Limited Dependent Variable Models

- Limited dependent variable (LDV) models might best be characterized as linear regression models subject to incomplete observation of the dependent variable.
- In order to be estimable, some information concerning the dependent variable must be available. This information may be of a purely qualitative nature, or a combination of qualitative and quantitative information.
- Given knowledge of the structure that limits observation of the dependent variable, the likelihood function for the observed data is determined, and estimates are obtained by the method of maximum likelihood (ML). For most LDV models, the ML estimators are the solution to a set of simultaneous nonlinear implicit functions.

The Likelihood Function

The likelihood function is defined to be any function proportional to the joint density function evaluated at the observed data. That is, $L(\theta) \propto f(X)$. This is often written as,

$$\begin{aligned} L(\theta) &= Kf(X) \\ &= K\Pi_{i=1}^{n} f(X_i) \qquad \text{with SI observations} \end{aligned}$$

for any constant (with respect to θ) K. Taking logs we have

$$\begin{aligned} \ln L(\theta) &= \ln(K) + \ln[f(X)] \\ &= \ln(K) + \sum_{i=1}^{n} \ln[f(X_i)] \qquad \text{with SI observations} \end{aligned}$$

The Likelihood Function

The log transformation is used because it is typically easier to differentiate a sum than a product. It is clear that:

- the maximized value of the likelihood function will differ from the maximized value of the log-likelihood function. That is, $\max_{\theta} L(\theta) \neq \max_{\theta} \ln L(\theta)$.
- the maximizing value of θ is invariant to the log transformation (since it is monotonic). That is, $\underset{\theta}{\operatorname{argmax}}\, L(\theta) = \underset{\theta}{\operatorname{argmax}} \ln L(\theta)$.
- the maximizing value of θ is invariant to the choice of K. The restriction $K = 1$ may be imposed without loss of generality.

The Score and Hessian

- Assuming that $f(X)$ is continuous and differentiable in θ, we can employ the methods of differential calculus to find a maximum. The score equations are a fundamental concept in this approach.
- The score equations are the vector of first partials of the log-likelihood function. That is, $S(\theta) = \partial \ln L(\theta)/\partial\theta$.
- A vector $\hat{\theta}$ is a root to the score equations if $S(\hat{\theta}) = 0$. Extreme values of the function $\ln L(\theta)$ correspond to roots of the score equations.
- The Hessian is the matrix of second partials and cross partials of the log-likelihood function. That is, $H(\theta) = \partial^2 \ln L(\theta)/\partial\theta\partial\theta' = \partial S(\theta)/\partial\theta'$.

Local and Global MLE

- A local maximum likelihood estimator (MLE) is defined as a root to the score equations that corresponds to a local maximum of $\ln L(\theta)$. Sufficient conditions to establish that $\hat{\theta}$ is a local MLE are:
  1. $S(\hat{\theta}) = 0$
  2. The Hessian matrix is negative definite at $\hat{\theta}$.
- A global maximum likelihood estimator (MLE) is defined as the value of θ that maximizes $\ln L(\theta)$ over the entire parameter space. Sufficient conditions to establish that $\hat{\theta}$ is a unique interior global MLE are:
  1. $S(\hat{\theta}) = 0$
  2. $H(\theta)$is negative definite for all θ in the parameter space.

Theorem: Properties of MLE

Let $\hat{\theta}$ denote a global MLE. Also, let $\Omega = \lim_{n\to\infty}[n^{-1}I(\theta)]^{-1}$, where $I(\theta)$ is the small sample information matrix. Then subject to certain regularity conditions (Amemiya Theorems 4.1.3 and 4.2.4),

1. $\hat{\theta} \to \theta$

   The global MLE is consistent.

2. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \Omega)$

   The global MLE is asymptotically normal and asymptotically efficient.

Theorem: Delta Method

Assume that:

1. $\hat{\theta} \to \theta_0$
2. $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \Omega)$

Let $g(\theta)$ denote a set of m continuous and differentiable transformations of the k-vector θ, with partial derivatives $G(\theta) = \partial g(\theta)/\partial\theta'$. Then, provided that $g(\theta_0)$ and $G(\theta_0)$ exist,

1. $g(\hat{\theta}) \to g(\theta_0)$
2. $\sqrt{n}[g(\hat{\theta}) - g(\theta_0)] \xrightarrow{D} N(0, G(\theta_0)\Omega G(\theta_0)')$

Theorem: Delta Method

A first-order Taylor's series about θ_0 gives

$$g(\hat{\theta}) - g(\theta_0) = G(\bar{\theta})(\hat{\theta} - \theta_0)$$

where $\bar{\theta} = \alpha\hat{\theta} + (1-\alpha)\theta_0$ for some $\alpha \in (0, 1)$.

