Details of methods and formulas for Stata ado file *selmlog.ado*

Martin Fournier and Marc Gurgand
(gurgand@pse.ens.fr)

November 22, 2005

Consider the following model

$$\begin{aligned} y_1 &= x\beta_1 + u_1 \\ y_j^* &= z\gamma_j + \eta_j, \qquad j = 1 \dots M \end{aligned} \tag{1}$$

where the disturbance u_1 is not parametrically specified and verifies $E(u_1|x,z) = 0$ and $V(u_1|x,z) = \sigma^2$. j is a categorical variable that describes the choice of an economic agent among M alternatives based on "utilities" y_j^*. The vector z represents the maximum set of explanatory variables for all alternatives and the vector x contains all determinants of the variable of interest. We assume that the model is non-parametrically identified from exclusion of some of the variables in z from the variables in x. Without loss of generality, the outcome variable y_1 is observed if and only if category 1 is chosen, which happens when:

$$y_1^* > \max_{j \neq 1}(y_j^*) \tag{2}$$

Define:

$$\begin{aligned} \varepsilon_1 &= \max_{j \neq 1}(y_j^* - y_1^*) \\ &= \max_{j \neq 1}(z\gamma_j + \eta_j - z\gamma_1 - \eta_1) \end{aligned} \tag{3}$$

Under definition (3), condition (2) is equivalent to:

$$\varepsilon_1 < 0$$

Assume that the (η_j)'s are independent and identically Gumbel distributed (the so-called IIA hypothesis). Their cumulative and density functions are respectively $G(\eta) = \exp(-e^{-\eta})$ and $g(\eta) = \exp(-\eta - e^{-\eta})$. As shown by McFadden (1973), this specification leads to the multinomial logit model with:

$$P(\varepsilon_1 < 0|z) = \frac{\exp(z\gamma_1)}{\sum_j \exp(z\gamma_j)}$$

Based on this expression, consistent maximum likelihood estimates of the (γ_j)'s can be easily obtained.

The problem is to estimate the parameter vector β_1 while taking into account that the disturbance term u_1 may not be independent of all (η_j)'s. This would introduce some correlation between the explanatory variables and the disturbance term in the outcome equation of model (1). Because of this, least squares estimates of β_1 would not be consistent.

1 Lee's model

Following Lee (1983), call $F_{\varepsilon_1}(.|\Gamma)$ the cumulative distribution function of ε_1. The cumulative $J_{\varepsilon_1}(.|\Gamma)$, defined by the following transform:

$$J_{\varepsilon_1}(.|\Gamma) = \Phi^{-1}(F_{\varepsilon_1}(.|\Gamma))$$

where Φ is the standard normal cumulative, has a standard normal distribution. Assume that u_1 and $J_{\varepsilon_1}(\varepsilon_1|\Gamma)$ are jointly distributed under the following hypothesis with $E(u_1|\varepsilon_1, \Gamma) = \sigma\rho_1.J_{\varepsilon_1}(\varepsilon_1|\Gamma))$ The expected value of the disturbance term u_1, conditional on category 1 being chosen, can now be written as:

$$E(u_1|\varepsilon_1 < 0, \Gamma) = -\sigma\rho_1 \frac{\phi(J_{\varepsilon_1}(0|\Gamma))}{F_{\varepsilon_1}(0|\Gamma)}$$

with ϕ the standard normal density. Under this hypothesis, a consistent estimator of β_1 is obtained by running least squares on the following equation:

$$y_1 = x_1\beta_1 - \sigma\rho_1 \frac{\phi(J_{\varepsilon_1}(0|\Gamma))}{F_{\varepsilon}1(0|\Gamma)} + w_1 \tag{4}$$

Two-step estimation of (4) is thus implemented by first estimating the $(\gamma_j)'$s in order to form $\phi(J_{\varepsilon_1}(0|\widehat{\Gamma}))/F_{\varepsilon_1}(0|\widehat{\Gamma})$ and then by including that variable in equation (4) to estimate consistently β_1 and $(\sigma\rho_1)$ by least squares. σ can then be recoverred.

