Discrete Dependent Variable Models

James J. Heckman
University of Chicago

This draft, April 10, 2006

Here's the general approach of this lecture:

$$\underbrace{\left[\begin{array}{c}\text{Economic model}\\ \text{(e.g. utility maximization)}\end{array}\right] \Rightarrow \left[\begin{array}{c}\text{Decision rule}\\ \text{(e.g. FOC)}\end{array}\right]}_{\text{Sec. 1 Motivation: Index function and random utility models}}$$

$$\Rightarrow \underbrace{\left[\begin{array}{c}\textbf{Underlying}\\ \textbf{regression}\\ \text{(e.g. solve the FOC for}\\ \text{a dependent variable)}\end{array}\right] \Rightarrow \left[\begin{array}{c}\textbf{Econometric model}\\ \text{(e.g. depending on}\\ \text{observed data, discrete}\\ \text{or limited dependent}\\ \text{variable model)}\end{array}\right]}_{\text{Sec. 2 Setup}}$$

$$\Rightarrow \underbrace{[\text{Estimation}]}_{\text{Sec. 4 Estimation}} \Rightarrow \underbrace{[\text{Interpretation}]}_{\text{Sec. 3 Marginal Effects}}$$

- We assume that we have an economic model and have derived implications of the model, e.g. FOCs, which we can test. Converting these conditions into an underlying regression usually involves little more than rearranging terms to isolate a dependent variable.

- Often this dependent variable is not directly observed, in a way that we'll make clear later. In such cases, we cannot simply estimate the underlying regression. Instead, we need to formulate an econometric model that allows us to estimate the parameters of interest in the decision rule/underlying regression using what little information we have on the dependent variable.

- We will present two models in part A which will help us bridge the gap between inestimable underlying regressions and an estimable econometric model.

- In part B, we will further develop the econometric model introduced in part A so that it is ready for estimation.

- In part C, we jump ahead to interpreting our results. In particular we will explain why, unlike in the linear regression models, the estimated $\widehat{\beta}$ does not give us the marginal effect of a change in the independent variables on the dependent variable. We jump ahead to this topic because it will give us some information we need when we estimate the model.

- Finally, part D will describe how to estimate the model.

1 Motivation

Discrete dependent variable models are often cast in the form of index function models or random utility models. Both models view the outcome of a discrete choice as a reflection of an underlying regression. The desire to inform econometric models with economic models suggests that the underlying regression be a marginal cost-benefit analysis calculation. The difference between the two models is that the structure of the cost-benefit calculation in index function models is simpler than that in random utility models.

1.1 Index function models

Since marginal benefit calculations are not observable, we model the difference between benefit and cost as an unobserved variable y^* such that:

$$y^* = \beta' x + \varepsilon,$$

where $\varepsilon \sim f(0, 1)$, with f symmetric. While we do not observe y^*, we do observe y, which is related to y^* in the sense that:

$$y = 0 \text{ if } y^* \leq 0$$

and

$$y = 1 \text{ if } y^* > 0.$$

In this formulation $\beta' x$ is called the index function. Note two things. First, our assumption that $var(\varepsilon) = 1$ could be changed

to $var(\varepsilon) = \sigma^2$ instead, by multiplying our coefficients by σ^2. Our observed data will be unchanged; $y = 0$ or 1, depending only on the sign of y^*, not its scale. Second, setting the threshold for y given y^*at 0 is likewise innocent if the model contains a constant term. (In general, unless there is some compelling reason, binomial probability models should not be estimated without constant terms.) Now the probability that $y = 1$ is observed is:

$$\begin{aligned} \Pr\{y = 1\} &= \Pr\{Y^* > 0\} \\ &= \Pr\{\beta' x + \varepsilon > 0\} \\ &= \Pr\{\varepsilon > -\beta' x\}. \end{aligned}$$

Then under the assumption that the distribution f of ε is symmetric, we can write:

$$\Pr\{y = 1\} = \Pr\{\varepsilon < \beta' x\} = F(\beta' x),$$

where F is the cdf of ε. This provides the underlying structural model for estimation by MLE or NLLS estimation.