- $\bar{\theta} \to \theta_0$ since $\hat{\theta} \to \theta_0$
- $g(\hat{\theta}) \to g(\theta_0)$ since $G(\bar{\theta}) \to G(\theta_0)$ and $\hat{\theta} \to \theta_0$
- $\sqrt{n}[g(\hat{\theta}) - g(\theta_0)] = G(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, G(\theta_0)\Omega G(\theta_0)')$ by the Generalized Slutsky theorem.

Application: Delta Method

Estimate the following model by OLS

$$\ln(SP) = \alpha + \beta SQFT + \delta BEDS + \epsilon$$

using the sold subsample of the MLS data set.

- Get an estimate of $g(\theta) = \beta\delta$ using the delta method.

$$\begin{aligned} g(\hat{\theta}) &= \hat{\beta}\hat{\delta} \\ &= (0.6729052)(-0.0770728) \\ &= -0.05187 \end{aligned}$$

Application: Delta Method

- Get an estimate of $Var[g(\hat{\theta})]$ using the delta method.

For this choice of $g(\theta)$, the vector of partials is:

$$\begin{aligned} G(\theta) &= \partial g(\theta)/\partial\theta' \\ &= [\partial g(\theta)/\partial\alpha \quad \partial g(\theta)/\partial\beta \quad \partial g(\theta)/\partial\delta] \\ &= [0 \quad \delta \quad \beta] \end{aligned}$$

The estimated covariance matrix is

$$\begin{aligned} G(\hat{\theta})\Sigma(\hat{\theta})G(\hat{\theta})' &= \hat{\delta}^2 Var(\hat{\beta}) + \hat{\beta}^2 Var(\hat{\delta}) + 2\hat{\delta}\hat{\beta}Cov(\hat{\beta},\hat{\delta}) \\ &= (-.0770728)^2(.00212656) + (.6729052)^2(.00260929) \\ &\quad + 2(.6729052)(-.0770728)(-.0017134) \\ &= 0.001372 \end{aligned}$$

Application: Delta Method

- Use an asymptotic t-test to test the null hypothesis $H_0 : \beta\delta = -0.1$.

If we denote the standard error by $s[g(\hat{\theta})] = \sqrt{G(\hat{\theta})\Sigma(\hat{\theta})G(\hat{\theta})'}$, then the sample statistic is:

$$\begin{aligned} \frac{[g(\hat{\theta}) - g(\theta_0)]}{s[g(\hat{\theta})]} &= \frac{(-0.05186 + 0.1)}{\sqrt{0.001372}} \\ &= 1.29965 \end{aligned}$$

This sample statistic has a p-value of 0.1959 and so is not significant at conventional levels.

- Note that the standard error cannot be estimated as $\sqrt{G(\theta_0)\Sigma(\hat{\theta})G(\theta_0)'}$ since the null hypothesis is not sufficient to determine the distinct values of β and δ.

The Newton-Raphson Algorithm

For many LDV problems, the score equations are a set of simultaneous nonlinear implicit functions. We need some method for solving these equations. A second order Taylor's series of $\ln L(\theta)$ about some initial guess θ_1 gives

$$\ln L(\theta) \approx \ln L(\theta_1) + (\theta - \theta_1)'S(\theta_1) + \frac{1}{2}(\theta - \theta_1)'H(\theta_1)(\theta - \theta_1)$$

Choosing θ to maximize this quadratic approximation to $\ln L(\theta)$, and denoting the maximizing value by θ_2 gives

$$S(\theta_1) + H(\theta_1)(\theta_2 - \theta_1) = 0$$

where $H(\theta_1)$ is negative definite. Solving for θ_2 gives

$$\theta_2 = \theta_1 - H(\theta_1)^{-1}S(\theta_1)$$

The Newton-Raphson Algorithm

With nonlinear score equations, θ_2 will generally fail to solve $S(\theta_2) = 0$, in which case the process is repeated. Updated values of the score equations, $S(\theta_2)$, and the Hessian, $H(\theta_2)$, are computed, and an updated value of θ, call it θ_3, is computed as

$$\theta_3 = \theta_2 - H(\theta_2)^{-1}S(\theta_2)$$

On the j^{th} iteration, the Newton-Raphson Algorithm choose

$$\theta_{j+1} = \theta_j - H(\theta_j)^{-1}S(\theta_j)$$

The algorithm continues until "convergence." That is, until $\theta_{j+1} - \theta_j \approx 0$.

