Qualitative dependent variables

We have been modelling **quantitative** variables but we are often interested in **qualitative** responses. There are analogues of the basic regression model for the case of such discrete variables—and analogues too of the simultaneous equations model.

Some possible situations

- Yes/no, 2 point scale, dummy variable, binary variable, dichotomous variable, ... as in Hannum who investigates the determinants of school enrolment (yes/no) for a cross-section of Chinese children.


- Rating a module—like this one—on a 5 point scale.

- As well as one-dimensional outcomes, multivariate outcomes can be considered.
Hannum also models whether all school-aged boys and whether all school-aged girls in a household are enrolled.

The prototype binary situation is coin-tossing where the probability of heads is constant and the tosses are independent. Say the probability of heads is $\pi$ (Greek $p$ for probability) and let

$$Y = \begin{cases} 1 & \text{if } H \text{ (with probability } \pi) \\ 0 & \text{if } T \text{ (with probability } 1 - \pi). \end{cases}$$

This distribution is called the **Bernoulli distribution**.

The expectation and variance of $Y$ are calculated as follows

$$E(Y) = 0 \cdot (1 - \pi) + 1 \cdot \pi = \pi$$

$$\text{var}Y = E(Y - EY)^2 = (0 - \pi)^2(1 - \pi) + (1 - \pi)^2\pi$$

$$= \pi(1 - \pi).$$

We extend this model by letting this probability vary across observations. Thus Hannum specifies the probability of an individual’s school enrolment, $\pi_i$, as depending on several variables. I will write
the formulae assuming one determinant, $x_i$, say family income or gender of the $i$-th individual.

A number of functional forms are available for $\pi_i$.

- The simplest is a linear relationship, 
  
  $$\pi_i = \alpha + \beta x_i$$

  with $\beta > 0$ because richer families are more likely to send their children to school. This is called the **Linear Probability Model**

- It is usually estimated by least squares

  in the LPM

  $$\pi_i = E_y_i = \alpha + \beta x_i$$

  can be written as

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

  Now regress $y$ on $x$ to get estimates of $\alpha$ and $\beta$!

I call $\alpha + \beta x_i$ the value of the **index** for the $i$-th individual. Usually (as in Hannum) there are several covariates/regressors/exogenous variables and the index is
\[ \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}. \]

**Problems with the Linear Probability Model**

The LPM is appealing because we can use familiar methods BUT the following assumptions of the linear regression model do NOT hold

- The average value of \( y \), for each value of \( x \), is given by a linear relationship. **However** the relationship satisfied by \( \pi \) **cannot** be linear.

- For each value of \( x \), the values of \( y \) are distributed about their mean value, following probability distributions which have the same variance, i.e. the errors are homoscedastic. **However** it’s a property of the Bernoulli random variable that \( var(y) = \pi(1 - \pi) \) so that if \( \pi \) varies across individuals then so does \( var(y) \).
We will consider only the first defect, the inappropriateness of a **linear** relationship.

- By choosing extreme values of $x$ (and the index) we can obtain values of $\pi$ that are less than 0 or greater than 1. Alas $\pi$ is a probability!

- To keep $\pi$ within the interval $[0, 1]$ we need a **nonlinear** relationship between the index and $\pi$.
  
  An S-shaped relationship often seems plausible. As the index increases from a low level, the probability initially does not change much but then rises rapidly and finally increases at a decreasing rate. The slope of this curve gives the change in probability given a unit change in the value of the index. The slope is not constant as in the linear probability model.
Distribution functions (of probability) are often S-shaped. The function shown is the distribution function of the standard normal.

\[ P(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt. \]

The choice of this function for the relationship between \(\pi\) and the index defines the probit model. For the \(i\)-th child in the sample the probability of being enrolled is

\[ \pi_i = \Phi(\alpha + \beta x_i) = \int_{-\infty}^{\alpha+\beta x_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt. \]

Also widely used is the logit specification where \(\pi_i\) the probability of success for the \(i\)-th individual depends on the index thus
\[ \pi_i = \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \]

\[ \Rightarrow 1 - \pi_i = \frac{e^{-(\alpha + \beta x_i)}}{1 + e^{-(\alpha + \beta x_i)}}. \]

This specification is also called **logistic regression**.

- The \( \pi \) function looks like this

This is so because the **odds** in favour of a success are
\[ \frac{\pi_i}{1 - \pi_i} = e^{\alpha + \beta x_i} \]

and so

\[
\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \beta x_i
\]

Probabilities are bounded between 0 and 1 but odds can take any positive value and the log-odds any value positive or negative.

Table 2 of Hannum (2005) presents 4 logistic regressions. The simplest (1) has an intercept and a gender dummy with ref. = male. The estimate of \( \alpha \) is 1.73 and the estimate of \( \beta \) is –0.62. So the estimated odds on enrolment for a male are

\[
e^{1.73} = 5.64
\]

and for a female

\[
e^{1.73 - 0.62} = 3.03.
\]

The logit and probit models could be fitted by nonlinear least squares but generally the method of **maximum likelihood** is used.

**EViews Probit Output**

Here are some EViews results for fitting a probit model to the Mroz data on female labour force participation.
The table is quite similar to that for regression.

One point to note is that the coefficients in the index formula

$$-1.234708 - 0.633676 KL6$$
$$+ 0.154801 WE - 0.044971 HW$$

do not give the effect of a unit change in the variable on the probability of LFP.

The relationship between $\pi$ and the index—and hence any variable—is nonlinear.
The effect of a unit change in $x$ on $\pi$ is given by (for the case of a single variable) the formula

$$\frac{d\pi}{dx} = \frac{d}{dx} \Phi(\alpha + \beta x) = \beta \phi(\alpha + \beta x)$$

(using the chain rule for differentiating a function of a function) where $\phi$, the derivative of $\Phi$, is the standard normal density function.

The plot of $\frac{d\pi}{dx}$ against $x$:

The derivative $\frac{d\pi}{dx}$ has the same sign as $\beta$ and so qualitative conclusions can be drawn immediately from the table of estimates. However $\frac{d\pi}{dx}$ depends on the value of the index. Any unit change in $x$ will change the index by the same amount ($\beta x$) but the change will not have much effect on $\pi$ when the index is already very large.
We obtain estimates of the **marginal effect**

\[ \beta \phi(\alpha + \beta x) \]

by replacing the parameters by estimates and specifying an \( x \) value of interest—such as \( \bar{x} \).