

As you recall from 710, Heckman showed that if we try to estimate β_2 by a regression using the available data, this is estimating the regression of y_{2i} on X_{2i}, conditional on $y_{1i} > 0$, which is

$$\begin{aligned} E\left(y_{2i} \mid X_i, y_{1i} = 1\right) &= X_{2i}'\beta_2 + \mathrm{E}\left(u_{2i} \mid X_i, y_{1i} > 0\right) \\ &= X_{2i}'\beta_2 + \mathrm{E}\left(u_{2i} \mid u_{1i} > -X_{1i}'\beta_1\right) \\ &= X_{2i}'\beta_2 + g\left(X_{1i}'\beta_1\right) \end{aligned}$$

for some function $g\left(z\right)$. When (u_{1i}, u_{2i}) are bivariate normal then $g(u)$ is a scaled inverse Mill's ratio. But when the errors are non-normal, the functional form of $g(z)$ is unknown and nonparametric. The one constraint it satisfies is

$$\lim_{z\to\infty} g(z) = \lim_{z\to\infty} \mathrm{E}\left(u_{2i} \mid u_{1i} > -z\right) = \mathrm{E}\left(u_{2i}\right) = 0$$

by normalization.

We can then write the regression for y_{2i} as

$$\begin{aligned} y_{2i} &= X_{2i}'\beta_2 + g\left(X_{1i}'\beta_1\right) + e_i \\ \mathrm{E}\left(e_i \mid X_i, y_{1i} > 0\right) &= 0 \end{aligned}$$

This is a partially linear single index model.

9.2 Two-Step Estimator

This method is developed in a working paper by Powell (1987) and in Li and Wooldridge (Econometric Theory, 2002)

Define $Z_i = X_{1i}'\beta_1$. If Z_i were observed the regression is

$$y_{2i} = X_{2i}'\beta_2 + g\left(Z_i\right) + e_i$$

which is a partially linear model, and can be estimated using Robinson's approach.

For the parially linear model, the intercept is absorbed by g, so it must be excluded from X_{2i}.

Since Z_i is not observed, we can use a two-step approach.

In step 1, β_1 is estimated by a semiparametric estimator, say $\hat{\beta}_1$. (A semiparametric binary choice estimator from the previous section, or a semiparametric Tobit estimator from the next section.) Set $\hat{Z}_i = X_{1i}'\hat{\beta}_1$.

In the second step, β_2 and g are estimated by Robinson's estimator (using the observations for which $y_{1i} = 1$)

Since the second step uses on the generated regressor $\hat{Z}_i = X_{1i}'\hat{\beta}_1$, the asymptotic distribution is affected.

From the text

$$\sqrt{n}\left(\hat{\beta}_2 - \beta_2\right) \to_d N\left(0, Q^{-1}\left(\Omega_1 + \Psi\Omega_2\Psi'\right)Q^{-1}\right)$$

the covariance terms defined in the text, as typical for two-step estimators.

9.3 Ichimura and Lee's Estimator

Ichimura and Lee (1991) proposed a joint estimator for $\beta = (\beta_1, \beta_2)$ based on the nonlinear regression

$$y_{2i} = X_{2i}'\beta_2 + g\left(X_{1i}'\beta_1\right) + e_i$$

for observations i such that y_{2i} is observed. Thus the first equation is ignored.

Their criterion is

$$S_n\left(\beta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - X_{2i}'\beta_2 - \hat{g}\left(X_{1i}'\beta_1, \beta_2\right)\right)^2 1_i(b)$$

where $1_i(b)$ is a trimming function

$$\hat{g}\left(X_{1i}'\beta_1, \beta\right) = \frac{\sum_{j\neq i} k\left(\frac{(X_{1i}-X_{1j})'\beta_1}{h}\right)(y_{2i}-X_{2i}'\beta_2)}{\sum_{j\neq i} k\left(\frac{(X_{1i}-X_{1j})'\beta_1}{h}\right)}$$

is a leave-one-out NW estimator of $\mathrm{E}\left(y_{2i}-X_{2i}'\beta_2 \mid X_{1i}'\beta_1\right)$. (Againl, this is computed only for the observations for which y_{2i} is observed.)

