Selection Models

- The Tobit model is a regression model subject to a common censoring threshold, where the dependent variable is censoring with probability one when it falls below the threshold. Maddala (1983) calls this non-stochastic censoring.
- The Selection model, which Amemiya (1985) classifies as a "Type 2 Tobit" model, is characterized by stochastic censoring. The censoring outcomes may differ for observations with identical values for the dependent variable in the regression equation.

Selection Models

The Selection model is a bivariate system where observation of the dependent variable in the "regression equation" is governed by the sign of the dependent variable in a distinct "selection equation." Specifically,

$$Y_i = X_i\beta + \epsilon_i$$
$$U_i = V_i\pi + \omega_i$$

where $(\epsilon_i, \omega_i) \sim iidN(0, \Sigma)$ and

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma\lambda \\ \rho\sigma\lambda & \lambda^2 \end{bmatrix}$$

Selection Models

These equations satisfy full classical conditions except:

1. U_i is unobserved. Instead, we observe a censoring indicator, J_i, where

$$\begin{array}{lll} J_i = 1 & \text{if} & U_i > 0 \\ J_i = 0 & \text{if} & U_i \leq 0 \end{array}$$

2. Y_i is observed only if $J_i = 1$. Its value is missing if $J_i = 0$.
3. The vector of regressors, X_i and V_i, are observed for both the $J_i = 1$ and $J_i = 0$ cases.

Subsample OLS

What would be the consequence of estimating the regression equation by OLS using the subsample of observations for which we have complete information?

Consider the mean of the subsample disturbance term.

$$\begin{array}{lll} E(\epsilon_i|J_i = 1) = E(\epsilon_i|U_i > 0) & & \text{def. of } J_i \\ = E(\epsilon_i|V_i\pi + \omega_i > 0) & & \text{def. of } U_i \\ = E(\epsilon_i|\omega_i > -V_i\pi) & & \text{algebra} \\ = E\left(\epsilon_i \Big| \frac{\omega_i}{\lambda} > -V_i\alpha\right) & & \text{more algebra} \end{array}$$

where $\alpha = \lambda^{-1}\pi$.

Subsample OLS

By properties of linear transformations,

$$\begin{aligned} E(\epsilon_i|J_i=1) &= \sigma E\left(\frac{\epsilon_i}{\sigma}\Big|\frac{\omega_i}{\lambda} > -V_i\alpha\right) \\ &= \sigma\rho\frac{\phi(-V_i\alpha)}{1-\Phi(-V_i\alpha)} && \text{censored cond. mean} \\ &= \sigma\rho\frac{\phi(V_i\alpha)}{\Phi(V_i\alpha)} && \text{symmetry} \\ &= \sigma\rho\gamma(V_i\alpha) && \text{notation} \end{aligned}$$

Subsample OLS

- The mean of the subsample disturbance term is observation-specific. Consequently, subsample OLS will provide biased estimates of β unless:
  1. $\rho = 0$
  2. $\gamma(V_i\alpha)$ is orthogonal to X_i.
- Note that condition 2 will typically not be satisfied when there are common regressors in X_i and V_i.

Heckman-Lee Estimation

The mean of the subsample disturbance term suggests the possibility of a simple two-stage estimation procedure. Letting $\kappa = \{i|J_i = 1\}$, the subsample regression equation may be written as

$$Y_i = X_i\beta + \sigma\rho\gamma(V_i\alpha) + \eta_i \qquad \text{for } i \in \kappa$$

where $\eta_i = \epsilon_i - \sigma\rho\gamma(V_i\alpha)$.

Heckman-Lee Estimation

Given the definition of η_i, we have:

1. $E(\eta_i|J_i = 1) = 0$
2. $Var(\eta_i|J_i = 1) = Var(\epsilon_i|J_i = 1)$

$$\begin{aligned} &= \sigma^2 Var\left(\frac{\epsilon_i}{\sigma}\Big|\frac{\omega_i}{\lambda} > -V_i\alpha\right) \\ &= \sigma^2[1 - \rho^2(V_i\alpha)\gamma(V_i\alpha) - \rho^2\gamma(V_i\alpha)^2] \end{aligned}$$

3. $Cov(\eta_i, \eta_j) = 0$ since (ϵ_i, ϵ_j) SI implies (η_i, η_j) SI

These properties indicate that the subsample regression function is a non-linear heteroskedastic regression model.

Heckman-Lee Estimation

- Let $\hat{\alpha}$ denote the estimate of α obtained from a first-stage Probit model using the qualitative information in (J_i, V_i) for all n observations.
- Since $\hat{\alpha}$ is a consistent estimator of α, the transformation $\gamma(V_i\hat{\alpha})$ is a consistent estimator of $\gamma(V_i\alpha)$ by the generalized Slutsky theorem.

Heckman-Lee Estimation

- Using the first-stage Probit estimates, the subsample regression model may be written as

$$\begin{aligned} Y_i &= X_i\beta + \sigma\rho\gamma(V_i\hat{\alpha}) + \{\eta_i - \sigma\rho[\gamma(V_i\hat{\alpha}) - \gamma(V_i\alpha)]\} \\ &= Z_i\theta + \{\eta_i - \sigma\rho[\gamma(V_i\hat{\alpha}) - \gamma(V_i\alpha)]\} \end{aligned}$$

for $i \in \kappa$, where $Z_i = [X_i \quad \gamma(V_i\hat{\alpha})]$ and $\theta' = [\beta' \quad \sigma\rho]$.

The term in braces is a composite error based upon random variation in η_i and sampling error in $\gamma(V_i\hat{\alpha})$.

