Logistic (RLOGIST) Example #5

SUDAAN Statements and Results Illustrated

- Modeling 2 interaction terms
- Test for “chunk interactions”
- EFFECTS
- SUBPOPX
- REFLEVEL

Input Data Set(s): SAMADULTED.SAS7bdat

Example

Using 2006 NHIS data, determine for white adults whether region of the country interacts with either gender or age group on the occurrence of not being able to afford prescription medications, controlling for education and marital status.

This example highlights the use of the EFFECTS statements in testing simultaneous (or “chunk”) interactions.

Solution

NHIS is an annual multipurpose health sample survey conducted by the National Center for Health Statistics (NCHS). For more information about the data used in this example, see ***Section 12.7***. The 2006 NHIS collected data on approximately 29,200 households; 29,900 families; 75,700 persons; 24,275 sample adults; and 9,800 sample children.

In the 2006 study, each sample adult was asked (variable AHCAFYR1):

“*During the PAST 12 MONTHS, was there any time when you needed prescription medicine but didn’t get it because you couldn’t afford it*?”

Possible answer codes are yes, no, don’t know, refused, and not ascertained. Only 0.96% of sample adults were coded as something other than yes or no. The constructed variable CANTAFMEDS is created from AHCAFYR1 and is coded as 1=*yes* (could not afford at least once in the past 12 months) or 0=*no* (event did not happen). All other responses are coded to missing.

Example 3 uses the same dataset and shows how to obtain various statistics and tests in a main effects model. This example uses the RLOGIST procedure (SAS-Callable SUDAAN) to model the probability that the dependent variable CANTAFMEDS is equal to 1, but fits the model with main effects *plus two interaction terms* (*sex*region* and *age*region*). The EFFECTS statement is used to simultaneously test the significance of these interaction terms.

For variance estimation purposes, the complex sampling plan is described as 300 pseudo-strata with two pseudo-PSUs per stratum. Sampling at the first PSU stage is assumed to be with replacement. Each unit of analysis (sample adult, sample child, person, etc.) is clustered within his/her PSU, and lower level sampling units are not identified.

In this example (see ***Exhibit ?***), we use the sample adult (age 18 and older) data file with 24,275 observations. The stratification and primary sampling unit variables are named STRAT_P and PSU_P, respectively, and appear on the NEST statement. The weight variable for the sample adult file is WTFA_SA and appears on the WEIGHT statement. The PROC statements specify DESIGN=WR (*i.e.*, unequal probability sampling of PSUs with replacement), and Taylor Series linearization is used for variance estimation. The subpopulation is defined as white (MRACRPI2=1) and at least 25 years old (AGE_P >= 25).

The MODEL statement of the RLOGIST code (***Exhibit 1***) identifies CANTAFMEDS as the dependent variable; it is coded as 1=*incurred event* (can't afford) and 0=*did not incur event*. Since the independent variables (SEX, AGE25_3, EDUC_3, REGION, and MARRY_3) are to be modeled as categorical, they all appear on the CLASS statement. The default Wald-*F* test is used for all tests of hypotheses.

The model terms are as follows:

- Sex (SEX: 1=Male, 2=Female);
- Age at three levels (AGE25_3: 1=25-44, 2=45-64, 3=65+);
- Education at three levels (EDUC_3: 1=HS or Less, 2=Some College, 3=College Grad);
- Region of the U.S. at four levels (REGION: 1=NE, 2=Midwest, 3=South, 4=West); and
- Marital status at three levels (MARRY_3: 1=Married, 2=Widowed, 3=Unmarried).
- Sex-by-Region interaction effect (specified SEX*REGION)
- Age-by-Region interaction effect (specified AGE25_3*REGION)

The SUBPOPX statement restricts the analysis to whites aged 25 years or older. The REFLEVEL statement defines the regression coefficient reference level for sex, region, and marital status to be the first level of each variable (REFLEVEL is used for continuity with ***Example 3***; it serves no key function in this example, and it could have been removed). Since age and education are not included on the REFLEVEL statement, the last level of each of these variables will be used as the reference level for estimating regression coefficients.

The EFFECTS statement tests the null hypothesis that the regression coefficients associated with the two interaction terms (*age*region*, *sex*region*) are simultaneously equal to 0. This is sometimes called a test for **chunk interactions**.

We include a PRINT statement, which is optional. The PRINT statement is used in this example to request the PRINT groups of interest (BETAS and TESTS groups only) and to specify a variety of formats for those printed statistics. Without the PRINT statement, default statistics are produced from each PRINT group, with default formats.

