The Binary Choice Model

The Binary Choice (BC) model may be interpreted as a classical regression model subject to qualitative observation of the dependent variable. Specifically, assume that

$$Y_i = X_i\beta - \sigma\epsilon_i$$

satisfies all CLRM assumptions except that Y_i is unobserved. Instead, we observe a binary variable, J_i, that equals one when Y_i is positive.

- The error structure is classical since $\epsilon_i \sim iid(0,1)$ implies that $-\sigma\epsilon_i \sim iid(0,\sigma^2)$.
- The latent model cannot be estimated since Y_i is unobserved.

The Binary Choice Model

In order to estimate the model, we need to determine the relationship between the observed dependent variable J_i and the regressors in X_i.

By definition of J_i

- $J_i = 1 \Leftrightarrow Y_i > 0 \Leftrightarrow X_i\beta - \sigma\epsilon_i > 0 \Leftrightarrow \epsilon_i < X_i\beta/\sigma$
- $J_i = 0 \Leftrightarrow Y_i \leq 0 \Leftrightarrow X_i\beta - \sigma\epsilon_i \leq 0 \Leftrightarrow \epsilon_i \geq X_i\beta/\sigma$

which implies that

- $P(J_i = 1) = P(\epsilon_i < X_i\beta/\sigma) = F(X_i\beta/\sigma)$
- $P(J_i = 0) = P(\epsilon_i \geq X_i\beta/\sigma) = 1 - F(X_i\beta/\sigma)$

where $F(\cdot)$ denotes the distribution function of ϵ_i and need not be normal in form.

The Binary Choice Model

The structure here is just a Bernoulli trial with observation specific probability $p_i = F(X_i\beta/\sigma)$. The marginal density of J_i is

$$\begin{aligned} f(J_i) &= p_i^{J_i}(1-p_i)^{(1-J_i)} \\ &= F(X_i\beta/\sigma)^{J_i}[1 - F(X_i\beta/\sigma)]^{(1-J_i)} \end{aligned}$$

or in log terms

$$\ln[f(J_i)] = J_i \ln[F(X_i\beta/\sigma)] + (1 - J_i)\ln[1 - F(X_i\beta/\sigma)]$$

The Binary Choice Model

Since the likelihood function is proportional to the joint density function of the J_i,

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

We will see that the $k+1$ parameters (β,σ) are not identified, but that the k standardized coefficients $\delta = \sigma^{-1}\beta$ are identified. In terms of δ, the log-likelihood function is

$$\ln L(\delta) = \sum_{i=1}^{n}\{J_i \ln[F(X_i\delta)] + (1 - J_i)\ln[1 - F(X_i\delta)]\}$$

The Binary Choice Model

While the latent regression is a classical linear regression model, the corresponding binary choice model is a heteroskedastic nonlinear regression model. To see this, note that

$$J_i = F(X_i\delta) + u_i$$

where

$$u_i = \begin{cases} 1 - F(X_i\delta) & \text{w/p} \quad F(X_i\delta) \\ -F(X_i\delta) & \text{w/p} \quad 1 - F(X_i\delta) \end{cases}$$

The Binary Choice Model

The error in this model is discrete, with two points of positive probability. It has mean

$$\begin{aligned} E(u_i) &= [1 - F(X_i\delta)]F(X_i\delta) - F(X_i\delta)[1 - F(X_i\delta)] \\ &= 0 \end{aligned}$$

and variance

$$\begin{aligned} E(u_i^2) &= [1 - F(X_i\delta)]^2 F(X_i\delta) + [-F(X_i\delta)]^2[1 - F(X_i\delta)] \\ &= [1 - F(X_i\delta)]F(X_i\delta)[1 - F(X_i\delta) + F(X_i\delta)] \\ &= [1 - F(X_i\delta)]F(X_i\delta) \end{aligned}$$

The Binary Choice Model

The binary choice model is:

- nonlinear in δ if $F(\cdot)$ is nonlinear. This is the case for all distribution functions except the uniform.
- heteroskedastic as long as $F(X_i\delta)$ is not constant.
- a sequence of Bernoulli trials with non-common probability $p_i = F(X_i\delta)$ when $X_i\delta$ is not a constant.
- a sequence of Bernoulli trials with common probability $p = F(X_i\delta)$ when $X_i\delta$ is a constant. This case occurs with an intercept-only model or if there is no sample variation in the regressors.

