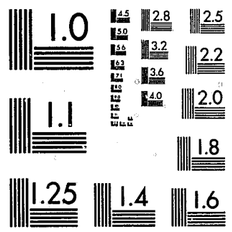

Rhodes Associates

Criminal Justice Research Series

DOCUMENTATION FOR SAS-COMPATIBLE LOGIT, TOBIT AND MVPROBIT PROCEDURES

by

David A. Lombardero
Frederick C. Nold

January 11, 1983

CJRS 5

This report and the computer programs it describes were developed under grant number 81-IJ-CX-0055 from the National Institute of Justice. Responsibility for defects and errors in the programs and documentation rests solely with the authors, not the NIJ. Rhodes Associates will be happy to receive comments and questions regarding the programs and documentation.

Rhodes Associates, 706 Cowper Street, 3rd floor, Palo Alto, California 94301 (415) 326 6246

Table of Contents

INTRODUCTION

This report provides documentation for three SAS-compatible statistical procedures: LOGIT, TOBIT and MVPROBIT. We embedded these routines in SAS to facilitate use of these statistical techniques by the research community and to allow users to take advantage of SAS's data management and manipulation capabilities. While these routines are not as efficient as similar self-contained programs which we developed;[1] additional computation-related costs are offset by ease of use and flexibility.[2]

---

[1]Several other individuals have made contributions to the self-contained versions of these programs. Paul A. David and, in particular, John G. Lewis contributed to a self-contained program for PROBIT analysis. Also, Takeshi Amemiya commented on our design of the Tobit routine and David Ross helped develop this documentation, as well as the program examples it contains.

[2]We wish to encourage use of these programs. We offer them at the cost of reproduction and shipping. This cost is $75 and should be sent to Rhodes Associates. We will prepare a tape containing the three programs and mail it along with a copy of this report. You should specify the type of tape you want and the system upon which the programs will be installed.

Instructions for Using SAS LOGIT Program

LOGIT is a program that produces maximum likelihood estimates of parameters of a Multinomial Logit model.[3] This statistical technique is appropriate for modelling a categorical dependent variable (the outcome of a trial) as a function of a linear combination of other variables.[4] Unlike MVPROBIT, which allows only two possible outcomes, LOGIT allows as many as eight possible outcomes.[5]

Setting Up PROC LOGIT

The LOGIT program is not a standard SAS procedure. In order to use PROC LOGIT, therefore, a special DD statement, which accesses the LOGIT program, must be included in the JCL:

```
// JOB
// EXEC SAS
//STEPLIB DD DSN=data.set.name,DISP=SHR
  :
  SAS statements
```

---

[3]An elementary discussion of the use of Logit analyses can be found in Chapter 10 of R. Pindych and D. Rubinfeld, Econometric Models and Economic Forecasts, Second Edition, McGraw Hill, 1981. Also see Takeshi Amemiya, "Qualitative Response Models: A Survey," Journal of Economic Literature, 1981, Vol. 19, 1483-1536.

[4]This statistical model is also appropriate for problems in which cardinal or ordinal dependent variables are reduced to categories such as "greater than zero" and "less than or equal to zero."

[5]Assessing the effect a particular independent variable has on the system of probabilities estimated in Logit models when there are three or more outcomes requires care. In addition to the nonlinear nature of the relationship between the probabilities and variables, one must treat the problem as a system recognizing that the probabilities of the various outcomes add to one.

This DD statement must immediately follow the // EXEC SAS statement.[6]

How LOGIT Works

LOGIT uses an iterative procedure to find maximum likelihood estimates of the parameters of a Multinomial Logit model. The procedure assigns the value zero to each of the parameters as initial values unless the user specifies other values with the INITIAL option (see below). A quasi Newton method, which uses only first derivatives, generates fairly accurate estimates of the parameters of the Multinomial Logit model which maximize the likelihood function. The program then switches to Newton's method to refine the solution. The estimated covariance matrix of the maximum likelihood estimators of the parameters is the inverse of the matrix of second derivatives of the negative of the log of the likelihood function evaluated at or near the final estimates.

One of the outcomes must be selected as the normalization outcome. The user may specify which outcome is selected by using the NORMOUTCOME parameter described below or allow the program to make the selection. The rule governing the selection by the program is given in the NORMOUTCOME section. The results, in terms of the probabilities estimated for the various outcomes for a particular set of independent variables, are the same, regardless of which outcome is selected for normalization.

---

[6]At Stanford's computing center data.set.name is WYL.XA.U76.DLSASLIB . If you have transferred this program from tape to your facility, the operating system (OS) data set name that the program is stored under at your facility must be used.

Invoking LOGIT from SAS

The syntax of the PROC LOGIT statement is:

```
PROC LOGIT options;
    VAR variable list;
    OUTCOME s;
```

Most applications do not require the special features that are available in LOGIT. In its simplest form, LOGIT is invoked as follows:

```
PROC LOGIT;
    VAR INCOME AGE SEX;
    OUTCOME BUYCAR;
```

The type of all independent variables must be numeric. If the VAR statement is omitted, LOGIT will use all numeric variables in the input data set (except the OUTCOME variable). This is the standard SAS default.

The OUTCOME variable may be of any type; its value can be either a number or a string of up to eight characters.

