Logistic (RLOGIST) Example #1

SUDAAN Statements and Results Illustrated

- EFFECTS
- RFORMAT, RLABEL
- REFLEVEL
- EXP option on MODEL statement
- Hosmer-Lemeshow Test

Input Data Set(s): BRFWGT.SAS7bdat

Example

Using data from the BRFSS, model the risk of acute drinking as a function of race, sex, age, and educational status. Estimate the odds ratios and their confidence intervals and evaluate the overall fit of the model using the Hosmer-Lemeshow Goodness of Fit Test.

This example also demonstrates the use of the EXP option in the context of a main-effects model.

Solution

This example uses PROC RLOGIST (SAS-Callable SUDAAN) to model the risk of acute drinking as a function of race, sex, age, and educational status. The data were extracted from the Behavioral Risk Factor Surveillance System (BRFSS), which is a multi-stage, random-digit-dialing telephone survey conducted in each state.

This example highlights the use of the REFLEVEL and EFFECTS statements, the estimation of default and user-defined odds ratios and their confidence limits, and the Hosmer-Lemeshow Test for goodness of fit.

This example was run in SAS-Callable SUDAAN, and the SAS program and *.LST files are provided. The main-effects model is specified on the MODEL statement. Each of the variables to be modeled as categorical also appear on the SUBGROUP and LEVELS statements. The default Wald-*F* test is used for all tests of hypotheses.

SAS data step statements are used to convert the outcome variable, _RFDRACU, from a 1-2 variable (1=not at risk, 2=at risk) to a 0-1 variable (ACUTEDRINK=0 if not at risk, 1 if at risk) for the RLOGIST procedure.

The REFLEVEL statement defines the reference level for income, education, and race to be the first level of each variable. Since sex and age group are not noted on the REFLEVEL statement, the last level of each of these variables will be used as the reference level (‘2’ for *sex* and ‘5’ for *agecat5*).

Finally, the EFFECTS statement forms a contrast comparing education level 4 vs. 2 (college vs. high school). The EXP option will exponentiate the contrast to provide the user-requested odds ratio for acute drinking among those with college vs. high school education (the default odds ratios compare each education group to the *less than high school* group).

We include two PRINT statements in the code sample below. The first requests default statistics in the *betas*, *tests*, *risk*, and *expcntrst* groups, and the second requests the results of the Hosmer-Lemeshow goodness of fit test (*hltest* group). Two PRINT statements allow us to set up different default print environments (SETENV statements) for different groups. The PRINT statements are used in this example to request the PRINT groups of interest and to specify a variety of formats for those printed statistics. Without the PRINT statement, only the default statistics are produced (*hltest* group is not default), with default formats.

The SETENV statements are optional. They set up default formats for printed statistics and manipulate the printout to the needs of the user.

The RFORMAT statements associate the SAS formats with the variables used in the CROSSTAB procedure. The RLABEL statement defines variable labels for use in the current procedure only. Without the RLABEL statement, SAS variable labels would be produced if already defined.