The SETENV statement is optional. It sets up default formats for printed statistics and further manipulates the printout to the needs of the user.

The RFORMAT statements associate the SAS formats with the variables used in the RLOGIST 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.

This example was run in SAS-Callable SUDAAN, and the SAS program and *.LST files are provided.

Exhibit 1. SAS-Callable SUDAAN Code

```
libname in "c:\10winbetatest\AmJEpid";

options nocenter pagesize=70 linesize=95;
proc format;
  value educ 1="1=HS or Less"
             2="2=Some College"
             3="3=College+";
  value age 1="25-44"
            2="45-64"
            3="65+";
  value sex 1="1=Male"
            2="2=Female";
  Value region 1="1=N.E."
               2="2=Midwest"
               3="3=South"
               4="4=West";
  value marry 1="1=Married"
              2="2=Widowed"
              3="3=Unmarried";
  value yesno 1="Yes"
              0="No";

Data samadult; set in.samadulted;
  if 0 le educ1 le 14 then educ_3=1;
  else if educ1=15 then educ_3=2;
  else if 16 le educ1 le 21 then educ_3=3;
  else educ_3=.;

  if 25 le age_p le 44 then age25_3=1;
  else if 45 le age_p le 64 then age25_3=2;
  else if age_p ge 65 then age25_3=3;

  if r_maritl in (1,2,3) then marry_3=1;
  else if r_maritl=4 then marry_3=2;
  else if r_maritl in (5,6,7,8) then marry_3=3;
  else marry_3=.;

  if ahcafyr1=1 then cantafmeds=1;
  else if ahcafyr1=2 then cantafmeds=0;
  else if ahcafyr1 in (7,8,9) then cantafmeds=.;

proc sort data=samadult; by strat_p psu_p;

PROC RLOGIST DATA=samadult DESIGN=WR FILETYPE=SAS;
 NEST STRAT_P PSU_P;
 WEIGHT WTFA_SA;

 SUBPOPX AGE_P>24 AND MRACRPI2=1 / NAME="WHITES AGED 25+";
 CLASS SEX AGE25_3 EDUC_3 REGION MARRY_3;

 REFLEVEL SEX=1 REGION=1 MARRY_3=1;
 MODEL CANTAFMEDS = SEX AGE25_3 EDUC_3 REGION MARRY_3 sex*region age25_3*region;
 EFFECTS sex*region age25_3*region / name="Chunk Interactions";

 setenv labwidth=24 colspce=2 colwidth=7 decwidth=4;
 print / betas=default tests=default sebetafmt=f8.4 t_betafmt=f6.2 waldffmt=f7.2
         dffmt=f7.0;

RLABEL age25_3="Age Group";
RLABEL cantafmeds="Can't Afford Rx Meds Past 12m";
RFORMAT sex sex.;
RFORMAT age25_3 age.;
RFORMAT educ_3 educ.;
RFORMAT region region.;
RFORMAT marry_3 marry.;
RTITLE "Modelling Can't Afford Rx Meds (Testing Chunk Interactions)";
RFOOTNOTE "Data Source: NCHS National Health Interview Survey (2006)" ;
```

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: WTFA_SA
    Stratification Variables(s): STRAT_P
    Primary Sampling Unit: PSU_P


Number of zero responses     : 14737
Number of non-zero responses :  1305

Independence parameters have converged in 7 iterations

Number of observations read       :  24275    Weighted count:220266693
Observations in subpopulation     :  16469    Weighted count:158409519
Observations used in the analysis :  16042    Weighted count:154637709
Denominator degrees of freedom    :    300


Maximum number of estimable parameters for the model is 20

File SAMADULT contains  600 Clusters
 596 clusters were used to fit the model
Maximum cluster size is  71 records
Minimum cluster size is   1 records


Sample and Population Counts for Response Variable CANTAFMEDS
Based on observations used in the analysis
0:  Sample Count    14737    Population Count 142746051
1:  Sample Count     1305    Population Count  11891658

R-Square for dependent variable CANTAFMEDS (Cox & Snell, 1989): 0.037109

-2 * Normalized Log-Likelihood with Intercepts Only :  8699.01
-2 * Normalized Log-Likelihood Full Model           :  8092.38
Approximate Chi-Square (-2 * Log-L Ratio)           :   606.63
Degrees of Freedom                                  :       19

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

Exhibit 2 indicates that there are 16,469 observations defined by the subpopulation on the SUBPOPX statement (white adults over 25 yrs of age). After deleting observations with missing values for *any* of the model variables (dependent as well as independent), there are 16,042 used in the analysis, of which 1,305 have incurred the event and 14,737 have not. There were 427 observations in the subpopulation that were not used in the analysis due to missing values on one or more model variables.

