Tobit Models

The following model has been classified as a "Type I Tobit" model by Amemiya (1985). The latent regression

$$Q_i = X_i\beta + \epsilon_i$$

satisfies full classical normal conditions except:

1. The latent dependent variable Q_i is unobserved. Instead, we observe a censoring indicator, J_i, where

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

   If this were the only information available, we would have a Probit model.

Tobit Models

2. The observed dependent variable, denoted Y_i, equals Q_i if $J_i = 1$. Its value is missing if $J_i = 0$. Quantitative information is available only from a subset of the sample space.
3. The vector of regressors, X_i, is observed for both the $J_i = 1$ and $J_i = 0$ cases.

In order to save space, programmers often code missing values as zero and report the product Y_iJ_i rather than the individual components Y_i and J_i. This works for $\lambda \geq 0$, but fails if $\lambda < 0$.

Tobit Models

- The parameter λ is a common censoring threshold. Regardless of what you hear or read, the restriction $\lambda = 0$ is not an identification condition.
- If λ is known, then the model may be expressed in equivalent form as a zero-threshold model. Subtracting λ from both sides of the latent regression gives

$$q_i = X_i\beta + \epsilon_i$$

  where $q_i = Q_i - \lambda$ and the intercept is now $\beta_1 - \lambda$. The censoring indicator, J_i, is determined as

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

Tobit Models

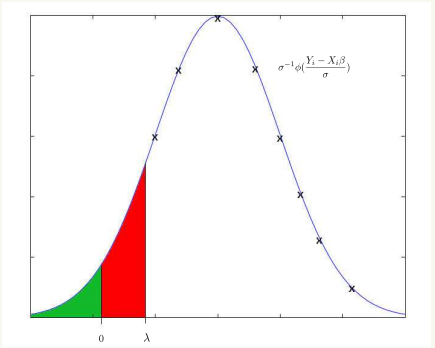
- The observed dependent variable, y_i, equals q_i if $J_i = 1$. The value of y_i is missing if $J_i = 0$.
- It is important to recognize, that when $\lambda \neq 0$, estimation of a zero threshold model requires that the dependent variable be constructed as $y_i = Y_i - \lambda$.
  - Monte Carlo studies show that failure to do this will bias the ML estimates (all coefficients, not just the intercept). In addition, the standard errors are typically biased upward and the t-statistics downward.
  - With HL estimation, only the intercept is affected, which is probably why the restriction $\lambda = 0$ was for so long mistakenly believed to be an identification condition. The HL estimator has other more significant problems however.

MLE of Common Threshold

- If λ is unknown, then it must be estimated. Let $\kappa = \{i|J_i = 1\}$ denote the set of observations that are not censored. The ML estimator of λ is $\hat{\lambda} = \min_{i \in \kappa}\{Y_i\}$. This is the smallest element in the "order statistic" for the observed subsample.
- The observed dependent variable is then transformed as $y_i = Y_i - \hat{\lambda}$ for $J_i = 1$ observations, and is missing for $J_i = 0$ observations. Again, failure to do this when λ is non-zero will bias the ML estimates.
- Conditional on $\hat{\lambda}$, the estimators of β and σ^2 are consistent, asymptotically normal, and asymptotically efficient.
- The ML estimator $\hat{\lambda}$ is biased in small samples, but is "super-consistent." It converges to the true value very quickly (at a rate faster than $\sqrt{n}$).

Nature of the Zero-Threshold Bias

The next few slides explain the nature of the bias that results when a zero-threshold ML estimator is used in conjunction with the unadjusted dependent variable Y_i.



Nature of the Zero-Threshold Bias

Assuming that the data is generated by the common threshold model with $\lambda > 0$:

- The quantitative information in the observed values of Y_i is measured by the density function $\sigma^{-1}\phi(\frac{Y_i - X_i\beta}{\sigma})$, for $Y_i > \lambda > 0$.
- The qualitative information in the censored values of Y_i is measured by the distribution function $\Phi(\frac{\lambda - X_i\beta}{\sigma})$. This is just $P(J_i = 0)$, and is depicted by the red and green areas under the density function.