Exhibit 1. SAS-Callable SUDAAN Code

```
libname in v604 "c:\10winbetatest\examplemanual\logistic";

options nocenter pagesize=70 linesize=85;
proc format;
  value educ 1="1=<HS"
             2="2=HS Grad"
             3="3=Some College"
             4="4=College Grad";
  value age 1="18-29"
            2="30-39"
            3="40-49"
            4="50-64"
            5="65+";
  value sex 1="1=Male"
            2="2=Female";
  value race 1="1=White"
             2="2=Black"
             3="3=Hispanic"
             4="4=Other";
  value inc 1="1 = <10K"
            2="2 = 10-20K"
            3="3 = 20-35K"
            4="4 = 35K+";

data one; set in.brfwgt; acutedrink=_rfdracu-1;
proc sort data=one; by _STSTR _PSU;

PROC RLOGIST DATA=ONE FILETYPE=SAS DESIGN=WR;
NEST _STSTR _PSU;
WEIGHT _FINALWT;

SUBGROUP EDUCAT SEX INCAT NRACE AGECAT5;
LEVELS   4      2   4     4     5;

REFLEVEL INCAT=1 EDUCAT=1 NRACE=1;
MODEL acutedrink = EDUCAT SEX INCAT NRACE AGECAT5;
EFFECTS EDUCAT=(0 -1 0 1) / EXP NAME="EDUCAT: Coll vs HS";

SETENV COLWIDTH=7 DECWIDTH=4 COLSPCE=1 TOPMGN=0;
PRINT / betas=default risk=default tests=default expcntrst=default
        t_betafmt=f7.2 waldffmt=f8.2 dffmt=f7.0 orfmt=f5.2 loworfmt=f5.2
        uporfmt=f5.2 exp_cntrstfmt=f13.2 low_cntrstfmt=f5.2 up_cntrstfmt=f5.2;

SETENV COLWIDTH=15 DECWIDTH=4 LABWIDTH=15 TOPMGN=0;
PRINT / HLTEST=default hlwaldpfmt=f17.4 hlchipfmt=f17.4;

RLABEL acutedrink="At Risk for Acute Drinking";
RFORMAT educat educ.;
RFORMAT incat inc.;
RFORMAT sex sex.;
RFORMAT nrace race.;
RFORMAT agecat5 age.;
RTITLE "Using LOGISTIC to Model At Risk for Acute Drinking";
```

Exhibit 2. First Page of SUDAAN Output (SAS *.LST File)

```
                                 S U D A A N
            Software for the Statistical Analysis of Correlated Data
           Copyright      Research Triangle Institute    February 2011
                                Release 11.0.0


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method,
Assuming a With Replacement (WR) Design
    Sample Weight: _FINALWT
    Stratification Variables(s): _STSTR
    Primary Sampling Unit: _PSU


Number of zero responses     :  4515
Number of non-zero responses :   558

Independence parameters have converged in 7 iterations

Number of observations read       :   5838    Weighted count: 18929149
Observations used in the analysis :   5073    Weighted count: 16055654
Denominator degrees of freedom    :   1959


Maximum number of estimable parameters for the model is 15

File ONE contains 1962 Clusters
1950 clusters were used to fit the model
Maximum cluster size is   4 records
Minimum cluster size is   1 records


Sample and Population Counts for Response Variable ACUTEDRINK
Based on observations used in the analysis
0:  Sample Count     4515    Population Count  13971453
1:  Sample Count      558    Population Count   2084201

R-Square for dependent variable ACUTEDRINK (Cox & Snell, 1989): 0.098829

-2 * Normalized Log-Likelihood with Intercepts Only :  3916.63
-2 * Normalized Log-Likelihood Full Model           :  3388.74
Approximate Chi-Square (-2 * Log-L Ratio)           :   527.90
Degrees of Freedom                                  :       14

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

Note from **Exhibit 2** that there are 558 cases at risk, and 4,515 cases not at risk for acute drinking. A total of 5,073 observations are used in the analysis.

Exhibit 3. Regression Coefficients: BETAS Group

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

by: Independent Variables and Effects.

--------------------------------------------------------------------------------
Independent                                Lower     Upper
  Variables and                            95%       95%                  P-value
  Effects               Beta               Limit     Limit     T-Test     T-Test
                        Coeff.    SE Beta  Beta      Beta      B=0        B=0
--------------------------------------------------------------------------------
Intercept               -3.6372    0.3542  -4.3320   -2.9425    -10.27     0.0000
Education Level
  1=<HS                  0.0000    0.0000   0.0000    0.0000      .         .
  2=HS Grad             -0.0477    0.2293  -0.4974    0.4019     -0.21     0.8351
  3=Some College        -0.1828    0.2481  -0.6693    0.3037     -0.74     0.4613
  4=College Grad        -0.3377    0.2474  -0.8229    0.1475     -1.37     0.1724
SEX
  1=Male                 1.2641    0.1290   1.0112    1.5171      9.80     0.0000
  2=Female               0.0000    0.0000   0.0000    0.0000      .         .
Income Code
  1 = <10K               0.0000    0.0000   0.0000    0.0000      .         .
  2 = 10-20K            -0.2830    0.2654  -0.8035    0.2374     -1.07     0.2863
  3 = 20-35K            -0.0947    0.2453  -0.5758    0.3864     -0.39     0.6994
  4 = 35K+               0.1359    0.2610  -0.3760    0.6479      0.52     0.6026
Race Code
  1=White                0.0000    0.0000   0.0000    0.0000      .         .
  2=Black               -0.8205    0.2050  -1.2226   -0.4185     -4.00     0.0001
  3=Hispanic            -0.8050    0.2716  -1.3377   -0.2723     -2.96     0.0031
  4=Other               -1.3565    0.4791  -2.2962   -0.4169     -2.83     0.0047
AGECAT5
  18-29                  2.1590    0.3260   1.5196    2.7984      6.62     0.0000
  30-39                  1.4208    0.3353   0.7631    2.0784      4.24     0.0000
  40-49                  0.9267    0.3543   0.2319    1.6214      2.62     0.0090
  50-64                  0.4838    0.3835  -0.2683    1.2359      1.26     0.2072
  65+                    0.0000    0.0000   0.0000    0.0000      .         .
--------------------------------------------------------------------------------
```

