How to estimate a model when the dependent variable is a proportion

Xiang Ao

May 12, 2009

1 QMLE Logit

In a regression model, when the dependent variable is a proportion, a different approach than OLS is proposed by Papke and Wooldridge (1996)

`http://www.stata.com/support/faqs/stat/logit.html`

`http://www.ats.ucla.edu/stat/Stata/faq/proportion.htm`

I'll try to summarize what they suggest.

OLS in this case is inappropriate. A proportion is bounded between 0 and 1, but the predicted values from an OLS regression can never be guaranteed to lie in the unit interval.

The transitional way of transforming the data by a logit function is not appropriate. It can be formulated as

$$\mathrm{E}(\log[y/(1-y)]|X] = X\beta$$

This model has two drawbacks:

- The model cannot be true if y takes 0 or 1.
- Even if the model is well defined. We cannot recover $\mathrm{E}[y|x]$ from it, which is ultimately what we want.

Papke and Wooldridge (1996) proposed the following model:
Their assumption is that for all i,

$$\mathrm{E}(y_i|x_i) = G(X_i\beta), \tag{1}$$

where $G()$ is a cumulative distribution function (cdf). Most popular ones are logistic function and normal cdf.

Then the log-likelihood function is given by

$$l_i(\mathbf{b}) = y_i \log[G(X_i'\mathbf{b})] + (1-y_i)\log[1-G(X_i'\mathbf{b})], \tag{2}$$

which looks the same as our familiar logit log-likelihood function. The difference is that in a logit model, y_i is a binary variable; here y_i can take any value between 0 and 1.

The powerful result from Papke and Wooldridge (1996) is that they proved that under assumption 1, $\hat{b}$ is consistent, regardless of the actual distribution of y is. This is why 2 is called Quasi-MLE or QMLE since y does not have to have a logistic distribution. As long as the mean of y is correctly specified the QMLE estimator is consistent.

Therefore, we should follow Stata's advice of using glm when we have a proportion as our dependent variable.

2 beta distribution

Another way to model a proportion is to assume beta distribution, that is, $y \sim beta(\alpha, \beta)$.

Then model the mean of the distribution.

Specifically,

$$\log[\mu/(1-\mu)]|X] = X\eta$$

here $\mu = \frac{\alpha}{\alpha+\beta}$.

In Stata, $betafit$ is the command which implements this.