Below are the frequency distributions for variables included on the CLASS statement (see ***Exhibit 3*** to ***Exhibit 9*)**. This default output is generated unless the NOFREQS option is included on the CLASS statement. Note that these distributions are computed for the specified subpopulation one variable at a time, and as such, only delete missing values for the variable at hand. So, for example, since sex, age, and region have no missing values, their distributions in the following tables sum to 16,469, the subpopulation total. However, the education and marital status distributions sum to just under 16,400 due to a small number of missing values for each of these variables.

Exhibit 3. CLASS Variable Frequencies (Sex)

```
Frequencies and Values for CLASS Variables
by: Sex.
-------------------------------------
Sex             Frequency       Value
-------------------------------------
Ordered
  Position:
  1                  7364      1=Male
Ordered
  Position:
  2                  9105    2=Female
-------------------------------------
```

Exhibit 4. CLASS Variable Frequencies (Age Group)

```
Frequencies and Values for CLASS Variables
by: AGE25_3.
----------------------------------
AGE25_3         Frequency    Value
----------------------------------
Ordered
  Position:
  1                  6626    25-44
Ordered
  Position:
  2                  6137    45-64
Ordered
  Position:
  3                  3706      65+
----------------------------------
```

Exhibit 5. CLASS Variable Frequencies (Education)

```
Frequencies and Values for CLASS Variables
by: EDUC_3.
-------------------------------------------
EDUC_3          Frequency             Value
-------------------------------------------
Ordered
  Position:
  1                  7660      1=HS or Less
Ordered
  Position:
  2                  2751    2=Some College
Ordered
  Position:
  3                  5852        3=College+
-------------------------------------------
```

Exhibit 6. CLASS Variable Frequencies (Region)

```
Frequencies and Values for CLASS Variables
by: Region.
--------------------------------------
Region          Frequency        Value
--------------------------------------
Ordered
  Position:
  1                  2801       1=N.E.
Ordered
  Position:
  2                  3930    2=Midwest
Ordered
  Position:
  3                  5867      3=South
Ordered
  Position:
  4                  3871       4=West
--------------------------------------
```

Exhibit 7. CLASS Variable Frequencies (Marital Status)

```
Frequencies and Values for CLASS Variables
by: MARRY_3.
----------------------------------------
MARRY_3         Frequency          Value
----------------------------------------
Ordered
  Position:
  1                  8974      1=Married
Ordered
  Position:
  2                  1763      2=Widowed
Ordered
  Position:
  3                  5633    3=Unmarried
----------------------------------------
```

The next table (***Exhibit 8***) contains the estimated regression coefficients from the main effects plus interaction model. There are 3 estimable parameters associated with the sex*region interaction and 6 estimable parameters associated with the age*region interaction. Since we are mainly interested in testing the simultaneous interaction effects, this table is included for completeness only.

## Exhibit 8.	Regression Coefficients Table

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable CANTAFMEDS: Can't Afford Rx Meds Past 12m
For Subpopulation: WHITES AGED 25+

Modelling Can't Afford Rx Meds (Testing Chunk Interactions)