Heckman-Lee Estimation

- For convenience, order the $J_i = 1$ observations first, then the $J_i = 0$ observations. This may be done without loss of generality with an SI sample.
- Partition the data as

$$Y = \begin{bmatrix} Y_1 \\ \cdot \end{bmatrix} \quad Z = \begin{bmatrix} Z_1 \\ Z_0 \end{bmatrix} = \begin{bmatrix} X_1 & \gamma(V_1\hat{\alpha}) \\ X_0 & \gamma(V_0\hat{\alpha}) \end{bmatrix}$$

- The Heckman-Lee estimator is defined as $\hat{\theta} = (Z_1'Z_1)^{-1}Z_1'Y_1$. This is just OLS applied to the observed subsample after constructing the auxiliary regressor $\gamma(V_1\hat{\alpha})$.

Heckman-Lee Estimation

- The HL estimator provides consistent estimates of β.
- The standard errors reported by the second stage OLS regression are incorrect since no correction is made for either the heteroskedastic nature of η_i or the sampling error in the constructed regressor $\gamma(V_i\hat{\alpha})$.
- Many ”canned” packages now present corrected standard errors. This point is of little consequence, as HL estimates should generally be used only as starting values for ML estimation.
- Small sample properties are not available for the HL estimator.

Heckman-Lee Estimation

- Recall that the magnitude of selection bias with subsample OLS approaches zero as either ρ approaches zero or the correlation between X and $\gamma(V_i\hat{\alpha})$ approaches zero.
- Monte Carlo studies have found that the "selection bias" in subsample OLS is only serious when both ρ is large in absolute value and variation in $\gamma(V_i\hat{\alpha})$ can be predicted with a high degree of accuracy using X. Unfortunately, these are precisely the circumstances when X and $\gamma(V_i\hat{\alpha})$ are most colinear, and consequently, HL is least precise. The HL estimator performs the worst under conditions when it is needed most!

Heckman-Lee Estimation

- The Monte Carlo work can be summarize as follows; when the selection bias of subsample OLS is modest, the differences in MSE between subsample OLS and HL are also modest, but when the selection bias of subsample OLS is larger, the imprecision of HL often dwarfs the bias of subsample OLS resulting in a significantly larger MSE with HL estimation.
- Is HL estimation the econometric equivalent of chopping off the hand to treat arthritis?

Heckman-Lee Estimation

- A final problem is concerned with getting distinct estimates of ρ and σ. The coefficient of $\gamma(V_i\hat{\alpha})$ is an estimate of the product $\sigma\rho$.
- If we denote the second-stage residuals as $e_i = Y_i - Z_i\hat{\theta}$ and the estimated coefficient of $\gamma(V_i\hat{\alpha})$ by $\widehat{\sigma\rho}$, then Heckman suggests an estimate of σ^2 based on the conditional variance of η_i as

$$\hat{\sigma}^2 = \sum_{i=1}^{n_1} \frac{e_i^2 + \widehat{\sigma\rho}^2(V_i\hat{\alpha})\gamma(V_i\hat{\alpha}) + \widehat{\sigma\rho}^2\gamma(V_i\hat{\alpha})^2}{n_1 - k_X - 1}$$

- Given $\hat{\sigma}^2$ and $\widehat{\sigma\rho}$, ρ may be estimated as $\hat{\rho} = \widehat{\sigma\rho}/\hat{\sigma}$. A practical difficulty with this estimator of ρ is that it often falls outside the interval $(-1, 1)$.

The Likelihood Function

Under the structure imposed by the Selection model, the joint density for (Y, U) pairs is

$$f(Y, U) = \sigma^{-1}\lambda^{-1}\phi(\sigma^{-1}\epsilon, \lambda^{-1}\omega)$$

where $\phi(\cdot, \cdot)$ denotes the bivariate standard normal with correlation ρ. The observational subscript has been omitted to simply notation.

- This result follows from the assumption that (ϵ, ω) is bivariate normal.
- Absent any censoring, this density could be used to find the likelihood function for a random sample of data.
- The likelihood function for the censored data will require both the joint density of (Y, J) and the marginal density of J.

The Likelihood Function

Consider first the marginal density of J. Since (Y, U) is bivariate normal, the marginal density function of U is univariate normal. Since J is a binary variable determined by the sign of U, we have:

$$\begin{aligned} f(J=0) &= P(U \leq 0) \\ &= F_U(0) \\ &= \Phi(-V\alpha) \\ &= 1 - \Phi(V\alpha) \end{aligned}$$

The Likelihood Function

Likewise,

$$\begin{aligned} f(J=1) &= P(U > 0) \\ &= 1 - F_U(0) \\ &= 1 - \Phi(-V\alpha) \\ &= \Phi(V\alpha) \end{aligned}$$

This is just a Bernoulli trial with observation specific probablility and may be summerized as:

$$f(J) = \Phi(V\alpha)^J [1 - \Phi(V\alpha)]^{(1-J)}$$

for $J \in \{0, 1\}$.