1.2 Random utility models

Suppose the marginal cost benefit calculation was slightly more complex. Let y_0 and y_1 be the net benefit or utility derived from taking actions 0 and 1, respectively. We can model this utility calculus as the unobserved variables y_0 and y_1 such that:

$$\begin{aligned} y_0 &= \beta' x_0 + \varepsilon_0, \\ y_1 &= \gamma' x_1 + \varepsilon_1. \end{aligned}$$

Now assume that $(\varepsilon_1 - \varepsilon_0) \sim f(0, 1)$, where f is symmetric. Again, although we don't observe y_0 and y_1, we do observe y where:

$$\begin{aligned} y &= 0 \text{ if } y_0 > y_{1,} \\ y &= 1 \text{ if } y_0 \leq y_1. \end{aligned}$$

In other words, if the utility from action 0 is greater than action 1, i.e., $y_0 > y_1$, then $y = 0$. $y = 1$ when the converse is true. Here the probability of observing action 1 is:

$$\begin{aligned}
\Pr\{y = 1\} &= \Pr\{y_0 \leq y_1\} = \Pr\{\beta' x_0 + \varepsilon_0 \leq \gamma' x_1 + \varepsilon_1\} \\
&= \Pr\{\varepsilon_1 - \varepsilon_0 \geq \beta' x_0 - \gamma' x_1\} \\
&= F(\gamma' x_1 - \beta' x_0).
\end{aligned}$$

2 Setup

The index function and random utility models provide the link between an underlying regression and an econometric model. Now we'll begin the process of flushing out the econometric model. First we'll consider different specifications for the distribution of ε and later, in part C, examine how marginal effects are derived from our probability model. This will pave the way for our discussion of how to estimate the model.

2.1 Why $\Pr\{y = 1\}$?

In both index function and random utility models, the probability of observing $y = 1$ has the structure: $\Pr\{y = 1\} = F(\beta' x)$. Why are we so interested in the probability that $y = 1$? Because the expected value of y given x is just that probability: $E[y] = 0 \cdot (1 - F) + 1 \cdot F = F(\beta' x)$.

2.2 Common specifications for $F(\beta' x)$

How do we specify $F(\beta' x)$? There are four basic specifications that dominate the literature.

(a) Linear probability model (LPM):

$$F(\beta' x) = \beta' x$$

(b) Probit:

$$F(x) = \Phi(\beta' x) = \int_{-\infty}^{\beta' x} \phi(t)dt = \int_{-\infty}^{\beta' x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

(c) Logit:

$$F(\beta' x) = \Lambda(\beta' x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

(d) Extreme Value Type I:

$$F(\beta' x) = W(\beta' x) = 1 - e^{-e^{\beta' x}}$$

2.3 Deciding which specification to use

Each specification has its advantages and disadvantages.

(1) **LPM.** The linear probability model is popular because it is extremely simple to estimate. This simplicity, however, comes at a cost. To see what we mean, set up the NLLS regression model.

$$y = E[y|x] + (y - E[y|x]) = F(\beta' x) + \varepsilon$$
$$= \beta' x + \varepsilon.$$

Because F is linear, this just collapses down to the CR model. Notice that the error term:

$\varepsilon = 1 - \beta' x$ with probability $F = \beta' x$ and

$-\beta' x$ with probability $1 - F = 1 - \beta' x$

This implies that:

$$\begin{aligned} var[\varepsilon|x] &= E[\varepsilon^2|x] - E^2[\varepsilon|x] = E[\varepsilon^2] \\ &= F \cdot (1-\beta'x)^2 + (1-F) \cdot (-\beta'x)^2 \\ &= F - 2F\beta'x + F[\beta'x]^2 + [\beta'x]^2 - F[\beta'x]^2 \\ &= F - 2F\beta'x + [\beta'x]^2 \\ &= \beta'x - 2[\beta'x]^2 + [\beta'x]^2 = \beta'x(1-\beta'x). \end{aligned}$$

So our first problem is that ε is heteroscedastic in a way that depends on β. Of course, absent any other problems, we could manage this with an FGLS estimator. A second more serious problem, however, is that since $\beta'x$ is not confined to the $[0,1]$ interval, the LPM leaves open the possibility of predicted probabilities that lie outside the $[0,1]$ interval, which is nonsensical, and of negative

variances:

$$\beta' x > 1 \Rightarrow E[y] = F = \beta' x > 1,$$
$$var[\varepsilon] = \beta' x(1 - \beta' x) < 0,$$
$$\beta' x < 0 \Rightarrow E[y] < 0,$$
$$var[\varepsilon] < 0.$$

This is a problem that is harder to correct. We could define $F = 1$ if $F(\beta' x) = \beta' x > 1$ and $F = 0$ if $F(\beta' x) = \beta' x < 0$, but this procedure creates unrealistic kinks at the truncation points for $(y, x \mid \beta' x = 0 \text{ or } 1)$.

(2) **Probit vs. Logit.** The probit model, which uses the normal distribution, is sometimes (inappropriately) justified by appealing to a central limit theorem,while the

logit model can be justified by the fact that it is similar to a normal distribution but has a much simpler form. The difference between the logit and normal distribution is that the logit has slightly heavier tails. The standard normal has mean zero and variance 1 while the logit has mean zero and variance equal to $\pi^2/3$.

