LOGITSE: A SAS® Macro for Logistic Regression Modeling in Complex Surveys

Jesse A. Canchola, University of California, San Francisco
Brian D. Marx, Louisiana State University, Baton Rouge
Joseph A. Catania, University of California, San Francisco

ABSTRACT

Traditional formulae for standard errors and subsequent statistical significance tests implemented in various popular statistical packages are based on the premise that the data are a simple random sample (SRS) of observations from a super-population. Equivalently, the observations are assumed to be independent and identically distributed (IID). For complex analytic surveys, these assumptions are almost always invalid leading to potentially incorrect inferences due to the failure to adjust relevant standard errors of the parameters. Here, we concentrate on binary responses where a logistic regression analysis would be meaningful and introduce a SAS® macro, LOGITSE, that takes the cluster-correlated nature of the complex survey design into account, thus providing for correct inference.

INTRODUCTION

This presentation assumes the SAS system available to the user is release 6.x or above. In this section we give an overview of LOGITSE and explain the options available and then give various useful SAS program examples to demonstrate the macro.

LOGITSE is a SAS® macro that fits logistic regression models sample survey data from complex multilevel designs. For more detailed statistical background see, for example, Skinner, Holt, and Smith (1989). Presently, this SAS/IML® macro uses the Taylor linearization method used to produce appropriate estimators for the standard errors of logistic regression coefficients. The program uses an iterative reweighted procedure to estimate regression coefficients. Final output from LOGITSE includes:

0. A summary of the working data set to be used in the analysis. This includes variable formats.

1. The following basic summaries:
   a. dependent variable specified;
   b. independent variables specified;
   c. number of observations used in the analysis;
   d. total number of observations in the data set;
   e. sum of the rescaled weights (**LOGITSE** does this automatically);
   f. sum of non-rescaled weights;
   g. coefficient of variation (CV) of weights;
   h. denominator degrees of freedom;
   i. deviance (-2 times the log likelihood);
   j. sampling type (with or without replacement);

2. The following estimation information:
   a. logistic regression coefficients and their adjusted standard errors;
   b. adjusted Wald statistics for maximum likelihood (ML) parameter estimates and their corresponding test for significance;
   c. odds ratios (OR), as well as inverse ORs, and their associated 100(1-α)% confidence intervals;
   d. overall complex survey design effect (deff)[1];
   e. estimated parameter design effects;
   f. complex and simple (i.e., adjusted and pre-adjusted) asymptotic covariance, correlation, and information matrices;
   g. complex and simple (i.e., weighted and unweighted) model fit information, including likelihood ratios, Akaike's information criteria (AIC), score test statistic and corresponding test for significance;
   h. a goodness-of-fit (GOF) test discussed by Hosmer and Lemeshow (1989) with an adjustment for weighting (Hudes, 1994);
   i. ranks covariate patterns based on significance of Pearson chi-square statistics due to deletion.

Since LOGITSE is a SAS macro, it has to be used within a SAS program. However, before this macro can be used in a SAS program, the macro itself should be run once in a SAS session. Afterwards, in any SAS session, supply the macro name, %LOGITSE, within the SAS program along with the relevant input parameters. For example the following two SAS/Windows 6.12 statements will correctly access the macro stored in a file LOGITSE.SAS stored in the directory "c:\winsas\sasmacro":

```
OPTIONS MAUTOSOURCE SASAUTOS = 'C:\WINSAS\SASMACRO';
```

The following command, with appropriate parameters, can be used to invoke the macro. All parameters have been assigned default values, so that they can be omitted if default values are acceptable. (Defaults are shown within {}). Parameters may be given in any order. The options are explained in detail below:

```
%LOGITSE(DATA =  SAS dataset                    {_LAST_}
      DEPVAR = dependent variable,              {Y}
      IVARS =  independent variables,           {X}
      WEIGHT= sampling or poststratification weights,
                                                {no weighting}
      ID = id variable,                         {ID}
      NEST = nesting variable,                  {no nesting}
      ORALPHA = significance level              {0.05}
      SAMPTYPE = with or without replacement {'WR'}
      TOTCNT = grouping variable,               {no grouping}
      PARMTEST = linear combination coefficients);
                                  {no linear combinations}
```

The macro calls for several arguments:

[1] The deff is the design effect or the relative efficiency, the ratio of variance from complex survey design to the variance from simple random sampling (SRS).

1) **DATA**: Input dataset has to be a SAS dataset. If you do not specify the name, then the most recently created SAS dataset will be assumed. Dataset must contain an outcome variable and covariates. All information corresponding to an outcome should constitute a single record in the dataset. Records corresponding to a cluster should be placed together. An ID variable should identify observations corresponding to each cluster.

The Macro will AUTOMATICALLY DELETE observations with MISSING values associated with the DEPVAR or IVARS, and negative WEIGHT. The DEFAULT dataset is DATA=_LAST_ or the last data set run in your SAS session.

2) **DEPVAR**: the BINARY response (MUST be numeric Zero='Failure'/One='Success').

3) **IVARS**: the independent variables (MUST be numeric). Recommend binary (zero/one) coding for categorical IVARS.

4) **WEIGHT**: weight associated with each observation, the Macro will AUTOMATICALLY DELETE observations with NEGATIVE weights and then RE-SCALE the weights. The DEFAULT Equal Weight for each observation.

5) **NEST**: the stratum identification variable. The DEFAULT is the Simple Random Sample Model, i.e. all Observations from one super stratum.

6) **ID**: the variable that identifies or codes the observations. The DEFAULT is ID=RESPID.

7) **ORALPHA**: (1-ORALPHA)100% is the coverage for Odds Ratio CI, Note: 0 < ORALPHA < 1, DEFAULT is ORALPHA=0.05.