Under the restriction $\lambda = 0$:

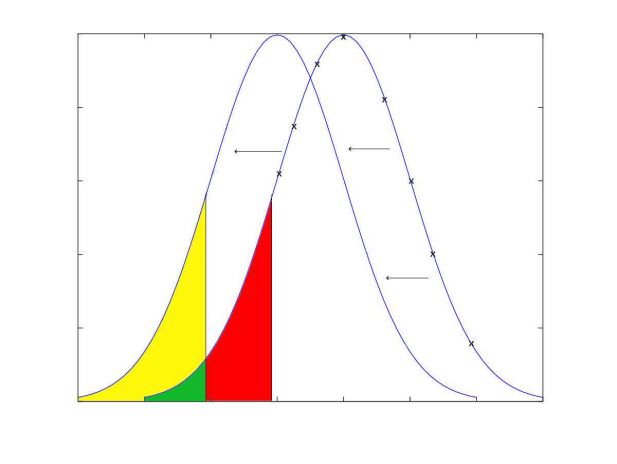
- $P(J_i = 0)$ is calculated as $\Phi(\frac{-X_i\beta}{\sigma})$, the area depicted in green.
- Since the data are generated with $P(J_i = 0) = \Phi(\frac{\lambda - X_i\beta}{\sigma}) > \Phi(\frac{-X_i\beta}{\sigma})$, the probability allotted by the zero-threshold model is insufficient to explain the magnitude of the observed sample proportion of $J_i = 0$ observations.

Nature of the Zero-Threshold Bias

- There are no values for Y_i in the interval $(0, \lambda)$. Under the structure of the common threshold model, these were all censored and stacked at zero. The probability of this interval is $\Phi(\frac{\lambda - X_i\beta}{\sigma}) - \Phi(\frac{-X_i\beta}{\sigma})$, the area depicted in red.
- The practical importance of estimating λ depends on the scale of the probability in this interval. If λ/σ is large, then $\Phi(\frac{\lambda - X_i\beta}{\sigma}) - \Phi(\frac{-X_i\beta}{\sigma})$ is large, and the impact of the zero-threshold restriction on the estimates of β is large.

Nature of the Zero-Threshold Bias

Why doesn't a simple reduction in the intercept fix this problem?



Nature of the Zero-Threshold Bias

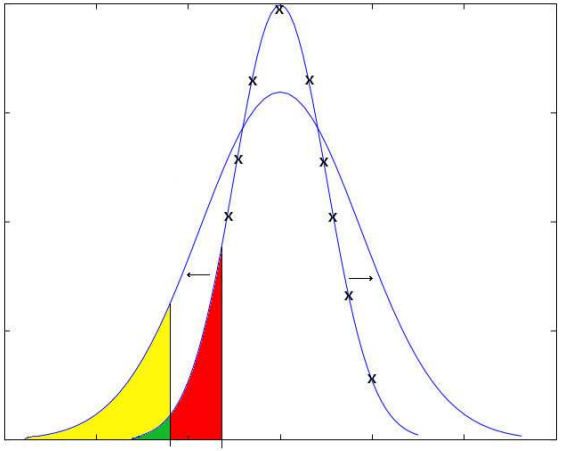
- We have just seen that if the restriction $\lambda = 0$ is invalid, $P(J_i = 0)$ is too low to explain the observed sample proportion of censored observations.
- If the intercept is reduced, leaving the other parameters unchanged, then the density function is shifted downward with an unchanged profile.
- This increases the area in the lower tail and better explains the observed sample proportion for $J_i = 0$, but uniformly degrades the fit of the Y_i. The observed values all seem improbably large now.
- Because the score equations of the Tobit ML are non-linear, this mis-specification spills over onto all of the parameters, not just the intercept.

Nature of the Zero-Threshold Bias

- With Heckman-Lee estimation, use of Y_i as the dependent variable in the second stage will only increase the estimated intercept by $\hat{\lambda}$ relative to that obtained with y_i. The other coefficients are unchanged.
- This is because the coefficient estimates in the first-stage Probit, $\hat{\delta}$, are not constrained to equal $\hat{\sigma}^{-1}\hat{\beta}$ from the second-stage OLS regression. Consequently, a shift in the mean of the quantitative distribution (second stage) does not feed back into the mean of the qualitative distribution (first stage).
- Failure to impose the restriction $\hat{\delta} = \hat{\sigma}^{-1}\hat{\beta}$ is the source of the inefficiency of Heckman-Lee relative to ML.

Nature of the Zero-Threshold Bias

Why do larger coefficients and larger standard errors result when a zero-threshold model is used with the unadjusted dependent variable?