(3) **Extreme Value Type I.** The extreme value type I distribution is the least common of the four models. It is important to note that this is an asymmetric pdf.

3 Marginal effects

Unlike in linear models such as the CR or Neo-CR models, the marginal effect of a change in x on $E[y]$ is not simply β. To see why, differentiate $E[y]$ by x:

$$\frac{\partial E[y]}{\partial x} = \frac{\partial F(\beta' x)}{\partial (\beta' x)} \frac{\partial (\beta' x)}{\partial x} = f(\beta' x)\beta.$$

These marginal effects look different in each of the four basic probability models.

1. **LPM.** Note that $f(\beta' x) = 1$, so $f(\beta' x)\beta = \beta$, which is the same as in the CR-type models, as expected.

2. **Probit.** Now, $f(\beta'x) = \phi(\beta'x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\beta'x)^2}{2}}$, so $f(\beta'x)\beta = \phi\beta$.

3. **Logit.** In this case:

$$\begin{aligned} f(\beta'x) &= \frac{\partial \Lambda(\beta'x)}{\partial(\beta'x)} = \frac{e^{\beta'x}}{1+e^{\beta'x}} - \frac{e^{\beta'x}}{(1+e^{\beta'x})^2} e^{\beta'x} \\ &= \frac{e^{\beta'x}}{1+e^{\beta'x}} \left(1 - \frac{e^{\beta'x}}{1+e^{\beta'x}}\right) \\ &= \Lambda(\beta'x)\left(1 - \Lambda(\beta'x)\right) \end{aligned}$$

Giving us the marginal effect $f(\beta'x)\beta = \Lambda(1-\Lambda)\beta$.

3.1 Converting probit marginal effects to logit marginal effects

To convert a probit coefficient estimate to a logit coefficient estimate, from the discussion above comparing the variances of probit and logit random variable, it would make sense to multiply the probit coefficient estimate by $\frac{\pi}{\sqrt{3}} \cong 1.8$ (since variance of logit is $\pi^2/3$ whereas variance of the normal is 1) . But Amemiya suggests a different conversion factor. Through trial and error he found that 1.6 works better at the center of the distribution, which demarcates the mean value of the regressors. At the center of the distribution, $F = 0.5$ and $\beta' x = 0$. Well $\Phi(0) = 0.3989$ while $\Lambda(0) = 0.25$. So we want to solve the equation, $0.3989\beta_{\text{probit}} = 0.25\beta_{\text{logit}}$ this gives us $\beta_{\text{logit}} = 1.6\beta_{\text{profit}}$.

4 Estimation and hypothesis testing

There are two basic methods of estimation, MLE and NLLS estimation. Since the former is far more popular, we'll spend most of our time on it.

4.1 MLE

Given our assumption that the ε are i.i.d., by the definition of independence, we can write the joint probability of observing $\{y_i\}_{i=1,\ldots,n}$ as

$$\Pr\{y_1, y_2, \ldots, y_n\} = \Pi_{y_i=0}[1 - F(\beta' x_i)] \cdot \Pi_{y_i=1}[F(\beta' x_i)].$$

Using the notational simplification $F(\beta' x_i) = F_i, f(\beta' x_i) = f_i, f'(\beta' x_i) = f_i'$ we can write the likelihood function as:

$$L = \Pi_i (1 - F_i)^{1-y_i} (F_i)^{y_i}.$$

Since we are searching for a value of β that maximizes the probability of observing what we have, monotonically increasing transformations will not affect our maximization result.

Hence we can take logs of the likelihood function; and since maximizing a sum is easier than maximizing a product, we take the log of the likelihood function:

$$\ln L = \sum_i \{(1 - y_i) \ln[1 - F_i] + y_i \ln F_i\} .$$

Now estimate $\widehat{\beta}$ by:

$$\widehat{\beta} = \arg \max_{\beta} \ln L.$$

Within the MLE framework, we shall now examine the following six (estimation and testing) procedures:

A. Estimating $\widehat{\beta}$;

B. Estimating asymptotic variance of $\widehat{\beta}$;

C. Estimating asymptotic variance of the predicted probabilities;

D. Estimating asymptotic variance of the marginal effects;

E. Hypothesis testing; and

F. Measuring goodness of fit

A. Estimating $\widehat{\beta}$

To solve $\max_{\beta} \ln L$ we need to examine the first and second order conditions.

