PROBIT

PROBIT (MILLS=*name for output inverse Mills ratio*, nonlinear options)
dependent variable list of independent variables ;

Function:

PROBIT obtains estimates of the linear probit model, where the dependent variable takes on only two values. Options allow you to obtain and save the inverse Mills ratio as a series so that the sample selection correction due to Heckman can be estimated (also see the SAMPSEL command).

Usage:

The basic PROBIT statement is like the OLSQ statement: first list the dependent variable and then the independent variables. If you wish to have an intercept term in the regression (usually recommended), include the special variable C or CONSTANT in your list of independent variables. You may have as many independent variables as you like subject to the overall limits on the number of arguments per statement and the amount of working space, as well as the number of data observations you have available.

The observations over which the regression is computed are determined by the current sample. If any of the observations have missing values within the current sample, PROBIT will print a warning message and will drop those observations. PROBIT also checks for complete or quasi-complete sample separation by one of the right hand side variables; such models are not identified.

The list of independent variables on the PROBIT command may include variables with explicit lags and leads as well as PDL (Polynomial Distributed Lag) variables. These distributed lag variables are a way to reduce the number of free coefficients when entering a large number of lagged variables in a regression by imposing smoothness on the coefficients. See the PDL section for a description of how to specify such variables.

The dependent variable need not be a strictly zero/one variable. Positive values are treated as one and zero or negative values are treated as zero.

Options:

MILLS= the name of a series used to store the inverse Mills ratio series evaluated at the estimated parameters. The default is @MILLS.

Nonlinear options - see the NONLINEAR section of this manual.

Examples:

Standard probit model:

```
PROBIT MOVE C WAGE1 WAGE2 COST1 COST2;
```

Heckman sample selection model (see the SAMPSEL command for ML estimation):

```
PROBIT(MILLS=RMILL) WORK C OCC1 OCC2 TENURE MSTAT AGE;
SELECT WORK;
```

PROBIT

```
OLSQ LWAGE C SCHOOL EXPER IQ UNION OCC1 OCC2 RMILL;
```

Computing fitted probabilities and inverse Mills ratios explicitly (see Method):

```
PROBIT MOVE C WAGE1 WAGE2 COST1 COST2;
FORCST XB;
MOVEP = CNORM(XB);
MILLSR = MOVE * DLCNORM(XB) + (1-MOVE) * (-DLCNORM(-XB));
```

Output:

The output of PROBIT 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.

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

Name	Type	Length	Variable Description
@LHV	list	1	Name of dependent variable.
@RNMS	list	#vars	Names of right hand side variables.
@IFCONV	scalar	1	1 if convergence achieved, 0 otherwise.
@YMEAN	scalar	1	Mean of the dependent variable.
@NOB	scalar	1	Number of observations.
@NPOS	scalar	1	Number of positive observations.
@SSR	scalar	1	Sum of squared residuals.
@SRSQ	scalar	1	Scaled R-squared.
@RSQ	scalar	1	R-squared.
@LOGL	scalar	1	Log of likelihood function.
@LR	scalar	1	Likelihood ratio test for zero slopes.
@NCOEF	scalar	1	Number of rhs variables (#vars).
@NCID	scalar	1	Number of identified coefficients.
@COEF	vector	#vars	Coefficient estimates.
@SES	vector	#vars	Standard errors.
@T	vector	#vars	T-statistics.
%T	vector	#vars	p-values for 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.
@FIT	series	#obs	Fitted probabilities.
@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 (number of time periods).
@LAGF	vector	#lags	Estimated lag coefficients, after "unscrambling".

Method:

PROBIT uses analytic first and second derivatives to obtain maximum likelihood estimates via the Newton-Raphson algorithm. This algorithm usually converges fairly quickly. TSP uses zeros for starting parameter values, unless @START is used to override this (see NONLINEAR in this manual).

Multicollinearity of the independent variables is handled with generalized inverses, like the regression procedures in TSP.

The numerical implementation involves evaluating the normal density and cumulative normal distribution functions. The cumulative normal distribution function is computed from an asymptotic expansion, since it has no closed form. See the reference under the CDF command for the actual method used to evaluate CNORM(). The ratio of the density to the distribution function is also known as the inverse Mills ratio. This is used in the derivatives and with the MILLS= option.

@MILLS is actually the expectation of the structural residual, where the model is given by

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N(0,1)$$

$$D = I(y>0)$$

@MILLS is the value of the following two expressions, depending on whether D=0 or 1:

$$E(\varepsilon|D=1) = \frac{NORM(-X\beta)}{1-CNORM(-X\beta)} = \frac{NORM(X\beta)}{CNORM(X\beta)} = DLCNORM(X\beta)$$

$$E(\varepsilon|D=0) = -\frac{NORM(-X\beta)}{CNORM(-X\beta)} = -DLCNORM(-X\beta)$$

Before estimation, PROBIT checks for univariate complete and quasi-complete separation of the data and flags this condition. (The model is not identified in this case.) Without this check, one or more RHS variables perfectly predict Y for some observations, and their coefficients would slowly iterate to + or - infinity.

The scaled R-squared is a measure of goodness of fit relative to a model with just a constant term; see Estrella(1998).

References:

Amemiya, Takeshi, "Qualitative Response Models: A Survey," **Journal of Economic Literature 19**, December 1981, pp. 1483-1536.

Estrella, Arturo, "*A New Measure of Fit for Equations with Dichotomous Dependent Variables*," **Journal of Business and Economic Statistics**, April 1998, pp.198-205.

Maddala, G. S., **Limited-dependent and Qualitative Variables in Econometrics**, Cambridge University Press, New York, 1983, pp. 22-27, 221-223, 231-234, 257-259, 365.