The Binary Choice Model

- The binary choice model was presented as a classical regression model subject to qualitative observation of the dependent variable. This was done primarily to illustrate the statistical relationships between the models.
- In reality, this process of censoring the quantitative information need never occur. In fact, the ”best” examples of binary choice models often have no underlying latent structure. For example, we conduct a random sample of individuals and ask them to respond to the question ”Do you like fish sticks?” The answer is qualitative in nature, as one cannot give a meaningful answer to the question ”How much do you like fish sticks?”

The Binary Choice Model

- Quantitative information gives more precise estimates. One should never deliberately censor the quantitative information in order to use a binary choice model. Perhaps the only potential excuse one could come up with for doing this would be if the qualitative information was somehow more reliable than the quantitative information.

The Linear Probability Model

The Binary Choice model represents a family of models indexed by the choice of distribution function $F(\cdot)$. The Linear Probability (LP) model is the specification obtained when $\epsilon_i \sim iidU(0,1)$. Since the Uniform distribution function is $F(X_i\delta) = X_i\delta$, for $0 < X_i\delta < 1$, the LP Model is a heteroskedastic linear regression. That is,

$$J_i = X_i\delta + u_i$$

where $\sigma^2(u_i) = (1 - X_i\delta)(X_i\delta)$. Hence, OLS estimates of δ are unbiased, but inefficient, due to the heteroskedastic error term.

The Linear Probability Model

The log-likelihood function of the LP model is

$$\ln L(\delta) = \sum_{i=1}^{n} \{J_i \ln(X_i\delta) + (1 - J_i) \ln(1 - X_i\delta)\}$$

The ML estimator of δ is the solution to $S(\hat{\delta}) = 0$. For the LP model, this is

$$S(\hat{\delta}) = \sum_{i=1}^{n} \left\{ \frac{J_i}{X_i\hat{\delta}} - \frac{(1 - J_i)}{(1 - X_i\hat{\delta})} \right\} X_i' = 0$$

or after a little algebra ...

The Linear Probability Model

$$\sum_{i=1}^{n} \left\{ \frac{(J_i - X_i\hat{\delta})}{X_i\hat{\delta}(1 - X_i\hat{\delta})} \right\} X_i' = 0$$

These score equations are identical to the ”normal equations” of the weighted least-squares estimator. That is, the ML estimator of δ is just OLS applied to

$$W_iJ_i = W_iX_i\delta + W_iu_i$$

where $W_i = [(1 - X_i\delta)(X_i\delta)]^{-0.5}$

The Linear Probability Model

There are theoretical and empirical difficulties with the LP model. In order to be a valid argument to the Uniform distribution function, $X_i\delta$ is restricted to the unit interval.

- The theoretical difficulty is that the parameter space and samples space are intertwined. Different values of δ result in different restrictions on the sample space of X.
- The empirical difficulty is that for values of X_i near the boundary of the sample space, sample variation in $\hat{\delta}$ may result in values of $X_i\hat{\delta}$ outside the unit interval.
- The advantage of the LP Model is that it provides a particularly simple estimation procedure.

The Logit Model

The Logit model is the specification of the BC model obtained when $\epsilon_i \sim iid\sec^2$. In this case, $F(X_i\delta) = \Psi(X_i\delta)$, where $\Psi(\cdot)$ denotes the Logistic distribution function. The log-likelihood function of the Logit model is

$$\ln L(\delta) = \sum_{i=1}^{n} \{J_i \ln[\Psi(X_i\delta)] + (1 - J_i) \ln[1 - \Psi(X_i\delta)]\}$$

and the score equations are

$$\partial \ln L(\delta)/\partial\delta = \sum_{i=1}^{n} \left\{ J_i \frac{\psi(X_i\delta)}{\Psi(X_i\delta)} - (1 - J_i)\frac{\psi(X_i\delta)}{1 - \Psi(X_i\delta)} \right\} X_i'$$

The Logit Model

Since the Logistic density and distribution functions satisfy $\psi(Z) = \Psi(Z)[1 - \Psi(Z)]$, the score equations may be written as

$$\partial \ln L(\delta)/\partial\delta = \sum_{i=1}^{n} \{J_i[1 - \Psi(X_i\delta)] - (1 - J_i)\Psi(X_i\delta)\} X_i'$$

and the ML estimator, $\hat{\delta}$, satisfies

$$\sum_{i=1}^{n} \left\{ J_i - \Psi(X_i\hat{\delta}) \right\} X_i' = 0$$

This is a set of k simultaneous nonlinear implicit functions and may be solved using the algorithms discussed earlier.