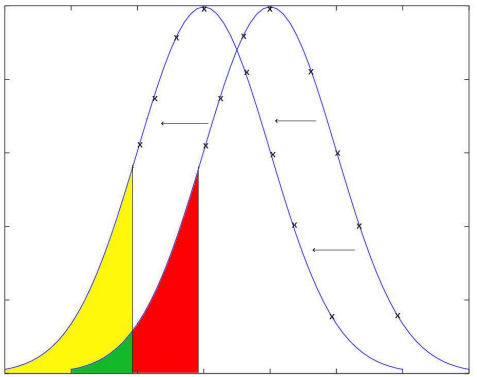


Nature of the Zero-Threshold Bias

- Since the zero-threshold model understates the probability of the $J_i = 0$ observations, the ML estimates are affected in ways that increase the ”tail areas” of the distribution.
- If the intercept tends to center the estimated density on the ”mode” of the quantitative portion of the sample distribution, how else can the ”tail areas” be increased?
  - The coefficients of the regressors tend to increase in absolute value, resulting in greater variation in the estimated values $X_i\hat{\beta}$.
  - Likewise, an increase in the value of $\hat{\sigma}$ directly increases the tail areas.
- In contrast to a decrease in the intercept, these changes only degrade the fit of the $Y_i > 0$ observations for cases that are not local to the intersection of the densities.

Nature of the Zero-Threshold Bias

Why does the transformation to y_i work?



Nature of the Zero-Threshold Bias

- When $\hat{\lambda}$ is subtracted from Y_i, the entire empirical distribution is shifted downward. The tail area is increased without degrading the fit for the quantitative observations.
- From this point on, we will dispense with the dual notation and presume that the transformed dependent variable has been properly constructed, using either a known value of λ or the ML estimate, so that a zero-threshold can be employed without loss of generality.

The Likelihood Function

To find the likelihood function we need the distribution of the observed dependent variable, Y_i. In what follows we assume that the missing values are ”stacked at zero.” Note that:

- $P(Y < 0) = 0$ since Y cannot take negative values by construction
- $P(Y = 0) = P(Q \leq 0)$ since all $Q \leq 0$ get ”stacked” at $Y = 0$
- and for $y > 0$, $P(Y \leq y) = P(Q \leq 0) + P(0 < Q \leq y) = P(Q \leq y)$.

Consequently, the distribution function is ...

The Likelihood Function

$$F_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ F_Q(y) & \text{for } y \geq 0 \end{cases}$$

The random variable Y_i is of mixed type. The distribution function is differentiable for $y \neq 0$, the derivative $F'_Y(y) > 0$ for $y > 0$, and $F_Y(0) - F_Y(0^-) = F_Q(0) > 0$ is a point of discontinuity. The corresponding density function is

$$f_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ F_Q(0) = 1 - \Phi(X_i\beta/\sigma) & \text{for } y = 0 \\ f_Q(y) = \sigma^{-1}\phi[(y - X_i\beta)/\sigma] & \text{for } y > 0 \end{cases}$$

The Likelihood Function

The log-likelihood function for an SI sample of observations is

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

where $Z_i = (Y_i - X_i\beta)/\sigma$.

Digression

The density function of a mixed random variable may always be written as a weighted average of discrete and continuous components. If "missing" is denoted -999 in the data, then

$$f_Y(y) = \tau f_1(y) + (1 - \tau) f_2(y)$$

where $\tau = 1 - \Phi(X_i\beta/\sigma)$,

$$f_1(y) = \begin{cases} 0 & \text{for } y \neq -999 \\ 1 & \text{for } y = -999 \end{cases}$$

$$f_2(y) = \begin{cases} 0 & \text{for } y \leq 0 \\ \sigma^{-1}\phi[(y - X_i\beta)/\sigma]/\Phi(X_i\beta/\sigma) & \text{for } y > 0 \end{cases}$$

Subsample OLS

What would be the consequence of estimating the model by OLS using only the subsample of observations ($J_i = 1$) for which we have complete information?

- This would correspond to maximizing the first term in the log-likelihood function, $\sum_{i=1}^{n} J_i[-\ln\sigma + \ln\phi(Z_i)]$, while ignoring the second, $\sum_{i=1}^{n}(1 - J_i)\ln[1 - \Phi(X_i\beta/\sigma)]$.
- It is clear that subsample OLS is not an ML estimator. In fact, subsample OLS provides biased estimates of β.