First Order Conditions (FOCs): A necessary condition for maximization is that the first derivative equal zero:

$$\frac{\partial \ln L}{\partial \beta} = \frac{\partial \ln L}{\partial (\beta' x)} \frac{\partial (\beta' x)}{\partial \beta} = \frac{\partial \ln L}{\partial (\beta' x)} x = 0.$$

If we write:

$$\frac{\partial F(\beta' x)}{\partial (\beta' x)} = f(\beta' x),$$

and we plug in:

$$\ln L = \sum_i \{(1 - y_i) \ln[1 - F_i] + y_i \ln F_i\},$$

then we just need to solve:

$$\sum_i \left[(1-y_i)\frac{-f_i}{1-F_i} + y_i\frac{f_i}{F_i}\right] x_i$$

$$= \sum_i \left[\frac{(y_i-1)f_iF_i + y_if_i(1-F_i)}{(1-F_i)F_i}\right] x_i = 0$$

$$\iff \sum_i \frac{(y_i-F_i)f_ix_i}{(1-F_i)F_i} = 0 \qquad \{\textbf{FOCs}\}$$

Now we look at the specific FOCs in three main models:

(1) **LPM.** Since $F_i = \beta' x_i$ and $f_i = 1 \forall i$, our FOC becomes:

$$\sum_i \frac{(y_i-F_i)f_ix_i}{(1-F_i)F_i} = \sum_i \frac{(y_i-\beta' x_i)x_i}{(1-\beta' x_i)\beta' x_i} = 0.$$

This is just a set of linear equations in x and y which we can solve explicitly for β in two ways.

(i) Least squares. The first solution gives us a result that is reminiscent of familiar least squares predictors.

(a) GLS. Solving for the β in the numerator, we get something resembling the generalized least squares estimator, where each x_i is weighted by the variance of ε_i.

$$\begin{aligned}\sum_i \frac{\beta' x_i^2}{(1-\beta' x_i)\beta' x_i} &= \sum_i \frac{y_i x_i}{(1-\beta' x_i)\beta' x_i} \\ \Rightarrow \beta &= \frac{\sum_i \frac{y_i x_i}{(1-\beta' x_i)\beta' x_i}}{\sum_i \frac{x_i^2}{(1-\beta' x_i)\beta' x_i}} = \frac{\sum_i \frac{y_i x_i}{var(\varepsilon_i)}}{\sum_i \frac{x_i^2}{var(\varepsilon_i)}}.\end{aligned}$$

(b) OLS. If we assume homoscedasticity, i.e:

$$(1 - \beta' x_i)\beta' x_i = var(\varepsilon_i) = var(\varepsilon) = \sigma^2 \ \forall \ i$$

Then the equation above collapses into the standard OLS estimator of β :

$$\beta = \frac{\frac{1}{var(\varepsilon)} \sum_i y_i x_i}{\frac{1}{var(\varepsilon)} \sum_i x_i^2} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}.$$

(ii) GMM. If we rewrite $y_i - \beta' x_i = \varepsilon_i$ then the FOC conditions resemble the generalized method of moments condition for solving the heteroscedastic linear LS model:

$$\sum_i \frac{\varepsilon_i x_i}{(1 - \beta' x_i)\beta' x_i} = 0 \Rightarrow \sum_i \frac{\varepsilon_i x_i}{var(\varepsilon_i)} = 0.$$

Again, if we assume homoskedasticity, we get the moment condition for solving the CR model:

$$\frac{1}{var(\varepsilon)} \sum_i \varepsilon_i x_i = \sum_i \varepsilon_i x_i = 0.$$

Note that each of these estimators is identical. Some may be more efficient than others in the presence of heteroscedasticity, but, in general, they are just different ways of motivating the LS estimator.

(2) **Probit.** Noting that $F_i = \Phi_i$, $f_i = \phi_i$, the FOC is just:

$$\begin{aligned}\sum_i \frac{(y_i - F_i) f_i x_i}{(1-F_i)F_i} &= \sum_i \frac{(y_i - \Phi_i)\phi_i x_i}{(1-\Phi_i)\Phi_i} \\ &= \sum_i \frac{y_i \phi_i x_i}{(1-\Phi_i)\Phi_i} - \sum_i \frac{\phi_i x_i}{(1-\Phi_i)}\end{aligned}$$

If we define (refer the results in the Roy Model handout):

$$\begin{aligned}\lambda_{0i} &= -E(z \mid z > \beta' x_i) = \frac{-\phi_i}{(1-\Phi_i)} \\ \lambda_{1i} &= E(z \mid z < \beta' x_i) = \frac{\phi_i}{\Phi_i}\end{aligned}$$

Then we can rewrite the FOC as:

$$\sum_i \lambda_i x_i = 0$$

where:

$\lambda_i = \lambda_{0i}$ if $y_i = 0$, and

λ_{1i} if $y_i = 1$.