The Likelihood Function

We will also need the joint density function for the (Y, J) pairs, which is obtained from the joint density for (Y, U) pairs as

$$f(Y,J) = \begin{cases} \int_0^\infty f(Y,U)dU & \text{for } J=1 \\ \\ \int_{-\infty}^0 f(Y,U)dU & \text{for } J=0 \end{cases}$$

The Likelihood Function

The joint density when $J = 1$ is

$$\begin{aligned} f(Y, J=1) &= \int_0^\infty f(Y,U)dU \\ &= \int_{-V\pi}^\infty \sigma^{-1}\lambda^{-1}\phi(\sigma^{-1}\epsilon, \lambda^{-1}\omega)d\omega \\ &= \int_{-V\alpha}^\infty \sigma^{-1}\phi(\sigma^{-1}\epsilon, \chi)d\chi \end{aligned}$$

where $\alpha = \lambda^{-1}\pi$ and χ is a variable of integration for $\lambda^{-1}\omega$. Factoring the joint density into the product of the conditional and marginal, we have

$$\begin{aligned} f(Y, J=1) &= \int_{-V\alpha}^\infty \sigma^{-1}\phi(\chi|\sigma^{-1}\epsilon)\phi(\sigma^{-1}\epsilon)d\chi \\ &= \sigma^{-1}\phi(\sigma^{-1}\epsilon) \int_{-V\alpha}^\infty \phi(\chi|\sigma^{-1}\epsilon)d\chi \end{aligned}$$

The Likelihood Function

Recall that if (u, v) is bivariate standard normal with correlation ρ, then $u|v \sim N(\rho v, 1-\rho^2)$. Thus,

$$f(Y, J=1) = \sigma^{-1}\phi(\sigma^{-1}\epsilon)\left[1 - \Phi\left(\frac{-V\alpha - \rho\sigma^{-1}\epsilon}{\sqrt{1-\rho^2}}\right)\right]$$

Letting $Z = \frac{\epsilon}{\sigma} = \frac{(Y - X\beta)}{\sigma}$ and $W = \frac{(V\alpha + \rho Z)}{\sqrt{1-\rho^2}} = A_1 V\alpha + A_2 Z$, where $A_1 = \frac{1}{\sqrt{1-\rho^2}}$ and $A_2 = \rho A_1$, we have

$$\begin{aligned} f(Y, J=1) &= \sigma^{-1}\phi(Z)[1 - \Phi(-W)] \\ &= \sigma^{-1}\phi(Z)\Phi(W) \end{aligned}$$

The Likelihood Function

Following the same line of reasoning, the joint density when $J = 0$ is

$$\begin{aligned} f(Y, J=0) &= \int_{-\infty}^{0} f(Y, U)dU \\ &= \int_{\infty}^{-V\pi} \sigma^{-1}\lambda^{-1}\phi(\sigma^{-1}\epsilon, \lambda^{-1}\omega)d\omega \\ &= \int_{-\infty}^{-V\alpha} \sigma^{-1}\phi(\sigma^{-1}\epsilon, \chi)d\chi \end{aligned}$$

where $\alpha = \lambda^{-1}\pi$, χ is a variable of integration for $\lambda^{-1}\omega$, and again factoring the joint density into the product of the conditional and marginal,

$$\begin{aligned} f(Y, J=0) &= \int_{-\infty}^{-V\alpha} \sigma^{-1}\phi(\chi|\sigma^{-1}\epsilon)\phi(\sigma^{-1}\epsilon)d\chi \\ &= \sigma^{-1}\phi(\sigma^{-1}\epsilon)\int_{-\infty}^{-V\alpha} \phi(\chi|\sigma^{-1}\epsilon)d\chi \end{aligned}$$

The Likelihood Function

Once again using the moments of the conditional normal,

$$f(Y, J=0) = \sigma^{-1}\phi(\sigma^{-1}\epsilon)\,\Phi\left(\frac{-V\alpha - \rho\sigma^{-1}\epsilon}{\sqrt{1-\rho^2}}\right)$$

Again, letting $Z = \frac{\epsilon}{\sigma} = \frac{(Y - X\beta)}{\sigma}$ and $W = \frac{(V\alpha + \rho Z)}{\sqrt{1-\rho^2}} = A_1 V\alpha + A_2 Z$, where $A_1 = \frac{1}{\sqrt{1-\rho^2}}$ and $A_2 = \rho A_1$, we have

$$\begin{aligned} f(Y, J=0) &= \sigma^{-1}\phi(Z)[\Phi(-W)] \\ &= \sigma^{-1}\phi(Z)[1 - \Phi(W)] \end{aligned}$$

Both cases may be summarized conveniently as:

$$f(Y, J) = \sigma^{-1}\phi(Z)\Phi(W)^J[1 - \Phi(W)]^{(1-J)}$$

The Likelihood Function

When J equals 0, we have only the qualitative information that U is non-positive. Since there is no information about Y, the relevant term in the likelihood function is the marginal density of J, which for the $J = 0$ case is

$$\begin{aligned} f(J=0) &= \Phi(V\alpha)^0[1 - \Phi(V\alpha)]^{(1-0)} \\ &= [1 - \Phi(V\alpha)] \end{aligned}$$

When J equals 1, we have the qualitative information that U is positive plus the quantitative information in the observed value of Y. The relevent term in the likelihood function is the joint density of (Y, J) evaluated at $J = 1$. This is

$$\begin{aligned} f(Y, J=1) &= \sigma^{-1}\phi(Z)\Phi(W)^1[1 - \Phi(W)]^{(1-1)} \\ &= \sigma^{-1}\phi(Z)\Phi(W) \end{aligned}$$

The Likelihood Function

Restoring the observation index and combining the relevant terms for the $J = 1$ and $J = 0$ cases, the log-likelihood function for a random sample is

$$\ln L(\beta, \sigma, \rho, \alpha) = \sum_{i=1}^{n} \{J_i[-\ln(\sigma) + \ln \phi(Z_i) + \ln \Phi(W_i)] + (1 - J_i) \ln[1 - \Phi(V_i\alpha)]\}$$

A Restricted Case

Note that if $\rho = 0$, W_i reduces to $V_i\alpha$, and the log-likelihood function of the selection model reduces to

$$\ln L(\beta, \sigma, \rho, \alpha) = \sum_{i=1}^{n} \{J_i[-\ln(\sigma) + \ln \phi(Z_i)]\} + \sum_{i=1}^{n} \{J_i \ln \Phi(V_i\alpha) + (1 - J_i) \ln[1 - \Phi(V_i\alpha)]\}$$

- The first term is the log-likelihood function of a classical regression model applied to the regression equation for the $J_i = 1$ subsample. This is commonly referred to as subsample OLS. The coefficient vector β and the parameter σ enter the likelihood function only through this first term.

