1 Dummy Dependent Variable

Have considered how to deal with discrte variables in terms of dummy varaibles -as explanatory variables. In some cases we may have a dummy dependent variable> For example if we want to look at transport mode choice, what determines whether individuals use a car. We have:

$$\begin{aligned} y_i &= 1 \text{ if choose car} \\ y_i &= 1 \text{ otherwise} \end{aligned}$$

Now if we simply estimate an OLS regression

$$y_i = \beta x_i + u_i$$

Then this is called the Linear Probability Model

$$\begin{aligned} E(u_i) &= 0 \\ E(y_i \mid x_i) &= \beta x_i \text{ which can be interpreted in probabilty terms} \end{aligned}$$

Clearly u_i can only take two values

$$\begin{aligned} \text{when } y_i &= 1 \text{ then } u_i = 1 - \beta x_i \\ \text{when } y_i &= 0 \text{ then } u_i = -\beta x_i \end{aligned}$$

which means the variance

$$var(u_i) = E(u_i u_i)^2 = -\beta x_i (1 - \beta x_i)^2 + (1 - \beta x_i)(\beta x_i)^2 = E(y_i)\left[1 - E(y_i)\right]$$

is not constant and will vary with y. So u is heteroscedastic. We could overcome this problem with WLS but there is a more important problem and readily available alternatives. The problem is that while $E(y_i \mid x_i)$ may be interpreted as a probability it can lie ouside 0 and 1.

One alternative is to use linear discriminant analysis rather than OLS. This minimises the the ratio

$$\frac{\text{Between group variance}}{\text{Within group variance}}$$

of

$$y_i = \alpha + \beta x_i$$

But as Maddala shows this is very similar to an alternative and better approach.

Take

$$y_i = \alpha + \beta x_i + u_i$$

Then

$$\begin{aligned} P_i &= \Pr ob(y_i = 1) = \Pr ob(u_i > -(\alpha + \beta x_i) \\ &= 1 - F\left[-\alpha - \beta x_i\right] \end{aligned}$$

where F is the cumulative distribution. Now

$$P_i = F\left[\alpha + \beta x_i\right]$$

as

$$1 - F(-z) = F(z)$$

which we can estimate using maximum likelihood (ML) methods

$$L = \underset{y_i=1}{\Pi} P_i \Pi (1 - P_i)$$

The method we use depends upn the assumption we make about the error term. The most common are

Logit: assume logistic distribution for u_i which means

$$P_i = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

or

$$\log\left[\frac{P_i}{1 - P_i}\right] = \alpha + \beta x_i$$

Note the interpretation of the coefficients differs from the LPM

Probit: assume a normal distribution for the u_i which means

$$P_i = \Phi(\alpha + \beta x_i) = \int_{-\infty}^{\frac{\alpha+\beta x_i}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp \frac{(-t^2)}{2} dt$$

These two are now very commonly available in econometrics and statistics packages. For more complex models it is customary to start with the linear probability model to get starting values.

The cumulative normal and logistic distributions are similar, so we would expect similar results. They are not, however, directly comparable and we need to make a constant adjustment. Amemiya suggests

$$1.6\widehat{\beta}_{\Phi} \approx \widehat{\beta}_{Logit}$$

Also

$$\begin{aligned} \widehat{\beta}_{LPM} &\approx 0.4\widehat{\beta}_{\Phi} \text{ except constant} \\ \widehat{\beta}_{LPM} &\approx 0.4\widehat{\beta}_{\Phi} + 0.5 \text{ for constant} \\ \widehat{\beta}_{LPM} &\approx 0.25\widehat{\beta}_{Logit} \text{ except constant} \\ \widehat{\beta}_{LPM} &\approx 0.25\widehat{\beta}_{Logit} + 0.5 \text{ for constant} \end{aligned}$$

This will work for probablilities between 30%and 70%, as over this range the logistic can easily be approximated by a straight line

In practice the LPM model will give acceptable results, but there is the issue of heteroscedasticity and nowadays it is easily to estimate logits and probits.

Note that these models differ form the usual ones in practice in that we cant interpret the coefficients directly -eg as elasticities. They are disaggreagte models and estimate a probability for eachobservation, so when trying to forecast we have to aggreate. For the linear regression model

$$y_i = \alpha + \beta x_i \text{ and } \overline{y} = \alpha + \beta \overline{x}$$

but for the logit model:

$$P_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \text{ but } \overline{P} \neq \frac{e^{\alpha+\beta \overline{x}}}{1+e^{\alpha+\beta \overline{x}}}$$

When interpreting the logit/probit results will often see them reported in a table which gives the average or the extreme values of the variables and then use the coefficients to give the probability. For example in mode choice you might indicate what an individual who has a really high proabability will look like in terms of the eplanatory variables and compare with one who has a very low probability.

Goodness of fit: Cant use conventional R^2 type of measure with limited dependent variable methods. Common to look at measure based on the likelihood ratio

$$\begin{aligned} \lambda &= \frac{L(\beta_0)}{L(\beta_0, ..., \beta_K)} \\ -2\log\lambda &\sim \chi^2_k \end{aligned}$$

Can also use this to test restrictions on subsets of coefficients.

Analogous to an R^2 is

$$\rho^2 = 1 - \frac{L^*(\beta_{0,...,}\ \beta_K)}{L*(\beta_0)}$$

which can be adjusted for degrees of freedom as well. Note that while this will lie between 0 and 1, in contrats to the R^2a perfect fit value is about 0.7 and a range of 0.2 to 0.4 can be considered a good fit. Might also consider the proportion of correct predictions

$$\frac{\text{no. correct predictions } (y_i = 1 \text{ and } P_i > 0.5)}{\text{no. observations}}$$

worth reporting, but has low discriminatory power. Maddala discusses some other measures

Another variant on these models is the TObitmodel which deals with the situation when the observed value is either 0 or some positive number. For example if we are looking at what determines smoking we have 0 if the person does not smoke an the number of cigs when they do. So

$$y_i^* = \beta x_i + u_i$$

but observe y_i^* only if it is greater than 0

$$\begin{aligned} y_i &= y_i^* = \beta x_i + u_i \text{ if } y_i^* > 0 \text{ and } u_i \sim IN(0, \sigma^2) \\ y_i &= 0 \qquad\qquad\qquad\qquad \text{if } y_i^* \leq 0 \end{aligned}$$

Can estimate using MLE

$$L = \prod_{y_i=0} \frac{1}{\sigma} f\left(\frac{y_i - \beta x_i}{\sigma}\right) \prod_{y_i=0} F\left(\frac{-\beta x_i}{\sigma}\right)$$