Using Special Features

LOGIT has two types of special features. The simplest features can be specified as options on the PROC LOGIT statement. The other more complicated features must be used in conjuction with a PARMCARDS4 statement.

1. The PROC LOGIT statement

The following options may appear in the PROC LOGIT statement:

(a) NOINT

This option eliminates all intercepts from the model.

(b) EXACTCOV

This option will cause the covariance matrix to be computed using estimates of the independent variable coefficients generated at the final iteration rather than those from the preceeding iteration. This requires additional computing time and ordinarily is not necessary, since values in the last two iterations are quite similar, but may be desirable if the covariance matrix is to be used for further computations.

(c) NORMOUTCOME=s

Because the probabilities must add to 1, the probabilities of any m - 1 of the m outcomes determine the probability of the remaining outcome. Consequently, no coefficients are generated for one of the outcomes; this outcome is referred to as the normalization outcome. The NORMOUTCOME parameter allows the user to select the normalization outcome s. (If s is a character string, it must be enclosed in single quotes.) If this parameter is omitted, LOGIT selects the normalization outcome as follows: If there is an OUTCOME with a numeric value of zero, that becomes the normalization state. Otherwise, the alphabetically last (or numerically greatest) outcome is selected.

(d) TOL=t

The value of t specifies the convergence criterion; the procedure will consider that it has found

sufficiently accurate estimates of the parameters of the model when the L_2 (euclidean) norm of the gradient is less than t. Ordinarily, the procedure selects a reasonable value of t that is based upon the number of parameters and the scale of the data. The value t may be stated in decimal form (e.g. .00001) or scientific notation of the form n.Dexp (e.g. 1.D-5).

(e) PRINT=n

This parameter controls the frequency of the printing of intermediate results and is primarily useful for diagnostic purposes. If n > 0, the initial estimate and each nth estimate thereafter are printed out. If n=0, or the PRINT statement is not used, only the results for the final quasi-Newton and each Newton iteration are printed out. The final estimates are printed regardless of the value of n.

(f) DATA=SASdsname

This parameter specifies the name of the input SAS data set. If omitted, the last SAS data set created is the default. (This is the same as for standard SAS procedures.)

Example 1: A Simple LOGIT Job

This SAS job produces the output shown in Illustration 1:

```
//LOGITEST JOB
// EXEC SAS
//STEPLIB DD DSN=WYL.XA.U76.DLSASLIB,DISP=SHR
//SAS.SYSIN DD *

DATA;
  INPUT OWNHOME $ INCOME NUMCHILD;
  CARDS;
     Y      25      2
     Y      20      1
     N      35      2
     Y      55      1
     Y      40      0
     N      30      2
     N      80      2
     Y      35      4
     N      75      2
     N      15      1

PROC LOGIT NORMOUTCOME='N';
  VAR INCOME NUMCHILD;
  OUTCOME OWNHOME;
TITLE PROBABILITY OF HOME OWNERSHIP AS A FUNCTION OF;
TITLE2 INCOME AND NUMBER OF CHILDREN;
```


Illustration 1

PROBABILITY OF HOME OWNERSHIP AS A FUNCTION OF
INCOME AND NUMBER OF CHILDREN

M U L T I N O M I A L L O G I T A N A L Y S I S OF OWNHOME

10 OF 10 OBSERVATIONS UTILIZED (OTHERS HAD MISSING OR INVALID DATA)

NUMBER OF POSSIBLE OUTCOMES IS 2
OUTCOMES (VALUES OF OWNHOME) ARE: Y N

		MEAN	VARIANCE	SCALE FACTOR
1	INCOME	41.0000000	444.0000000	72.9931504
2	NUMCHILD	1.7000000	1.0100000	3.4813790

CONVERGENCE CRITERIA SATISFIED

FINAL ITERATION (NO. 6); FUNCTION VALUE IS 6.487379 |GRADIENT| = 1.551D-02

		COEFFICIENTS	GRADIENT
1	(Y ,INCOME)	-0.028688	-1.053985D-02
2	(Y ,NUMCHILD)	-0.140259	-3.033125D-03
3	(Y ,INTERCPT)	1.397386	-1.096352D-02

HESSIAN (SCALED INVERSE COVARIANCE MATRIX) EVALUATED AT |GRADIENT| = 1.551D-02

NEWTON STEP NO. 1; FUNCTION VALUE IS 6.487272 |GRADIENT| = 5.231D-05

		COEFFICIENTS	GRADIENT
1	(Y ,INCOME)	-0.028299	3.540236D-05
2	(Y ,NUMCHILD)	-0.144108	1.951006D-05
3	(Y ,INTERCPT)	1.393302	3.319913D-05

R E S U L T S O F L O G I T A N A L Y S I S -2*LN(L) = 12.9745

NORMALIZATION OUTCOME IS N -- NO COEFFICIENTS ARE GENERATED FOR THIS VALUE OF OWNHOME

COEFFICIENTS FOR OWNHOME = Y

			ESTIMATED COEF	EST. STD. ERR.	EST. T-VALUE
1	INCOME	(VARIABLE 1)	-0.02830	0.03357	-0.84299
2	NUMCHILD	(VARIABLE 2)	-0.14411	0.63971	-0.22527
3	INTERCPT	(VARIABLE 3)	1.39330	1.78653	0.77989