The Newton-Raphson Algorithm

There are several questions of interest concerning this algorithm.

1. When will an iteration yield an increase in the value of the function?
   - Substituting $S(\theta_j) = -H(\theta_j)(\theta_{j+1} - \theta_j)$ into the quadratic approximation for $\ln L(\theta_{j+1})$ gives

$$\ln L(\theta_{j+1}) \approx \ln L(\theta_j) - \frac{1}{2}(\theta_{j+1} - \theta_j)'H(\theta_j)(\theta_{j+1} - \theta_j)$$

   - Hence, we can expect an improvement in the value of the function, $\ln L(\theta_{j+1}) > \ln L(\theta_j)$, if $H(\theta_j)$ is negative definite.

The Newton-Raphson Algorithm

2. Does convergence of the algorithm yield a root to the score equations?
   - Note that $\theta_{j+1} - \theta_j \approx 0$ implies $-H(\theta_j)^{-1}S(\theta_j) \approx 0$.
   - If $H(\theta_j)$ is nonsingular, and failure of this condition would be obvious in computation of θ_{j+1}, then $S(\theta_j) \approx 0$ and θ_j is a root to the score equations.

The Method of Scoring Algorithm

The "method of scoring" replaces the observed second derivative matrix in the NR algorithm, $H(\theta)$, with the expected second derivative matrix, $E_X[H(\theta)] = -I(\theta)$. Hence, θ_{j+1} is determined as

$$\theta_{j+1} = \theta_j + I(\theta_j)^{-1}S(\theta_j)$$

The Method of Scoring Algorithm

Since $I(\theta) = E_X[S(\theta)S(\theta)']$, it is often claimed that one advantage of this algorithm is that it can be utilized without the burden of computing the matrix of second derivatives. This fails to recognize that:

1. it is often easier to compute $E_X[H(\theta)]$ than $E_X[S(\theta)S(\theta)']$.
2. without an examination of $H(\theta)$, one cannot be certain whether the root is a local or global MLE.

Berndt, Hall, Hall, and Hausman

The Berndt, Hall, Hall, and Hausman Algorithm (BHHH) replaces the population moment matrix $E_X[S(\theta)S(\theta)']$ used in the "method of scoring" algorithm, with the corresponding sample moment matrix

$$\sum_{i=1}^{n} \left[\frac{\partial \ln f(X_i)}{\partial \theta}\right] \left[\frac{\partial \ln f(X_i)}{\partial \theta'}\right]$$

where $\ln f(X_i)$ is the log-likelihood function of the i^{th} observation.

Berndt, Hall, Hall, and Hausman

Note that $\partial \ln f(X_i)/\partial\theta'$ is a row vector of length k. Stacking these rows into an $n \times k$ matrix gives $V(\theta) = [\partial \ln f(X_i)/\partial\theta']$ for $i = 1 \ldots n$. The sample moment matrix may be written as

$$V(\theta)'V(\theta) = \sum_{i=1}^{n} \left[\frac{\partial \ln f(X_i)}{\partial \theta}\right] \left[\frac{\partial \ln f(X_i)}{\partial \theta'}\right]$$

Berndt, Hall, Hall, and Hausman

In matrix notation, the BHHH algorithm chooses θ_{j+1} as

$$\theta_{j+1} = \theta_j + [V(\theta_j)'V(\theta_j)]^{-1}S(\theta_j)$$
$$\theta_{j+1} = \theta_j + [V(\theta_j)'V(\theta_j)]^{-1}V(\theta_j)'i$$

where i is an $n \times 1$ vector of ones. Computationally, the BHHH update, $\theta_{j+1} - \theta_j$, is the result of an OLS regression of i on the columns of $V(\theta_j)$.