The Logit Model

The matrix of second derivatives for the Logit model is negative definite for all values of δ. Specifically,

$$\begin{aligned} \frac{\partial^2 \ln L(\delta)}{\partial\delta\partial\delta'} &= \sum_{i=1}^{n} X_i' \left\{ \frac{\partial[J_i - \Psi(X_i\delta)]}{\partial\delta'} \right\} \\ &= -\sum_{i=1}^{n} X_i'\{\psi(X_i\delta)\}X_i \\ &= -X'DX \end{aligned}$$

where D is a diagonal matrix with diagonal elements $d_{ii} = \psi(X_i\delta) > 0$.

The Logit Model

Since:

- X has full column rank k, and
- D is diagonal with strictly positive diagonal elements,

the matrix $X'DX$ is positive definite, and the Hessian matrix, $H(\delta) = -X'DX$, is negative definite for any δ. Thus, and any root to the score equations is a unique interior global MLE.

The Probit Model

The Probit model is the specification of the BC model obtained when $\epsilon_i \sim iidN(0,1)$. In this case, $F(X_i\delta) = \Phi(X_i\delta)$, where $\Phi(\cdot)$ denotes the standard normal distribution function. The log-likelihood function of the Probit model is

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

and the score equations are

$$\partial \ln L(\delta)/\partial\delta = \sum_{i=1}^{n} \left\{ J_i \frac{\phi(X_i\delta)}{\Phi(X_i\delta)} - (1 - J_i)\frac{\phi(X_i\delta)}{1 - \Phi(X_i\delta)} \right\} X_i'$$

The Probit Model

Let $\gamma(Z) = \phi(Z)/\Phi(Z)$ denote the "inverse Mill's ratio." Then the score equations may be written as

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

and the ML estimator, $\hat{\delta}$, satisfies

$$\sum_{i=1}^{n} \left\{ J_i\gamma(X_i\hat{\delta}) - (1 - J_i)\gamma(-X_i\hat{\delta}) \right\} X_i' = 0$$

This is a set of k simultaneous nonlinear implicit functions and may be solved using the algorithms discussed earlier.

The Probit Model

Will a root to the score equations correspond to a global MLE? The answer is yes. The matrix of second derivatives for the Probit model is negative definite for all values of δ.

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

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

The Probit Model

Likewise, if we let $Z = -X_i\delta$, we have

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

The Hessian matrix for the probit model is then

$$\begin{aligned}\frac{\partial^2 \ln L(\delta)}{\partial\delta\partial\delta'} &= \sum_{i=1}^{n} X_i' \left\{J_i\partial\gamma(X_i\delta)/\partial\delta' - (1-J_i)\partial\gamma(-X_i\delta)/\partial\delta'\right\} \\ &= -\sum_{i=1}^{n} X_i'\{d_{ii}\}X_i \\ &= -X'DX\end{aligned}$$

The Probit Model

where D is an $n \times n$ diagonal matix with diagonal elements

$$d_{ii} = J_i[X_i\delta + \gamma(X_i\delta)]\gamma(X_i\delta) + (1-J_i)[\gamma(-X_i\delta) - X_i\delta]\gamma(-X_i\delta)$$

for $i = 1, ..., n$.

Note that

- $d_{ii} = [X_i\delta + \gamma(X_i\delta)]\gamma(X_i\delta)$ when $J_i = 1$

  and
- $d_{ii} = [\gamma(-X_i\delta) - X_i\delta]\gamma(-X_i\delta)$ when $J_i = 0$

Since $\gamma(\cdot) > 0$ for all arguments, the sign of d_{ii} in each case is determined by the sign of the term in brackets.

The Probit Model

These terms may be written as conditional expectations of a standard normal. Specifically, if $K = X_i\delta$ and $Z \sim N(0,1)$, then

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

and likewise

$$[\gamma(-X_i\delta) - X_i\delta] = [E(Z|Z > K) - K] > 0$$

Hence, $d_{ii} > 0$ for all i.