Subsample OLS

To see the bias of subsample OLS, we need the subsample mean of the disturbance term.

$$\begin{aligned} E(\epsilon_i|J_i=1) &= E(\epsilon_i|Q_i>0) && \text{def. of } J_i \\ &= E(\epsilon_i|X_i\beta+\epsilon_i>0) && \text{def. of } Q_i \\ &= E(\epsilon_i|\epsilon_i>-X_i\beta) && \text{algebra} \\ &= E\left(\epsilon_i\Big|\frac{\epsilon_i}{\sigma}>\frac{-X_i\beta}{\sigma}\right) && \text{more algebra} \end{aligned}$$

Subsample OLS

$$\begin{aligned} E(\epsilon_i|J_i=1) &= \sigma E\left(\frac{\epsilon_i}{\sigma}\Big|\frac{\epsilon_i}{\sigma}>\frac{-X_i\beta}{\sigma}\right) && \text{mean of linear trans.} \\ &= \sigma\frac{\phi(-X_i\beta/\sigma)}{1-\Phi(-X_i\beta/\sigma)} && \text{censored cond. mean} \\ &= \sigma\frac{\phi(X_i\beta/\sigma)}{\Phi(X_i\beta/\sigma)} && \text{symmetry} \\ &= \sigma\gamma(X_i\delta) && \text{notation} \end{aligned}$$

Subsample OLS will provide biased estimates of β since the mean of the subsample disturbance term is observation specific and correlated with the regressors in X_i.

Heckman-Lee Estimation

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

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

where $\eta_i=\epsilon_i-\sigma\gamma(X_i\delta)$.

Heckman-Lee Estimation

Given the definition of η_i, we have:

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

$$\begin{aligned} &= \sigma^2 Var\left(\frac{\epsilon_i}{\sigma}\Big|\frac{\epsilon_i}{\sigma}>\frac{-X_i\beta}{\sigma}\right) \\ &= \sigma^2[1-(X_i\delta)\gamma(X_i\delta)-\gamma(X_i\delta)^2] \end{aligned}$$

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 nonlinear heteroskedastic regression model.

Heckman-Lee Estimation

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

Heckman-Lee Estimation

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

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

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

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

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

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

- 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 $\gamma(X_1\hat{\delta})$.

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 sample variation in constructed regressor $\gamma(X_i\hat{\delta})$.
- 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

- There are Monte Carlo studies that find OLS often has a MSE advantage over HL. This is because HL basically has its own built in source of near multicolinearity.
- The regressors used in HL are $Z_1 = [X_1 \quad \gamma(X_1\hat{\delta})]$. Note that X_1 and $X_1\hat{\delta}$ are perfectly colinear. It is only the nonlinear transformation $\gamma(\cdot) = \phi(\cdot)/\Phi(\cdot)$ that prevents perfect colinearity with HL. It may be shown, however, that over most of its range, $\gamma(\cdot)$ may be closely approximated by a linear function, so despite the absence of perfect colinearity, there is a high degree of near colinearity.
- Consequently, HL gives consistent estimates that are relatively imprecise, and OLS often has a MSE advantage despite its bias.

Heckman-Lee Estimation

- A final problem concerns estimation of σ. The estimate provided by the coefficient of $\gamma(X_i\hat{\delta})$ is frequently negative! The reason for this will become clearer after the discussion of Selection models.
- If we denote the second-stage residuals as $e_i = Y_i - Z_i\hat{\theta}$, and the estimated coefficient of $\gamma(X_i\hat{\delta})$ by $\hat{s}$, then Heckman suggests a "better" estimate of σ^2 based on the conditional variance of η_i is

$$\hat{\sigma}^2 = \sum_{i=1}^{n_1} \frac{e_i^2 + \hat{s}^2(X_i\hat{\delta})\gamma(X_i\hat{\delta}) + \hat{s}^2\gamma(X_i\hat{\delta})^2}{n_1 - k_X - 1}$$

Score Equations

The k score equations for β are

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

where $Z_i = (Y_i - X_i\beta)/\sigma$.