The results on Page 1 of the output (**Exhibit 3**, above) show the vector of estimated regression coefficients, their estimated standard errors and 95% confidence limits, and *t*-tests and *p*-values for testing whether each individual regression coefficient is equal to zero.

Note that the reference cells for education, income, and race are the first levels of each of these variables (because of their specification on the REFLEVEL statement), while the reference cells for the other categorical covariates in the model (sex and age group) are the default last levels of those variables. The reference level regression coefficients are estimated as 0, but are retained on the beta vector.

The *p*-values for betas deriving from categorical covariates contain the significance levels for comparing each group to the reference cell. So we see that males are significantly different from females (*p*=0.0000), and the positive beta estimate (1.2641) tells us that the log-odds of acute drinking are increased for males vs. females. Other significant pairwise comparisons are the reduced log-odds for each race group compared to White, and the increased log-odds for each of the three youngest age groups compared to those who are 65+.

Exhibit 4. ANOVA Table (TESTS Group)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

by: Contrast.

---------------------------------------------------
Contrast               Degrees
                       of                   P-value
                       Freedom     Wald F   Wald F
---------------------------------------------------
OVERALL MODEL               15      64.14    0.0000
MODEL MINUS
  INTERCEPT                 14      17.67    0.0000
INTERCEPT                    .          .       .
EDUCAT                       3       1.16    0.3250
SEX                          1      96.07    0.0000
INCAT                        3       1.64    0.1784
NRACE                        3       9.36    0.0000
AGECAT5                      4      23.50    0.0000
EDUCAT: Coll vs HS           1       3.01    0.0830
---------------------------------------------------
```

This ANOVA table (**Exhibit 4**, above) provides a test for each model term (just main effects in this model), as well as the contrast defined by the EFFECTS statement. The main effects of sex, race, and age on the risk of acute drinking are statistically significant, but education and income are not. The comparison of education level 4 (college) vs. 2 (high school grad) was not statistically significant at α=0.05 (*p*=0.0830).

Exhibit 5. Odds Ratios: RISK Group

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

by: Independent Variables and Effects.

--------------------------------------------
Independent
  Variables and                Lower   Upper
  Effects                      95%     95%
                       Odds    Limit   Limit
                       Ratio   OR      OR
--------------------------------------------
Intercept               0.03    0.01    0.05
Education Level
  1=<HS                 1.00    1.00    1.00
  2=HS Grad             0.95    0.61    1.49
  3=Some College        0.83    0.51    1.35
  4=College Grad        0.71    0.44    1.16
SEX
  1=Male                3.54    2.75    4.56
  2=Female              1.00    1.00    1.00
Income Code
  1 = <10K              1.00    1.00    1.00
  2 = 10-20K            0.75    0.45    1.27
  3 = 20-35K            0.91    0.56    1.47
  4 = 35K+              1.15    0.69    1.91
Race Code
  1=White               1.00    1.00    1.00
  2=Black               0.44    0.29    0.66
  3=Hispanic            0.45    0.26    0.76
  4=Other               0.26    0.10    0.66
AGECAT5
  18-29                 8.66    4.57   16.42
  30-39                 4.14    2.14    7.99
  40-49                 2.53    1.26    5.06
  50-64                 1.62    0.76    3.44
  65+                   1.00    1.00    1.00
--------------------------------------------
```

