Latent variables

I introduced the logit and probit models as convenient specifications for the nonlinear relationship between \( \pi \) the probability of a success and the index \( \alpha + \beta x \).

Another way of rationalising them is based on the idea of an underlying \textit{continuous latent variable} that is only \textit{partially observed}.

- Suppose we have a regression specification
  \[
y_i^* = \alpha + \beta x_i + \varepsilon_i
  \]
  but that the \textit{continuous} dependent variable is \textit{not} fully observed.
  What \textit{is} observed is the indicator \( y_i \)
  \[
y_i = \begin{cases} 
   1 & \text{if } y_i^* > 0 \\
   0 & \text{otherwise}.
  \end{cases}
  \]
  If exceeds the threshold (0) then 1 is recorded— if it does not, 0 is recorded.
  The diagram shows the part of the \( y^* \) density (centred at \( \alpha + \beta x \)) which gives
positive values. The area under the curve gives $\pi$, the probability that $y^* > 0$. The complementary probability $1 - \pi$ is given by the area to the left of 0.

After some algebra it turns out that

- $\varepsilon_i$ is $N(0, 1)$ produces the **probit model**.
- If $\varepsilon_i$ has the **logistic** distribution with distribution function $\frac{1}{1+e^{-t}}$ we get the **logit model**.

The idea of a latent variable with thresholds can be developed in a number of ways.

The **Tobit** model (a probit-like model introduced in econometrics by James Tobin) is a regression model with the observation ‘mechanism’
\[ y_i = \begin{cases} y_i^* \text{ if } y_i^* > 0 \\ 0 \text{ otherwise} \end{cases} \]

This kind of mechanism is called **censoring**.

The Mroz variable WHRS has many zeros because many of the women work 0 hours in a year. Censored regression (Tobit) seems more appropriate than ordinary regression. Models that take into account censoring are usually estimated by maximum likelihood. EViews has a censored estimation option.

The **ordered probit** or **ordered logit** models are used to model situations where there are more than 2 responses but these reflect an underlying order.

On the latent variable interpretation of probit/logit the 2 categories can be thought of as low and high. If we have 3 categories they may be based on high, medium or low values of a continuous latent variable.

Consider the case of a 3 point scale with the relationship between \( y \) and the underlying continuous variable \( y^* \):
\[
    y = \begin{cases} 
    1 & \text{if } y^* < c_1 \\
    2 & \text{if } c_1 \leq y^* < c_2 \\
    3 & \text{if } y^*_i \geq c_2 
    \end{cases}
\]

If \( y \) is a score on a student questionnaire response, we may be interested in how this response variable depends on the student’s major, background knowledge, etc. as put into a regression specification

\[
    y^*_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i
\]

This line of thought produces the **ordered probit** or **ordered logit** models, according as \( \varepsilon \) is normal or logistic. EViews has an ORDERED choice option.

These ordered models may not be appropriate because there may **no natural ordering** of the responses. E.g. individual’s choice of modes of transport—car, foot, bike. A standard model here is the **multinomial logit** model.

Hannum (2005) Table 5 p. 293 presents an analysis using the multinomial logit of the factors underlying the reasons for a child’s leaving school—“sick, handicapped,
economic difficulty,..., inconvenient transportation, other”.

There may also be structured choices as when somebody chooses from a menu between the categories meat and fish and then chooses within these categories. For tree-like decision processes the conditional logit (or probit) is used.