REGRESS Example #1

SUDAAN Statements and Results Illustrated

- TEST
- SUBPOPX
- REFLEVEL
- COND_EFF
- LSMEANS

Input Data Set(s): NHANES_C_3.SAS7bdat

Example

Using the continuous NHANES 1999-2004 data, determine the effect of race/ethnicity upon body mass index (BMI) among female adults, adjusting for age, education, health status, and marital status.

This example also highlights the new confidence limits for predicted and conditional marginals introduced in SUDAAN 11.0.

Solution

Continuous NHANES data (1999-2004) were derived from a home interview, a physical examination, and nutrient and laboratory values.

The dependent variable in this example is BMI. Height and weight (components of BMI) were measured in the mobile examination center (MEC) physical examination and in the home examination for adults 20 years and older who were unable to go to the MEC. Since we wish to use both the MEC and home-examined females in the analysis, we restrict the analysis to females aged 20 years and older. Further, we include only three of the four race/ethnicity groups in the analysis: non-Hispanic white, non-Hispanic black, and Mexican-American. The SUBPOPX statement defines the analysis subpopulation as the intersection of these three restrictions on age, gender, and race/ethnicity.

Since we analyze both MEC and home-examined subjects, we use the weight variable WTMEC6YR on the WEIGHT statement. Because we analyze data from six years of NHANES (1999-2004), we use SDMVSTRA and SDMVPSU as the stratification and PSU variables, respectively, on the NEST statement. The first stage sampling of NHANES is approximated as unequal probability sampling of primary sampling units (PSUs) with replacement; hence, we use DESIGN=WR on the PROC statement.

Before proceeding to the linear regression model, we use PROC DESCRIPT to estimate mean BMI by race/ethnicity and by each of the model covariates (***Exhibit 1***). The SUBPOPX statement restricts the DESCRIPT analysis to women who have a value for all variables in the linear regression model so that the DESCRIPT and REGRESS analyses use exactly the same women.

The VAR statement specifies BMXBMI as the dependent variable. The TABLES statement includes all of the covariates and the independent variable race/ethnicity so that mean BMI is estimated for the subpopulations defined by race/ethnicity and the covariates. We interpret the DESCRIPT output only to get a general idea of the univariate relationships with BMI; we do not perform statistical tests of significance at this stage.

Exhibit 1. SAS-Callable SUDAAN Code: DESCRIPT

```
libname in "c:\11winbetatest\BreslowDay XTAB\Manual Example";

options pagesize=70 linesize=90;
proc format;
  value yesno 1="1=Yes"
              2="2=No";
  value health 1="Excellent"
               2="Very Good/Good"
                 3="Fair/Poor";
  value race 1="NH White"
             2="NH Black"
               3="Mexican American";
  value educ 1="HS or Less"
             2="Some College"
               3="College+";

data one; set in.nhanes_c_3;
  if 20 le ridageyr le 29 then age_decade=20;
  else if 30 le ridageyr le 39 then age_decade=30;
  else if 40 le ridageyr le 49 then age_decade=40;
  else if 50 le ridageyr le 59 then age_decade=50;
  else if 60 le ridageyr le 69 then age_decade=60;
  else if 70 le ridageyr le 79 then age_decade=70;
  else if 80 le ridageyr le 89 then age_decade=80;

  if hsd010=1 then health3=1;
  else if hsd010 in (2,3) then health3=2;
  else if hsd010 in (4,5) then health3=3;

  if dmdmartl in (1,6) then married=1;
  else if dmdmartl in (2,3,4,5) then married=2;
  else married=.;

  if dmdeduc2 in(1,2,3) then educ3=1;
  else if dmdeduc2=4 then educ3=2;
  else if dmdeduc2=5 then educ3=3;
  else educ3=.;

  age_c = ridageyr - 47.35;
  age_c_sq = age_c*age_c;
proc sort data=one; by sdmvstra sdmvpsu;

proc descript data=one filetype=sas design=wr;
  NEST sdmvstra sdmvpsu;
  WEIGHT wtmec6yr;

  subpopx ridageyr ge 20 and riagendr=2 and ridreth2 in(1,2,3) and educ3 in(1,2,3)
          and (1 le hsd010 le 5) and (1 le dmdmartl le 6) and bmxbmi>0 /
          name="Female Aged 20+, 3 Ethnic Groups, in linear regression";

  class age_decade ridreth2 educ3 health3 married;
  tables age_decade ridreth2 educ3 health3 married;
  var bmxbmi;

  setenv colwidth=10 decwidth=4 labwidth=32;
  print nsum mean semean / nsumfmt=f6.0 style=nchs;
  rformat health3 health.;
  rformat married yesno.;
  rformat ridreth2 race.;
  rformat educ3 educ.;
  rtitle "Mean BMI by Age Decade, Race/Eth, Education, Health Status, and Marital
          Status";
  rfootnote "NHANES 1999-2004";
```

