8.1.8 Sample selection

A common problem involves estimation of β for the model

$$Y^* = X\beta + \varepsilon$$

however sample selection results in Y being observed only for individuals receiving treatment (when $D = 1$). The data are censored but not at a fixed value (as in a Tobit problem; see chapter 5). Treating sample selection D as an exogenous variable is inappropriate if the unobservable portion of the selection equation, say V_D, is correlated with unobservables in the outcome equation ε.

Heckman [1974, 1976, 1979] addressed this problem and proposed the classic two stage approach. In the first stage, estimate the selection equation via probit. Identification in this model does not depend on an exclusion restriction (Z need not include variables appropriately excluded from X) but if instruments are available they're likely to reduce collinearity issues.

To fix ideas, identification conditions include

Condition 8.4 (X, D) *are always observed,* Y_1 *is observed when* $D = 1$ *(*$D^* > 1$*),*

Condition 8.5 (ε, V_D) *are independent of* X *with mean zero,*

Condition 8.6 $V_D \sim N(0, 1)$*,*

Condition 8.7 $E[\varepsilon \mid V_D] = \gamma_1 V_D$.[11]

The two-stage procedure estimates θ from a first stage probit.

$$D^* = Z\theta - V_D$$

These estimates $\widehat{\theta}$ are used to construct the inverse Mills ratio $\lambda_i = \frac{\phi(Z_i\widehat{\theta})}{\Phi(Z_i\widehat{\theta})}$ which is utilized as a covariate in the second stage regression.

$$Y_1 = X\beta + \gamma\lambda + \eta$$

where $E[\eta \mid X, \lambda] = 0$. Given proper specification of the selection equation (including normality of V_D), Heckman shows that the two-step estimator is asymptotically consistent (if not efficient) for β, the focal parameter of the analysis.[12]

---

[11] Bivariate normality of (ε, V_D) is often posed, but strictly speaking is not required for identification.

[12] It should be noted that even though Heckman's two stage approach is commonly employed to estimate treatment effects (discussed later), treatment effects are not the object of the sample selection model. In fact, since treatment effects involve counterfactuals and we have no data from which to identify population parameters for the counterfactuals, treatment effects in this setting are unassailable.

A semi-nonparametric alternative

Concern over reliance on normal probability assignment to unobservables in the selection equation as well as the functional form of the outcome equation, has resulted in numerous proposals to relax these conditions. Ahn and Powell [1993] provide an alternative via their semi-nonparametric two stage approach. However, nonparametric identification involves an exclusion restriction or, in other words, at least one instrument. That is, (at least) one variable included in the selection equation is properly omitted from the outcome equation. Intuitively, this is because the selection equation could be linear and the second stage would then involve colinear regressors. Ahn and Powell propose a nonparametric selection model coupled with a partial index outcome (second stage) model. The first stage selection index is estimated via nonparametric regression

$$\widehat{g}_i = \frac{\sum_{j=1}^{n} K\left(\frac{w_i - w_j}{h_1}\right) D_j}{\sum_{j=1}^{n} K\left(\frac{w_i - w_j}{h_1}\right)}$$

The second stage uses instruments Z, which are functions of W, and the estimated selection index.

$$\widehat{\beta} = \left[\widehat{S}_{XX}\right]^{-1} \widehat{S}_{XY}$$

where

$$\begin{aligned} \widehat{S}_{XX} &= \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \widehat{\omega}_{ij} \left(z_i - z_j\right) \left(x_i - x_j\right)^T \\ \widehat{S}_{XY} &= \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \widehat{\omega}_{ij} \left(z_i - z_j\right) \left(y_i - y_j\right) \end{aligned}$$

and

$$\widehat{\omega}_{ij} = \frac{1}{h_2} K\left(\frac{\widehat{g}_i - \widehat{g}_j}{h_2}\right) D_i D_j$$

Ahn and Powell show the instrumental variable density-weighted average derivative estimator for β achieves root-n convergence (see the discussion of nonparametric regression and Powell, Stock, and Stoker's [1989] instrumental variable density-weighted average derivative estimator in chapter 6).

8.1.9 Duration models

Sometimes the question involves how long to complete a task. For instance, how long to complete an audit (internal or external), how long to turn around a distressed business unit or firm, how long to complete custom projects, how long will a recession last, and so on. Such questions can be addressed via duration models.