LOGIT

_____

LOGIT (CASE=*var*, COND, NCHOICE=*n*, NREC=*var*, SUFFIX=*list,* nonlinear options)
dependent variable   independent variables ;
or
dependent variable   conditional variables | multinomial(alternative) variables ;

_____

Function:

LOGIT is used to estimate a conditional and/or multinomial logit model. The explanatory variables in the model may vary across alternatives (choices) for each observation or they may be characteristics of the observation, or both. There is no limit on the number of alternatives.

Usage:

There are three types of logit model: those where the regressors are the same across all choices for each observations, i.e., they are characteristics of the chooser, those where the regressors are characteristics of the specific choice, and mixed models, which have regressors of both kinds. In the first case (multinomial logit), a separate coefficient for each regressor is estimated for all but one of the choices. In the second case (conditional logit), the regressors change across the choices, and a single coefficient is estimated for each set of regressors.

1. Binary or multinomial logit -- just like OLSQ or PROBIT:

> LOGIT *dependent variable   multinomial variables   (chooser characteristics)*;
>   LOGIT Y C X1 X2 ... XK;

Y can be 0/1 or 1/2 or any integral values. If Y takes on more than 2 values, the model is multinomial logit. The names of the coefficients are determined by appending the values of Y for each choice to the names of the explanatory variables. The coefficients are normalized by setting the coefficients for the lowest choice to 0. If Y is 0/1, the coefficients C1, X11, X21,..., XK1 would be estimated, with C0, X10, X20,..., XK0 normalized to zero. If Y is 1/2/3, the coefficients C2, X12, X22,..., XK2 and C3, X13, X23,..., XK3 would be estimated, with C1, X11, X21,..., XK1 normalized to zero.

> LOGIT (NCHOICE=2) Y C X1 X2 ... XK;

causes TSP to check the range of Y to make sure there are only 2 choices. The model estimated has K+1 coefficients and K+1 variables.

The multinomial logit procedure checks for "univariate complete and quasi-complete separation", which prevents identification of the coefficients.

2. Conditional logit:

> LOGIT (COND,NCHOICE=n) *dependent variable   conditional variables   (choice characteristics)* ;

For example,

> LOGIT (COND,NCHOICE=2) Y X Z;

looks for the variables X1,X2,Z1,Z2 corresponding to the 2 choices. The coefficients X and Z would be estimated. C is not allowed as a conditional variable (since it does not vary across choices, it is not identified). For a choice-specific

LOGIT

set of dummies, use C as a multinomial variable in mixed logit. In this case, the example shown becomes

    LOGIT (COND,NCHOICE=2) Y X Z | C;

The CASE option allows you to use data organized with one choice per observation rather than with one case per observation. For example,

    LOGIT (CASE=V) Y X Z;

where the variable V is a case ID which is equal for adjacent observations which belong to the same case. In this case, the variable names X and Z are used directly (not X1,X2, etc.). There need not be an equal number of observations per case. Only the first Y for each case is examined for a valid choice number.

For multiple observations per case datasets, you can also use

    LOGIT (NREC=W) Y X Z;

In this case, W specifies the number of records per case and you do not need to supply an ID variable with the CASE option.

3. Mixed logit:

The general form of the command is now

    LOGIT (COND,NCHOICE=n) *dependent variable   conditional variables | multinomial variables* ;

For example,

    LOGIT (COND,NCHOICE=3) Y ZA ZB | XA XB XC;

Y takes on the values 1,2, and 3. TSP looks for the conditional variables ZA1,ZA2,ZA3, ZB1,ZB2,ZB3 corresponding to the 3 choices. XA,XB,XC are the multinomial variables. The coefficients ZA,ZB, XA2,XB2,XC2, XA3,XB3,XC3 would be estimated, with XA1,XB1,XC1 normalized to zero.

Options:

The standard options for nonlinear estimation are available: see the NONLINEAR section in this manual. Note that HITER=N, HCOV=N are the defaults for the Hessian approximation and standard error computations. In addition, the following options are specifically for the LOGIT procedure:

CASE=*case series* for multiple observations per case. This variable holds a case identification which is equal for adjacent observations that belong to the same case. Note that any such variable may be used; it does not necessarily have to be the case identification.

COND/**NOCOND** for conditional or mixed models versus pure multinomial estimation (see the Usage section). If CASE= or NREC= is used, COND is assumed and need not be specified.

NCHOICE=*number of choices* can be supplied when the number is equal for all observations. The program then checks to make sure the data satisfy this constraint. This is not used with CASE= or NREC=, since then it is valid to take the first choice every time.

NREC=*choice count series* for multiple observations per case. This variable specifies the number of observations in each case. Usually the number is repeated in each observation, but only the count in the first observation for each case is used. You cannot say NREC=3, but you can say NCHOICE=3.