COVARIANCE OF COEFFICIENTS (INDEXED AS ABOVE)

	1	2	3
1	1.1269D-03	-1.5658D-03	-4.2072D-02
2	-1.5658D-03	4.0922D-01	-6.3537D-01
3	-4.2072D-02	-6.3537D-01	3.1917D+00

2. The PARMCARDS4 statement

LOGIT recognizes certain statements that look like SAS statements but which SAS itself cannot process. These statements must be preceded by the SAS statement PARMCARDS4. (This is similar to the use of PROC BMDP.) The special LOGIT statements must appear in the order below (but all need not be present). They must be followed by a separate SAS statement consisting of four semicolons (;;;;).

(a) EXCLUDE s X_p ... X_q;
where s is an outcome other than the normalization outcome and X_p ... X_q is a list of one or more of the independent variables (the order of the X's is immaterial). The effect of this statement is to eliminate the terms corresponding to X_p ... X_q from the likelihood function for outcome s and is equivalent to setting the coefficients of these variables to zero. This feature may be utilized when X_p ... X_q are known not to affect the probability that the outcome is s. A different EXCLUDE statement is required for each outcome for which variable exclusion is to occur, but the order is not significant.

(b) INITIAL; or INITIAL (format);
This statement is utilized when the user wants to specify starting values for the coefficients. The string (format), if present, should be a valid FORTRAN format of at most 64 characters containing no semicolons.

If the format is not specified, a default format of (5D16.8) (scientific notation, 16 characters per coefficient) is used. The coefficients must be in card format starting right after the START statement. The coefficients must appear in the order
$\beta_{11}, \beta_{12} \cdots \beta_{1n}, \beta_{10}, \beta_{21} \cdots \beta_{2n}, \beta_{20}, \ldots, \beta_{m-1,1} \cdots \beta_{m-1,0}$
where the β's denote the coefficients, the first subscript the outcome and the second subscript the independent variable. The zeroth independent variable is the intercept. Any coefficients that are not utilized, either because they are covered by an EXCLUDE statement or because NOINT was specified, must be omitted.

(c) OUTPUT; or OUTPUT (format);
This statement, which is analogous to the START statement, is utilized to save the estimated coefficients and the covariance matrix for future use. The coefficients followed by successive rows of the covariance matrix, are written to FORTRAN file 20 (FT20F001). The user must include a JCL DD statement to define the attributes of file FT20F001, and (format) must be suitable for writing to that file.

Remember to include a separate ;;;; statement at the end of all the special PARMCARDS4 statements.

Example 2: A More Complex LOGIT Job

This SAS job produces the output in Illustration 2. The data read by the job, contained in WYL.XA.U76.EDUC, is listed for reference:

```
//LOGITST2  JOB
// EXEC SAS
//STEPLIB DD DSN=WYL.XA.U76.DLSASLIB,DISP=SHR
//SURDATA3 DD DSN=WYL.XA.U76.EDUC,DISP=SHR

DATA SURVEY;
  INFILE SURDATA3;
  INPUT COLDEG $ SATSCORE INCOME HSGRADES;

PROC LOGIT NORMOUTCOME='NONE' NOINT
     DATA=SURVEY;
  VAR SATSCORE INCOME HSGRADES;
  OUTCOME COLDEG;
PARMCARDS4;
  EXCLUDE GRAD HSGRADES;
;;;;
```

WYL.XA.U76.EDUC

```
1.   POSTGRAD  1210  22  3.4
2.   GRAD       980  35  3.6
3.   NONE      1020  25  3.2
4.   NONE       820  30  2.9
5.   GRAD      1100  44  3.1
6.   NONE      1140  30  2.8
7.   NONE       720  35  3.2
8.   POSTGRAD  1080  28  3.9
9.   NONE      1100  40  2.4
10.  GRAD      1340  32  3.1
11.  GRAD      1270  18  3.7
12.  GRAD       990  24  3.8
13.  POSTGRAD  1010  38  3.5
14.  NONE       950  42  3.0
15.  GRAD      1280  38  3.7
16.  NONE       920  16  2.6
17.  GRAD      1410  26  3.1
18.  POSTGRAD  1420  22  3.8
19.  GRAD      1150  48  3.9
20.  POSTGRAD   970  36  4.0
```

Note: In this program example, DATA=SURVEY is included as an option in the PROC LOGIT statement to demonstrate the DATA= option. However, since SURVEY is the most recently created SAS data set, it would have been used by default.

