ARE213 **Econometrics**

Fall 2004 **UC Berkeley Department of Agricultural and Resource Economics**

Limited Dependent Variable Models II: Selection Models (W 17.4.1)

1. The Model

In this lecture we study *selection models*. Typically they consist of two equations, one outcome equation describing the relation between an outcome of interest Y_i and a vector of covariates X_i, and the second, the selection equation, describing the relation between a binary participation decision D_i and another vector of covariates Z_i. There are various forms of these models. Here we consider a specific case, originally studied by Heckman (1979).

$$Y_i = X_i'\beta + \varepsilon_i, \tag{1}$$

$$D_i = 1\{Z_i'\gamma + \eta_i > 0\}. \tag{2}$$

The parametric form of the model assumes that

$$\left(\begin{array}{c} \varepsilon_i \\ \eta_i \end{array} \right) \Big| X_i, Z_i \;\sim\; \mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma_\varepsilon^2 & \rho \cdot \sigma_\varepsilon \\ \rho \cdot \sigma_\varepsilon & 1 \end{array} \right) \right). \tag{3}$$

The variance for η_i is normalized to one since we only observe the sign of $X_i'\gamma + \eta_i$. For a random sample from the population we observe D_i, Z_i, and X_i. Only for observations with $D_i = 1$ do we observe Y_i.

This model is known as the Heckman selection model, or the type II Tobit model (Amemiya), or the probit selection model (Wooldridge). Variations include the case where $Z_i'\gamma + \eta_i$ is observed if $D_i = 1$, so that the selection equation is not probit but tobit. That case is referred to as type II tobit by Amemiya and the tobit selection model by Wooldridge.

The classic example is a wage equation, where we only observe the age if the individual decided to work ($D_i = 1$). Unlike in the Tobit case non participation does not imply that

Y_i is negative. In fact, the Tobit model is a special case of this model with $\varepsilon_i = \eta_i$, $\gamma = \beta$, and thus $D_i = 1\{Y_i \geq 0\}$. In the wage example we think that those who participate in the labor market get relatively high wages compared to those who decided not to participate. If the selection equation had hours worked, with the actual number of observed if hours is positive, we would have the tobit selection model.

Another example is that of people buying life insurance (see Wooldridge). We are interested in the relation between the price people pay for life insurance and their characteristics. However, we only observe the price of life insurance for those who purchase it. We do not know what price people who choose not to purchase life insurance would have paid, had they done so. The selection equation models the decision to purchase life insurance. Here we may be concerned that those who did purchase the life insurance (and thus who had relatively high values of η) paid different prices from those who did not. Specifically, those who purchase life insurance may be less healthy, and that may imply they pay relatively high prices for life insurance, conditional on covariates if these do not adequately control for health status.

The first key issue is that the disturbances in the two equations are potentially correlated. If they are known not to be correlated, ($\rho = 0$), then conditional on Z_i and $D_i = 1$ it would still be the case that ε_i is independent of X_i, and so we could just do least squares on the complete observations.

Second, is whether Z_i and X_i are the same. If not, we have exclusion restrictions. In particular it is important whether we have variables in Z_i that are not in X_i. That is, variables that affect participation but not the outcome directly.

2. Maximum Likelihood Estimation

The likelihood function is tricky to derive. First consider observations with $D_i = 0$. For these units we only know that $Z_i'\gamma + \eta_i < 0$. Although we (possibly) observe X_i, we know nothing about Y_i. Hence the likelihood contribution for these units is just the probabilty that $D_i = 0$:

$$\Pr(D_i = 0|Z_i, X_i) = \Pr(\eta_i < -Z_i'\gamma) = \Phi(-Z_i'\gamma) = 1 - \Phi(Z_i'\gamma).$$

Next consider the observations with $D_i = 1$. Rather than look at the probability that $D_i = 1$ times the conditional density of Y_i given $D_i = 1$, we look at the marginal density of Y_i (normal with mean $X_i'\beta$ and variance σ_ε^2, times the conditional probability of $D_i = 1$ given Y_i. So, the first factor is

$$\frac{1}{\sigma_\varepsilon} \cdot \phi((Y_i - X_i'\beta)/\sigma_\varepsilon).$$

The conditional distribution of η_i given Y_i and X_i is normal with mean $(\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta)$, and variance $1 - \rho^2/\sigma_\varepsilon^2$. Thus the probability that $D_i = 1$ given Y_i, X_i, and Z_i is

$$\Pr(D_i = 1|X_i, Z_i, Y_i) = \Pr(Z_i'\gamma + \eta_i > 0|Y_i, X_i, Z_i)$$

$$= \Pr(\eta_i > -Z_i'\gamma|Y_i, X_i, Z_i)$$

$$= \Pr\left(\eta_i - (\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta) > -Z_i'\gamma - (\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta) \middle| Y_i, X_i, Z_i\right)$$

$$= \Pr\left(\frac{\eta_i - (\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta)}{\sqrt{1 - \rho^2/\sigma_\varepsilon^2}} > \frac{-Z_i'\gamma - (\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta)}{\sqrt{1 - \rho^2/\sigma_\varepsilon^2}} \middle| Y_i, X_i, Z_i\right)$$

$$= 1 - \Phi\left(\frac{-Z_i'\gamma - (\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta)}{\sqrt{1 - \rho^2/\sigma_\varepsilon^2}}\right)$$

$$= \Phi\left(\frac{Z_i'\gamma + (\rho/\sigma_\varepsilon) \cdot (Y_i - X_i'\beta)}{\sqrt{1 - \rho^2/\sigma_\varepsilon^2}}\right).$$