Score Equations

The score equation for σ is

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

Score Equations

The ML estimator solves

$$\sum_{i=1}^{n} \hat{\sigma}^{-1} \left\{ J_i \hat{Z}_i - (1 - J_i)\gamma(-X_i\hat{\beta}/\hat{\sigma}) \right\} X_i' = 0$$

and

$$\sum_{i=1}^{n} \hat{\sigma}^{-1} \left\{ J_i[\hat{Z}_i^2 - 1] + (1 - J_i)(X_i\hat{\beta}/\hat{\sigma})\gamma(-X_i\hat{\beta}/\hat{\sigma}) \right\} = 0$$

where $\hat{Z}_i = (Y_i - X_i\hat{\beta})/\hat{\sigma}$. This is a set of $k+1$ simultaneous nonlinear implicit functions for $\hat{\beta}$ and $\hat{\sigma}$ which may be solved with the BHHH algorithm.

A Global MLE

The Hessian matrix for $\ln L(\beta, \sigma)$ is not negative definite for all (β, σ). Nevertheless, a root to the score equations will be a unique global MLE. To see why, we must first express the log-likelihood function in terms of an alternative set of parameters. Specifically, let $\tau = \sigma^{-1}$ and $\delta = \sigma^{-1}\beta$. The log-likelihood function is now

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

where, under this set of parameters, $Z_i = (\tau Y_i - X_i\delta)$.

A Global MLE

The score equations for δ and τ are

$$\frac{\partial \ln L(\delta, \tau)}{\partial \delta} = \sum_{i=1}^{n} \{ J_i Z_i - (1 - J_i)\gamma(-X_i\delta) \} X_i'$$

and

$$\frac{\partial \ln L(\delta, \tau)}{\partial \tau} = \sum_{i=1}^{n} \{ J_i[\tau^{-1} - Z_i Y_i] \}$$

where $Z_i = (\tau Y_i - X_i\delta)$.

A Global MLE

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 $Z = -X_i\delta$, we have

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

A Global MLE

The Hessian matrix for $\ln L(\delta, \tau)$ is composed of the following blocks

$$\begin{aligned}\frac{\partial^2 \ln L(\delta,\tau)}{\partial\delta\partial\delta'} &= \sum_{i=1}^{n} X_i' \left\{ J_i(\partial Z_i/\partial\delta') - (1-J_i)[\partial\gamma(-X_i\delta)/\partial\delta'] \right\} \\ &= \sum_{i=1}^{n} X_i' \left\{ -J_i - (1-J_i)[\gamma(-X_i\delta) - X_i\delta]\gamma(-X_i\delta) \right\} X_i \\ &= -(X_1'X_1 + X_0'BX_0)\end{aligned}$$

where B is an $n_0 \times n_0$ diagonal matrix with diagonal elements $[\gamma(-X_i\delta) - X_i\delta]\gamma(-X_i\delta)$ corresponding to the $J_i = 0$ observations.

A Global MLE

$$\begin{aligned}\frac{\partial^2 \ln L(\delta,\tau)}{\partial\delta\partial\tau} &= \sum_{i=1}^{n} X_i' \left\{ J_i(\partial Z_i/\partial\tau) \right\} \\ &= \sum_{i=1}^{n} X_i' \left\{ J_i Y_i \right\} \\ &= X_1'Y_1\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln L(\delta,\tau)}{\partial\tau\partial\tau} &= \sum_{i=1}^{n} \left\{ J_i(\partial\tau^{-1}/\partial\tau) - Y_i(\partial Z_i/\partial\tau) \right\} \\ &= \sum_{i=1}^{n} J_i(-\tau^{-2} - Y_i^2) \\ &= -(n_1\tau^{-2} + Y_1'Y_1)\end{aligned}$$

A Global MLE

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

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

and the $n \times (k+1)$ matrices

$$W = \begin{bmatrix} X_1 & -Y_1 \\ X_0 & 0 \end{bmatrix} \text{ and } V = \begin{bmatrix} 0 & \tau^{-1}J \end{bmatrix}$$

A Global MLE

Using this notation, the Hessian may be written as

$$\frac{\partial^2 \ln L(\delta,\tau)}{\partial\theta\partial\theta'} = -(W'CW + V'V)$$

- The matrix $W'CW$ is positive definite if W has full column rank (which is satisfied by assumption) and if the diagonal elements of C are strictly positive.
- These elements of C are 1 and $[\gamma(-X_i\delta) - X_i\delta]\gamma(-X_i\delta)$ for the $J_i = 1$ and $J_i = 0$ cases, respectively. As seen during the discussion of the Probit model, the term in brackets is positive, since it may be written as $E(Z|Z > K) - K$, where $Z \sim N(0,1)$ and $K = X_i\delta$.