Note that, unlike in the LPM, these FOC are a set of nonlinear equations in β. They cannot be easily solved explicitly for β. So β has to be estimated using the numerical methods outlined in the Asymptotic Theory Notes.

(3) **Logit.** Here $F_i = \Lambda_i$ and $f_i = \Lambda_i(1 - \Lambda_i)$, so the FOC becomes:

$$\begin{aligned} \sum_i \frac{(y_i - F_i) f_i x_i}{(1 - F_i) F_i} &= \sum_i \frac{(y_i - \Lambda_i)\Lambda_i(1 - \Lambda_i) x_i}{(1 - \Lambda_i)\Lambda_i} = 0 \\ \iff \sum_i (y_i - \Lambda_i) x_i &= 0. \end{aligned}$$

Interestingly, note that we can write $y_i - \Lambda_i = \varepsilon_i$ so that the FOC can be written $\sum_i (y_i - \Lambda_i) x_i = \sum_i \varepsilon_i x_i = 0$, which is similar to the moment conditions for the LPM. Like the probit model, however, the FOC for the logit model are nonlinear in β and must therefore be solved using numerical methods.

Second Order Condition (SOC): Together, the FOCs and the

SOC that the second derivative or Hessian be negative definite are necessary and sufficient conditions for maximization. To verify the second order condition, let:

$$\frac{\partial f(\beta' x)}{\partial(\beta' x)} = f'(\beta' x),$$

So that we need to check:

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial(\beta' x)}\left[\frac{\partial \ln L}{\partial(\beta' x)} x\right]\frac{\partial(\beta' x)}{\partial\beta} \\
&= \frac{\partial^2 \ln L}{\partial(\beta' x)\partial(\beta' x)'} xx' \\
&= \sum_i \frac{\partial}{\partial(\beta' x_i)}\left[\frac{(y_i - F_i) f_i x_i}{(1 - F_i) F_i}\right] x_i' < 0.
\end{aligned}$$

(1) LPM. We can prove that the LPM satisfies the SOC $\forall \beta \in B$:

$$\sum_i \frac{\partial}{\partial(\beta' x_i)} \left[\frac{(y_i - \beta' x_i) x_i}{(1 - \beta' x_i)\beta' x_i} \right] x_i'$$

$$= \sum_i \left[\frac{-x_i}{(1 - \beta' x_i)\beta' x_i} - \frac{(y_i - \beta' x_i) x_i}{(1 - \beta' x_i)^2 (\beta' x_i)^2} (1 - 2\beta' x_i) \right] x_i'$$

$$= \sum_i \left[\frac{-\beta\beta' x_i^3 - y_i x_i + 2 y_i \beta' x_i^2}{(1 - \beta' x_i)^2 (\beta' x_i)^2} \right] x_i'$$

$$= \sum_i \left[\frac{-(y_i - \beta' x_i)^2}{(1 - \beta' x_i)^2 (\beta' x_i)^2} \right] x_i x_i' < 0$$

(Using fact $y_i \in \{0, 1\} \Rightarrow y_i^2 = y_i$)

(2) Probit. The same can be said about the probit model,

and the proof follows from the results in the Roy model. First, note that $\phi'(\beta' x) = -\beta' x \phi(\beta' x)$. Taking the derivative of the first derivative we need to show:

$$\sum_i \frac{\partial}{\partial(\beta' x_i)}[\lambda_i x_i] x_i' = \sum_i \frac{\partial}{\partial(\beta' x_i)}[\lambda_i] x_i x_i' < 0.$$

We can simplify this expression using results for the truncated normal (see results on truncated normal in Roy

Model handout):

$$\begin{aligned}
\frac{\partial \lambda_{0i}}{\partial(\beta' x_i)} &= \frac{\partial}{\partial(\beta' x)}\left[\frac{-\phi_i}{1-\Phi_i}\right] \\
&= \frac{-\beta' x_i \phi_i}{1-\Phi_i} - \frac{\phi_i^2}{(1-\Phi_i)^2} = -\beta' x_i \lambda_{0i} - \lambda_{0i}^2 \\
&= -\lambda_{0i}(\beta' x_i + \lambda_{0i}) < 0 \\
\frac{\partial \lambda_{1i}}{\partial(\beta' x_i)} &= \frac{\partial}{\partial(\beta' x_i)}\left[\frac{\phi_i}{\Phi_i}\right] = \frac{-\beta' x_i \phi_i}{\Phi_i} - \frac{\phi_i^2}{\Phi_i^2} \\
&= -\beta' x_i \lambda_{1i} - \lambda_{1i}^2 = -\lambda_{1i}(\beta' x_i + \lambda_{1i}) < 0
\end{aligned}$$