Combining all these parts leads to the following log likelihood function:

$$L(\beta, \gamma, \sigma_\varepsilon^2, \rho) = \sum_{i=1}^{N} (1 - D_i) \cdot \ln(1 - \Phi(Z_i'\gamma))$$

$$+D_i \cdot \left(\ln \Phi \left(\left(Z_i'\gamma + \frac{\rho}{\sigma_\varepsilon}(Y_i - X_i'\beta) \right) \cdot (1 - \rho/\sigma_\varepsilon)^{-1/2} \right) + \ln \phi \left((Y_i - X_i'\beta) / \sigma_\varepsilon \right) - \ln \sigma_\varepsilon \right).$$

Maximizing this is messy, with the terms for observations with $D_i = 1$ consisting of the logarithm of a sum. It is possible, but in addition to the difficulty of calculating the derivatives, the computational problem tends to be somewhat badly behaved, so that iterative methods do not always converge to the maximum likelihood estimator.

3. Heckman Two-step Estimator

Heckman proposed a different estimator. First note that because of the normality assumption we have

$$\mathbb{E}[\varepsilon_i|\eta] = \delta \cdot \eta_i, \tag{4}$$

where $\delta = \rho \cdot \sigma_\varepsilon$. Thus we have

$$\mathbb{E}[Y_i|X_i, \eta_i] = X_i'\beta + \delta \cdot \eta_i.$$

In addition,

$$\mathbb{E}[\eta_i|X_i, Z_i, D_i = 1] = \mathbb{E}[\eta_i|X_i, Z_i, Z_i'\gamma + \eta_i > 0]$$

$$= \mathbb{E}[\eta_i|\eta_i > -Z_i'\gamma] = \lambda(Z_i'\gamma),$$

where

$$\lambda(a) = \phi(a)/\Phi(a),$$

is the invese Mill's ratio. Thus

$$\mathbb{E}[Y_i|X_i, \eta_i] = X_i'\beta + \delta \cdot \lambda(Z_i'\gamma).$$

Heckman's idea is the following. First estimate γ by probit maximum likelihood. This works well (much easier than doing the full selection model by maximum likelihood). Then calculate for each unit with $D_i = 1$ the inverse Mill's ratio $\hat{\lambda}_i = \lambda(Z_i'\hat{\gamma})$, and regress Y_i on X_i and $\hat{\lambda}_i$.

This is a relatively straightforward way of getting a point estimate for β. However, there are some disadvantages relatively to maximum likelihood. First, the estimator is not necessarily efficient, whereas maximum likelihood is. Second, getting the variance is not easy in general. You have to take account of the fact that $\hat{\gamma}$ in the inverse Mill's ratio is estimated. In one simple case it is not so difficult. If we just want to test whether there is a selection problem, and get the test statistic under the null of no selection ($\delta = 0$), we do not have to take account of the fact that γ is estimated, and we can simply use ols standard errors.

4. The Case Without Exclusion Restrictions

Formally, we do not need exclusion restrictions, and Z_i can be identical to X_i. In practice you are likely to get close to perfect collinearity, and will end up with large standard errors. The identification in this case comes purely from the functional form. That is, λ_i is a nonlinear function of X_i, and the conditional expectation of Y given X ends up being nonlinear in X_i, with the nonlinear part interpreted as selection bias. Typically we are not so sure about the functional form that we would be comfortable just interpreting nonlinearities as evidence for endogeneity of the covariates.

With exclusion restrictions (variables in Z_i that are not in X_i), the sensitivity is much less of an issue. In that case there is variation in λ_i conditional on X_i, so the selection bias coefficient is separately identified. In fact for these cases there are identification results that do not rely on normality. However, just as in instrumental variables settings exclusion restrictions are often difficult to motivate. Why should a variable Z_i affect the decision to participate, if it is not related to the outcome of interest. Examples of exclusion restrictions that have been used in the female wage equation are presence and age of children. However, without sufficient additional control variables, these may affect wages directly through human capital accumulation arguments.

Even in this case identification is controversial though. See the paper by Little for a skeptical view from a statistician on these models.

References

Amemiya, T. (1985), *Advanced Econometrics,* Harvard University Press, Cambridge, MA.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica,* Vol. 47, No. 1 153-162.

Little, R., (1985), "A Note about Models for Selectivity Bias," *Econometrica,* Vol. 53, No. 6, 1469-1474.