A Global MLE

- The matrix $V'V$ is positive semi-definite since the rank of V is 1.
- Consequently, the Hessian matrix is negative definite since it is the sum of negative definite and negative semi-definite matrices.
- Since $\ln L(\delta, \tau)$ is globally concave in δ and τ, any root to the score equations for δ and τ is a unique global MLE.

The Original Parameters

- We have just established that any root to the score equations for δ and τ is a unique global MLE. So what?
- To see the importance of this result, note that the transformations $\tau = \sigma^{-1}$ and $\delta = \sigma^{-1}\beta$ are one-to-one transformations from the parameter space of (β, σ) to that of (δ, τ).
- A similar result holds for the inverse transformations $\sigma = \tau^{-1}$ and $\beta = \tau^{-1}\delta$. Any unique point in the space of (β, σ) corresponds to a unique point in the space of (δ, τ), and vice versa.
- The transformations are said to be "one to one and onto."

The Original Parameters

Proposition 1: If $(\hat{\delta}, \hat{\tau})$ is a root to the score equations for δ and τ, then $(\hat{\beta} = \hat{\tau}^{-1}\hat{\delta}, \hat{\sigma} = \hat{\tau}^{-1})$ is a root to the score equations for β and σ.

By the chain rule,

$$\begin{aligned}\frac{\partial \ln L(\beta, \sigma)}{\partial \beta'} &= \frac{\partial \ln L(\delta, \tau)}{\partial \delta'}\frac{\partial \delta}{\partial \beta'} \\ &= \frac{\partial \ln L(\delta, \tau)}{\partial \delta'}[\sigma^{-1} I_k] \\ &= \sigma^{-1}\frac{\partial \ln L(\delta, \tau)}{\partial \delta'}\end{aligned}$$

The Original Parameters

Likewise,

$$\begin{aligned}\frac{\partial \ln L(\beta, \sigma)}{\partial \sigma} &= \frac{\partial \ln L(\delta, \tau)}{\partial \delta'}\frac{\partial \delta}{\partial \sigma} + \frac{\partial \ln L(\delta, \tau)}{\partial \tau}\frac{\partial \tau}{\partial \sigma} \\ &= \frac{\partial \ln L(\delta, \tau)}{\partial \delta'}[-\sigma^{-2}\beta] + \frac{\partial \ln L(\delta, \tau)}{\partial \tau}[-\sigma^{-2}] \\ &= -\sigma^{-2}\left[\frac{\partial \ln L(\delta, \tau)}{\partial \delta'}\beta + \frac{\partial \ln L(\delta, \tau)}{\partial \tau}\right]\end{aligned}$$

These results imply that if $(\hat{\delta}, \hat{\tau})$ is a root to the score equations for δ and τ, then $(\hat{\beta}, \hat{\sigma})$ is a root to the score equations for β and σ, and vice versa.

The Original Parameters

Proposition 2: If $(\hat{\delta}, \hat{\tau})$ is a global MLE of (δ, τ), then $(\hat{\beta}, \hat{\sigma})$ is the global MLE of (β, σ).

We have seen that $\ln L(\delta, \tau) = \ln L(\beta, \sigma)$ for any (δ, τ) satisfying $\delta = \sigma^{-1}\beta$ and $\tau = \sigma^{-1}$. Assume, contrary to the theorem, that there exists a vector $(\tilde{\beta}, \tilde{\sigma})$ such that $\ln L(\tilde{\beta}, \tilde{\sigma}) > \ln L(\hat{\beta}, \hat{\sigma})$. Then it must be that for $\tilde{\delta} = \tilde{\sigma}^{-1}\tilde{\beta}$ and $\tilde{\tau} = \tilde{\sigma}^{-1}$ we have $\ln L(\tilde{\delta}, \tilde{\tau}) > \ln L(\hat{\delta}, \hat{\tau})$ since the transformations are bijective. This contradicts the statement that $(\hat{\delta}, \hat{\tau})$ is a global MLE of (δ, τ). Consequently, no such $(\tilde{\beta}, \tilde{\sigma})$ can exist, and $(\hat{\beta}, \hat{\sigma})$ is the global MLE.