Digression

The results on the previous slide should be intuitive, but to illustrate more directly, for any random variable Z,

$$\begin{aligned}E(Z|Z \leq K) &= \int_{-\infty}^{K} Z f(Z|Z \leq K)dZ \\ &\leq \int_{-\infty}^{K} K f(Z|Z \leq K)dZ \qquad \text{since } Z \leq K \\ &\leq K\int_{-\infty}^{K} f(Z|Z \leq K)dZ \\ &\leq K\end{aligned}$$

Hence, $K - E(Z|Z \leq K) \geq 0$, with the equality holding only when Z is degenerate at K, which of course is not the case with the censored normal.

The Probit Model

Since:

- X has full column rank k, and
- D is diagonal with strictly positive diagonal elements,

the matrix $X'DX$ is positive definite, and the Hessian matrix, $H(\delta) = -X'DX$, is negative definite for any δ. Thus, any root to the score equations is a unique interior global MLE.

Identification

- The preceding discussion of binary choice models focused on estimation of the standardized coefficients, δ, which are identified if X has full column rank.
- Recall, however, that the binary choice model can be expressed in terms of the underlying parameters β and σ. Is it possible to estimate these $k+1$ parameters? The answer is no.
- Let $\theta_1' = [\beta_1' \ \sigma_1]$ and $\theta_2' = [\beta_2' \ \sigma_2]$. There are an infinity of $\theta_1 \neq \theta_2$ such that $\sigma_1^{-1}\beta_1 = \sigma_2^{-1}\beta_2$. 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 and σ, there is an identification problem.

Identification

- The restriction $\sigma = 1$ is an "identification condition" for the BC model. That is, imposition of the constraint $\sigma = 1$ does not alter the maximized value of the log-likelihood function. Consequently, 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.

Identification

Proposition 1: If $\hat{\delta}$ maximizes $\ln L(\delta)$, then $\hat{\beta}_0 = \sigma_0\hat{\delta}$ maximizes $\ln L(\beta, \sigma_0)$ for any fixed value σ_0.

Assume, contrary to the statement of the theorem, that $\hat{\beta}_0$ does not maximize $\ln L(\beta, \sigma_0)$. This implies that there exists $\tilde{\beta}$ such that $\ln L(\tilde{\beta}, \sigma_0) > \ln L(\hat{\beta}_0, \sigma_0)$. Let $\tilde{\delta} = \sigma_0^{-1}\tilde{\beta}$. Then since $\ln L(\beta, \sigma) = \ln L(\delta)$ for all $\delta = \sigma^{-1}\beta$, this implies that $\ln L(\tilde{\delta}) > \ln L(\hat{\delta})$, which contradicts the statement that $\hat{\delta}$ maximizes $\ln L(\delta)$. Hence, $\hat{\beta}_0$ maximizes $\ln L(\beta, \sigma_0)$ for any fixed value σ_0.

Identification

Proposition 2: The maximized value of the log-likelihood function, $\ln L(\hat{\beta}_0, \sigma_0)$, is invariant to the choice of σ_0.

The maximizing values, $\hat{\beta}_0$, must satisfy $\sigma_0^{-1}\hat{\beta}_0 = \hat{\delta}$ for all σ_0. Since the parameters β and σ enter $\ln L(\beta, \sigma)$ only through $\delta = \sigma^{-1}\beta$, this implies that $\ln L(\hat{\beta}_0, \sigma_0) = \ln L(\hat{\delta})$ for all σ_0. Since $\hat{\delta}$ is the unique global maximizer of $\ln L(\delta)$, this implies that $\ln L(\hat{\beta}_0, \sigma_0)$ is invariant to the value of σ_0.

Identification

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

1. Identification conditions are generally not unique. The restriction $\sigma = 2$ is also an identification condition for the BC model. So why is the condition $\sigma = 1$ typically adopted? Probably because the maximizing values of δ and β coincide when $\sigma = 1$.
2. While the maximizing value of δ is invariant to the choice of σ, the maximizing value of β is not. Variation in σ results in proportional variation in β.

Identification

3. The "truth" of the identification condition is irrelevant. Data generated by different values of σ 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 σ are "observationally equivalent." They cannot be distinguished on the basis of the observed data alone.

In the case of Binary Choice models, the intuition for this result is simple. An increase in σ implies greater variation in the latent dependent variable Y, but not in the observed dependent variable J. All positive values of Y, regardless of their magnitude, are translated into the single value $J = 1$.