A Restricted Case

- The second term is the log-likelihood function of a Probit model applied to the selection equation. The coefficient vector α enters the likelihood function only through this second term.
- Consequently, when $\rho = 0$, the two equations may be estimated independently. The ML estimators of β and σ are given by subsample OLS, and the ML estimator of α is the Probit MLE. It is important to remember that these results are valid only when $\rho = 0$.
- This gives an intuitive explanation for the Selection model as a regression equation and Probit equation linked through error correlation.

Identification

- The likelihood function above is expressed in terms of the standardized coefficients, α. The k_V coefficients in α are identified so long as V has full column rank.
- The $k_V + 1$ elements of (π', λ) are not identified. This is because they enter the likelihood function only through the ratios $\lambda^{-1}\pi$.
- Let $\theta_1' = [\beta_1' \ \sigma_1 \ \rho_1 \ \pi_1' \ \lambda_1]$ and $\theta_2' = [\beta_1' \ \sigma_1 \ \rho_1 \ \pi_2' \ \lambda_2]$. There are an infinity of $\theta_1 \neq \theta_2$ such that $\lambda_1^{-1}\pi_1 = \lambda_2^{-1}\pi_2$. Since $\ln L(\theta_1) = \ln L(\theta_2)$ at any such points, θ_1 and θ_2 are observationally equivalent, and the parameter vector θ is not identified. When the likelihood function is expressed in terms of β, σ, ρ, π, and λ, there is an identification problem.

Identification

- The restriction $\lambda = 1$ is an "identification condition" for the Selection model. That is, imposition of the constraint does not alter the maximized value of the log-likelihood function. The constraint is non-binding and results in no loss in generality. There is simply no information in the data to test the validity of the restriction.
- The motivation for this result is the same as with the Probit model. An increase in λ implies greater variation in the latent dependent variable U, but not in the observed selection indicator J. All positive values of U, regardless of their magnitude, are translated into the single value $J = 1$.
- In fact, α may be estimated consistently, though inefficiently, with a simple Probit model of J on V.

Identification

Proposition 1: If $(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\alpha})$ maximizes $\ln L(\beta, \sigma, \rho, \alpha)$, and $\hat{\pi}_0 = \lambda_0 \hat{\alpha}$, then $(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0)$ maximizes $\ln L(\beta, \sigma, \rho, \pi, \lambda_0)$ for any fixed value λ_0.

Assume, contrary to the statement of the theorem, that $(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0)$ does not maximize $\ln L(\beta, \sigma, \rho, \pi, \lambda_0)$. This implies that there exists $(\tilde{\beta}, \tilde{\sigma}, \tilde{\rho}, \tilde{\pi})$ such that $\ln L(\tilde{\beta}, \tilde{\sigma}, \tilde{\rho}, \tilde{\pi}, \lambda_0) > \ln L(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0, \lambda_0)$. Let $\tilde{\alpha} = \lambda_0^{-1} \tilde{\pi}$. Then since $\ln L(\beta, \sigma, \rho, \pi, \lambda) = \ln L(\beta, \sigma, \rho, \alpha)$ for all $\alpha = \lambda^{-1}\pi$, this implies that $\ln L(\tilde{\beta}, \tilde{\sigma}, \tilde{\rho}, \tilde{\alpha}) > \ln L(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\alpha})$, which contradicts the statement that $(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\alpha})$ maximizes $\ln L(\beta, \sigma, \rho, \alpha)$. Hence, $(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0)$ maximizes $\ln L(\beta, \sigma, \rho, \pi, \lambda_0)$ for any fixed value λ_0.

Identification

Proposition 2: The maximized value of the log-likelihood function, $\ln L(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0, \lambda_0)$, is invariant to the choice of λ_0.

The maximizing values, $\hat{\pi}_0$, must satisfy $\lambda_0^{-1} \hat{\pi}_0 = \hat{\alpha}$ for all λ_0. Since the parameters $(\beta, \sigma, \rho, \pi, \lambda)$ enter $\ln L(\beta, \sigma, \rho, \pi, \lambda)$ only through $\alpha = \lambda^{-1}\pi$, this implies that $\ln L(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0, \lambda_0) = \ln L(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\alpha})$ for all λ_0. Since $(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\alpha})$ is the unique global maximizer of $\ln L(\beta, \sigma, \rho, \alpha)$, this implies that $\ln L(\hat{\beta}, \hat{\sigma}, \hat{\rho}, \hat{\pi}_0, \lambda_0)$ is invariant to the choice of λ_0.

Identification

Proposition 2 shows that the restriction $\lambda = 1$ does not affect the maximized value of the log-likelihood function, and consequently, is an identification condition for the Selection model. There are a few additional points about identification conditions that are worth considering.

1. As with the Binary Choice model, this identification condition is not unique.
2. While the maximizing value of α is invariant to the choice of λ, the maximizing value of π is not. Variation in λ results in proportional variation in π.

Identification

3. The "truth" of the identification condition is irrelevant. Data generated by different values of λ will have the same maximized value of the log-likelihood function regardless of the identification condition adopted by the econometrician. Models with different values of λ are "observationally equivalent." They cannot be distinguished on the basis of the observed data alone.