SUFFIX=*list* is an alternative way of selecting a subset of the choices. For example, NCHOICE=2 is equivalent to SUFFIX=(1,2). A SELECT command can also be used to do this; this option has not seen much testing.

Examples:

      NOSUPRES @DPDX ;
      LOGIT Y C X ;

See the usage section for other examples of how to use this procedure.

Output:

The printed output of the LOGIT procedure is similar to that of the other nonlinear estimation procedures in TSP. A title is followed by a table of the frequency distribution of the choices. Then the starting values and iteration-by-iteration printout is printed; the amount is controlled by the PRINT and SILENT options. This is followed by a message indicating final convergence status, the value of the likelihood and a table of parameter estimates and their asymptotic standard errors and t-statistics. The variance-covariance matrix of the estimates is also printed if it has been unsuppressed. The @DPDX or @DPDZ matrices (described below) are printed unless they have been SUPRESed. The following are also stored in data storage:

Name	Type	Length	Variable Description
@RNMS	list	#params	Parameter names.
@LOGL	scalar	1	Log of likelihood function.
@IFCONV	scalar	1	Convergence status (1 = success).
@LR	scalar	1	Likelihood ratio test for zero slopes.
@SRSQ	scalar	1	Scaled R-squared for multinomial logit.
@RSQ	scalar	1	Squared correlation between Y and @FIT for binary logit.
@SSR	scalar	1	Sum of squared residuals (Y-@FIT) for binary logit.
@GRAD	vector	#params	Gradient of likelihood function at maximum.
@COEF	vector	#params	Estimated values of parameters.
@SES	vector	#params	Standard Errors of estimated parameters.
@T	vector	#params	T-statistics.
%T	vector	#params	p-values for T-statistics.
@VCOV	vector	#par*#par	Estimated variance-covariance of estimated parameters.
@DPDX	matrix	#vars* #choices	Mean of probability derivatives for multinomial variables. This matrix is invariant to the set of coefficients which are normalized to zero (as are the differences between sets of coefficients).
@DPDZ	matrix	#vars* #choices	Mean of probability derivatives for conditional variables (#vars = #condvars* #choices). This matrix consists of NCHOICE submatrices of dimension (NCOND x NCHOICE) stacked vertically, and it is block-symmetric. Not calculated if NCHOICE varies by case.
@FIT	matrix or series	#obs* #choices or #obs	Matrix of fitted probabilities when NCHOICE>2 and there is one observation per case. (NOB x 1) series when NCHOICE=2 or when there are multiple observations per case. If NCHOICE=2, @FIT will contain the probabilities for the highest choice only (just like binary PROBIT).

Method:

If C_t is the choice set for the tth observations, and observation t chooses the ith alternative out of C_t, then the expression

3

LOGIT

for the choice probability is

$$\Pr(i \text{ is chosen from } C_t | Z) = \exp(Z_{ti}b) / \sum_j \exp(Z_{tj}b)$$

and the likelihood function is

$$\text{Log } L = -\sum_t \log \left[\sum_j \exp((Z_{tj} - Z_{ti})b) \right]$$

The coefficient vector to be estimated is b. If some of the Zs do not vary across the choices, these equations would apply to an expanded Z vector formed by taking the Kronecker product of the fixed Zs with an identity matrix of order of the number of choices (less one for a normalization). The actual implementation does not expand the Zs, but treats the conditional and multinomial variables differently to conserve space.

Newton's method is used to maximize this likelihood function with respect to the parameter vector b. The global concavity of the likelihood function makes estimates fairly straightforward to obtain with this method. Zero starting values are the default, unless @START is supplied. See the NONLINEAR section in this manual for more information about TSP's nonlinear optimization procedures in general.

The evaluation of the EXP() functions in the likelihood function and derivatives has been written to avoid floating overflows and zero divides. When these conditions might occur, the appropriate limit is taken instead. This may result in some slight inaccuracy in the likelihood function, but it is certainly preferable to halting the estimation. Observations subject to these problems can be identified by exact 0 and 1 values in @FIT.

Before estimation, LOGIT 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:

Albert, A., and J.A. Anderson, "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," in **Biometrika**, 71, (1984).

Amemiya, Takeshi, **Advanced Econometrics,** Harvard University Press, 1985, Chapter 9.

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 Variables and Qualitative Variables in Econometrics,** Cambridge University Press, 1983, Chapters 2 and 3.

McFadden, Daniel S., "Conditional Logit Analysis of Qualitative Choice Behavior," in Zarembka, P., ed., **Frontiers in Econometrics**, Academic Press, 1973.

McFadden, Daniel S., "Quantal Choice Analysis: A Survey," **Annals of Economic and Social Measurement**, Vol. 5 (1976), pp. 363-390.

Train, Kenneth, **Qualitative Choice Analysis,** The MIT Press, Cambridge, MA, 1986.