So that we can write the SOC as:

$$-\sum_i \lambda_i(\beta' x_i + \lambda_i) x_i x_i' < 0,$$

Where:

$$\lambda_i = \lambda_{0i} = \frac{-\phi_i}{(1-\Phi_i)}, \quad \text{if } y_i = 0, \text{ and}$$

$$\lambda_{1i} = \frac{\phi_i}{\Phi_i}, \quad \text{if } y_i = 1$$

(3) **Logit.** Taking the derivative of the FOC for logit, we get the SOC :

$$\sum_i \frac{\partial[y_i - \Lambda_i)x_i]}{\partial(\beta' x_i)} x_i' = -\sum_i \Lambda_i(1-\Lambda_i)x_i x_i' < 0$$

which clearly holds $\forall\ \beta \in B$. Note that since the Hessian does not include y_i, the Newton-Raphson method of numerical optimization, which uses H in its iterative algo-

rithm, and the method of scoring, which uses $E[H]$, are identical in the case of the logit model. Why? Because $E[H]$ is taken with respect to the distribution of y.

We've shown that the LPM, probit and logit models are globally concave. So the Newton-Raphson method of optimization will converge in just a few iterations for these three models unless the data are very badly conditioned.

B. Estimating the Asy Cov matrix for $\widehat{\beta}$

Recall the following two results from the MLE notes:

(a) $\sqrt{T}(\widehat{\beta} - \beta_0) \to N(0, -I(\beta_0)^{-1})$

where $I(\beta_0) = \text{plim}\left(\frac{1}{T} \left.\frac{\partial^2 \ln L}{\partial\beta\partial\beta'}\right|_{\beta_0}\right)$

$$\text{(b)}\quad \lim_{T\to\infty} -\frac{1}{T}\frac{\partial \ln L}{\partial \beta}\frac{\partial \ln L'}{\partial \beta}\bigg|_{\widehat{\beta}} = -E\left(\frac{1}{T}\frac{\partial \ln L}{\partial \beta}\frac{\partial \ln L'}{\partial \beta}\right)$$

$$= E\left[\frac{\partial^2 \ln L}{\partial \beta \partial \beta'}\right] = \text{plim}\left(\frac{1}{T}\,\frac{\partial^2 \ln L}{\partial \beta \partial \beta'}\bigg|_{\beta_0}\right) = \lim_{T\to\infty}\frac{1}{T}\frac{\partial^2 \ln L}{\partial \beta \partial \beta'}\bigg|_{\widehat{\beta}}.$$

We have three possible estimators for Asy.Var$[\widehat{\beta}]$ based on these two facts.

(1) Asy.Var$[\widehat{\beta}] = -\widehat{H}^{-1}$ where

$$\widehat{H} = \sum_i \frac{\partial}{\partial(\beta' x_i)}\left[\frac{(y_i - F_i)f_i}{(1-F_i)F_i}\right] x_i x_i'\bigg|_{\beta}.$$

(2) Asy. Var$[\widehat{\beta}] = -E[H]^{-1}$ where $E[H] = E\left[\dfrac{\partial^2 \ln L}{\partial \beta \partial \beta'}\right]$.

– In any model where H does not depend on y_i, $E[H] = \widehat{H}$ since the expectation has taken over the distribution of y. So in models such as logit the first and second estimators are identical. In the probit model, $\widehat{H}$ depends on y_i so $\widehat{H} \neq E[H]$. Amemiya ("Qualitative Response Models: A Survey," Journal of Economic Literature, 19, 4, 1981, pp. 481-536) showed that:

$$E[H]|_{\text{probit}} = \sum_i \lambda_{0i}\lambda_{1i}x_i x_i' = \sum_i \frac{-\phi_i^2}{(1-\Phi_i)}x_i x_i'.$$

(3) Berndt, Hall, Hall and Hausman took the following estimator from T.W. Anderson (1959) which we call the

TWA estimator:

$$\text{Asy.Var}[\widehat{\beta}] = \widehat{H}^{-1},$$

where

$$\widehat{H} = \sum_i \left(\frac{(y_i - F_i) f_i}{(1 - F_i) F_i} \right)' x_i x_i' \left(\frac{(y_i - F_i) f_i}{(1 - F_i) F_i} \right)$$

Notice there is no negative sign before the $\widehat{H}^{-1}$, as the two negative signs cancel each other out. Note that the three estimators listed here are the basic three variants on the gradient method of iterative numerical optimization explained in the numerical optimization notes.

