ECON 5360 Class Notes
Qualitative Dependent Variable Models

Here we consider models where the dependent variable is discrete in nature.

1 Linear Probability Model

Consider the linear probability (LP) model

$$y_i = \beta' x_i + \mu_i$$

where $E(\mu_i) = 0$. The conditional expectation

$$E(y_i|x_i) = \beta' x_i$$

is interpreted as the probability of an event occurring given x_i. There are a couple of drawbacks to the LP model that limits its use:

1. Heteroscedasticity. Given that $y_i = \{0, 1\}$, the error term can take on two values with probability

μ_i	$f(\mu_i)$
$1 - \beta' x_i$	$\beta' x_i$
$-\beta' x_i$	$1 - \beta' x_i$

   so that the variance is

$$\begin{aligned} var(\mu_i) &= \beta' x_i (1 - \beta' x_i)^2 + (1 - \beta' x_i)(-\beta' x_i)^2 \\ &= \beta' x_i (1 - \beta' x_i) \\ &= E(y_i)[1 - E(y_i)]. \end{aligned}$$

2. Predictions outside [0,1]. The predicted probabilities from the LP model, $\hat{y}_i = \beta' x_i$, can be less than zero and greater than one.

2 Binomial Probit and Logit Models

The drawbacks of the LP model are solved by letting the probability of an event (i.e., $y = 1$) be given by a well-defined cumulative density function

$$Prob(y_i = 1|x) = \int_{-\infty}^{x'\beta} f(t)dt = F(x'\beta). \tag{1}$$

In this manner, the predicted probabilities will always be bounded between zero and one. If $F(x'\beta)$ is the cdf for a standard normal random variable, we get the probit model. If

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

then we get the logit model. Estimates from the logit and probit models often give similar results. The logit model is less computationally intense because $F(x'\beta)$ has a closed form, however, the logistic pdf $f(\cdot)$ has fatter tails than the standard normal pdf. Because $y_i = \{0, 1\}$ is discrete, while (1) implies continuity, we replace y_i with the latent variable y_i^*. This produces

$$y_i^* = \beta' x_i + \mu_i.$$

y_i^* can be interpreted as an unobservable index function that measures individual i's propensity to choose $y = 1$. For example, y_i^* could be the net benefits (benefits less costs) of selecting option A. Alternatively, y_i^* could be interpreted as the difference in utility derived from choosing option A less the utility of choosing option B. Therefore, we assume

$$\begin{aligned} \text{if } y_i^* &> 0 \text{ then } y_i = 1 \\ \text{if } y_i^* &\leq 0 \text{ then } y_i = 0. \end{aligned}$$

The choice of zero as a threshold is innocuous if the vector x_i includes a constant term.

2.1 Estimation

The parameters of the model are estimated via maximum likelihood. The relevant probability can be written as

$$Prob(y_i = 1|x) = Prob(y_i^* > 0|x) = Prob(\beta' x_i + \mu_i > 0|x) = Prob(\mu_i > -\beta' x_i|x).$$

Assuming a symmetric, mean-zero pdf for μ_i, we have

$$Prob(\mu_i > -\beta' x_i|x) = Prob(\mu_i < \beta' x_i|x).$$

It will be convenient to standardize μ_i, which gives

$$Prob(\frac{\mu_i}{\sigma} < (\frac{\beta}{\sigma})'x_i|x) = \Phi((\frac{\beta}{\sigma})'x_i),$$

where $\Phi(\cdot)$ and σ are the cdf and standard deviation for μ_i, respectively. Therefore, the parameters are only identifiable up to a scalar σ, which is commonly set to unity (i.e., $\sigma = 1$). The likelihood function is given by

$$L = \prod_{i=1}^{n} \left[\Phi_i^{y_i}\{1-\Phi_i\}^{1-y_i}\right]$$

and the log-likelihood function is given by

$$\ln L(\beta) = \sum_{i=1}^{n}\{y_i \ln(\Phi_i) + (1-y_i)\ln(1-\Phi_i)\}. \tag{2}$$

Maximization of (2) will require nonlinear optimization methods, such as Newton's algorithm.

2.2 Marginal Effects

The estimated coefficients, $\hat{\beta}_{ML}$, are problematic in two senses:

1. The true βs are not identified. Recall, that all we can really estimate is β/σ.

2. Aside from problem #1, we know that

$$\hat{\beta}_k = \frac{\partial y_i^*}{\partial x_{i,k}}.$$

Because y_i^* is an unobservable index function, it is difficult to interpret this derivative.