8) **SAMPTYPE**: With Replacement use 'WR', WithOut Replacement use 'WOR'. Capitol letters and quotes NEEDED. DEFAULT is SAMPTYPE='WR'.

9) **TOTCNT**: A variable in the data set that assigns the Total number of PSUs in given stratum for each observation. TOTCNT MUST BE PRESENT with 'WOR' option & is NOT needed with 'WR' option.

10) **PARMTEST**: A data set of linear combinations of the coefficients of the IVARS. These are often useful for understanding interactions among variables. A linear combo contains as many constants as the number of IVARS. The DEFAULT SUPPRESSES parmtest output.

TYPICAL MACRO CALLS

Example 1: TYPICAL WR EXAMPLE, USING STANDARD DEFAULTS:

```
%logitse(data=survey1, depvar=hivtest, ivars= age sex race,
        weight=rsweight, nest=areacode, id=respid);
```

Example 2: EXAMPLE, USING LINEAR COMBINATIONS

```
data cmatrix;
input col1-col3;
 1 -1  0     *contrast between race1 and race2*
 2  1  0     *linear combination (2*A+1*B+0)=0*
%logitse(data=survey1, depvar=hivtest,
        ivars= race1 race2 race3,
        weight=rsweight, nest=areacode, id=respid,
        oralpha=0.05, samptype='WOR', totcnt=stratcnt,
        parmtest=cmatrix);
```

To obtain standard logistic regression output in an SRS setting (only one stratum in the NEST variable), then standard (weighted or unweighted) PROC LOGISTIC output, including various diagnostics and summaries, can be produced by using DEFAULTS. Here are two examples.

Example 3: WEIGHTED EXAMPLE:

```
%logitse(data=surveya, depvar=hivtest, ivars= age sex race,
        weight=rsweight, id=respid);
```

Example 4: UNWEIGHTED EXAMPLE:

```
%logitse(data=surveya, depvar=hivtest, ivars= age sex race);
```

CAUTIONARY NOTES:

If the Total Count of PSUs in each stratum is large, or unknown and assumed to be large, the variances can be computed using the With Replacement (WR) technique. However, should the ratio: {number PSUs SAMPLED}/{TOTAL number PSUs in given Stratum} exceeds .05 or 5%, then the Without Replacement (WOR) technique should be considered.

Should UNDEFINED errors arise, check the iteration history of PROC LOGISTIC. Lack of convergence of the parameter estimates is a consequence of severe collinearity among the independent variables or a (quasi) complete separation of the dependent variable in the independent variables space. See: Lesaffre and Marx (1993). Communications in Statistics: Theory and Methods 22(7): 1933-1952.

ILLUSTRATIVE EXAMPLE

We revisit data from a national telephone probability AIDS behavior survey conducted in 1992. Subjects were asked whether (Y=1) or not (Y=0) they had an HIV test. Explanatory variables of interest include gender (0=female, 1=male) and two indicator variables defining race, one for each of the Black and Hispanic races (Whites as baseline). Additionally, the age of the subject is used as a continuous regressor.

We are unable to use standard approaches because subjects were sampled within 22 specific areacodes across the contiguous United States. This defines our nesting variable, i.e., we assume subjects within an areacode are correlated.

Following is the output produced by the LOGITSE macro. This output was generated by the program listed as "Example 1" above. One can interpret this as an ordinary logistic regression where adjustments have been made to the standard errors due to the correlations of subjects within areacodes. Some details will be given in the following section.

Regression analysis using **LOGITSE**: (Release 2.2)

```
                          The SAS System
 Logistic Regression Output for Correlated Data Sampled in Clusters

                      Contents of Raw Data Set

                         CONTENTS PROCEDURE

Data Set Name: WORK.ONE                  Observations:         5735
Member Type:   DATA                      Variables:            9
Engine:        V608                      Indexes:              0
Created:       12:30 Tue, Jun 6, 1997    Observation Length:   72
Last Modified: 12:30 Tue, Jun 6, 1997    Deleted Observations: 0
Protection:                              Compressed:           NO
Data Set Type:                           Sorted:               NO
Label:

           -----Engine/Host Dependent Information-----

                Data Set Page Size:       4096
                Number of Data Set Pages: 103
                File Format:              607
```

First Data Page: 1
Max Obs per Page: 56
Obs in First Data Page: 36

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Format	Informat	Label
8	AGE1	Num	8	56	F4.	4.	Pre-Screening Interview: Age,in years
2	AREACOD1	Num	8	8	F4.	4.	Respondent's telephone areacode
6	BLACK1	Num	8	40	F4.	4.	Ethnicity: Dummy variable for Blacks
5	GENDER1	Num	8	32	F4.	4.	Pre-Screening Interview: Gender
7	HISP1	Num	8	48	F4.	4.	Ethnicity: Dummy variable for Hispanics
9	POSTHRC1	Num	8	64	F8.4	8.4	Post-strat. weight: High Risk Cities
3	PSU	Num	8	16	F4.	4.	Primary Sampling Unit
1	RESPID	Num	8	0	F5.	5.	Respondent's ID Number
4	RHIVTST1	Num	8	24	F4.	4.	Q15A: Ever had HIV antibody blood test?

The Dependent Variable: RHIVTST1

The Independent Variables:
GENDER1 BLACK1 HISP1 AGE1

The Total Number of Observations in Raw Dataset= 5735
The Number of Observations use in Analysis (N)= 5735
The Number of Nesting Levels (L)= 22
The Number of Distinct Covariate Patterns= 192
The Sum of the Non-Rescaled Weights= 5297.65
The Sum of the Rescaled Weights= 5735
The Coefficient of Variation of the Weights (CV)= 82.05
The Overall Design Effect= 1.67
The Denominator DF (N-L)= 5713
The Sampling Type is: WR

Note: The Overall Design Effect = 1 + square of (CV/100), where CV={stdev(Weight)/ave(Weight)}*100%. The Overall Design Effect is unitless, always greater than or equal to 1.000.

