help for **selmlog**

Selection bias correction based on the multinomial logit model (version 1.3)

selmlog *depvar varlist* [**if***exp*] [**in***range*],**select**(*depvar_m=varlist_m*) [**lee dmf**(*#*) **dhl**(*#* [*all*]) **showmlogit wls bootstrap**(*number_of_replications* [*sample_size*]) **mloptions**(*mlogit options*) **gen**(*variable generic name*)]

Description

selmlog estimates linear regression models on a selected subset of observations, where selectivity is modelled as a multinomial logit (as opposed, for instance, to univariate probit as in the Heckman model). Estimation is run by step (multi ogit, then linear regression with selectivity correction).

It applies a set of methods reviewed in Bourguignon, Fournier and Gurgand (2004).

In the equation of interest, *depvar* is regressed on *varlist* .

In the selection equation, *depvar_m* is a variable that identifies the multiple choices and *varlist_m* the corresponding explanatory variables: refer to the **mlogit** command for this syntax. *depvar_m* should not take negative valu

The outcome variable *depvar* is observed for only one value of *depvar_m* . It is important that depvar should have missing values for any other value of *dep*

In the output, **selmlog** adds to varlist a series of variables labelled *_m[depvar_m value]* , except for the **dhl** option.

These variables are consistent estimators of conditional expected values of the residuals derived from the multinomial logit model. Their formula depends on the bias correction method chosen in the option command.

The coefficients on these variables are functions of the covariance between the residual in the regression and the residuals (or some function of the residuals) from the multinomal logit model.

With the **dhl** option, they are the coefficients on polynomials of the selection probabilities and have no structural interpretation (in particular, they do not correspond to well defined correlations).

With the **dhl** option the variables are labeled with four indexes: *_m[depvar_m value i][order][depvar_m value j][order]* for all combinaisons of *depvar_m* va and j, with *order* the polynomial orders running from 0 to the use ed maximum (when *order* is 0 for one of the probability, the *_m* variable is then *_m[depva value i][order]*

Except for the **dhl** option, the implied standard error of the residual of the regression equation is also reported, as well as implied correlation coefficients (note that they are not restricted to [-1,1]).

Options

lee performs the Lee (1983) correction method.

dmf(0) performs the Dubin-McFadden (1984) correction method.

dmf(1) performs the Dubin-McFadden (1984) correction method, waving the restriction (imposed in Dubin-McFadden (1984)) that all correlation coefficients sum-up to zero.

dmf(2) performs a variant of the Dubin-McFadden (1984) correction method suggested in Bourguignon, Fournier and Gurgand (2004).

dhl performs corrections based on Dahl (2002) using selection probabilities in polynomial form.

The **dhl** options include the order of the polynomials on the selection probabilities. With this number alone, the correction term includes only the probability to be selected on the observed outcome. If this number is followed by **all**, probabilities are included in polynomial form, with interactions, up to the specified order.

showmlogit reports the multinomial logit estimated in the first-step.

wls applies weighted least squares in the second step regression to account for heteroskedasticity present in the model due to selectivity.

This option can achieve (asymptotic) efficiency, but, in some instances, some of the estimated variances may be negative. Waiving this option then allows to estimate the model however. This option is not available with **dhl**.

bootstrap uses bootstrap to estimate the parameter standard errors. The user must specify the number of replications. The default *sample_size* is the size the sample in use.

If this option is waived, the reported variances take no account of the two-step nature of the procedure and implied residual variance and correlations have no reported standard errors.

mloptions contains the list of stata mlogit options that need to be executed during the first-stage estimation.

gen outputs the *_m* variables used in the estimation but with the provided *generic name* instead of *_m*.

Methods and formulas

The regression of interest is y = xb + u, with V(u)=s2.

y is observed only if category 1 (say) is chosen among K alternatives. This happens when y*1 > max(y*j), with y*j = zaj + vj, for j=1 to K.

When the residuals (vj) are assumed independent and identically Gumbel distributed, this leads to the multinomial logit model.

Let Pj be the probability that category j is chosen. All methods considered here include a selectivity correction term of the form y = xb + h(P1...PK) + e.

The Lee method assumes: h(P1...PK)=-s*c*normd(invnorm(P1))/P1, where c is a covariance parameter. The program generates _m1= normd(invnorm(P1))/P1 and estimates (-s*c). s2 and c are then recovered.

The Dubin-McFadden method (**dmf(0)**) assumes: h(P1...PK)=s*r2*(_m2)+...+s*rK*(_mK) where mj=Pj*log(Pj)/(1-Pj)+ log(P1), j>2, and rj is the correlation coefficient between u and (vj-v1). The program estimates (s*r2) to (s*rK). s2 and r2 to rK recovered.

The Dubin-McFadden first variant (**dmf(1)**) assumes: h(P1...PK)=s*r1*(_m1)+...+s*rK*(_mK) where _m1=log(P1) and _mj=Pj*log(Pj)/(1-Pj), j>2, and rj is the correlation coefficient between u and (vj). The program estimates (s*r1) to (s*rK). s2 a rK are then recovered.

For the second Dubin-McFadden variant (**dmf(2)**), define the transformed normally distributed residuals: v*j = invnorm(G(vj)), for j=1 to K, where G(.) is the cumulative of the Gumbel distribution. Bourguignon, Fournier and Gurgand (2004) sho (P1...PK)=s*r1*_m1 + ... + s*rK*_mK. The program estimates (s*r1) to (s*rK). s2 and r1 to rK are then recovered.

The _m1 to _mK variables involve numerical integrals that are computed using Gauss-Laguerre quadrature. The abscissas and weight factors used in the program are from Davis and Polonsky (1964).

In this program, the second step regressions are estimated by linear least squares and the standard error s2 is not estimated separately from the correlation coefficients.

Neither are the latter constrained between -1 and 1. However, implied s2 and correlations are presented in the output.

In the second step model, the residual e is heteroscedastic. Weights for the weighted least squares estimates are detailed in Bourguignon, Fournier and Gurgand (2004), appendix 2. s2 and correlations can be recovered based on the formulas there.

For the Dahl method (**dhl**) when the option *all* is absent, h(P1...PK)=f(P1) where f() is a polynom, the order of which is user-supplied.

With the option "all", h(P1...PK) is a polynomial function of all probabilities interacted up to the user-supplied order.

References

Bourguignon F., Fournier M. and Gurgand M., Selection Bias Corrections Based on the Multinomial Logit Model: Monte-Carlo comparisons, mimeo Delta, 2004 (download from http\\:www.pse.ens.fr\senior\gurgand\selmlog13.htm).

Dahl G. B., "Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets", Econometrica, vol. 70, 2367-2420, 2003.

Davis Ph. and Polonsky I., "Numerical Interpolation Differentiation and Integration" in Abramovitz M. and Stegun I.A. (Eds.), Handbook of Mathematical Functions, National Bureau of Standards - Applied mathematics series 55, 1964.

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

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

Authors

Marc Gurgand, PSE-CNRS and CREST-INSEE (France) & Martin Fournier, CEFC (Hong-Kong) and Universite d'Auvergne (France). Contact: Marc Gurgand, gurgand@pse.ens.fr