ML Estimation

ML estimation will require the score equations for β, σ, ρ, and α. Most of the derivatives involved are similar to those we encountered with the Tobit model. The following intermediate result will be useful when finding the score equation for ρ.

Given that $W_i = (V_i\alpha + \rho Z_i)(1-\rho^2)^{-0.5}$,

$$\begin{aligned}\frac{\partial W_i}{\partial \rho} &= Z_i(1-\rho^2)^{-0.5} + (V_i\alpha + \rho Z_i)[(-0.5)(1-\rho^2)^{-1.5}(-2\rho)] \\ &= Z_i(1-\rho^2)^{-0.5} + (V_i\alpha + \rho Z_i)\rho(1-\rho^2)^{-1.5} \\ &= (A_1 Z_i + A_1 A_2 W_i)\end{aligned}$$

ML Estimation

The k_X score equations for β are

$$\begin{aligned}\frac{\partial \ln L(\beta,\sigma,\rho,\alpha)}{\partial \beta} &= \sum_{i=1}^{n}\left\{J_i\frac{\partial \ln \phi(Z_i)}{\partial \beta} + J_i\frac{\partial \ln \Phi(W_i)}{\partial \beta}\right\} \\ &= \sum_{i=1}^{n}\left\{J_i\sigma^{-1}Z_i - J_i\sigma^{-1}\frac{\phi(W_i)}{\Phi(W_i)}A_2\right\}X_i' \\ &= \sum_{i=1}^{n}\left\{J_i\sigma^{-1}[Z_i - \gamma(W_i)A_2]\right\}X_i'\end{aligned}$$

where $Z_i = (Y_i - X_i\beta)/\sigma$, $W = \frac{(V\alpha+\rho Z)}{\sqrt{1-\rho^2}} = A_1 V\alpha + A_2 Z$, $A_1 = \frac{1}{\sqrt{1-\rho^2}}$, and $A_2 = \rho A_1$.

ML Estimation

The score equation for σ is

$$\begin{aligned}\frac{\partial \ln L(\beta,\sigma,\rho,\alpha)}{\partial \sigma} &= \sum_{i=1}^{n}\left\{J_i\left[-\frac{\partial \ln(\sigma)}{\partial \sigma} + \frac{\partial \ln \phi(Z_i)}{\partial \sigma} + \frac{\partial \ln \Phi(W_i)}{\partial \sigma}\right]\right\} \\ &= \sum_{i=1}^{n}\left\{J_i\left[-\sigma^{-1} + \sigma^{-1}Z_i^2 - \frac{\phi(W_i)}{\Phi(W_i)}A_2\sigma^{-1}Z_i\right]\right\} \\ &= \sum_{i=1}^{n}\left\{J_i[-\sigma^{-1} + \sigma^{-1}Z_i^2 - \gamma(W_i)A_2\sigma^{-1}Z_i]\right\}\end{aligned}$$

ML Estimation

The score equations for ρ is

$$\begin{aligned}\frac{\partial \ln L(\beta,\sigma,\rho,\alpha)}{\partial \rho} &= \sum_{i=1}^{n}\left\{J_i \frac{\partial \ln \Phi(W_i)}{\partial \rho}\right\} \\ &= \sum_{i=1}^{n}\left\{J_i \frac{\phi(W_i)}{\Phi(W_i)}\frac{\partial W_i}{\partial \rho}\right\} \\ &= \sum_{i=1}^{n}\left\{J_i \gamma(W_i)\frac{\partial W_i}{\partial \rho}\right\} \\ &= \sum_{i=1}^{n}\{J_i \gamma(W_i)(A_1 Z_i + A_1 A_2 W_i)\}\end{aligned}$$

ML Estimation

The k_V score equations for α are

$$\begin{aligned}\frac{\partial \ln L(\beta,\sigma,\rho,\alpha)}{\partial \alpha} &= \sum_{i=1}^{n}\left\{J_i \frac{\partial \ln \Phi(W_i)}{\partial \alpha} + (1-J_i)\frac{\partial \ln[1-\Phi(V_i\alpha)]}{\partial \alpha}\right\} \\ &= \sum_{i=1}^{n}\left\{J_i \frac{\phi(W_i)}{\Phi(W_i)}A_1 - (1-J_i)\frac{\phi(V_i\alpha)}{1-\Phi(V_i\alpha)}\right\}V_i' \\ &= \sum_{i=1}^{n}\{J_i[\gamma(W_i)A_1 - (1-J_i)\gamma(-V_i\alpha)]\}V_i'\end{aligned}$$

For the Selection model, the score equations are a set of $(k_X + k_V + 2)$ simultaneous nonlinear implicit functions for β, σ, ρ, and α.

Bounding ρ

The parameter space of ρ is the open interval (-1,1). How do we impose this restriction?

- Many algorithms do so in an ad-hoc manner; check the restriction, if violated, reset ρ just inside the nearest boundary. This approach can lead to ”cycling” on the boundary of the parameter space.
- A better approach employs the transformation $\rho = \frac{\exp(\zeta)-1}{\exp(\zeta)+1}$, where the parameter ζ is estimated instead of ρ. The restriction is imposed by the form of the transformation since real values of ζ generate values of ρ on the interval (-1,1).

Bounding ρ

- Neither of these methods is necessary in the neighborhood of the maximum. We will see that insuring that starting values are in the neighborhood of the maximum is important with a Selection model.
- The advantage of this transformation is that it prevents the algorithm from ”crashing” when the starting values are poor. This is small comfort since poor starting values with a Selection model can result in convergence to a local MLE.
- More generally, this transformation provides a method of restricting a parameter to the interval (-1,1), something that might be useful in other optimization problems.