Equality to 1.000 occurs only with uniform weights. Thus as the Weights become more variable due to complex design sampling, then the Overall Design Effect grows larger than unity.

For example, an Overall Design Effect of 1.23 indicates for every 100 obs under simple random sampling, 123 obs are needed under the complex sampling process to achieve comparable variances.

Odds Ratio with
95 % Confidence Limits

OddRatio	Lower OR	Upper OR
GENDER1	1.4382	1.2460	1.6600
BLACK1	1.0922	0.9330	1.2784
HISP1	1.4134	1.1964	1.6697
AGE1	0.9915	0.9831	0.9998

Inverse Odds Ratio with
95 % Confidence Limits

inv(OR)	Lower invOR	Upper invOR
GENDER1	0.6953	0.6024	0.8026
BLACK1	0.9156	0.7822	1.0718
HISP1	0.7075	0.5989	0.8359
AGE1	1.0086	1.0002	1.0171

Analysis of Maximum Likelihood Parameter Estimates
(with adjusted standard errors)

	DF	Beta	Adj SE	Wald	Pr>ChiSq	deff	Unadj SE
INTERCPT	1	-0.8965	0.1622	30.5564	0.0000	1.7816	0.1215
GENDER1	1	0.3634	0.0732	24.6533	0.0000	1.5512	0.0588
BLACK1	1	0.0882	0.0803	1.2041	0.2725	1.2835	0.0709
HISP1	1	0.3460	0.0850	16.5520	0.0000	1.1540	0.0792
AGE1	1	-0.0086	0.0043	3.9872	0.0458	1.5333	0.0035

Note: deff is the Design Effect or Relative Efficiency of the ratio of the variance of the Complex Design to the variance from Simple Random Sampling.

deff values greater than 1.0 are generally expected.

For example, a deff=1.23 indicates for every 100 obs under simple random sampling, 123 obs are needed under complex sampling to achieve comparable estimated coefficient variances.

Adjusted Covariance Matrix for Beta

	INTERCPT	GENDER1	BLACK1	HISP1	AGE1
INTERCPT	0.0263	-0.0035	-0.0047	-0.0048	-0.0006
GENDER1	-0.0035	0.0054	0.0005	0.0002	0.0000
BLACK1	-0.0047	0.0005	0.0065	0.0031	0.0000
HISP1	-0.0048	0.0002	0.0031	0.0072	0.0001
AGE1	-0.0006	0.0000	0.0000	0.0001	0.0000

Adjusted Correlation Matrix for Beta

	INTERCPT	GENDER1	BLACK1	HISP1	AGE1
INTERCPT	1.0000	-0.2935	-0.3621	-0.3488	-0.9132
GENDER1	-0.2935	1.0000	0.0836	0.0242	0.0671
BLACK1	-0.3621	0.0836	1.0000	0.4554	0.1276
HISP1	-0.3488	0.0242	0.4554	1.0000	0.1434
AGE1	-0.9132	0.0671	0.1276	0.1434	1.0000

The ROW(S) is(are) Requested LINEAR COMBINATION(S)

INTERCPT	GENDER1	BLACK1	HISP1	AGE1
0	2	-1	0	-1
0	0	0	1	-1

Odds Ratios for requested LINEAR COMBINATION(S) with
95 % Confidence Limits

OR	Low OR	Up OR	invOR	Low invOR	Up invOR
1.9101	1.3931	2.6190	0.5235	0.3818	0.7178
1.4255	1.2079	1.6824	0.7015	0.5944	0.8279

SUMMARY

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	6893.093	6836.197	.
SC	6899.747	6869.469	.
-2 LOG L	6891.093	6826.197	64.896 with 4 DF (p=0.0001)
Score	.	.	65.183 with 4 DF (p=0.0001)

Collapsed Table Based on Estimated Probability Percentiles
Each Grouping with N/10 Equal Observations

SAS

Hosmer and Lemeshow Goodness-of-Fit Test

		RHIVTST1 = 1		RHIVTST1 = 0	
Group	Total	Observed	Expected	Observed	Expected
1	574	135	127.98	439	446.02
2	574	145	136.21	429	437.79
3	574	164	142.17	410	431.83
4	574	168	149.06	406	424.94
5	574	171	161.03	403	412.97
6	574	206	172.10	368	401.90
7	574	203	178.86	371	395.14
8	574	184	185.01	390	388.99
9	574	189	196.71	385	377.29
10	569	239	223.02	330	345.98

Goodness-of-fit Statistic = 26.425 with 8 DF (p=0.0009)

Preferred Hosmer-Lemeshow DECILE Goodness-of-Fit Test
Palermo-Hudes Version
Deciles of Risk
Observed & Expected Frequencies

DECILE	TOTAL	OBS0	EXP0	OBS1	EXP1
1	660.87	532.97	513.89	127.89	146.97
2	579.29	433.48	441.95	145.81	137.34
3	611.65	446.64	460.25	165.01	151.40
4	491.07	366.06	364.06	125.02	127.01
5	531.80	394.07	381.82	137.73	149.99
6	599.88	406.20	420.39	193.68	179.49
7	600.43	398.27	413.32	202.16	187.11
8	656.64	444.76	444.94	211.88	211.70
9	595.86	414.32	392.76	181.54	203.09
10	407.51	243.75	247.13	163.77	160.39
	5735.00	4080.52	4080.52	1654.48	1654.48

CHISQR	DF	PVAL
13.8804	8	0.0849

Covariate Patterns & Associated Weights having
Significant Pearson Chi-Square Statistic Due to Deletion
i.e. OUTLIER Greater Than 3.841
Pearson (Chi) and Deviance (Dev) Residuals Identify Poor Fits