Example

A binary variable, J_i, is determined by the following structure:

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

where $Y_i \sim iidN(\mu, \sigma^2)$. The underlying values of Y_i and the censoring threshold α are unobserved. The log-likelihood function for this model is

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

where $\delta = (\alpha - \mu)/\sigma$ is a standardized censoring threshold.

Example

The score equation is

$$\partial \ln L(\delta)/\partial\delta = \sum_{i=1}^{n}\left\{J_i\frac{\phi(\delta)}{\Phi(\delta)} - (1-J_i)\frac{\phi(\delta)}{1-\Phi(\delta)}\right\}$$

The ML estimator, $\hat{\delta}$, satisfies

$$\sum_{i=1}^{n}\left\{J_i\frac{\phi(\hat{\delta})}{\Phi(\hat{\delta})} - (1-J_i)\frac{\phi(\hat{\delta})}{1-\Phi(\hat{\delta})}\right\} = 0$$

Multiplying both sides of the above by $\Phi(\hat{\delta})[1-\Phi(\hat{\delta})]/\phi(\hat{\delta})$, the score equations reduce to ...

Example

$$\sum_{i=1}^{n}\left\{J_i - \Phi(\hat{\delta})\right\} = 0$$

Solving for the ML estimator gives $\hat{\delta} = \Phi^{-1}(\bar{J})$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. (Just read the table backward.)

Example

Any choices for α, μ, and σ that satisfy $\hat{\delta} = (\alpha-\mu)/\sigma$ will also satisfy the score equations. In this example, a pair of identification conditions are necessary.

- If we choose $\mu = 0$ and $\sigma = 1$, then $\hat{\alpha} = \hat{\delta}$. We estimate the censoring threshold.
- Alternatively, we could choose $\alpha = 0$ and $\sigma = 1$, in which case $\hat{\mu} = -\hat{\delta}$. We estimate the mean of the population.
- Finally, we could choose $\alpha = 1$ and $\mu = 0$, in which case $\hat{\sigma} = \hat{\delta}^{-1}$. We estimate the standard deviation of the population.

Example

Obviously, there are many other possibilities. These alternatives illustrate that the parameter estimate and its interpretation are contingent on the identification conditions adopted. Since only $\hat{\delta}$ is uniquely identified, perhaps the standardized censoring threshold is the least confusing parameter to estimate and interpret.

As a final illustration, it is worth noting that not all parameter restrictions are identification conditions. For example, if we impose $\alpha = 0$ and $\mu = 0$, we get $\hat{\delta} = 0$, which will generally not be equal to the ML estimator $\Phi^{-1}(\bar{J})$. These restrictions are binding on the maximization process, and consequently, are not identification conditions for this model.

Marginal Effects

- The vector of coefficients in a binary choice model, δ, are the marginal effects of the regressors on the latent variable $\sigma^{-1}Y$. That is, $\partial\sigma^{-1}Y_i/\partial X_i' = \delta$. Note that these marginal effects are invariant to the specific value of X_i.
- Since the latent model generally has no meaningful interpretation, it is common to report an alternative set of marginal effects, $\partial F(X_i\delta)/\partial X_i' = f(X_i\delta)\delta$. This gives the marginal effect of a regressor on the probability that $J_i = 1$. This provides meaningful quantitative information that is easy to interpret.
- In the context of binary choice, the term "marginal effect" refers to this later concept in virtually every application.

Marginal Effects

- The marginal effects must be estimated as $f(X_i\hat{\delta})\hat{\delta}$. This is observation specific; it depends on the individual X_i.
- Reporting each of these would involve k marginal effects for each of n observations. Way too many! For this reason, marginal effects are typically reported as $f(\bar{X}\hat{\delta})\hat{\delta}$, where $\bar{X}$ is the vector of sample means for the regressors. In some cases, sample medians are used in place of some of the sample means.

Marginal Effects

- Tests of hypotheses about δ use of the asymptotic distribution of ML estimators. Specifically, $\hat{\delta} \overset{a}{\sim} N(\delta, n^{-1}\Omega)$, where $\Omega = \text{plim}[n^{-1}I(\delta)]^{-1}$.
- The sample moment matrix $n^{-1}[V(\hat{\delta})'V(\hat{\delta})]$ is a consistent estimator of the population moment matrix $\text{plim}[n^{-1}I(\delta)]$.
- The resulting estimate of the asymptotic covariance matrix is $[V(\hat{\delta})'V(\hat{\delta})]^{-1}$. This estimate is constructed at every iteration of a BHHH algorithm.