Illustration 2

STATISTICAL ANALYSIS SYSTEM

MULTINOMIAL LOGIT ANALYSIS OF COLDEG

LOGIT CONTROL CARD: EXCLUDE GRAD HSGRADES;

20 OF 20 OBSERVATIONS UTILIZED (OTHERS HAD MISSING OR INVALID DATA)

NUMBER OF POSSIBLE OUTCOMES IS 3
OUTCOMES (VALUES OF COLDEG) ARE: GRAD POSTGRAD NONE

		MEAN	VARIANCE	SCALE FACTOR
1	SATSCORE	1094.0000000	33444.0000000	633.5045383
2	INCOME	31.4500000	74.4475000	29.8892957
3	HSGRADES	3.3350000	0.2042750	1.5656628

CONVERGENCE CRITERIA SATISFIED

FINAL ITERATION (NO. 6); FUNCTION VALUE IS 19.30999 |GRADIENT| = 1.278D-02

		COEFFICIENTS	GRADIENT
1	(GRAD ,SATSCORE)	0.001609	1.797989D-03
2	(GRAD ,INCOME)	-0.044093	-6.970026D-03
3	(POSTGRAD,SATSCORE)	-0.002062	6.038188D-03
4	(POSTGRAD,INCOME)	-0.131802	6.786470D-03
5	(POSTGRAD,HSGRADES)	1.802093	5.377658D-03

HESSIAN (SCALED INVERSE COVARIANCE MATRIX) EVALUATED AT |GRADIENT| = 1.278D-02

NEWTON STEP NO. 1; FUNCTION VALUE IS 19.30988 |GRADIENT| = 7.817D-05

		COEFFICIENTS	GRADIENT
1	(GRAD ,SATSCORE)	0.001593	-3.069204D-05
2	(GRAD ,INCOME)	-0.043585	-1.577295D-05
3	(POSTGRAD,SATSCORE)	-0.002090	4.248883D-05
4	(POSTGRAD,INCOME)	-0.132215	3.554569D-05
5	(POSTGRAD,HSGRADES)	1.813591	4.302418D-05

STATISTICAL ANALYSIS SYSTEM

RESULTS OF LOGIT ANALYSIS -2*LN(L) = 38.6198

NORMALIZATION OUTCOME IS NONE -- NO COEFFICIENTS ARE GENERATED FOR THIS VALUE OF COLDEG

COEFFICIENTS FOR COLDEG = GRAD

		ESTIMATED COEF	EST. STD. ERR.	EST. T-VALUE
1	SATSCORE (VARIABLE 1)	0.00159	0.00161	0.98827
2	INCOME (VARIABLE 2)	-0.04358	0.05145	-0.84713

COEFFICIENTS FOR COLDEG = POSTGRAD

		ESTIMATED COEF	EST. STD. ERR.	EST. T-VALUE
3	SATSCORE (VARIABLE 1)	-0.00209	0.00321	-0.65130
4	INCOME (VARIABLE 2)	-0.13222	0.07817	-1.69140
5	HSGRADES (VARIABLE 3)	1.81359	1.22260	1.48339

COVARIANCE OF COEFFICIENTS (INDEXED AS ABOVE)

	1	2	3	4	5
1	2.5983D-06	-7.8959D-05	1.8386D-06	-5.2075D-05	-1.5432D-05
2	-7.8959D-05	2.6471D-03	-5.3643D-05	1.6515D-03	3.9282D-04
3	1.8386D-06	-5.3643D-05	1.0301D-05	1.6063D-05	-3.2199D-03
4	-5.2075D-05	1.6515D-03	1.6063D-05	6.1104D-03	-5.7516D-02
5	-1.5432D-05	3.9282D-04	-3.2199D-03	-5.7516D-02	1.4947D+00

Instruction for Using TOBIT and MVPROBIT Programs

Both Tobit and Probit[7] models rest on an underlying random variable which is normally distributed. The TOBIT program described below is appropriate for a normally distributed dependent variable whose mean is a linear combination of independent variables but where the observations have been truncated from below at zero. In other words, the dependent variable is not permitted to assume negative values; if its value would be negative, the value zero is observed instead.

The MVPROBIT program is appropriate for a dependent variable which can take on only 2 values, such as 0 and 1. The probability that the dependent variable takes on a particular value is assumed to be a standard normal cumulative distribution function (mean 0 and variance 1) whose argument is a linear combination of the explanatory variables.

The TOBIT and MVPROBIT programs calculate maximum likelihood estimators of the parameters for Tobit and Probit analysis along with the corresponding asymptotic covariance matrix.

Setting up TOBIT or MVPROBIT

Neither TOBIT or MVPROBIT is a standard SAS procedure. In order to use these procedures, therefore, a special DD statement, which accesses the TOBIT (MVPROBIT) program, must be included in the JCL:

---

[7] For a discussion of Tobit models, see Takeshi Amemiya, "Regression Analysis when the Dependent Variable Is Truncated Normal," Econometrica, 1973, Vol. 41, 997-1016, and for Probit models, see Takeshi Amemiya, "Qualitative Response Models: A Survey," Journal of Economic Literature, 1981, Vol. 19, 1483-1536.

```
// JOB
// EXEC SAS
//STEPLIB DD DSN=data.set.name,DISP=SHR
:  SAS statements
```

This DD statement must immediately follow the // EXEC SAS statement.[8]

How TOBIT Works

TOBIT uses an iterative procedure to find maximum likelihood estimates of the parameters of the model. The user may select the initial estimates by using the INITIAL option (see below). If the initial estimates are not specified, the program starts with the intercept set to the mean of Y, the outcome variable, the variance equal to the sample variance of Y, and the remaining parameters set to zero. The program uses a quasi-Newton method to obtain a fairly accurate estimate of the values of the parameters of the model which maximize the likelihood function.[9] The program then switches to Newton's method to refine the solution. The estimated covariance matrix is the inverse of the matrix of second derivatives of the negative of the log of the likelihood function evaluated at or near the final estimates.