This works, but it ignores the first equation of the system. It is a semiparametric extension of a NLLS Heckit estimator based on the equation

$$y_{2i} = X_{2i}'\beta_2 + \sigma_{12}\lambda\left(X_{1i}'\beta_1\right) + e_i.$$

Such estimators ignore the first equation. This is convenient as it simplifies estimation, but ignoring relevant information reduces efficiency. My view is that identification of β_2 versus $g(X_1'\beta_1)$ is based on (dubious) exclusion restrictions plus assuming linearity of the first part.

9.4 Powell's Estimator

An alternative creative estimator was propsoed by Powell (1987, unpublished working paper), reviewed in his chapter 41 of the Handbook of Econometrics.

As for Ichimura and Lee, we ignore the first equation and only consider observations for which y_{2i} are observed.

Take two observations i and j

$$\begin{aligned} y_{2i} &= X_{2i}'\beta_2 + g\left(Z_i\right) + e_i \\ y_{2j} &= X_{2j}'\beta_2 + g\left(Z_j\right) + e_j \end{aligned}$$

and their pairwise difference

$$y_{2i} - y_{2j} = \left(X_{2i} - X_{2j}\right)'\beta_2 + g\left(Z_i\right) - g\left(Z_j\right) + e_i - e_j$$

Now focus on observations for which $Z_i \simeq Z_j$. For these observations, $g\left(Z_i\right) - g\left(Z_j\right) \simeq 0$ and β_2 can be estimated by a regression of $y_{2i} - y_{2j}$ on $X_{2i} - X_{2j}$.

This is made operational by using a kernel for $Z_i - Z_j$, and by replacing Z_i with $\hat{Z}_i = X_{1i}'\hat{\beta}_1$, where $\hat{\beta}_1$ is a first stage estimate of β_1, yielding

$$\begin{aligned}\hat{\beta}_2 &= \left(\sum_i \sum_{j\neq i} k\left(\frac{(X_{1i}-X_{1j})'\hat{\beta}_1}{h}\right)(X_{2i}-X_{2j})(X_{2i}-X_{2j})'\right)^{-1} \\ &\quad \cdot \sum_i \sum_{j\neq i} k\left(\frac{(X_{1i}-X_{1j})'\hat{\beta}_1}{h}\right)(X_{2i}-X_{2j})(y_{2i}-y_{2j})\end{aligned}$$

Unfortunately Powell didn't publish the original paper. This type of estimator effectly identifies β_2 from a small subset of observations, so is unlikely to be precise. The good side is that the nonparametric function g does not need to be estimated in any sense.

9.5 Estimation of the Intercept

While the conditional equation

$$y_{2i} = X_{2i}'\beta_2 + g(Z_i) + e_i$$

excludes an intercept (it is absorbed in g) the original equation of interest

$$y_{2i}^* = \mu + X_{2i}'\beta_2 + u_{2i}$$

say, contains an intercept. Its value can be relevant in practice. That is, the parameters of interest for policy evaluation may be (μ, β_2), not (β_2, g).

To estimate μ, Heckman (AER, 1990) suggest using the observation that the function g satisfies $g(-\infty) = 0$. Thus

$$\mu = \mathrm{E}\left(y_{2i} - X_{2i}'\beta_2 \mid y_{1i} = 1, X_{1i}'\beta_1 = \infty\right) \simeq \mathrm{E}\left(y_{2i} - X_{2i}'\beta_2 \mid y_{1i} = 1, X_{1i}'\beta_1 > \gamma_n\right)$$

where $\gamma_n \to \infty$ is a bandwidth. This can be estimated by

$$\hat{\mu} = \frac{\sum_{i=1}^n 1\left(X_{1i}'\hat{\beta}_1 > \gamma_n\right)(y_{2i} - X_{2i}'\beta_2)}{\sum_{i=1}^n 1\left(X_{1i}'\hat{\beta}_1 > \gamma_n\right)}$$

where the sample is only for those observations for which y_{2i} is observed (those for which $y_{1i} = 1$).

Notet that this estiamtor depends on the first-step estimate $\hat{\beta}_1$.

Andrews and Schafgans (1998, Review of Economic Studies) suggest that a better estimator is obtained by relacing the indicator variable by a DF kernel. They find that the asymptotic distribution has a non-standard rate, depending on the distribution of $X_{1i}'\beta_1$.