NOTE: Casewise Type I Error Rate is 0.05
Simultaneous Type I Error Rate May Be Greater Than 0.05
See Hosmer & Lemeshow (1989, Equation 5.11)

OBS	RESPID	RHIVTST1	GENDER1	BLACK1	HISP1	AGE1	POSTHRC1
1	10090	1	0	0	0	22	5.5593
2	632	1	0	0	0	18	5.5593
3	4108	1	0	0	0	37	3.9492
4	11483	1	0	0	0	43	3.7415
5	1259	1	0	0	0	21	4.4474
6	7703	1	0	0	0	31	3.9492
7	11257	1	0	0	0	36	3.6142
8	614	1	0	0	0	36	3.3833
9	10628	1	1	0	0	29	5.1445
10	5288	1	1	0	0	26	5.0833

.
.
.
{abbreviated}

SOME DETAILS ON ADJUSTED SE's

We have considered the logistic model

$$\log\left\{\frac{\pi(X)}{1-\pi(X)}\right\} = X\beta,$$

where the rows of X are the explanatory information for the i^{th} subject (i = 1,...,n) and $\pi(x)$ is the (unknown) probability response vector, P(Y=1 | X) = $\pi(x)$. Estimation of unknown β is usually obtained by the (iteratively reweighted) method of scoring. For details, the reader is referred to Hosmer and Lemeshow (1989). Upon convergence, denote the MLE of β as $\hat{\beta}$ and

$\hat{\pi}(X) = \frac{e^{(X\hat{\beta})}}{1+e^{(X\hat{\beta})}}$. The estimated (large sample) information matrix is $\hat{\Phi} = X'\hat{V}\hat{W}X$, where $\hat{V} = diag\{\hat{\pi}(1-\hat{\pi})\}$ and $\hat{W}$ is a diagonal matrix of estimated sampling weights. The (adjusted) covariance for $\hat{\beta}$ is derived from the linearized column of vectors of $\hat{Z} = W\#(y - \hat{\Pi})\# X$ (# denotes element-wise multiplication) and utilizes $COV(\hat{Z})$. We express the adjusted (with replacement) covariance matrix as $\hat{Var}(\hat{\beta}) = \hat{\Phi}^{-1}\hat{S}\hat{\Phi}^{-1}$. The formula for $\hat{S}$ is given by $\hat{S}_{wr} = \sum_{h=1}^{H} n_h Cov(\hat{Z}_h)$, where n_h is the sample size associated with the h^{th} stratum. The matrix $\hat{S}$ for without replacement has further adjustments for finite sampling. The reader is referred to Binder (1983).

LIMITATIONS & CONCLUSIONS

LOGITSE is written using SAS/IML®. To facilitate the use of LOGITSE, the SAS® Macro Language has been used for passing parameters. For this reason LOGITSE may not be useable if SAS/IML® is not available at the user's installation. Since SAS/IML® allocates memory dynamically, there is no constraint within LOGITSE regarding number of variables or number of observations. It is only limited by the amount of memory available on your system.

The SAS® LOGITSE macro v2.2 provides a reasonable method for adjusting standard errors for complex survey designs.

ACKNOWLEDGMENTS

We thank Lance Pollack, UCSF, for his critical suggestions for the improvement of the original macro output. We also thank Estie Hudes and Lisa Palermo for allowing the integration of their Hosmer-Lemeshow goodness-of-fit macro into LOGITSE. Of course, any remaining errors, omissions, and commissions are ours and ours alone.

Funding was provided by the US National Institute of Mental Health (MH52022 & MH54320).

Users may obtain the LOGITSE macro and all relevant files via anonymous FTP at psg111.ucsf.edu under the LOGITSE subdirectory using their e-mail address as the password.

AUTHOR ADDRESSES:
Jesse A. Canchola, MS and Joseph A. Catania, PhD
Center for AIDS Prevention Studies
University of California , San Francisco
74 New Montgomery Street, Suite 600
San Francisco. CA 94105.
415 / 597-9354; 415 / 597-9395 (FAX)
e-mail: Jesse_Canchola@psgcaps.ucsf.edu

Brian D. Marx, PhD
Louisiana State University
Dept. of Experimental Statistics
161 Agricultural Admin. Bldg.
Baton Rouge, LA 70803
504 / 388-8366; 504 / 388-2526 (FAX)
e-mail: brian@multico.stat.lsu.edu

REFERENCES

Hudes, ES, Palermo, L (1993), "The LACKFIT Option in SAS LOGISTIC Procedure: A Problem and a Proposed Solution," *Proceedings of the First Annual Western Users of SAS Software, Regional Users Group Conference, Santa Monica, October 20-22,* SAS Institute Inc., Cary, NC.

Binder (1983), *International Statistical Review,* 51: 279-292.

Kish, L (1989), *Sampling Methods for Agricultural Surveys*, 182-186, Rome: Food and Agriculture Organization of the United Nations. 261 pp.

Lesaffre and Marx (1993), *Communications in Statistics: Theory and Methods* 22(7): 1933-1952.

SAS Institute Inc. (1995), *Logistic Regression Examples Using the SAS System, Version 6, First Edition,* pp. 81-86, Cary, NC: SAS Institute Inc. 163 pp.

SAS Institute (1985), *SAS/IML User's Guide, Version 6, First Edition*, Cary, NC: SAS Institute Inc. 501 pp.

Stokes, Maura E., Davis, Charles S., and Koch, Gary G. (1995), *Categorical Data Analysis Using the SAS System,* pp. 413-423, Cary, NC: SAS Institute Inc. 499 pp.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989*), Analysis of Complex Surveys*, Wiley, New York. 309 pp.