---

[8] At Stanford's computing center data.set.name is WYL.XA.U76.DLSASLIB . If you have transferred this program from tape to your facility, the OS data set name that the program is stored under at your facility must be used.

[9] Internally, TOBIT scales the input data and coefficients in order to make the various parameters associated with each variable approximately equal in weight for purposes of assessing convergence. This improves numerical stability and makes convergence more rapid.

How to Invoke TOBIT from SAS

The syntax of the PROC TOBIT statement is:

```
PROC TOBIT options;
     VAR variable list;
     OUTCOME  Y;
```

The type of the dependent variable and the independent variables must be numeric.[10] TOBIT can be used for problems where the truncation point is some number other than zero by pre-processing the data. For example, if the truncation point is τ, create a new variable $W = Y - \tau$, and use TOBIT with W as the dependent variable. Then add τ to intercept. Problems where the truncation is from above--where we observe $Y = \min(z,\theta)$--can be transformed by multiplication by -1 and subtraction of θ from the estimated intercept.

How MVPROBIT Works

MVPROBIT uses the same computational methods as TOBIT. Unless the INITIAL option is used, the first iteration begins with all the parameters set to zero.

How to Invoke MVPROBIT from SAS

The syntax for the PROC MVPROBIT statement is:

```
PROC MVPROBIT options;
     VAR variable list;
     FREQ  r  s;
```

Variables named in variable list are the independent variables while r and s are the names of the variables giving the frequencies of

---

[10] If the VAR statement is omitted, all numeric variables in the input data set, except the OUTCOME variable, are utilized. This is the standard SAS default.

outcomes for each observation. Each observation may represent several trials, so r gives the number of times that a particular outcome was observed and s the number of times that the second outcome occurred. If an observation represents one trial, then either r = 1 and s = 0 or vice versa. The type of all variables, including frequencies, must be numeric.[11]

Missing Data

TOBIT ignores any observation for which the value of any independent variable is missing or for which the value of the dependent variable is missing or negative.

MVPROBIT ignores any observation for which any data is missing or either of the frequencies is negative.

Options for TOBIT and MVPROBIT

The following options or parameters may appear in the PROC TOBIT or PROC MVPROBIT statement:

(a) NOINT

This option will cause the intercept term, β_0, to be omitted.

(b) EXACTCOV

This option will cause the covariance matrix to be computed at the final estimate of the parameters rather than merely at a close estimate. This requires additional

[11] If the VAR statement is omitted, all numeric variables in the input data set, except the FREQ variables, are utilized. This is the standard SAS default.

computing time and ordinarily is not necessary, but may be desirable if the covariance matrix is to be used for further computations.

(c) INITIAL

If this option is specified, MVPROBIT (TOBIT) will expect the user to supply initial estimates of the coefficients, $\underset{\sim}{\beta}$ ($\underset{\sim}{\beta}$, σ for TOBIT, where σ is the standard deviation of the underlying normal distribution). Estimates are supplied through the use of a PARMCARD statement followed by the initial values. The values are input in the order $\beta_1,\ldots,\beta_n$, β_0, ($\beta_1,\ldots,\beta_n$, β_0, σ for TOBIT). The format utilized is 5D15.0 (floating point, one every 15 columns), unless the INFORMAT parameter is specified.

(d) OUTPUT

This option causes MVPROBIT (TOBIT) to write the final parameter estimates and covariance matrix to disk, tape or cards on FORTRAN file 20 (FT20F001). The parameters β_1, $\beta_2,\ldots,\beta_n$, β_0 (and σ if TOBIT is used) are written first; then the covariance matrix is written one row (or column, since the matrix is symmetric) at a time. The format used is (1P5D15.7), which provides 5 values per line in scientific notation with 7 decimal places, unless the OUTFORMAT parameter is specified. Note that the user must define the FT20F001 file using JCL. The JCL statement should appear just after the //STEPLIB DD statement. If the default format is used, record length must be at least 75 characters.

(e) TOL=t

The value of t specifies the convergence criterion; the procedure will consider that it has found sufficiently accurate estimates of the parameters of the model when the L_2 (euclidean) norm of the gradient is less than t. Ordinarily, the procedure selects a reasonable value of t that is based upon the number of parameters and the scale of the data. The value t may be stated in decimal form (e.g. .00001) or scientific notation of the form n.Dexp (e.g. 1.D-5).

(f) PRINT=n

This parameter controls the frequency of the printing of intermediate results and is primarily useful for diagnostic purposes. If n > 0, the initial estimate and each nth estimate thereafter are printed out along with the final estimates. If n = 0, or PRINT is not used, only the results for the final quasi-Newton and each Newton iteration are printed out.

(g) INFORMAT=s

This parameter defines the input format when the INITIAL option is selected. The parameter s must be a string of at most 64 characters that comprise a valid FORTRAN format for reading cards, e.g., (8F10.3).

