SAMPSEL

SAMPSEL (nonlinear options) *probit dep var probit indep vars*
| *regression dep var regression indep vars* ;

Function:

SAMPSEL estimates a generalized Tobit or sample selection model where both the regression and the latent variable which predicts selection are linear regression functions of the exogenous variables. Either a censored (regression variables not observed for non-selected observations) or truncated (all variables not observed) model may be estimated.

Usage:

The model estimated is the Tobit type II model described by Amemiya or the censored regression model with a stochastic threshold described by Maddala (see the references). It can be written as

$y_{2i} = X_{2i}\, b + e_i$ if $y_{1i}>0$ (regression equation)

$= 0$ otherwise

$y_{1i} = X_{1i}\, d + u_i$ (probit equation)

e_i and u_i are assumed to be joint normally distributed with variances SIGMA squared and unity (a normalizing assumption) and correlation coefficient RHO.

To use the procedure to estimate this model, supply the name of a zero/one variable which tells whether the observation was observed or not ($y_{1i}>0$) as the probit dependent variable, the regressors X_{1i} as the probit independent variables, y_{2i} as the regression dependent variable, and X_{2i} as the regression independent variables. Missing values for the regression variables are allowed for those observations for which $y_{1i}=0$.

If y_{1i} is always greater than zero, the truncated (conditional) model is estimated (Bloom and Killingsworth). This is flagged with the message "LATENT SELECTION VARIABLE". The identifying condition that there be variables other than the constant in the probit equation is not checked.

Options:

Standard nonlinear options -- see NONLINEAR.

HITER=N, HCOV=N is the default.

Example:

```
SAMPSEL (PRINT, MAXIT=50, HCOV=NBW) IY C Z | Y C X ;
```

SAMPSEL

Output:

The output of SAMPSEL begins with an equation title and the name of the dependent variable. Starting values and diagnostic output from the iterations will be printed. Final convergence status is printed. This is followed by the mean of the dependent variable, number of positive observations, sum of squared residuals, R-squared, and a table of right hand side variable names, estimated coefficients, standard errors and associated t-statistics.

SAMPSEL also stores some of these results in data storage for later use. The table below lists the results available after a SAMPSEL command.

Name	Type	Length	Variable Description
@LHV	list	1	Name of dependent variable.
@RNMS	list	#vars	Names of right hand side variables.
@YMEAN	scalar	1	Mean of the probit dependent variable.
@NOB	scalar	1	Number of observations.
@NPOS	scalar	1	Number of positive observations in probit equation.
@SSR	scalar	1	Sum of squared residuals (regression equation).
@LOGL	scalar	1	Log of likelihood function.
@IFCONV	scalar	1	1 if convergence achieved, 0 otherwise.
@NCOEF	scalar	1	Number of rhs variables.
@NCID	scalar	1	Number of identified coefficients.
@COEF	vector	#vars	Coefficient estimates.
@SES	vector	#vars	Standard errors.
@T	vector	#vars	T-statistics.
@GRAD	vector	#vars	Gradient of log L at convergence.
@VCOV	matrix	#vars* #vars	Variance-covariance of estimated coefficients.
@DPDX	matrix	#vars*2	Mean of probability derivatives.
@RES	series	#obs	Residuals.
@MILLS	series	#obs	Inverse Mills ratios.

If the regression includes a PDL variable, the following will also be stored:

@SLAG	scalar	1	Sum of the lag coefficients.
@MLAG	scalar	1	Mean lag coefficient.
@LAGF	vector	#lags	Estimated lag coefficients, after "unscrambling".

@RES for the regression equation is stored for the observed sample only.

Method:

The method used is maximization of the likelihood function by means of a gradient method with the Hessian approximation given by the HITER option. Since this likelihood function is known to often have multiple local optima, the method of Nawata (1994, 1995, 1996) is used to find the global optimum. A grid search is done on RHO to find local optima. The (27) grid points used are 0 .1 .2 ... .8 .85 .9 .95 .99 .9999 , -.1 -.2 ... -.9999 . At each grid point, RHO is held fixed, and LogL is maximized with respect to the other parameters. For each local optimum, final iterations are done which allow RHO to vary. These iterations refine the optima to full precision and make it possible to choose the global optimum.

Sometimes the global optimum shows RHO = 1.0000 or -1.0000 . In these cases, the actual estimate of RHO is slightly

less than 1 in absolute value, and the residual covariance matrix is nearly singular. The standard error of RHO and its covariance with other parameters is set to zero in these cases. It is not clear how to interpret the model in these cases. They appear to occur when the probit equation is dominating the likelihood function, and the intercepts become distorted. Often if more variables are added to the regression equation to improve its fit, the estimate of RHO will be less extreme.

References:

Amemiya, Takeshi, **Advanced Econometrics**, Harvard University Press, Cambridge, Massachusetts, 1985, Chapter 13.

Bloom, David E., and Killingsworth, Mark R., "Correcting for Truncation Bias caused by a Latent Truncation Variable", **Journal of Econometrics**, 1985, pp. 131-135.

Griliches, Z., Hall, B. H., and Hausman, J. A., "Missing Data and Self- selection in Large Panels," **Annales de l'Insee**, Avril-Sept 1978, pp. 137- 176.

Maddala, G. S., **Limited-Dependent and Qualitative Variables in Econometrics**, Cambridge University Press, Cambridge, 1983, Chapter 6.

Nawata, Kazumitsu,"Estimation of Sample Selection Bias Models by the Maximum Likelihood Estimator and Heckman's two-step Estimator," **Economics Letters** 45, 1994, pp. 33-40.

Nawata, Kazumitsu,"Estimation of Sample Selection Models by the Maximum Likelihood Method," **Mathematics and Computers in Simulation** 39, 1995, pp. 299-303.

Nawata, Kazumitsu, and Nagase, Nobuko, "Estimation of Sample Selection Bias Models," **Econometric Reviews** 15, 1996, pp. 387-400.

Olsen, R.J., "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator," **International Economic Review** 23, 1982, pp. 223-240.