Waksberg, J. (1978), "Sampling methods for random digit dialing," *Journal of the American Statistical Association*, 73, 40-46.

APPENDIX

```
/* ************************************************************
*  Macro LOGITSE 2.2       January 1997        Brian D. Marx    *
************************************************************; */
%MACRO LOGITSE(DATA=ONE,DEPVAR=,IVARS=,WEIGHT=1,
               NEST=1, ID=RESPID, ORALPHA=0.05, SAMPTYPE='WR',
```

```
               TOTCNT=STRATCNT, PARMTEST=-999);
OPTIONS LS=68 PS=69 NODATE
NONUMBER CENTER; TITLE5 ' ';TITLE6 ' ';TITLE7 ' ';TITLE8 '
';TITLE9 ' ';
TITLE2 'Logistic Regression Output for Correlated Data Sampled in
Clusters';
TITLE3 ' ';

%IF &NEST=1 %THEN %DO;
DATA &DATA; SET &DATA; _NEST_=1; RUN;
%LET NEST=_NEST_;%END;

%IF &WEIGHT=1 %THEN %DO;
DATA &DATA; SET &DATA; _WGHT_=1; RUN;
%LET WEIGHT=_WGHT_;
%END;

PROC CONTENTS DATA=&DATA;
TITLE4 'Contents of Raw Data Set';

PROC MEANS DATA=&DATA N NOPRINT;
OUTPUT OUT=NNSIZE N=NN; TITLE1 ' '; TITLE2 ' ';TITLE3 ' ';
TITLE4 ' ';RUN;
DATA NSIZE; SET NNSIZE (KEEP=NN); RUN;

DATA NOMISS; SET &DATA; ONEN=1;
  ARRAY MISS{*} &IVARS &DEPVAR;
        DO I=1 TO DIM(MISS);
             IF MISS{I} LE . THEN DELETE;
             IF &WEIGHT LE 0 THEN DELETE;
        END; DROP I; RUN;
PROC MEANS DATA=NOMISS N SUM CV NOPRINT;
VAR &WEIGHT; OUTPUT OUT=
SUMWT N=SAMPSIZE SUM=SUMWT CV=CV1; TITLE ' ';RUN; QUIT;

DATA CV1; SET SUMWT(KEEP=CV1);RUN;

DATA NOMISS(DROP=SAMPSIZE SUMWT);
IF _N_=1 THEN SET SUMWT; SET NOMISS;
RSWEIGHT=&WEIGHT /SUMWT *SAMPSIZE;
ONE=1; RUN;

PROC DATASETS; DELETE NNSIZE SUMWT; RUN; QUIT;

PROC SORT DATA=NOMISS; BY &NEST; RUN; QUIT;
PROC SORT DATA=NOMISS OUT=STRATUM NODUPKEY; BY &NEST; RUN; QUIT;
DATA _NULL_;
SET STRATUM END=LASTREC;
MACROVAR='STRTM'||LEFT(PUT(_N_,3.));
CALL SYMPUT(MACROVAR, STRTM);
IF LASTREC THEN CALL SYMPUT ('NSTRAT',PUT(_N_,3.)); RUN; QUIT;

PROC SORT DATA=NOMISS OUT=COVPAT NODUPKEY; BY &IVARS; RUN; QUIT;
DATA _NULL_;
SET COVPAT END=LASTREC;
IF LASTREC THEN CALL SYMPUT ('NCOVPAT',PUT(_N_,8.)); RUN; QUIT;

FILENAME ROUTED 'NEWOUT';
PROC PRINTTO PRINT=ROUTED NEW;RUN;

PROC LOGISTIC  DATA=NOMISS DESCENDING;
MODEL &DEPVAR=&IVARS/ LACKFIT LINK=LOGIT;
OUTPUT OUT=VPRED P=PHAT RESCHI=CHI RESDEV=DEV H=HAT;
WEIGHT RSWEIGHT; RUN; QUIT;

PROC PRINTTO ; RUN; QUIT;

DATA SPRED; SET VPRED;
OUTLIER=(CHI**2)/(1-HAT); IF OUTLIER<3.841 THEN DELETE;
PROC SORT DATA=SPRED OUT=NDSPRED NODUPKEY;
BY DESCENDING OUTLIER RSWEIGHT &DEPVAR &IVARS; RUN; QUIT;

PROC DATASETS; DELETE SPRED; RUN; QUIT;

DATA GOF; INFILE ROUTED; LG=68;
INPUT SUMMARY $VARYING68. LG @@; DROP LG COUNTRR;
RETAIN COUNTRR 0;
IF INDEX(SUMMARY,'Model Fit') THEN COUNTRR=COUNTRR+1;
IF COUNTRR >0 THEN COUNTRR=COUNTRR+1;
IF (1=< COUNTRR<=11) THEN OUTPUT;
DATA HL; INFILE ROUTED;
RETAIN COUNTER 0; LG=68;
INPUT SAS $VARYING68. LG @@; DROP LG COUNTER;
IF INDEX(SAS,'Hosmer') THEN COUNTER=COUNTER+1;
IF COUNTER>0 THEN COUNTER=COUNTER+1;
IF 1<= COUNTER<=20 THEN OUTPUT;
PROC RANK DATA=VPRED GROUPS=10
        OUT=DATRANK(KEEP=&DEPVAR RSWEIGHT PHAT DECILE ONEN);
        VAR PHAT; RANKS DECILE; RUN; QUIT;
PROC SORT DATA=DATRANK; BY DECILE; RUN; QUIT;
DATA DATRANKK; SET DATRANK; DECILE=DECILE+1;
PROC MEANS DATA=DATRANKK SUM N NOPRINT;
        VAR &DEPVAR PHAT ONEN;
        WEIGHT RSWEIGHT; BY DECILE;
        OUTPUT OUT=DATSUM N=RNOBS SUM=OBS1 EXP1 TOTAL; RUN; QUIT;
DATA TABLE (KEEP=TOTAL DECILE OBS0 EXP0 OBS1 EXP1
        DROP= RNOBS)
        CHISQ(KEEP=CHISQR DF PVAL);
        SET DATSUM END=LAST;