C. Estimating the Asy Cov matrix for predicted probabilities, $F(\widehat{\beta}'x)$.

For simplicity, let $F(\widehat{\beta}'x) = \widehat{F}$. Recall the delta method: if g is twice continuously differentiable and $\sqrt{T}(\theta_T - \theta_0) \xrightarrow{d} N(0, \sigma^2)$, then:

$$\sqrt{T}(g(\theta_T) - g(\theta_0)) \xrightarrow{d} N(0, [g'(\theta_0)]^2\sigma^2).$$

Applying this to $\widehat{F}$ we get

$$\sqrt{T}\left(F(\widehat{\beta}) - F(\beta_0)\right) \xrightarrow{d} N(0, [\widehat{F}'(\beta_0)]^2 Var[\widehat{\beta}]),$$

where β_0 is the true parameter value. So a natural estimator for the asymptotic covariance matrix for the predicted probabilities is:

Asy.Var$[\widehat{F}] = \left(\frac{\partial \widehat{F}}{\partial \widehat{\beta}}\right)' V \left(\frac{\partial \widehat{F}}{\partial \widehat{\beta}}\right)$ where V =Asy.Var$[\widehat{\beta}]$.

Since: $\dfrac{\partial \widehat{F}}{\partial \widehat{\beta}} = \dfrac{\partial \widehat{F}}{\partial(\widehat{\beta}'x)} \dfrac{\partial(\widehat{\beta}'x)}{\partial \widehat{\beta}} = (\widehat{f})x$, we can write the estimator as:
Asy.Var$[\widehat{F}] = (\widehat{f})^2 x' V x$.

D. Estimating the Asy Cov matrix for marginal effects, $f(\widehat{\beta}'x)\beta$.

To recap, the marginal effects are given by:

$$\frac{\partial E[y]}{\partial x} = \frac{\partial F}{\partial x} = \frac{\partial F}{\partial(\beta' x)} \frac{\partial(\beta' x)}{\partial x} = f\beta.$$

To simplify notation, let $f(\widehat{\beta}'x)\widehat{\beta} = \widehat{f}\widehat{\beta} = \widehat{\gamma}$. Again, using the delta method as motivation, a sensible estimator for the asymptotic variance of $\gamma(\widehat{\beta})$ would be:

$$\text{Asy. Var}[\widehat{\gamma}] = \left(\frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}}\right) V \left(\frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}}\right)',$$

where V is as above. We can be more explicit in defining our estimator by noting that:

$$\begin{aligned}
\frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}} &= \frac{\partial(\widehat{f}\widehat{\beta})}{\partial \widehat{\beta}} = \widehat{f}\frac{\partial \widehat{\beta}}{\partial \widehat{\beta}} + \frac{\partial \widehat{f}}{\partial(\beta' x)}\frac{\partial(\widehat{\beta}'x)}{\partial \widehat{\beta}}\widehat{\beta} \\
&= \widehat{f}I + \frac{\partial \widehat{f}}{\partial(\widehat{\beta}'x)}\widehat{\beta}'x,
\end{aligned}$$

This gives us:

$$\text{Asy.Var}[\widehat{\widehat{f}\widehat{\beta}}] = \left(\widehat{f}I + \frac{\partial \widehat{f}}{\partial(\widehat{\beta}'x)}\widehat{\beta}'x\right) V \left(\widehat{f}I + \frac{\partial \widehat{f}}{\partial(\widehat{\beta}'x)}\widehat{\beta}'x\right)'.$$

This equation still does not tell us much. It may be more interesting to look at what the estimator looks like under different specifications of F.

(1) **LPM.** Recall $F = \beta'x$, $f = 1$, and $f' = 0$, so:

$$\text{Asy.Var}[\widehat{\widehat{f}\widehat{\beta}}]_{LPM} = V = \text{Asy.Var}[\widehat{\beta}]$$

(2) **Probit.** Here $F = \Phi$, $f = \phi$ and $f' = -\beta'x\phi$, leaving us with:

$$\text{Asy.Var}[\widehat{\widehat{f}\widehat{\beta}}]_{probit} = \widehat{\phi}^2 \left(I - \left(\widehat{\beta}'x\right)\widehat{\beta}'x\right) V \left(I - \left(\widehat{\beta}'x\right)\widehat{\beta}'x\right)'$$

(1) **Logit.** Now $F = \Lambda$, $f = \Lambda(1-\Lambda)$, and $f' = \Lambda(1-\Lambda)[1-2\Lambda]$, so:

$$\begin{aligned}\text{Asy. Var}[\widehat{f}\widehat{\beta}]_{\text{logit}} &= \left[\widehat{\Lambda}(1-\widehat{\Lambda})\right]^2 \left(I + (1-2\widehat{\Lambda})\widehat{\beta}'x\right) \\ &\quad \times V\left(I + (1-2\widehat{\Lambda})\widehat{\beta}'x\right)'\end{aligned}$$

E. Hypothesis testing

Suppose we want to test the following set of restrictions, $H_0 : R\beta = q$. If we let p be the number of restrictions in R, i.e., rank (R), then MLE provides us with three test statistics (refer also the Asymptotic Theory notes).