A simple solution is to calculate

$$\hat{\delta}_{i,k} = \frac{\partial Prob(y_i = 1)}{\partial x_{i,k}} = \phi((\frac{\beta}{\sigma})'x_i)\frac{\beta_k}{\sigma} \tag{3}$$

where $\phi(\cdot)$ is the pdf for μ_i. The advantage of the estimated marginal effect, $\hat{\delta}_{i,k}$, is that it only depends on β/σ (so that it is identifiable) and it is easy to interpret. Note that $\hat{\delta}_{i,k}$ depends on the entire vectors for x_i and β. The standard errors for $\hat{\delta}_{i,k}$ can be calculated using the delta method, which is based on a first-order Taylor approximation. We have

$$asy.var.(\hat{\delta}) = \left(\frac{\partial\hat{\delta}}{\partial\hat{\beta}'}\right) V \left(\frac{\partial\hat{\delta}}{\partial\hat{\beta}'}\right)'$$

where V is the variance-covariance matrix for $\hat{\beta}_{ML}$.

2.3 Goodness of Fit

Unfortunately, the standard R^2 measure of goodness of fit does not have the same interpretation (i.e., percentage of variation in Y explained by the variation in X) in binary choice models. Many alternatives have been suggested, of which a few are:

- McFadden's pseudo R^2. This measure,

$$\tilde{R}^2 = 1 - \frac{\ln L_U}{\ln L_R},$$

  is bounded between zero and one but is difficult to interpret between the limits. It is not uncommon to see low $\tilde{R}^2$ values (e.g., less than 0.25) for models that seemingly explain the data well.

- Likelihood ratio statistic. The standard likelihood ratio statistic is

$$LR = -2(\ln L_R - \ln L_U)$$

  and is asymptotically distributed chi-square.

- Table of hits and misses. In the binary case, a 2 x 2 table can be created to summarize the number of correct and incorrect predictions. Typically, predicted probabilities greater than 0.5 (i.e., $\Phi(\hat{\beta}' x_i) > 0.5$) are associated with $\hat{y}_i = 1$. The main diagonal gives the number of correct predictions and the off-diagonal gives the number of incorrect predictions.

2.4 Fixed and Random Effects Models for Panel Data

Sometimes we may have a panel of cross sectional - time series data intended to explain a single binary choice. Consider the following extension of the binary models above

$$\begin{aligned} y_{it}^* &= x_{it}'\beta + \mu_i + \nu_{it} \\ y_{it} &= 1 \text{ if } y_{it}^* > 0. \end{aligned}$$

As before, whether we treat μ_i as a fixed or random effect depends upon the correlation (or lack thereof) between x_{it} and μ_i. Both fixed and random effects versions of this binary choice model are available. However, there are additional complications above and beyond the standard quantitative dependent-variable cases. In particular, in the RE case, the likelihood function will involve integration over the μ_i and, in the FE case, it is not possible to remove the fixed-effects term μ_i by subtracting group means. See section 21.5.1 in Greene for more details.

2.5 Bivariate Probit Model

The bivariate binary choice model takes the (SUR-like) form

$$\begin{aligned} y_1^* &= x_1'\beta_1 + \epsilon_1 \\ y_2^* &= x_2'\beta_2 + \epsilon_2 \end{aligned}$$

where $E(\epsilon_1) = E(\epsilon_2) = 0$ and

$$var(\epsilon) = var[(\epsilon_1 \ \epsilon_2)'] = E(\epsilon\epsilon') = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

As before, y_1^* and y_2^* are unobserved index functions such that

$$\begin{aligned} y_1 &= 1 \text{ if } y_1^* > 0 \text{ and } y_1 = 0 \text{ otherwise;} \\ y_2 &= 1 \text{ if } y_2^* > 0 \text{ and } y_2 = 0 \text{ otherwise.} \end{aligned}$$

There are four possible outcomes in the binary case. For example, the probability that $y_{i1} = y_{i2} = 0$ is

$$P_{00} = Prob(y_{1i} = 0, y_{2i} = 0|x_1, x_2) = \int_{-\infty}^{-x_1'\beta_1} \int_{-\infty}^{-x_2'\beta_2} \phi(\epsilon_1, \epsilon_2; \rho) d\epsilon_1 d\epsilon_2$$

where $\phi(\epsilon_1, \epsilon_2; \rho)$ represents the bivariate pdf. If $\phi(\epsilon_1, \epsilon_2; \rho)$ is specified as the bivariate normal pdf, then we have the bivariate probit model. The log likelihood function is

$$\ln L(\beta_1, \beta_2, \rho) = \sum_{y_1=0, y_2=0} \ln P_{i,00} + \sum_{y_1=1, y_2=0} \ln P_{i,10} + \sum_{y_1=0, y_2=1} \ln P_{i,01} + \sum_{y_1=1, y_2=1} \ln P_{i,11}$$

which is maximized through nonlinear optimization methods by choosing $\theta = (\beta_1, \beta_2, \rho)$. A potentially useful test is the Lagrange multiplier test to see whether $\rho = 0$ so that the probit models can be estimated separately (see Greene section 21.6.2). Marginal effects can be calculated (although they are a bit more complicated than the univariate probit case) and the delta method can be used to calculate standard errors.