        RETAIN CHISQR G 0;
        IF DECILE<=.Z THEN DELETE;
                ELSE DO;
                G=G+1;
        OBS0=TOTAL-OBS1;
        EXP0=TOTAL-EXP1;
        C1=((OBS1-EXP1)**2)/EXP1;
        C0=((OBS0-EXP0)**2)/EXP0;
        CHISQR=CHISQR+C1+C0;
        IF LAST THEN DO;
                DF=G-2;
                PVAL=1.0000-PROBCHI(CHISQR,DF);
                OUTPUT CHISQ;
                END;
        OUTPUT TABLE;
        END;
PROC DATASETS; DELETE DATRANKK DATRANK DATSUM; RUN; QUIT;

PROC IML; RESET NOLOG NONAME;
USE NOMISS; READ ALL VAR{&IVARS} INTO X [COLNAME=XNAMES];
READ ALL VAR{&DEPVAR} INTO Y [COLNAME=YNAME];
READ ALL VAR{&WEIGHT} INTO W; READ ALL VAR{&NEST} INTO NEST;
CLOSE NOMISS;
CREATE DEPNAME FROM YNAME; APPEND FROM YNAME;
CREATE INAMES FROM XNAMES; APPEND FROM XNAMES;
USE VPRED VAR{PHAT}; READ ALL VAR{PHAT}; CLOSE VPRED;
N=NROW(X); ONEN=J(N,1,1); PWSUM=ONEN'*W;
W=W*N/PWSUM; WSUM=ONEN'*W;
/*
D1=-2#W#Y#LOG(PHAT);D2=2#(Y-ONEN)#W#LOG(ONEN-PHAT); D=D1+D2;
DT=SQRT(D); DEV=DT'*DT;
CREATE DEVV FROM DEV; APPEND FROM DEV;
*/
CREATE WSUMM FROM WSUM; APPEND FROM WSUM;
CREATE PWSUMM FROM PWSUM; APPEND FROM PWSUM;
XINT=ONEN||X;
PQ=PHAT-PHAT#PHAT; WPQ=W#PQ;
INFORMW=XINT'*((WPQ)#XINT); IINFORMW=INV(INFORMW);
ZWORK=LOG(PHAT/(ONEN-PHAT))+((Y-PHAT)/PQ);
BBETA=IINFORMW*XINT'*(WPQ#ZWORK);
FREE INFORMW ZWORK WPQ PQ;
CREATE IINFORM FROM IINFORMW; APPEND FROM IINFORMW;
CREATE BETA FROM BBETA; APPEND FROM BBETA;
Z=(W#(Y-PHAT))#XINT;
FREE BBETA IINFORMW  PHAT XINT W;
NESTZ=NEST||Z; CREATE NEW1 FROM NESTZ; APPEND FROM NESTZ;
FREE NESTZ NEST Z PWSUM WSUM; QUIT;

PROC DATASETS; DELETE NOMISS VPRED; RUN; QUIT;

PROC CORR COV OUTP=COV DATA=NEW1 NOPRINT; BY COL1; RUN; QUIT;
DATA FREQQ; SET COV; IF _TYPE_='N'; DROP COL1;
DATA COVV; SET COV; IF _TYPE_='COV'; DROP COL1;

PROC DATASETS LIBRARY=WORK; DELETE COV NEW1; RUN; QUIT;

PROC IML; RESET NOLOG NONAME;
USE DEPNAME; READ ALL VAR _CHAR_ INTO DVNAME; CLOSE DEPNAME;
USE INAMES; READ ALL VAR _CHAR_ INTO IVNAMES; CLOSE INAMES;
USE CV1; READ ALL INTO CV1; CLOSE CV1;
USE BETA; READ ALL INTO BETA; CLOSE BETA;
USE IINFORM; READ ALL INTO BREAD; CLOSE IINFORM;
USE FREQQ; READ ALL INTO F; CLOSE FREQQ;
USE COVV; READ ALL INTO RAW; CLOSE COVV;
USE WSUMM; READ ALL INTO WSUM; CLOSE WSUMM;
USE PWSUMM; READ ALL INTO PWSUM; CLOSE PWSUMM;
USE NSIZE; READ ALL INTO PNSIZE; CLOSE NSIZE;
NOVARS=NCOL(BREAD); FREQS=F[,1]; NOBS=J(1,NROW(FREQS),1)*FREQS;
DENDF=NOBS-&NSTRAT;PCOOK=J(NOVARS,NOVARS,0);
DESEFF=1+(CV1/100)**2;
SAMPTYPE=&SAMPTYPE;
START LOOPA;
%IF &NSTRAT>1 %THEN %DO;
IF SAMPTYPE='WOR' THEN DO;
USE STRATUM; READ ALL VAR{&TOTCNT} INTO TOTCNT;
   %DO H=1 %TO &NSTRAT;
        TOP=((&H-1)*NOVARS)+1; BOT=TOP+NOVARS-1;
        COOK&H=FREQS[&H,]#(1-
(FREQS[&H,]/TOTCNT[&H,]))#RAW[TOP:BOT,];
        MEAT=PCOOK+COOK&H; PCOOK=MEAT;
   %END;
END;
ELSE DO;
   %DO H=1 %TO &NSTRAT;
        TOP=((&H-1)*NOVARS)+1; BOT=TOP+NOVARS-1;
        COOK&H=FREQS[&H,]#RAW[TOP:BOT,];
        MEAT=PCOOK+COOK&H; PCOOK=MEAT;
   %END;
END;
SANDWICH=BREAD*MEAT*BREAD;
%END;
%ELSE %DO;
SANDWICH=BREAD;
DENDF=NOBS-NOVARS;
%END;
FINISH; RUN LOOPA;
FREE/WSUM PWSUM IVNAMES DVNAME SANDWICH DESEFF
     CV1 PNSIZE SAMPTYPE BREAD BETA DENDF NOVARS NOBS;
NAMES='INTERCPT'||IVNAMES;
PRINT 'The Dependent Variable:' DVNAME;
PRINT 'The Independent Variables:',IVNAMES;
```

```
PRINT 'The Total Number of Observations in Raw Dataset=' PNSIZE;
PRINT 'The Number of Observations use in Analysis (N)=' NOBS;
PRINT 'The Number of Nesting Levels (L)=' &NSTRAT [FORMAT=4.0];
