Modeling fractions

Herman J. Bierens

December 12, 2007

1 Modeling a single fraction

Let Y be a dependent variable which is bounded between zero and one, $Y \in (0,1)$, for example if Y is a fraction. A possible way to model the distribution of Y conditional on a vector X of predetermined variables, including 1 for the constant term, is to assume that

$$Y = \frac{\exp(\beta' X + U)}{1 + \exp(\beta' X + U)} = \frac{1}{1 + \exp(-\beta' X - U)}, \tag{1}$$

where U is an unobserved error term. Then

$$\ln[Y/(1-Y)] = \beta' X + U, \tag{2}$$

which, under standard assumptions on the error term U, can be estimated by OLS.

Note that model (1) is of the form

$$Y = F(\beta' X + U),$$

where $F(x) = (1 + \exp(-x))^{-1}$ is the logistic distribution function. This distribution function is convenient because its inverse F^{-1} has a closed form: If $y = F(x)$ then $x = F^{-1}(y) = \ln[y/(1-y)]$.

Of course, if Y is a percentage rather than a fraction, i.e., $Y \in (0,100)$, then (2) has to be modified to

$$\ln[Y/(100-Y)] = \beta' X + U.$$

As to the interpretation of the coefficients in the vector β, let $\beta = (\beta_1, ..., \beta_m)'$ and $X = (X_1, ..., X_m)'$, where $X_m = 1$ for the constant term. It follows from (1) that for $i = 1, ..., m-1$,

$$\frac{\partial Y}{\partial X_i} = \frac{\exp(-\beta' X - U)\beta_i}{(1+\exp(-\beta' X - U))^2} = Y^2 \exp(-\beta' X - U)\beta_i = \frac{\partial Y}{\partial U}.\beta_i \qquad (3)$$

Thus, β_i measures the marginal effect of X_i on Y relative to the marginal effect of the error term U on Y:

$$\beta_i = \frac{\partial Y/\partial X_i}{\partial Y/\partial U}.$$

2 Dynamic fraction models

In the presence of a lagged dependent variables model (2) can be modified to

$$\ln[Y_t/(1-Y_t)] = \alpha \ln[Y_{t-1}/(1-Y_{t-1})] + \beta' X_t + U_t, \qquad (4)$$

or

$$\ln[Y_t/(1-Y_t)] = \gamma Y_{t-1} + \beta' X_t + U_t, \qquad (5)$$

for example.

Because
$$\frac{d \ln[Y_{t-1}/(1-Y_{t-1})]}{dY_{t-1}} = \frac{1}{Y_{t-1}} - \frac{1}{1-Y_{t-1}}$$

it follows from (3) that in the case of model (4),

$$\begin{aligned} \frac{\partial Y_t}{\partial Y_{t-1}} &= \frac{\partial Y_t}{\partial \ln[Y_{t-1}/(1-Y_{t-1})]} \times \frac{d \ln[Y_{t-1}/(1-Y_{t-1})]}{dY_{t-1}} \\ &= \left(\frac{1}{Y_{t-1}} - \frac{1}{1-Y_{t-1}}\right) \frac{\partial Y_t}{\partial U_t}.\alpha \end{aligned}$$

so that the previous interpretation of α in terms of relative marginal effects no longer applies. On the other hand, in the case of model (5) we have

$$\gamma = \frac{\partial Y_t/\partial Y_{t-1}}{\partial Y_t/\partial U_t}.$$

Nevertheless, I would prefer model (4) over model (5) because the dynamic properties of model (4) are standard. In particular, if $|\alpha| < 1$ then under some further regularity conditions it follows by backwards substitution that

$$\ln[Y_t/(1-Y_t)] = \sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} + \sum_{j=0}^{\infty} \alpha^j U_{t-j}.$$

Consequently, with $X_{i,t-j}$ component i of X_{t-j},

$$\frac{\partial Y_t}{\partial X_{i,t-j}} = \frac{\exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right) \alpha^j \beta_i}{\left(1 + \exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right)\right)^2}$$

$$\frac{\partial Y_t}{\partial U_{t-j}} = \frac{\exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right) \alpha^j}{\left(1 + \exp\left(-\sum_{j=0}^{\infty} \alpha^j \beta' X_{t-j} - \sum_{j=0}^{\infty} \alpha^j U_{t-j}\right)\right)^2}$$

Hence

$$\beta_i = \frac{\partial Y_t/\partial X_{i,t-j}}{\partial Y_t/\partial U_{t-j}}, \ \alpha = \frac{\partial Y_t/\partial U_{t-j-1}}{\partial Y_t/\partial U_{t-j}}$$

for $j = 0, 1, 2, \ldots.$

3 Modeling multiple fractions

Now suppose that we have multiple fractions $Y_0, \ldots, Y_k$ as dependent variables, i.e., each $Y_i \in (0,1)$, and $\sum_{i=0}^{k} Y_i = 1$. For example, let Y_i be the fraction of people smoking a particular brand $i \in \{1, \ldots, k\}$ of cigarettes, with $Y_0 = 1 - \sum_{i=1}^{k} Y_i$ the fraction of all other people. The question now is how to model the joint distribution of $Y_0, \ldots, Y_k$ conditional on a vector X of predetermined variables such that the conditions $Y_i \in (0,1)$, and $\sum_{i=0}^{k} Y_i = 1$ hold, and how to estimate the parameters of this model. A possible way is the following.

Let for $i = 1, \ldots, k$,

$$Y_i = \frac{\exp(\beta_i' X + U_i)}{1 + \sum_{j=1}^{k} \exp(\beta_j' X + U_j)},$$

and

$$Y_0 = \frac{1}{1 + \sum_{j=1}^{k} \exp(\beta_j' X + U_j)},$$

where the U_i's are unobserved error terms. Then

$$\ln(Y_i/Y_0) = \beta_i'X + U_i,\ i = 1, ..., k, \tag{6}$$

which is a system of seemly unrelated regressions (SUR). If the vector X is common and there are no restrictions on the parameters β_i, SUR estimation is equivalent to estimating each equation separately by OLS. Otherwise one has to conduct SUR estimation in the usual (textbook) way. The latter can be done in EasyReg via the GMM module.

Note that model (6) also applies if $Y_0, ..., Y_k$ are percentages which add up to 100%.