Bounding ρ

To use the transformation $\rho = \frac{\exp(\zeta)-1}{\exp(\zeta)+1}$ we need score equations for β, σ, α, and ζ.

- One method of finding these score equations would be to go back and express the log likelihood function of the Selection model in terms of β, σ, α, and ζ, and then differentiate. Fortunately, this is not necessary.
- We already have the score equations for β, σ, α, and ρ. The numerical value of the score equations are the same whether expressed in terms of ρ or ζ.

Bounding ρ

- Using the chain rule,

$$\frac{\partial \ln L(\beta,\sigma,\zeta,\alpha)}{\partial \zeta} = \frac{\partial \ln L(\beta,\sigma,\rho,\alpha)}{\partial \rho}\frac{\partial \rho}{\partial \zeta}$$

- The inverse transformation is $\zeta = \ln(1+\rho) - \ln(1-\rho)$. This implies that

$$\frac{\partial \zeta}{\partial \rho} = \frac{1}{1+\rho} + \frac{1}{1-\rho} = \frac{2}{(1+\rho)(1-\rho)}$$

and by the inverse function rule

$$\frac{\partial \rho}{\partial \zeta} = \frac{(1+\rho)(1-\rho)}{2}$$

Bounding ρ

- The score equation for ζ is then

$$\frac{\partial \ln L(\beta,\sigma,\zeta,\alpha)}{\partial \zeta} = \frac{\partial \ln L(\beta,\sigma,\rho,\alpha)}{\partial \rho}\frac{(1+\rho)(1-\rho)}{2}$$

This equality expresses the score equation for ζ in terms of β, σ, α, and ρ.

Bounding ρ

An iterating algorithm for $(\beta,\sigma,\alpha,\zeta)$ may be summarized as follows:

- Given the current value of $(\beta,\sigma,\alpha,\zeta)$, find the corresponding value of $(\beta,\sigma,\alpha,\rho)$ and use it to evaluate the score equations for β, σ, α, and ρ.
- Convert the numerical value of the score equation for ρ to that of ζ by multiplying by $\frac{(1+\rho)(1-\rho)}{2}$.
- Apply the BHHH algorithm to get an updated value for $(\beta,\sigma,\alpha,\zeta)$.
- Repeat until convergence.

Bounding ρ

The ML estimator of ρ is $\hat{\rho} = \frac{\exp(\hat{\zeta})-1}{\exp(\hat{\zeta})+1}$.

- Standard errors for $\hat{\rho}$ may be obtained using the "delta method."
- Alternatively, once a root is found, $\hat{\beta}, \hat{\sigma}, \hat{\alpha}$ and $\hat{\rho}$ may be used as starting values for a BHHH algorithm using the score equations for β, σ, α, and ρ.
- Note that we NEVER need to program the score equation for ζ. All of this is done by simply re-weighting the numerical value of the score equation for ρ.

Concavity

- The log-likelihood function for the Selection model is not globally concave. Absent further information we cannot be sure that a root to the score equations is a global MLE. "Canned" software packages fail to control for this problem and will report a local MLE without any warning. This includes *Stata*!
- We will see, however, that the log-likelihood function is globally concave conditional on ρ. The score equations for β, σ, and α may be used in conjunction with a grid search over the bounded space of ρ to locate the neighborhood of the global MLE. The resulting estimates are then used as starting values for a fully simultaneous estimation of β, σ, α, and ρ.

Standardized Parameters

- In order to show that the log-likelihood function is globally concave conditional on ρ, it is again convenient to express the likelihood function in terms of the alternative parameters $\delta = \sigma^{-1}\beta$ and $\tau = \sigma^{-1}$.
- As with the Tobit model, if $\ln L(\alpha, \delta, \tau|\rho)$ is globally concave, then any root to the score equations of $\ln L(\alpha, \beta, \sigma|\rho)$ will be a global MLE since the transformations are one-to-one and onto.

Standardized Parameters

- Expressed in terms of the new parameter set, the log-likelihood function is

$$\ln L(\alpha, \delta, \tau|\rho) = \sum_{i=1}^{n} \{J_i[\ln(\tau) + \ln \phi(Z_i) + \ln \Phi(W_i)] + (1 - J_i) \ln[1 - \Phi(V_i\alpha)]\}$$

where $Z = (\tau Y - X\delta)$, $W = A_1 V\alpha + A_2 Z$, $A_1 = \frac{1}{\sqrt{1-\rho^2}}$, and $A_2 = \rho A_1$.

Score Equations

The k_V score equations for α are

$$\begin{aligned}\frac{\partial \ln L(\alpha,\delta,\tau|\rho)}{\partial\alpha} &= \sum_{i=1}^{n}\left\{J_i\frac{\partial \ln \Phi(W_i)}{\partial\alpha} + (1-J_i)\frac{\partial \ln[1-\Phi(V_i\alpha)]}{\partial\alpha}\right\}\\ &= \sum_{i=1}^{n}\left\{J_i\frac{\phi(W_i)}{\Phi(W_i)}A_1 - (1-J_i)\frac{\phi(V_i\alpha)}{1-\Phi(V_i\alpha)}\right\}V_i'\\ &= \sum_{i=1}^{n}\left\{J_i\gamma(W_i)A_1 - (1-J_i)\gamma(-V_i\alpha)\right\}V_i'\end{aligned}$$