PRINT 'The Number of Distinct Covariate Patterns=' &NCOVPAT
[FORMAT=8.0];
PRINT 'The Sum of the Non-Rescaled Weights=' PWSUM [FORMAT=8.2];
PRINT 'The Sum of the Rescaled Weights=' WSUM;
PRINT 'The Coefficient of Variation of the Weights (CV)=' CV1
[FORMAT=8.2];
PRINT 'The Overall Design Effect=' DESEFF [FORMAT=8.2];
PRINT 'The Denominator DF (N-L)=' DENDF;
PRINT 'The Sampling Type is:' SAMPTYPE,;
PRINT 'Note: The Overall Design Effect = 1 + square of (CV/100),
where CV={stdev(Weight)/ave(Weight)}*100%. The Overall Design
Effect is unitless, always greater than or equal to 1.000.';
PRINT 'Equality to 1.000 occurs only with uniform weights.
 Thus as the Weights
 become more variable due to complex
 design sampling, then the Overall Design Effect grows larger
 than unity.';
PRINT 'For example, an Overall Design Effect of 1.23 indicates
for every
 100 obs under simple random sampling, 123 obs are needed
under the complex sampling process to achieve comparable
variances.',,,;
DF=J(NOVARS,1,1); VART=DIAG(SANDWICH);VART=VART*DF;
SE=SQRT(VART);
INVSE=1/SE; COVADJ=INVSE*INVSE';
VARSRS=DIAG(BREAD); VARSRS=VARSRS*DF; DEFF=VART/VARSRS;
OLDSE=SQRT(VARSRS);
CORBETA=COVADJ#SANDWICH;
ORAT=EXP(BETA); PIVOT=PROBIT(1-&ORALPHA/2);
COVERAGE=100*(1-&ORALPHA);
LORAT=EXP(BETA-(PIVOT#SE));
UORAT=EXP(BETA+(PIVOT#SE));
ORSUMMRY=ORAT||LORAT||UORAT; ORSUMMRY=ORSUMMRY[2:NOVARS,];
NAMEOR='OddRatio'||'Lower OR'||'Upper OR';
PRINT 'Odds Ratio with', COVERAGE '% Confidence Limits';
PRINT ORSUMMRY [ROWNAME=IVNAMES] [COLNAME=NAMEOR] [FORMAT=8.4],,;
IORAT=EXP(-BETA);
ILORAT=EXP(-BETA-(PIVOT#SE));
IUORAT=EXP(-BETA+(PIVOT#SE));
IORSUMM=IORAT||ILORAT||IUORAT; IORSUMM=IORSUMM[2:NOVARS,];
INAMEOR='inv(OR)'||'Lower invOR'||'Upper invOR';
PRINT 'Inverse Odds Ratio with', COVERAGE '% Confidence Limits';
PRINT IORSUMM [ROWNAME=IVNAMES] [COLNAME=INAMEOR] [FORMAT=8.4],,;
WALD=BETA/SE;PVAL=2*PROBNORM(-(ABS(WALD)));WALD=WALD#WALD;
SUMMARY=BETA||SE||WALD||PVAL||DEFF||OLDSE;
NAMESUM='Beta'||'Adj SE'||'Wald'||'Pr>ChiSq'||'deff'||'Unadj SE';
NAMEDF='DF';
PRINT 'Analysis of Maximum Likelihood Parameter Estimates',
'(with adjusted standard errors)' ,;
PRINT DF [ROWNAME=NAMES COLNAME=NAMEDF FORMAT=1.0]
SUMMARY [COLNAME=NAMESUM FORMAT=8.4],;
PRINT 'Note: deff is the Design Effect or Relative Efficiency of
 the ratio of the variance of the Complex Design to the
 variance from Simple Random Sampling.' ;
PRINT 'deff values greater than 1.0 are generally expected.';
PRINT 'For example, a deff=1.23 indicates for every
 100 obs under simple random sampling, 123 obs are needed
 under complex sampling to achieve comparable
 estimated coefficient variances.',,,,;

PRINT 'Adjusted Covariance Matrix for Beta' ,,
SANDWICH [ROWNAME=NAMES
COLNAME=NAMES FORMAT=8.4],,,;

PRINT 'Adjusted Correlation Matrix for Beta' ,, CORBETA
[ROWNAME=NAMES
COLNAME=NAMES FORMAT=8.4],,,;

%IF &PARMTEST ^= -999 %THEN %DO;
USE &PARMTEST; READ ALL INTO CONTRAST; CLOSE &PARMTEST;
ZEROS=J(NROW(CONTRAST),1,0); CONTRAST=ZEROS||CONTRAST;
CHECK=CONTRAST*J(NCOL(CONTRAST),1,1); CHECK=SUM(ABS(CHECK));
IF (NCOL(CONTRAST)^=NROW(BETA)) THEN PRINT
'ERROR: THE Number of COLUMNS in the PARMTEST MATRIX does NOT
Equal
 the Number of IVARS';
IF (NCOL(CONTRAST)=NROW(BETA)) THEN DO;
CBETA=CONTRAST*BETA; CVAR=CONTRAST*SANDWICH*CONTRAST';
CVAR=DIAG(CVAR)*J(NROW(CONTRAST),1,1);CSE=SQRT(CVAR);
NAMECOR='OR'||'Low OR'||'Up OR'||'invOR'||'Low invOR'||'Up
invOR';
CORAT=EXP(CBETA);
CLORAT=EXP(CBETA-(PIVOT#CSE));
CUORAT=EXP(CBETA+(PIVOT#CSE));