Exhibit 2. First Page of DESCRIPT Output (*.lst file)

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


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method,
Assuming a With Replacement (WR) Design
    Sample Weight: WTMEC6YR
    Stratification Variables(s): SDMVSTRA
    Primary Sampling Unit: SDMVPSU


Number of observations read    :  29402    Weighted count :281175748
Number of observations skipped :   1724
(WEIGHT variable nonpositive)
Observations in subpopulation  :   4162    Weighted count : 58735489
Denominator degrees of freedom :     44
```

Exhibits 2 indicates that SUDAAN read in 29,402 adults from the data set with a positive value for the weight variable WTMEC6YR, and an additional 1,724 with a zero value for WTMEC6YR. These 1,724 subjects did not participate in the examination component of NHANES, only in the home interview component. The 29,402 examined adults make inference to an estimated adult (aged 17 years and older) population of 281,175,748 (sum of the weight variable WTMEC6YR over the 29,402 adults).

There were 4,162 adults in the subpopulation defined as female, aged 20 years and older, either non-Hispanic white, non-Hispanic black, or Mexican-American, and having non-missing values for each of the following variables: BMI, education, self-rated health status, and marital status. These 4,162 sample adults make inference to an estimated population of 58,735,489.

The denominator degrees of freedom (DDF) is calculated for continuous NHANES as 87 “pseudo-PSUs” minus 43 “pseudo-strata” = 44.

Exhibit 3 through ***Exhibit 7*** contain the frequency distributions for all variables contained on the CLASS statement.

Exhibit 3. Frequencies for CLASS Variable AGE_DECADE

```
Frequencies and Values for CLASS Variables
----------------------------------
AGE_DECADE      Frequency    Value
----------------------------------
Ordered
  Position:
  1                   825       20
Ordered
  Position:
  2                   713       30
Ordered
  Position:
  3                   672       40
Ordered
  Position:
  4                   512       50
Ordered
  Position:
  5                   651       60
Ordered
  Position:
  6                   424       70
Ordered
  Position:
  7                   365       80
----------------------------------
```

Exhibit 4. Frequencies for CLASS Variable Race/Ethnicity

```
Frequencies and Values for CLASS Variables
---------------------------------------------
Linked NH3
  Race/Ethn-
  icity -
  Recode        Frequency               Value
---------------------------------------------
Ordered
  Position:
  1                  2416            NH White
Ordered
  Position:
  2                   852            NH Black
Ordered
  Position:
  3                   894    Mexican American
---------------------------------------------
```

Exhibit 5. Frequencies for CLASS Variable EDUC3

```
Frequencies and Values for CLASS Variables
-----------------------------------------
EDUC3           Frequency           Value
-----------------------------------------
Ordered
  Position:
  1                  2166      HS or Less
Ordered
  Position:
  2                  1207    Some College
Ordered
  Position:
  3                   789        College+
-----------------------------------------
```

Exhibit 6. Frequencies for CLASS Variable HEALTH3