The Original Parameters

- The first two propositions suggest that the global MLE of (β, σ) may be obtained by maximizing $\ln L(\delta, \tau)$ to obtain the global MLE, $(\hat{\delta}, \hat{\tau})$, and then solving for $(\hat{\beta}, \hat{\sigma})$ as $(\hat{\tau}^{-1}\hat{\delta}, \hat{\tau}^{-1})$. This method does not provide standard errors for $(\hat{\beta}, \hat{\sigma})$, however.
  - Standard errors can be obtained by using $(\hat{\tau}^{-1}\hat{\delta}, \hat{\tau}^{-1})$ as starting values in a BHHH algorithm to maximize $\ln L(\beta, \sigma)$. For a sufficiently tight convergence criterion, this will yield the global MLE of (β, σ) and its standard errors, since $(\hat{\tau}^{-1}\hat{\delta}, \hat{\tau}^{-1})$ must be in the neighborhood of the global.
  - Alternatively, one could use the "delta method", which requires the Jacobian of the transformations and the covariance matrix of $(\hat{\delta}, \hat{\tau})$, to estimated the covariance matrix of $(\hat{\beta}, \hat{\sigma})$.

The Original Parameters

Proposition 3: If $(\hat{\beta}, \hat{\sigma})$ is a root to the score equations for β and σ, it is unique.

If there exists a second distinct root to the score equations for β and σ, then because the transformations are one-to-one and onto, the relationship between the score equations obtained in Proposition 1 above implies that there must be a second distinct root to the score equations for δ and τ. This is ruled out by the global concavity of $\ln L(\delta, \tau)$.

The Original Parameters

- The most important implication of Proposition 3 is that there is never any need to mess with the score equations for δ and τ. They were necessary only from a theoretical perspective. Just use the score equations for β and σ, and if you get a root it's the global MLE.
- The results above are a reminder that the usual conditions for a unique global MLE are sufficient conditions, not necessary conditions. If the sufficient conditions are not satisfied, the problem may sometimes be solved by other means.
- This problem also illustrates why many statisticians prefer to estimate σ^{-1}, called the "precision," rather than the variance or standard deviation.

Positive Variance

With the Tobit model, σ must be strictly positive.

- Many algorithms impose this restriction in an ad-hoc manner; check the restriction, if violated, reset σ to some small positive value. This approach can lead to ”cycling” on the boundary of the parameter space.
- A better approach employs the transformation $\sigma = \exp(\lambda)$, where the parameter λ is estimated instead of σ. The restriction is imposed by the transformation since real values of λ generate positive values of σ.
- Neither of these methods is necessary in the neighborhood of the maximum. The value of these methods is preventing the algorithm from ”crashing” when the starting values are poor.

Positive Variance

To use the transformation $\sigma = \exp(\lambda)$ we need score equations for β and λ.

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

Positive Variance

- Using the chain rule,

$$\frac{\partial \ln L(\beta,\lambda)}{\partial \lambda} = \frac{\partial \ln L(\beta,\sigma)}{\partial \sigma}\frac{\partial \sigma}{\partial \lambda}$$

Since $\partial\sigma/\partial\lambda = \partial \exp(\lambda)/\partial\lambda = \exp(\lambda) = \sigma$,

$$\frac{\partial \ln L(\beta,\lambda)}{\partial \lambda} = \frac{\partial \ln L(\beta,\sigma)}{\partial \sigma}\sigma$$

This equality expresses the score equation for λ in terms of β and σ.

Positive Variance

An iterating algorithm for (β,λ) may be summarized as follows:

- Given the current value of (β,λ), find the corresponding value of (β,σ) and use it to evaluate the score equations for β and σ.
- Convert the numerical value of the score equation for σ to that of λ by multiplying by σ.
- Apply the BHHH algorithm to get an updated value for (β,λ).
- Repeat until convergence.

Positive Variance

The ML estimator of σ is $\exp(\hat{\lambda})$.

- Standard errors for $\exp(\hat{\lambda})$ may be obtained using the ”delta method.”
- Alternatively, once a root is found, $\hat{\beta}$ and $\exp(\hat{\lambda})$ may be used as starting values for a BHHH algorithm using the score equations for β and σ. Since the Tobit model has a unique root, these starting values will be in the neighborhood of the global maximum for a sufficiently tight convergence criteria.
- Note that we NEVER need to program the score equation for λ. All of this is done by simply re-weighting the numerical value of the score equation for σ.