CIORAT=EXP(-CBETA);
CILORAT=EXP(-CBETA-(PIVOT#CSE));
CIUORAT=EXP(-CBETA+(PIVOT#CSE));
CORSUMM=CORAT||CLORAT||CUORAT||CIORAT||CILORAT||CIUORAT;
PRINT 'The ROW(S) is(are) Requested LINEAR COMBINATION(S)';
PRINT CONTRAST [COLNAME=NAMES FORMAT=2.0] ,;

PRINT 'Odds Ratios for requested LINEAR COMBINATION(S) with',
COVERAGE '% Confidence Limits';
PRINT CORSUMM  [COLNAME=NAMECOR] [FORMAT=8.4],,;
END;%END;
FREE/NOBS;
QUIT;

PROC PRINT NOOBS  DATA=GOF; LABEL SUMMARY=' ';
TITLE1 '  '; RUN; QUIT;

PROC PRINT NOOBS  DATA=HL; LABEL SAS=' ';
TITLE1 '  '; TITLE2 ' '; TITLE3 ' '; TITLE4 ' ';
TITLE5 'Collapsed Table Based on Estimated Probability
Percentiles';
TITLE6 'Each Grouping with N/10 Equal Observations';
RUN; QUIT;

PROC PRINT
 DATA=TABLE NOOBS;
 VAR DECILE TOTAL OBS0 EXP0 OBS1 EXP1 ;
        SUM    TOTAL OBS0 EXP0 OBS1 EXP1 ;
TITLE1 '  '; TITLE2 ' '; TITLE3 ' ';  TITLE4  ' ';
TITLE5 'Preferred Hosmer-Lemeshow DECILE Goodness-of-Fit Test';
TITLE6 'Palermo-Hudes Version';
TITLE7 ' ';
TITLE8 "Deciles of Risk";
TITLE9 "Observed & Expected Frequencies";
        RUN; QUIT;

PROC PRINT DATA=CHISQ NOOBS;
        VAR CHISQR DF PVAL;
        FORMAT PVAL 8.4;
        TITLE '  ';
        RUN; QUIT;

PROC PRINT DATA=NDSPRED; FORMAT &WEIGHT RSWEIGHT CHI DEV HAT
OUTLIER 8.4;
VAR &ID &DEPVAR &IVARS &WEIGHT RSWEIGHT CHI DEV HAT OUTLIER;
TITLE1 ' ';
TITLE2 'Covariate Patterns & Associated Weights having';
TITLE3 'Significant Pearson Chi-Square Statistic Due to
Deletion';
TITLE4 'i.e. OUTLIER Greater Than 3.841';
TITLE5 'Pearson (Chi) and Deviance (Dev) Residuals Identify Poor
Fits';
TITLE6 ' ';
TITLE7 'NOTE: Casewise Type I Error Rate is 0.05';
TITLE8 'Simultaneous Type I Error Rate May Be Greater Than 0.05';
TITLE9 'See Hosmer & Lemeshow (1989, Equation 5.11)';
RUN; QUIT;

PROC SORT DATA=NDSPRED; BY &IVARS; RUN; QUIT;
PROC PRINT DATA=NDSPRED; FORMAT &WEIGHT RSWEIGHT CHI DEV HAT
OUTLIER 8.4;
VAR &ID &DEPVAR &IVARS &WEIGHT RSWEIGHT CHI DEV HAT OUTLIER;
TITLE1 ' ';   TITLE2 ' ';
TITLE3 'Identical Outlier Summary as Above';
TITLE4 'Sorted by Covariate Pattern';

FOOTNOTE1 'Copyright (c) 1994-1997 by the Regents of the';
FOOTNOTE2 'University of California, San Francisco. All Rights
Reserved.';
FOOTNOTE3 'Created and Written by Brian D. Marx';
RUN; QUIT;

PROC DATASETS LIBRARY=WORK; DELETE DEPNAME INAMES BETA IINFORM
NSIZE
HL TABLE CHISQ GOF STRATUM COVV FREQQ NDSPRED WSUMM DEVV; RUN;
QUIT;
%MEND LOGITSE;

/*************************************************************
*************************************************************
                  EXAMPLE OF MACRO CALL
*************************************************************
*************************************************************
PROVIDE THE DATA SET HERE
*************************************************************;
LIBNAME DATA "C:\DATADIR";
DATA ONE;
SET DATA.SASDATA1;
*************************************************************
DEFINE THE PARMTEST MATRIX HERE
*************************************************************;
DATA CMATRIX;
INPUT COL1-COL4;
CARDS;
2 -1 0  0
0  0 1 -1 ;
*************************************************************
DEFINE THE MACRO PARAMETERS HERE
*************************************************************;
%LOGITSE(DATA=ONE,DEPVAR=RHIVTST1,
               IVARS= GENDER1 BLACK1 HISP1 AGE1,
               WEIGHT=POSTHRC1,
               NEST=ONES,
               ID=RESPID,
               ORALPHA=0.05,
               SAMPTYPE='WR',
              PARMTEST=CMATRIX); */
```