```
Frequencies and Values for CLASS Variables
-------------------------------------------
HEALTH3         Frequency             Value
-------------------------------------------
Ordered
  Position:
  1                   489         Excellent
Ordered
  Position:
  2                  2701    Very Good/Good
Ordered
  Position:
  3                   972         Fair/Poor
-------------------------------------------
```

Exhibit 7. Frequencies for CLASS Variable MARRIED

```
Frequencies and Values for CLASS Variables
----------------------------------
MARRIED         Frequency    Value
----------------------------------
Ordered
  Position:
  1                  2409    1=Yes
Ordered
  Position:
  2                  1753     2=No
----------------------------------
```

Exhibit 8. DESCRIPT results for BMI within Age Decades

```
Variance Estimation Method: Taylor Series (WR)
For Subpopulation: Female Aged 20+, 3 Ethnic Groups, in linear regression

Mean BMI by Age Decade, Race/Eth, Education, Health Status, and Marital Status
--------------------------------------------------------------------
Variable                           Sample
   AGE_DECADE                      Size           Mean      SE Mean
--------------------------------------------------------------------
Body Mass Index (kg/m**2)
   Total                             4162      28.3652       0.1862
   20                                 825      26.9669       0.3162
   30                                 713      27.9631       0.3900
   40                                 672      29.0189       0.4314
   50                                 512      29.2927       0.4209
   60                                 651      29.3872       0.2814
   70                                 424      28.2664       0.2671
   80                                 365      26.5070       0.4845
--------------------------------------------------------------------
NHANES 1999-2004
```

As seen from ***Exhibit 8***, mean BMI for women seems to increase with age until the decade of the sixties, but then begins to decrease with age. Thus, the linear regression model will contain terms for linear and quadratic age.

Exhibit 9. DESCRIPT Results for BMI within Race/Ethnicity Classes

```
Variance Estimation Method: Taylor Series (WR)
For Subpopulation: Female Aged 20+, 3 Ethnic Groups, in linear regression

Mean BMI by Age Decade, Race/Eth, Education, Health Status, and Marital Status
--------------------------------------------------------------------
Variable
   Linked NH3 Race/Ethnicity -     Sample
     Recode                        Size           Mean      SE Mean
--------------------------------------------------------------------
Body Mass Index (kg/m**2)
   Total                             4162      28.3652       0.1862
   NH White                          2416      27.8207       0.2067
   NH Black                           852      31.3851       0.2973
   Mexican American                   894      29.3901       0.3884
--------------------------------------------------------------------
NHANES 1999-2004
```

Non-Hispanic white women (RIDRETH2=1) appear to have a lower BMI than women in the other two race/ethnicity groups (***Exhibit 9***).

Exhibit 10. DESCRIPT results for BMI within Education Levels

```
Variance Estimation Method: Taylor Series (WR)
For Subpopulation: Female Aged 20+, 3 Ethnic Groups, in linear regression

Mean BMI by Age Decade, Race/Eth, Education, Health Status, and Marital Status
-------------------------------------------------------------------
Variable                           Sample
   EDUC3                           Size           Mean      SE Mean
-------------------------------------------------------------------
Body Mass Index (kg/m**2)
   Total                             4162      28.3652       0.1862
   HS or Less                        2166      29.0473       0.1984
   Some College                      1207      28.5871       0.3067
   College+                           789      26.8514       0.3323
-------------------------------------------------------------------
NHANES 1999-2004
```

Mean BMI seems inversely related to years of education for adult females (***Exhibit 10***).

Exhibit 11. DESCRIPT results for BMI within Health Status

```
Variance Estimation Method: Taylor Series (WR)
For Subpopulation: Female Aged 20+, 3 Ethnic Groups, in linear regression