Marginal Effects

- In order to test hypotheses about the marginal effects, we need the asymptotic distribution of $f(\bar{X}\hat{\delta})\hat{\delta}$. This is normally done using the delta method.
- Let $g(\delta)$ denote $f(\bar{X}\delta)\delta$ and $G(\delta)$ denote the $k \times k$ matrix of partials $\partial g(\delta)/\partial\delta'$, then by the delta method theorem

$$\sqrt{n}[g(\hat{\delta}) - g(\delta)] \xrightarrow{D} N[0, G(\delta)\Omega G(\delta)']$$

That is, $g(\hat{\delta}) \overset{a}{\sim} N[g(\delta), n^{-1}G(\delta)\Omega G(\delta)']$

Probit Marginal Effects

More specific results will require we specify the form of the density used when constructing the marginal effects, $f(\cdot)$. With the Probit model, $g(\delta) = \phi(\bar{X}\delta)\delta$ and

$$\begin{aligned} G(\delta) &= \frac{\partial \phi(\bar{X}\delta)\delta}{\partial \delta'} \\ &= \phi(\bar{X}\delta)I_k - \delta\{(\bar{X}\delta)\phi(\bar{X}\delta)\}\bar{X} \\ &= \phi(\bar{X}\delta)[I_k - (\bar{X}\delta)\delta\bar{X}] \end{aligned}$$

Note that $\bar{X}\delta$ is a scalar while $\delta\bar{X}$ is a $k \times k$ matrix. The covariance matrix is then estimated as $G(\hat{\delta})[V(\hat{\delta})'V(\hat{\delta})]^{-1}G(\hat{\delta})'$.

Logit Marginal Effects

With the Logit model, $g(\delta) = \psi(\bar{X}\delta)\delta$ and

$$\begin{aligned} G(\delta) &= \frac{\partial \psi(\bar{X}\delta)\delta}{\partial \delta'} \\ &= \psi(\bar{X}\delta)I_k + \delta\psi(\bar{X}\delta)[1 - 2\Psi(\bar{X}\delta)]\bar{X} \\ &= \psi(\bar{X}\delta)\{I_k + [1 - 2\Psi(\bar{X}\delta)]\delta\bar{X}\} \end{aligned}$$

The covariance matrix is again estimated as $G(\hat{\delta})[V(\hat{\delta})'V(\hat{\delta})]^{-1}G(\hat{\delta})'$.

Minimum Chi-Square Methods

Under certain conditions, minimum Chi-Square estimation provides a GLS alternative to full ML estimation of the binary choice model. Minimum Chi-Square estimation requires:

1. Repeated observations on J_i for each distinct value of X_i.
2. For each distinct value of X_i, the sample proportion of observations for which $J_i = 1$ must fall within the open interval $(0, 1)$.

Minimum Chi-Square Methods

Essentially, the information in (J_i, X_i) for $i = 1, ..., n$, may be reduced to $(\hat{p}_j, x_j, n_j)$ for $j = 1, ..., m$.

- The x_j denote the m distinct value of X_i.
- The set $\alpha(j)$ denotes $\{i|X_i = x_j\}$ for $j = 1...m$.
- $0 < \hat{p}_j < 1$ denotes the sample proportion of observations within the $i \in \alpha(j)$ subsample for which $J_i = 1$.
- n_j denotes the number of observations in the $i \in \alpha(j)$ subsample.

Minimum Chi-Square Methods

The binary choice model assumes that

$$P(J_i = 1) = F(X_i\delta) \qquad \text{for } i = 1, ..., n$$

where $\{J_i\}$ is a sequence of SI random variables, and $F(\cdot)$ is a distribution function. For the $i \in \alpha(j)$ subsample, $P(J_i = 1)$ is common and may be denoted p_j. Written in terms of the x_j, the model is

$$p_j = F(x_j\delta) \qquad \text{for } j = 1, ..., m$$

Applying the inverse distribution function gives

$$F^{-1}(p_j) = x_j\delta \qquad \text{for } j = 1, ..., m$$

Minimum Chi-Square Methods

For any given value of x_j, the process is a Bernoulli trial with probability p_j.

- The ML estimator $\hat{p}_j$ is an efficient estimator of p_j. It has mean p_j and variance $p_j(1 - p_j)/n_j$.
- It is also consistent, asymptotically efficient, and asymptotically normal.