Score Equations

The k_X score equations for δ are

$$\begin{aligned}\frac{\partial \ln L(\alpha,\delta,\tau|\rho)}{\partial\delta} &= \sum_{i=1}^{n}\left\{J_i\frac{\partial \ln \phi(Z_i)}{\partial\delta} + J_i\frac{\partial \ln \Phi(W_i)}{\partial\delta}\right\}\\ &= \sum_{i=1}^{n}\left\{J_iZ_i - J_i\frac{\phi(W_i)}{\Phi(W_i)}A_2\right\}X_i'\\ &= \sum_{i=1}^{n}\left\{J_i[Z_i - \gamma(W_i)A_2]\right\}X_i'\end{aligned}$$

Score Equations

The score equation for τ is

$$\begin{aligned}\frac{\partial \ln L(\alpha,\delta,\tau|\rho)}{\partial\tau} &= \sum_{i=1}^{n}\left\{J_i\left[\frac{\partial \ln(\tau)}{\partial\tau} + \frac{\partial \ln \phi(Z_i)}{\partial\tau} + \frac{\partial \ln \Phi(W_i)}{\partial\tau}\right]\right\}\\ &= \sum_{i=1}^{n}\left\{J_i\left[\tau^{-1} + Z_iY_i + A_2\frac{\phi(W_i)}{\Phi(W_i)}Y_i\right]\right\}\\ &= \sum_{i=1}^{n}\left\{J_i[\tau^{-1} + Z_iY_i + A_2Z_i\gamma(W_i)Y_i]\right\}\end{aligned}$$

There is never any need to uses these score equations for estimation purposes. Use those expressed in terms of the original parameters instead.

The Hessian Matrix

We have seen that

$$\frac{\partial\phi(Z)\Phi(Z)^{-1}}{\partial Z} = -[Z + \frac{\phi(Z)}{\Phi(Z)}]\frac{\phi(Z)}{\Phi(Z)}$$

Letting $Z = W_i$, we have

$$\frac{\partial\gamma(W_i)}{\partial W_i} = -[W_i + \gamma(W_i)]\gamma(W_i)$$

and letting $Z = -V_i\alpha$, we have

$$\frac{\partial\gamma(-V_i\alpha)}{\partial V_i\alpha} = [-V_i\alpha + \gamma(-V_i\alpha)]\gamma(-V_i\alpha)$$

The Hessian Matrix

The Hessian matrix of $\ln L(\alpha, \delta, \tau|\rho)$ is composed of:

$$\frac{\partial^2 \ln L(\alpha, \delta, \tau|\rho)}{\partial\alpha\partial\alpha'} = \sum_{i=1}^{n} V_i' \left\{ J_i A_1 \frac{\partial\gamma(W_i)}{\partial\alpha'} - (1 - J_i)\frac{\partial\gamma(-V_i\alpha)}{\partial\alpha'} \right\}$$

$$= \sum_{i=1}^{n} V_i' \{ -J_i A_1^2 [W_i + \gamma(W_i)]\gamma(W_i) - (1 - J_i)[\gamma(-V_i\alpha) - V_i\alpha]\gamma(-V_i\alpha)\} V_i$$

$$= -(A_1^2 V_1' D V_1 + V_0' C V_0)$$

where D is an $n_1 \times n_1$ diagonal matrix with diagonal elements $[W_i + \gamma(W_i)]\gamma(W_i)$ and where C is an $n_0 \times n_0$ diagonal matrix with diagonal elements $[\gamma(-V_i\alpha) - V_i\alpha]\gamma(-V_i\alpha)$.

The Hessian Matrix

$$\frac{\partial^2 \ln L(\alpha, \delta, \tau|\rho)}{\partial\alpha\partial\delta'} = \sum_{i=1}^{n} V_i' \left\{ J_i A_1 \frac{\partial\gamma(W_i)}{\partial\delta'} \right\}$$

$$= \sum_{i=1}^{n} V_i' \{ J_i A_1 A_2 [W_i + \gamma(W_i)]\gamma(W_i)\} X_i$$

$$= (A_1 A_2 V_1' D X_1)$$

The Hessian Matrix

$$\frac{\partial^2 \ln L(\alpha, \delta, \tau|\rho)}{\partial\alpha\partial\tau} = \sum_{i=1}^{n} V_i' \left\{ J_i A_1 \frac{\partial\gamma(W_i)}{\partial\tau} \right\}$$

$$= \sum_{i=1}^{n} V_i' \{ -J_i A_1 A_2 [W_i + \gamma(W_i)]\gamma(W_i)\} Y_i$$

$$= -(A_1 A_2 V_1' D Y_1)$$

The Hessian Matrix

$$\frac{\partial^2 \ln L(\alpha, \delta, \tau|\rho)}{\partial\delta\partial\delta'} = \sum_{i=1}^{n} X_i' \left\{ J_i \frac{\partial Z_i}{\partial\delta'} - J_i A_1 \frac{\partial\gamma(W_i)}{\partial\delta'} \right\}$$

$$= \sum_{i=1}^{n} X_i' \left\{ -J_i - J_i A_2^2 [W_i + \gamma(W_i)]\gamma(W_i) \right\} X_i$$

$$= -(X_1' X_1 + A_2^2 X_1' D X_1)$$

The Hessian Matrix

$$\frac{\partial^2 \ln L(\alpha,\delta,\tau|\rho)}{\partial\delta\partial\tau} = \sum_{i=1}^{n} X_i' \left\{ J_i \frac{\partial Z_i}{\partial \tau} - J_i A_1 \frac{\partial \gamma(W_i)}{\partial \tau} \right\}$$

$$= \sum_{i=1}^{n} X_i' \left\{ J_i + J_i A_2^2 [W_i + \gamma(W_i)]\gamma(W_i) \right\} Y_i$$