by: Independent Variables and Effects.
---------------------------------------------------------------------------------
Independent Variables                                Lower     Upper
  and Effects                                        95%       95%                 P-value
                        Beta                         Limit     Limit     T-Test    T-Test
                        Coeff.      SE Beta          Beta      Beta      B=0       B=0
---------------------------------------------------------------------------------
Intercept               -4.9427     0.2931    -5.5195   -4.3659   -16.86    0.0000
Sex
  1=Male                 0.0000     0.0000     0.0000    0.0000      .         .
  2=Female               0.2773     0.2496    -0.2140    0.7685     1.11    0.2676
Age Group
  25-44                  1.4713     0.3088     0.8636    2.0789     4.76    0.0000
  45-64                  0.8360     0.3135     0.2191    1.4530     2.67    0.0081
  65+                    0.0000     0.0000     0.0000    0.0000      .         .
EDUC_3
  1=HS or Less           0.8955     0.0799     0.7382    1.0527    11.21    0.0000
  2=Some College         0.8881     0.1024     0.6867    1.0896     8.68    0.0000
  3=College+             0.0000     0.0000     0.0000    0.0000      .         .
Region
  1=N.E.                 0.0000     0.0000     0.0000    0.0000      .         .
  2=Midwest              0.2399     0.3986    -0.5444    1.0242     0.60    0.5477
  3=South                0.1513     0.3733    -0.5832    0.8858     0.41    0.6855
  4=West                 0.3384     0.3878    -0.4248    1.1015     0.87    0.3836
MARRY_3
  1=Married              0.0000     0.0000     0.0000    0.0000      .         .
  2=Widowed              0.3195     0.1691    -0.0133    0.6523     1.89    0.0599
  3=Unmarried            0.8020     0.0704     0.6636    0.9405    11.40    0.0000
Sex, Region
  1=Male, 1=N.E.         0.0000     0.0000     0.0000    0.0000      .         .
  1=Male, 2=Midwest      0.0000     0.0000     0.0000    0.0000      .         .
  1=Male, 3=South        0.0000     0.0000     0.0000    0.0000      .         .
  1=Male, 4=West         0.0000     0.0000     0.0000    0.0000      .         .
  2=Female, 1=N.E.       0.0000     0.0000     0.0000    0.0000      .         .
  2=Female, 2=Midwest    0.2020     0.2859    -0.3606    0.7647     0.71    0.4804
  2=Female, 3=South      0.3497     0.2851    -0.2113    0.9108     1.23    0.2209
  2=Female, 4=West       0.2399     0.2976    -0.3457    0.8255     0.81    0.4208
Age Group, Region
  25-44, 1=N.E.          0.0000     0.0000     0.0000    0.0000      .         .
  25-44, 2=Midwest      -0.2537     0.4150    -1.0704    0.5629    -0.61    0.5414
  25-44, 3=South        -0.1061     0.3661    -0.8265    0.6143    -0.29    0.7721
  25-44, 4=West         -0.5359     0.4270    -1.3761    0.3043    -1.26    0.2104
  45-64, 1=N.E.          0.0000     0.0000     0.0000    0.0000      .         .
  45-64, 2=Midwest       0.2802     0.3951    -0.4972    1.0576     0.71    0.4787
  45-64, 3=South         0.4822     0.3762    -0.2582    1.2226     1.28    0.2009
  45-64, 4=West          0.3339     0.4031    -0.4595    1.1272     0.83    0.4082
  65+, 1=N.E.            0.0000     0.0000     0.0000    0.0000      .         .
  65+, 2=Midwest         0.0000     0.0000     0.0000    0.0000      .         .
  65+, 3=South           0.0000     0.0000     0.0000    0.0000      .         .
  65+, 4=West            0.0000     0.0000     0.0000    0.0000      .         .
---------------------------------------------------------------------------------
Data Source: NCHS National Health Interview Survey (2006)
```

Next, we report the results produced by the EFFECTS statement to test chunk interactions.

Exhibit 9. ANOVA Table (Model Terms, User-Specified Contrasts)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable CANTAFMEDS: Can't Afford Rx Meds Past 12m
For Subpopulation: WHITES AGED 25+

Modelling Can't Afford Rx Meds (Testing Chunk Interactions)

by: Contrast.
------------------------------------------------------
Contrast                  Degrees
                          of                 P-value
                          Freedom    Wald F   Wald F
------------------------------------------------------
OVERALL MODEL                  20    229.97    0.0000
MODEL MINUS INTERCEPT          19     24.95    0.0000
INTERCEPT                       .         .       .
SEX                             .         .       .
AGE25_3                         .         .       .
EDUC_3                          2     65.88    0.0000
REGION                          .         .       .
MARRY_3                         2     65.15    0.0000
SEX * REGION                    3      0.54    0.6544
AGE25_3 * REGION                6      1.85    0.0896
Chunk Interactions              9      1.70    0.0886
------------------------------------------------------
Data Source: NCHS National Health Interview Survey (2006)
```

We see from ***Exhibit 9*** that each interaction effect alone (each adjusted for the other interaction effect and all other main effects in the model) is not statistically significant (*p*=0.6544 and 0.0896). We also see that the combined interaction effect (labeled "Chunk Interactions") is not statistically significant (*p*=0.0886). It has nine df, since there are three df associated with sex*region and six df associated with age*region. The non-significant effect is not surprising, since none of the nine pairwise interaction effects from the regression coefficient table was statistically significant.

The conclusion is that among white adults, region of the country does not significantly interact with either gender or age group on the occurrence of not being able to afford prescription medications. These interaction effects can be removed from the model.