Using $\hat{p}_j$ as an estimator of p_j, the model may be written as

$$F^{-1}(\hat{p}_j) = x_j\delta + \eta_j \qquad \text{for } j = 1, ..., m$$

where $\eta_j = [F^{-1}(\hat{p}_j) - F^{-1}(p_j)]$ is an unobservable error term.

Minimum Chi-Square Methods

- The dependent variable, $F^{-1}(\hat{p}_j)$, may be constructed given the sample proportions, $\hat{p}_j$, and the form of the distribution function,$F(\cdot)$.
- The $\hat{p}_j$ are resticted to $(0, 1)$ since the inverse distribution function generally fails at the interval endpoints.
- The model is linear in δ, but the error term η_j is generally heteroskedastic.
- OLS is unbiased but inefficient. Efficient estimates are obtained with GLS.

Minimum Chi-Square Estimation

The form of the inverse distribution function and the correction for heteroskedasticity are model specific.

- The Linear Probability model assumes that $F(\cdot)$ is the distribution function of a uniform on the unit interval. Thus, $p_j = F(x_j\delta) = x_j\delta$ for $0 < x_j\delta < 1$.
- Solving for $x_j\delta$ in terms of p_j gives $x_j\delta = F^{-1}(p_j) = p_j$ for $0 < p_j < 1$.
- This implies that $\eta_j = (\hat{p}_j - p_j)$, $E(\eta_j) = 0$, and $Var(\eta_j) = p_j(1 - p_j)/n_j$.
- The minimum chi-square estimator is WLS using the multiplicative weighting factor $\sqrt{\frac{n_j}{\hat{p}_j(1-\hat{p}_j)}}$.

Minimum Logit Chi-Square

- The distribution function for the Logit model is $p_j = \Psi(x_j\delta) = \exp(x_j\delta)/[1+\exp(x_j\delta)]$ for $x_j\delta \in R$.
- Solving for $x_j\delta$ in terms of p_j gives $x_j\delta = \Psi^{-1}(p_j) = \ln(p_j) - \ln(1-p_j)$ for $0 < p_j < 1$.
- A linear approximation about p_j gives

$$\eta_j \approx \frac{\partial \Psi^{-1}(p_j)}{\partial p_j}(\hat{p}_j - p_j) = \left[\frac{1}{p_j(1-p_j)}\right](\hat{p}_j - p_j)$$

Minimum Logit Chi-Square

- This implies $E(\eta_j) \approx 0$, and

$$Var(\eta_j) \approx \left[\frac{1}{p_j(1-p_j)}\right]^2 Var(\hat{p}_j) = \frac{1}{p_j(1-p_j)n_j}$$

- The minimum logit chi-square estimator is WLS using the multiplicative weighting factor $\sqrt{\hat{p}_j(1-\hat{p}_j)n_j}$.

Minimum Normit Chi-Square

- The distribution function for the Probit model is $p_j = \Phi(x_j\delta)$ for $x_j\delta \in R$.
- Solving for $x_j\delta$ in terms of p_j gives $x_j\delta = \Phi^{-1}(p_j)$ for $0 < p_j < 1$.
- By definition of a continuous density function

$$\frac{\partial \Phi(x_j\delta)}{\partial x_j\delta} = \phi(x_j\delta) = \phi[\Phi^{-1}(p_j)]$$

- The inverse function rule gives

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

Minimum Normit Chi-Square

- A linear approximation about p_j gives

$$\eta_j \approx \frac{\partial \Phi^{-1}(p_j)}{\partial p_j}(\hat{p}_j - p_j) = \left[\frac{1}{\phi[\Phi^{-1}(p_j)]}\right](\hat{p}_j - p_j)$$

- This implies $E(\eta_j) \approx 0$, and

$$Var(\eta_j) \approx \left[\frac{1}{\phi[\Phi^{-1}(p_j)]}\right]^2 Var(\hat{p}_j) = \left[\frac{1}{\phi[\Phi^{-1}(p_j)]}\right]^2 \frac{p_j(1-p_j)}{n_j}$$

- The minimum normit chi-square estimator is WLS using the multiplicative weighting factor $\phi[\Phi^{-1}(\hat{p}_j)]\sqrt{\frac{n_j}{\hat{p}_j(1-\hat{p}_j)}}$.