The table above (**Exhibit 5**) provides the default estimated odds ratios and their 95% confidence limits. For example, the odds of being at risk for acute drinking for males vs. females are 3.54. In other words, the odds are increased more than threefold for males compared to females. This effect is statistically significant at α=0.05, and hence the 95% confidence interval does not contain the null value of 1.0.

The default odds ratios for education compare each group to <HS. None of these odds ratios are significantly different from the null value of 1.0. The user-requested odds ratio for College Grad vs. HS grad (**Exhibit 6**) is provided on the following page.

Exhibit 6. Customized Odds Ratios: EXPCNTRST Group

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

by: Contrast.

----------------------------------------------------

Contrast                               Lower   Upper
                                       95%     95%
                       EXP(Contrast)   Limit   Limit
----------------------------------------------------
EDUCAT: Coll vs HS              0.75    0.54    1.04
----------------------------------------------------
```

Exhibit 6 displays the results requested by the EXP option on the EFFECTS statement (*expcntrst* group). The user-requested odds ratio for College Grad vs. HS grad is 0.75, representing a 25% reduction in odds for the college grad group. Recall that the *p*-value for this comparison was 0.0830 (just missing statistical significance at α=0.05). Hence, the 95% confidence limits contain the null value of 1.0.

The code "PRINT / HLTEST=DEFAULT" in **Exhibit 1** requests the default Hosmer-Lemeshow test statistics. This output is shown below in **Exhibit 7**.

Exhibit 7. Hosmer-Lemeshow Test: HLTEST Group

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

Hosmer-Lemeshow Goodness-of-Fit Test Statistics
-----------------------------------------------------------------------
                   H-L Chi-Square      H-L ChiSq DF   H-L ChiSq P-value
-----------------------------------------------------------------------
                           5.4259            8.0000              0.7112
-----------------------------------------------------------------------
```

Exhibit 7. Hosmer-Lemeshow Test: HLTEST Group-cont.

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

Hosmer-Lemeshow Goodness-of-Fit Test Statistics
-----------------------------------------------------------------------
                       H-L Wald F            H-L DF    H-L Wald P-value
-----------------------------------------------------------------------
                           0.4566            9.0000              0.9039
-----------------------------------------------------------------------
```

Exhibit 7. Hosmer-Lemeshow Test: HLTEST Group-cont.

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable ACUTEDRINK: At Risk for Acute Drinking

Using LOGISTIC to Model At Risk for Acute Drinking

Hosmer-Lemeshow Goodness-of-Fit Test Statistics
-------------------------------------------------------------------
                                    H-L                H-L
                  H-L               Satterthwaite      Satterthwaite
                  Satterthwaite F   Adjusted DF        P-value
-------------------------------------------------------------------
                           0.4168         8.5577             0.9204
-------------------------------------------------------------------
```

The default output for the Hosmer-Lemeshow includes goodness-of-fit tests using the Chi-Square, Wald F, and Satterthwaite F (**Exhibit 7**). A Pearson’s chi-square statistic is calculated from the 2 x G tables of observed and predicted responses (predicted from the logistic regression model), where G represents the number of groups specified. All three tests are testing the null hypothesis that the observed number of events in a given group is equal to the predicted number of events in the same group, across all G groups. This is equivalent to saying that the weighted totals of the residuals for each of the G groups are simultaneously equal to 0. When the null hypothesis is not rejected (*i.e.*, large *p*-values), we conclude the model is a good fit for the data.

In this example, all three tests yield large *p*-values. Thus, we fail to reject the null hypothesis and conclude that the model is a good fit for the data.

Users can override the default number of groups by including the `/HLGROUPS=xx` option in the model statement, where xx is the number of groupings, from 1 to the number of unique covariate patterns (see Hosmer-Lemeshow text for details).