(1) **Wald test**

$$W = \left(R\widehat{\beta} - q\right)' [R \text{ Est.Asy.Var}(\widehat{\beta})R'](R\widehat{\beta} - q) \sim \chi^2(p).$$

– Example. Suppose H_0: the last L coefficients or elements of β are 0. Define $R = [0, I_L]$ and $q = 0$; and let $\widehat{\beta}_L$ be the last L elements of $\widehat{\beta}$. Then we get $W = \widehat{\beta}_L' V_L^{-1} \widehat{\beta}_L$.

(2) **Likelihood ratio test**

$$LR = -2[\ln L_R(\widehat{\beta}) - \ln L(\widehat{\beta})] \sim \chi^2(p)$$

where $\ln L_R(\widehat{\beta})$ and $\ln L(\widehat{\beta})$ are the log likelihood function evaluated with and without the restrictions on $\widehat{\beta}$, respectively.

Example. To test H_0: all slope coefficients except that on the constant term are 0, let

$$\begin{aligned}
\ln L_R(\widehat{\beta}) &= \sum_i \{y_i \ln F_i + (1 - y_i) \ln(1 - F_i)\} \\
&= n \sum_i \{(y_i/n) \ln F_i + ([1 - y_i]/n) \ln(1 - F_i)\} \\
&= n\{P \ln P + (1 - P) \ln(1 - P)\}
\end{aligned}$$

where P is the proportion of observations with $y = 1$.

(3) **Score or Lagrange multiplier test**

Write out the Lagrangian for the MLE problem given the restriction $\beta = \beta_R$: $L = \ln L - \lambda(\beta - \beta_R)$. The first order condition is $\dfrac{\partial \ln L}{\partial \beta} = \lambda$. So the test statistic is $LM = \lambda_R' V \lambda_R$, where λ_R is just λ evaluated at β_R.

- Example. In the logit model, suppose we want to test H_0: all slopes are 0. Then $LM = nR^2$, where R^2 is the uncentered coefficient of determination in the regression of $(y_i - P)$ on x_i, where P is the proportion of $y = 1$ observations in the sample. (Don't worry about how this is derived.)

F. Measuring goodness of fit

There are three basic ways to describe how well a limited dependent variable model fits the data.

(1) Log likelihood function, $\ln L$. The most basic way to describe how successful the model is at fitting the data is to report the value of $\ln L$ at $\widehat{\beta}$. Since the hypothesis that all other slopes in the model are zero is also interesting, $\ln L$ computed with only a constant term ($\ln L_0$), which should also be reported. Comparing $\ln L_0$ to $\ln L$ gives us an idea of how much the likelihood improves on adding the explanatory variables.

(2) Likelihood ratio index, LRI. An analog to the R^2 in the CR model is the likelihood ratio index, $LRI = 1 - (\ln L / \ln L_0)$. This measure has an intuitive appeal in that it is bounded by 0 and 1 since $\ln L$ is a small negative number while $\ln L_0$ is a large negative number, making $\ln L / \ln L_0 < 1$. If $LRI = 1$, $F_i = 1$ whenever $y_i = 1$ and $F_i = 0$ whenever $y_i = 0$, giving us a perfect fit. $LRI = 0$ when the fit is miserable, i.e. $\ln L = \ln L_0$. Unfortunately, values between 0 and 1 have no natural interpretation like they do in the R^2 measure.

(3) Hit and miss table. A useful summary of the predictive ability of the model is a 2×2 table of the hits and misses of a prediction rule: $\widehat{y}_i = 1$ if $F(\widehat{\beta})'x) > F^*$, and 0 otherwise.

	$y_i = 0$	$y_i = 1$
Hits	# of obs. where $\widehat{y}_i = 0$	# of obs. where $\widehat{y}_i = 1$
Misses	# of obs. where $\widehat{y}_i = 1$	# of obs. where $\widehat{y}_i = 0$

The usual value for $F^* = 0.5$. Note, however, that 0.5 may seem reasonable but is arbitrary.