3 Multinomial (Unordered) Logit

Consider explaining J different unordered choices (e.g., religion choice – Protestant, Catholic, Islam, Hindu, etc.). The multinomial logit model is derived by letting

$$Prob(y_i = j) = \frac{\exp(\beta_j' x_i)}{\sum_{k=1}^{J} \exp(\beta_k' x_i)}$$

for $j = 1, 2, .., J$.[1] Typically the normalization $\beta_1' = 0$ is made, which gives

$$P_{ij} = Prob(y_i = j) = \frac{\exp(\beta_j' x_i)}{1 + \sum_{k=2}^{J} \exp(\beta_k' x_i)}$$

and clearly satisfies the condition that $P_{i1} + P_{i2} + \cdots + P_{iJ} = 1$ for all $i = 1, ..., n$. This model gets the name multinomial logit because we assume that the binary probability

$$\frac{P_{ij}}{P_{i1} + P_{ij}} = \frac{\exp(\beta_j' x_i)}{1 + \exp(\beta_j' x_i)}$$

is given by the logistic cdf for $j = 2, ..., J$. The likelihood function is

$$L(\beta_2, \beta_3, ..., \beta_J) = \left[\prod_{y_i=1} P_{i1}\right] \left[\prod_{y_i=2} P_{i2}\right] \cdots \left[\prod_{y_i=J} P_{iJ}\right]. \tag{4}$$

Maximization of (4) will produce $J - 1$ coefficient vectors. The marginal effects are

$$\delta_{ij} = \frac{\partial P_{ij}}{\partial x_i} = P_{ij}[\beta_j - \bar{\beta}]$$

where $\bar{\beta}$ is the average coefficient vector. Therefore, the marginal effects for the j^{th} choice depends upon i and the parameters for all the choices $(\beta_2, \beta_3, ..., \beta_J)$. Standard errors can be calculated through the delta method (Greene, p. 722).

As a final note, the log-odds ratio is

$$\ln(P_{ij}) - \ln(P_{i1}) = x_i' \beta_j$$

and only depends upon β_j. This is called the independence of irrelevant alternatives (IIA) and it is a feature of the multinomial logit model. To test for IIA, Hausman and McFadden provide the following test statistic

$$HM = (\hat{\beta}_R - \hat{\beta}_U)'[\hat{V}_R - \hat{V}_U]^{-1}(\hat{\beta}_R - \hat{\beta}_U) \overset{asy}{\sim} \chi^2(K).$$

where the R subscript denotes the model with the other choices omitted and U denotes the full model. Should the HM statistic indicate a rejection of the null hypothesis of IIA, then the disturbances may not be independent and homoscedastic. In this case, one alternative to the multinomial logit model is a multivariate model (such as the bivariate case described above), which allows for correlations across alternatives. Another alternative is the nested logit model, where we break the choices into subgroups, where the IIA may hold within a group but not across subgroups. An example is the choice of community and type of housing, which

[1] When x_{ij} consists of choice-specific, as opposed to individual-specific charateristics, the appropriate model is the conditional logit model. The conditional logit differs from the multinomial logit model in that the coefficients do not vary across choices.

can be nested by first considering the choice of community and then the choice of housing type, conditional on the chosen community. See section 21.7.4 in Greene for more details.

4 Ordered Probit

Consider the following index function model

$$y_i^* = \beta' x_i + \mu_i$$

used to explain the ordered choices $y_i = \{1, 2, ..., m\}$. One example is choice of educational level, such as high school degree ($y_i = 1$), undergraduate degree ($y_i = 2$) or graduate degree ($y_i = 3$). We assume that

$$\begin{aligned}
\text{if } y_i^* &< \alpha_1 \text{ then } y_i = 1 \\
\text{if } \alpha_1 &< y_i^* < \alpha_2 \text{ then } y_i = 2 \\
\text{if } \alpha_2 &< y_i^* < \alpha_3 \text{ then } y_i = 3 \\
&\vdots \\
\text{if } y_i^* &> \alpha_{m-1} \text{ then } y_i = m
\end{aligned}$$

where the α_j are the threshold values for $j = 1, 2, ..., m-1$. Probabilities are given by

$$\begin{aligned}
P_1 &= \Phi(\alpha_1 - \beta' x_i) \\
P_2 &= \Phi(\alpha_2 - \beta' x_i) - \Phi(\alpha_1 - \beta' x_i) \\
P_3 &= \Phi(\alpha_3 - \beta' x_i) - \Phi(\alpha_2 - \beta' x_i) \\
&\vdots \\
P_m &= 1 - \sum\nolimits_{j=1}^{m-1} P_j.
\end{aligned}$$

The likelihood function is given by

$$L(\beta) = \left[\prod_{y_i=1} P_{1,i}\right]\left[\prod_{y_i=2} P_{2,i}\right] \cdots \left[\prod_{y_i=m} P_{m,i}\right]. \tag{5}$$

Maximum likelihood estimates are found by taking natural logs of (5) and then maximizing by choosing β. Note that although there are m different choices, there is only a single coefficient vector β. The ordered probit model results if Φ_i is specified as the standard normal cdf. Again, this will result in a nonlinear optimization problem.