(h) OUTFORMAT=s

This parameter defines the output format when the OUTPUT option is selected. The parameter must be a string of

at most 64 characters that comprise a valid FORTRAN format for writing to file 20 (FT20F001). That file's characteristics must be defined by the user in a JCL statement. The OUTFORMAT parameter need not be specified if the default format is acceptable.

(i) DATA=SASdsname

This parameter specifies the name of the input data set. If omitted, the last data set created is the default. (This is the same as for other SAS procedures.)

(j) STEPMAX=d

This parameter limits the total variation at each iteration in the parameter estimates for the model. Ordinarily, the value selected by the program ($5+5\sqrt{n}$, where n is the number of parameters in the model), is suitable; however, in rare instances, it may be too large, resulting in computational problems such as underflows. If this occurs, the user should specify a smaller value of STEPMAX. This option should not be used except for ill-behaved problems since it will probably slow down convergence. The value d may be stated in decimal form (e.g. 10.0) or scientific notation of the form n.Dexp (e.g. 1.D+1).

Options for TOBIT only -- Forecasting

(k) TESTDATA=SASdsname

This parameter specifies the name of the data set containing the test observations for which the expected values of the outcome variable and the variance of prediction and mean squared error of forecast of the estimate are to be computed. The variance of prediction is the estimate of variability for the estimate of the expected value of the outcome variable, and is appropriate for constructing confidence intervals for the expected value of the outcome variable. The mean squared error of forecast is the estimate of variability appropriate for normalizing comparisons of forecast values to corresponding realizations of the outcome variable. The TESTDATA data set must contain all of the independent variables in the VAR list, but not necessarily in the same order, and may contain the outcome variable and other variables which are not used in the analysis. If no TESTDATA data set is specified, no predicted values are computed. In the output, the value of -1.0 for the observed value of the dependent variable signifies that it was missing.

(l) PRINTTESTDATA

This option causes the values of the independent variables to be printed along with the forecasts.

(m) TESTOUTFILE=n or TOF=n

This parameter causes the forecast results to be written to FORTRAN file |n|. The items written are observation number, observed Y, predicted Y, variance of prediction, and mean squared error of forecast. Note that the observation number is the absolute position in the TESTDATA data set; observations which are unusable are counted. The format is (I6, 4D16.8). If n < 0, then the independent variables are written on a new record immediately following each forecast record. The format is (6X, 4D16.8). The user must include a DD statement for the output file.

(n) NOPRINTFORECAST or NPF

This option causes the printed forecast output to be suppressed if a TESTOUTFILE is also specified.

Example 3: PROC MVPROBIT and PROC TOBIT

```
//TOBTST1 JOB
// EXEC SAS
//STEPLIB DD DSN=WYL.XA.U76.DLSASLIB,DISP=SHR
//SAS.SYSIN DD *

DATA CRIME;
  INPUT AGE INCOME OWNHOME BURGED80 NOTBURGD AMOUNT;
  CARDS;
22   20     0    0   1      0
47   32     0    0   1      0
39   64     1    0   1      0
28   52     0    1   0     60
62   38     1    0   1      0
31    9     0    0   1      0
34   22     0    0   1      0
35   28     1    0   1      0
41   17     0    0   1      0
26   23     0    0   1      0
21   18     0    0   1      0
```

```
26   37    1   1   0    50
54   15    0   0   1     0
43   35    1   1   0   575
60   30    0   0   1     0
58   75    1   0   1     0
39   48    1   0   1     0
27   15    0   0   1     0
22   60    1   1   0   115
36   11    1   0   1     0
22   20    0   0   1     0
47   32    0   0   1     0
39   64    1   0   1     0
31   32    0   1   0    45
62   38    1   0   1     0
31   17    0   0   1     0
34   22    0   0   1     0
33   28    1   0   1     0
41   17    0   0   1     0

PROC MVPROBIT;
  VAR AGE INCOME OWNHOME;
  FREQ BURGED80 NOTBURGD;
  TITLE PROBABILITY OF BEING BURGLARIZED AS A FUNCTION OF;
  TITLE2 AGE, INCOME AND HOME OWNERSHIP;
  TITLE3 ILLUSTRATION OF PROC MVPROBIT;

DATA RESERVED;
  INPUT AGE INCOME OWNHOME BURGED80 NOTBURGD AMOUNT;
  CARDS;
26   23    .   0   1     0
21   18    0   0   1     0
26   37    1   1   0    30
34   13    0   0   1     0
43   33    1   1   0   617
.    30    0   0   1     0
38   73    1   0   1     .
39   48    1   0   1     0
27   13    0   0   1     0
31   60    1   1   0   113
36   11    1   0   1     0

PROC TOBIT DATA=CRIME TESTDATA=RESERVED PRINTTESTDATA TOL=1.D-4 ;
  VAR AGE INCOME OWNHOME;
  OUTCOME AMOUNT;
  TITLE PREDICTED LOSS FROM BURGLARY AS A FUNCTION OF;
  TITLE2 AGE, INCOME AND HOME OWNERSHIP;
  TITLE3 ILLUSTRATION OF PROC TOBIT;
```