$$= (X_1' Y_1 + A_2^2 X_1' D Y_1)$$

The Hessian Matrix

$$\frac{\partial^2 \ln L(\alpha,\delta,\tau|\rho)}{\partial\tau^2} = \sum_{i=1}^{n} \left\{ J_i \left[\frac{\partial \tau^{-1}}{\partial \tau} + Y_i \frac{\partial Z_i}{\partial \tau} + A_2 Y_i \frac{\partial \gamma(W_i)}{\partial \tau} \right] \right\}$$

$$= \sum_{i=1}^{n} \left\{ -J_i \left[\tau^{-2} + Y_i^2 + A_2^2 Y_i^2 [W_i + \gamma(W_i)]\gamma(W_i) \right] \right\}$$

$$= -\left(n_1 \tau^{-2} + Y_1' Y_1 + A_2^2 Y_1' D Y_1 \right)$$

The Hessian Matrix

Let $\theta' = [\alpha' \ \delta' \ \tau]$ denote the vector of $k_V + k_X + 1$ parameters. Also, define the $n \times n$ diagonal matrix

$$B = \begin{bmatrix} D & 0 \\ 0 & C \end{bmatrix}$$

the $n \times (k_V + k_X + 1)$ matrices

$$G = \begin{bmatrix} A_1 V_1 & -A_2 X_1 & A_2 Y_1 \\ V_0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} 0 & \tau^{-1} J \end{bmatrix}$$

and the $n_1 \times (k_V + k_X + 1)$ matrix

$$H = \begin{bmatrix} 0 & X_1 & -Y_1 \end{bmatrix}$$

The Hessian Matrix

Using this notation

$$\frac{\partial^2 \ln L(\alpha,\delta,\tau|\rho)}{\partial\theta\partial\theta'} = -(G'BG + H'H + T'T)$$

- Neither H or T have full column rank, so both $H'H$ and $T'T$ are positive semidefinite.
- Since B is a diagonal matrix, $G'BG$ is positive definite if G has full column rank (which is satisfied by assumption) and if the diagonal elements of B are strictly positive.
- The diagonal elements of B are $[W_i + \gamma(W_i)]\gamma(W_i)$ and $[\gamma(-V_i\alpha) - V_i\alpha]\gamma(-V_i\alpha)$ for the $J_i = 1$ and $J_i = 0$ cases, respectively.

The Hessian Matrix

- Since $\gamma(Z) > 0$ for all Z, the signs of the diagonal elements of B are determined by the terms in brackets. For the $J_i = 1$ terms,

$$[W_i + \gamma(W_i)] = K - E(Z|Z \leq K) > 0$$

where $Z \sim N(0,1)$ and $K = W_i$, while for the $J_i = 0$ terms

$$[\gamma(-V_i\alpha) - V_i\alpha] = E(Z|Z > K) - K > 0$$

where $Z \sim N(0,1)$ and $K = V_i\alpha$. Hence, $G'BG$ is positive definite.

The Hessian Matrix

- Since the sum of a positive definite and positive semidefinite matrix is positive definite, $G'BG + H'H + T'T$ is positive definite and $-(G'BG + H'H + T'T)$ is negative definite.
- Thus, the Hessian matrix for $\ln L(\alpha, \delta, \tau|\rho)$ is negative definite.
- Since the transformatons from (α, δ, τ) to (α, β, σ) are one-to-one and onto, any root to the score equations for (α, β, σ) is a unique global MLE conditional on ρ.
- Consequently, a grid search over ρ may be used to map the profile of the ”concentrated” log-likelihood function and determine the neighborhood of the global MLE if one exists.

Grid Search over ρ

This grid search procedure may be summarized as follows:

1. Let $G = \{\rho_j\}$ for $j = 1, ..., J$ denote a grid of ρ values over the interval (-1,1). Typically, one starts with a course grid of equally spaced values.
2. For each $\rho_j \in G$, find the values of $\hat{\beta}_j$, $\hat{\sigma}_j$, and $\hat{\alpha}_j$ that maximize $\ln L(\beta, \sigma, \alpha, \rho_j)$.
3. Use the value of $(\hat{\beta}_j, \hat{\sigma}_j, \hat{\alpha}_j, \rho_j)$ that maximizes the ”concentrated” log-likelihood function, $\ln L(\hat{\beta}_j, \hat{\sigma}_j, \hat{\alpha}_j, \rho_j)$, over $\rho_j \in G$ as starting values for a fully simultaneous maximization of $\ln L(\beta, \sigma, \alpha, \rho)$.

For a sufficiently fine grid, these starting values will be in the neighborhood of the global maximum if one exists.

Grid Search over ρ

Potential outcomes of this procedure include:

1. There is no root to the score equations for β, σ, α, and ρ. The concentrated log-likelihood function is monotonically increasing (or decreasing) over the space of ρ. Distinguishing between this case and that of a root where $|\rho|$ is extremely close to 1 is problematic since the space of ρ is the open interval (-1,1).
2. There is a local MLE (a root) but no global MLE.
3. There are multiple roots, one of which is the global MLE.
4. There is a unique root which is a global MLE.

Grid Search over ρ

The diagrams on the next slide illustrate these cases. They were generated using Monte Carlo methods by resampling the errors, (ϵ, ω), for a Selection model with fixed structure.

- The values of β, σ, α, and ρ are fixed.
- The regressors X and V are generated and fixed when resampling the errors.
- The errors are resampled, the values of Y and J are recomputed, and the likelihood function is maximized.

This process shows that any of these cases can occur with a properly specified model. The absence of a root or the presence of multiple roots is NOT a sign of mis-specification.

Profiles of Maximized Log-Likelihood