2 Dubin and Mc Fadden's model

Dubin and Mc Fadden (1984) use the following linearity assumption: $E(u_1|\eta_1 \ldots \eta_M) = \sigma\frac{\sqrt{6}}{\pi} \sum\limits_{j=1\ldots M} r_j(\eta_j - E(\eta_j))$, where r_j is a correlation coefficient between u_1 and η_j. With the multinomial logit model:

$$\begin{aligned} E(\eta_1 - E(\eta_1)|y_1^* &> \max_{s\neq 1}(y_s^*), \Gamma) = -\ln(P_1), \\ E(\eta_j - E(\eta_j)|y_1^* &> \max_{s\neq 1}(y_s^*), \Gamma) = \frac{P_j \ln(P_j)}{1 - Pj}, \qquad \forall j > 1 \end{aligned}$$

Model (1) can thus be estimated by least squares based on:

$$y_1 = x_1\beta_1 + \sigma\frac{\sqrt{6}}{\pi}\sum_{j=2\ldots M} r_j\left(\frac{P_j\ln(P_j)}{1-Pj}\right) - r_1\ln(P_1) + w_1 \tag{5}$$

This is dmf(1) option in the program.

dmf(0) option uses the following retriction: $\sum_{j=1\ldots M} r_j = 0$. The model then becomes:

$$y_1 = x_1\beta_1 + \sigma\frac{\sqrt{6}}{\pi}\sum_{j=2\ldots M} r_j\left(\frac{P_j\ln(P_j)}{1-Pj} + \ln(P_1)\right) + w_1 \tag{6}$$

To implement dmf(2) option, define the following standard normal variables:

$$\eta_j^* = J(\eta_j) = \Phi^{-1}(G(\eta_j)), \qquad j = 1\ldots M$$

For every j, assume that the expected values of u_1 and η_j^* are linearly related. This holds in particular under the classical assumption that u_1 is normal and (u_1, η_j^*) is bivariate normal for any category j. If r_j^* is the correlation between u_1 and η_j^*, u_1 may be expressed as the following linear combination: $E(u_1|\eta_1\ldots\eta_M) = \sigma\sum_{j=1\ldots M} r_j^*\eta_j^*$. In this setup, conditional expectations are more involved. Note for convenience:

$$m(P_j) = \int J(v - \log P_j)g(v)dv, \qquad \forall j$$

The following results can be derived:

$$\begin{aligned}
E(\eta_1^*|y_1^* &> \max_{s\neq 1}(y_s^*), \Gamma) = m(P_1) \\
E(\eta_j^*|y_1^* &> \max_{s\neq 1}(y_s^*), \Gamma) = m(P_j)\frac{P_j}{P_j - 1}, \qquad \forall j > 1
\end{aligned}$$

The outcome equation in (1) conditional on choosing $j = 1$ is now:

$$y_1 = x_1\beta_1 + \sigma\left[r_1^* m(P_1) + \sum_{j=2\ldots M} r_j^* m(P_j)\frac{P_j}{(P_j - 1)}\right] + w_1 \tag{7}$$

The integrals $m(P_j)$ have no closed form, but they can be computed numerically after the multinomial logit estimation. This is not a source of computational complexity, however, as it must be done only once for each observation.

3 Dahl's model

Following Dahl (2002) we consider a selectivity correction term of the general form $\mu(P_1, \ldots, P_M)$.

The estimated equation becomes:

$$y_1 = x_1\beta_1 + \mu(P_1, \ldots, P_M) + w_1 \tag{8}$$

The function μ takes the form of a polynomial in $P_1, \ldots, P_M$ with the order provided in the command. With dhl(# all), all $P_1, \ldots, P_M$ are included. Otherwise, with dhl(#), only a polynomial (of order #) in P_1 is used.

References

Dahl G. B., 2002, "Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets", *Econometrica*, vol. 70, 2367-2420.

Dubin J.A. & McFadden D.L., 1984, "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption", *Econometrica*, vol. 52, 345-362.

Lee L.F., 1983, "Generalized Econometric Models with Selectivity", *Econometrica*, vol. 51, 507-512.

McFadden D.L., 1973, "Conditional Logit Analysis of Qualitative Choice Behavior", in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press.