PROBABILITY OF BEING BURGLARIZED AS A FUNCTION OF
AGE, INCOME AND HOME OWNERSHIP
ILLUSTRATION OF PROC MVPROBIT

M U L T I V A R I A T E P R O B I T A N A L Y S I S OF BURGED80 AGAINST AGE INCOME OWNHOME INTERCPT

29 OF 29 OBSERVATIONS UTILIZED (OTHERS HAD MISSING OR INVALID DATA)

		MEAN	VARIANCE	SCALE FACTOR
1	AGE	37.6206897	147.5457788	24.2936847
2	INCOME	31.6896552	291.8692033	34.1683598
3	OWNHOME	0.4137931	0.2425634	0.9850246

CONVERGENCE CRITERIA SATISFIED

FINAL ITERATION (NO. 8); FUNCTION VALUE IS 9.461441 |GRADIENT| = 5.930D-03

	COEFFICIENTS	GRADIENT
1 (AGE)	-0.077640	-4.504860D-03
2 (INCOME)	0.035921	-1.596771D-03
3 (OWNHOME)	0.394176	-8.975401D-04
4 (INTERCPT)	0.276199	-3.393397D-03

HESSIAN (SCALED INVERSE COVARIANCE MATRIX) EVALUATED AT |GRADIENT| = 5.930D-03

NEWTON STEP NO. 1; FUNCTION VALUE IS 9.461439 |GRADIENT| = 2.460D-06

	COEFFICIENTS	GRADIENT
1 (AGE)	-0.077615	1.933753D-06
2 (INCOME)	0.035892	7.018190D-07
3 (OWNHOME)	0.394174	3.097021D-07
4 (INTERCPT)	0.276816	1.312641D-06

F I N A L R E S U L T S O F P R O B I T A N A L Y S I S OF BURGED80

-2*LN(L) = 18.9229 NORM OF GRADIENT = 2.460D-06

		ESTIMATED COEF	EST. STD. ERR.	EST. T-VALUE
1	AGE	-0.07761	0.04459	-1.74082
2	INCOME	0.03589	0.02315	1.55063
3	OWNHOME	0.39417	0.84130	0.46853
4	INTERCPT	0.27682	1.33057	0.20804

COVARIANCE OF COEFFICIENTS

	AGE	INCOME	OWNHOME	INTERCPT
AGE	1.9878D-03	-2.7458D-04	-1.2616D-02	-4.8983D-02
INCOME	-2.7458D-04	5.3576D-04	-8.3156D-03	-6.3398D-03
OWNHOME	-1.2616D-02	-8.3156D-03	7.0778D-01	3.5726D-01
INTERCPT	-4.8983D-02	-6.3398D-03	3.5726D-01	1.7704D+00

PREDICTED LOSS FROM BURGLARY AS A FUNCTION OF
AGE, INCOME AND HOME OWNERSHIP
ILLUSTRATION OF PROC TOBIT

M U L T I V A R I A T E T O B I T A N A L Y S I S OF AMOUNT AGAINST AGE INCOME OWNHOME INTERCPT

29 OF 29 OBSERVATIONS UTILIZED (OTHERS HAD MISSING OR INVALID DATA)

		MEAN	VARIANCE	SCALE FACTOR
1	AGE	37.6206897	147.5457788	24.2936847
2	INCOME	31.6896552	291.8692033	34.1683598
3	OWNHOME	0.4137931	0.2425684	0.9850246
	AMOUNT	29.1379310	11288.0499405	106.2452349

CONVERGENCE CRITERIA SATISFIED

FINAL ITERATION (NO. 4); FUNCTION VALUE IS 37.55815 |GRADIENT| = 2.585D-03

		COEFFICIENTS	GRADIENT
1	AGE	-11.833081	4.09159D-04
2	INCOME	3.948644	-2.06005D-03
3	OWNHOME	170.675498	-1.39912D-03
4	INTERCPT	-111.855203	2.16207D-04
5	SIGMA	320.092953	5.16956D-04

HESSIAN (SCALED INVERSE COVARIANCE MATRIX) EVALUATED AT |GRADIENT| = 2.585D-03

NEWTON STEP NO. 1; FUNCTION VALUE IS 37.45279 |GRADIENT| = 4.301D-04

		COEFFICIENTS	GRADIENT
1	AGE	-13.279466	2.84015D-04
2	INCOME	6.007582	2.41822D-05
3	OWNHOME	193.631619	-2.72924D-05
4	INTERCPT	-162.027195	1.74290D-04
5	SIGMA	329.278479	-2.69527D-04

HESSIAN (SCALED INVERSE COVARIANCE MATRIX) EVALUATED AT |GRADIENT| = 4.301D-04

NEWTON STEP NO. 2; FUNCTION VALUE IS 37.44881 |GRADIENT| = 3.042D-05

		COEFFICIENTS	GRADIENT
1	AGE	-13.867213	2.05928D-05
2	INCOME	6.354846	3.54444D-06
3	OWNHOME	199.033305	-2.85664D-07
4	INTERCPT	-169.348862	1.24051D-05
5	SIGMA	338.311908	-1.82940D-05

PREDICTED LOSS FROM BURGLARY AS A FUNCTION OF
AGE, INCOME AND HOME OWNERSHIP
ILLUSTRATION OF PROC TOBIT

F I N A L R E S U L T S O F T O B I T A N A L Y S I S OF AMOUNT

-2*LN(L) = 74.8976 NORM OF GRADIENT = 3.042D-05

		ESTIMATED COEF	EST. STD. ERR.	EST. T-VALUE
1	AGE	-13.86721	11.20484	-1.23761
2	INCOME	6.35485	6.82276	0.93142
3	OWNHOME	199.03330	229.86436	0.86587
4	INTERCPT	-169.34886	374.10233	-0.45268
5	SIGMA	338.31191	117.62554	2.87618

COVARIANCE OF COEFFICIENTS

	AGE	INCOME	OWNHOME	INTERCPT	SIGMA
AGE	1.2555D+02	-1.7057D+01	-7.1362D+02	-2.5653D+03	-6.0667D+02
INCOME	-1.7057D+01	4.6550D+01	-5.9466D+02	-1.1680D+03	2.9847D+02
OWNHOME	-7.1362D+02	-5.9466D+02	5.2838D+04	1.3716D+04	3.5488D+03
INTERCPT	-2.5653D+03	-1.1680D+03	1.3716D+04	1.3995D+05	-7.2846D+03
SIGMA	-6.0667D+02	2.9847D+02	3.5488D+03	-7.2846D+03	1.3836D+04

OBS #	OBSD AMOUNT	PRED AMOUNT	VAR OF PRED	MSE FORECAST	AGE	INCOME	OWNHOME
2	.0	26.961326	7462.6269	8293.4882	21.0000	18.0000	.0
3	30.000000	92.467990	27074.330	31878.841	26.0000	37.0000	1.00000
4	.0	6.9814251	1716.1125	1796.4611	34.0000	13.0000	.0
5	617.00000	25.359434	6985.5617	7573.8766	43.0000	33.0000	1.00000
7	-1.0000000	118.93986	34616.255	45239.262	38.0000	73.0000	1.00000
8	.0	56.213901	16293.275	17639.973	39.0000	48.0000	1.00000
9	.0	13.448700	3509.2084	3750.6231	27.0000	13.0000	.0
10	113.00000	125.72349	36493.177	43812.739	31.0000	60.0000	1.00000
11	.0	19.721575	5321.3454	6228.4171	36.0000	11.0000	1.00000

APPENDIX

Descriptions of the LOGIT, TOBIT and MVPROBIT Models

LOGIT

Let $X_1, X_2, \ldots, X_n$ be independent variables and let R be a discrete dependent variable that can take on values $s_1, s_2, \ldots, s_m$ (referred to as outcomes). Using independent observations on $X_1, X_2, \ldots, X_n$ and R PROC LOGIT computes the maximum likelihood estimates for the Multinomial Logit model of the outcomes where the probability of outcomes depend on the coefficients, β_{ij}, and are given by

$$P\{R = s_i\} = \alpha_i / \sum_{j=1}^{m} \alpha_j \quad ,$$

where $\alpha_i = \exp\left(-\beta_{i0} - \sum_{j=1}^{n} \beta_{ij} X_j\right)$, $(i \neq m)$; $\alpha_m = 1$. The index m refers to the outcome selected for normalization. This selection can be made arbitrarily. Estimates of the standard errors and the asymptotic covariance matrix of the $\{\beta_{ij}\}$ are also computed.

LOGIT uses a two-stage computational algorithm to maximize the likelihood function. The procedure is iterative and starts with $\beta_{ij} = 0$ for all i and j as the initial value unless the INITIAL option is used.

TOBIT

Let $X_1, X_2, \ldots, X_n$ be independent variables and Z be a variable that depends upon the X's. Furthermore, assume that wherever $Z < 0$, the observed value of the dependent variable is 0. Thus, the observed dependent variable is $Y = \max(Z, 0)$.

Using the data for the X's and the Y's, TOBIT computes maximum likelihood estimators for the parameters, $\beta_1, \beta_2, \ldots, \beta_n$, β_0 and σ, when the density of Z is given by

$$Z \sim N(\beta_0 + \sum_{i=1}^{n} \beta_i x_i, \sigma^2) \quad ,$$

where $N(\mu, \sigma^2)$ is the univariate normal density with mean and variance σ^2. TOBIT also computes the asymptotic covariance matrix for these estimators. In addition, PROC TOBIT has an option which predicts the mean and variance of prediction and mean squared error of forecast of Y for a set of observation on the X's specified by the user.[12]

MVPROBIT

Let $X_1, X_2, .., X_n$ be independent variables and let S be a dichotomous variable that represents the observed outcome of a trial (for example, 1 = success, 0 = failure). Using independent observations on $X_1, X_2, \ldots, X_n$ and S, PROC MVPROBIT computes maximum likelihood estimators for the binomial model of the outcome where the probability of the outcome depends on β_i's and is given by

$$\Pr\{s = 1\} = F(\beta_0 + \sum_{i=1}^{n} \beta_i X_i) \quad ,$$

[12] Formulas are available in Amemiya (1973) op.cit. and section 7.1 of Continuous Univariate Distribution 1, N.L. Johnson and S. Kotz, 1970, Houghton Mifflin Company, Boston.

where F is the univariate normal cumulative distribution function with mean 0 and variance 1. (Note that the variance can be assumed to be 1 without loss of generality since the variance is incorporated as a scale factor into the β_i's.)

END