Mean BMI by Age Decade, Race/Eth, Education, Health Status, and Marital Status
-------------------------------------------------------------------
Variable                           Sample
   HEALTH3                         Size           Mean      SE Mean
-------------------------------------------------------------------
Body Mass Index (kg/m**2)
   Total                             4162      28.3652       0.1862
   Excellent                          489      25.3628       0.2600
   Very Good/Good                    2701      28.3645       0.2021
   Fair/Poor                          972      30.8015       0.3044
-------------------------------------------------------------------
NHANES 1999-2004
```

Women with a better self-rated health status appear to have a lower mean BMI (***Exhibit 11***).

Exhibit 12. DESCRIPT results for BMI within Marital Status

```
Variance Estimation Method: Taylor Series (WR)
For Subpopulation: Female Aged 20+, 3 Ethnic Groups, in linear regression

Mean BMI by Age Decade, Race/Eth, Education, Health Status, and Marital Status
-------------------------------------------------------------------
Variable                           Sample
   MARRIED                         Size           Mean      SE Mean
-------------------------------------------------------------------
Body Mass Index (kg/m**2)
   Total                             4162      28.3652       0.1862
   1=Yes                             2409      28.1776       0.1860
   2=No                              1753      28.6759       0.3339
-------------------------------------------------------------------
NHANES 1999-2004
```

Married women (MARRIED=1) appear to have the same mean BMI as unmarried women (***Exhibit 12***).

Next, we use the REGRESS procedure to perform the linear regression of BMI on race/ethnicity and several covariates. Age will be treated as a continuous variable in the model, specifically centered age (AGE_C), defined as age in years less weighted mean age, and the square of centered age (AGE_C_SQ):

```
age_c = ridageyr - 47.35;
age_c_sq = age_c*age_c;
```

How did we determine the weighted mean age for this subpopulation to be 47.35 yrs? ***Exhibit 13*** and ***Exhibit 14*** contain the DESCRIPT code and results.

Exhibit 13. DESCRIPT Code to Obtain Weighted Mean Age

```
proc descript data=one filetype=sas design=wr nomarg;
  NEST sdmvstra sdmvpsu;
  WEIGHT wtmec6yr;

  subpopx ridageyr ge 20 and riagendr=2 and ridreth2 in(1,2,3) and educ3 in(1,2,3)
          and (1 le hsd010 le 5) and (1 le dmdmartl le 6) and bmxbmi>0 /
          name="Females Aged 20+, 3 Ethnic Groups, in Linear Regression";
  var ridageyr;
  print nsum wsum mean semean / nsumfmt=f6.0 wsumfmt=f8.0 style=nchs;
```

Exhibit 14. DESCRIPT Results for Weighted Mean Age

```
Variance Estimation Method: Taylor Series (WR)
For Subpopulation: Females Aged 20+, 3 Ethnic Groups, in Linear Regression

Weighted Mean Age for the SUBPOP Used in Regression Analysis
-----------------------------------------------------------------------------
Variable                              Sample   Weighted
   SUDAAN Reserved Variable One       Size     Size           Mean    SE Mean
-----------------------------------------------------------------------------
Age at Screening Adjudicated -
  Recode
   1                                    4162   58735489     47.352      0.407
-----------------------------------------------------------------------------
NHANES 1999-2004
```

Exhibit 15 contains the SAS-callable SUDAAN code for the linear regression model in REGRESS. The NEST and WEIGHT statements and DESIGN=WR are the same as in the previous DESCRIPT program. The SUBPOPX statement restricts the analysis to women aged 20 and older of the three specified race/ethnic groups.

The MODEL statement includes the following categorical variables: race/ethnicity at three levels (RIDRETH2); educational level (EDUC3) at three levels; health status (HEALTH3) at three levels; and marital status (MARRIED) at two levels. These categorical variables are on the CLASS statement.

Based on the previous DESCRIPT analysis of mean BMI by age decade, both continuous AGE_C and AGE_C_SQ (square of AGE_C) variables are in the model. Note that SUDAAN does not form the interaction of two continuous variables (i.e., AGE_C*AGE_C is not a valid term in the model). Thus, AGE_C_SQ was formed in the SAS data step before running REGRESS.

The REFLEVEL statement defines the reference group for each of the categorical variables. All categorical variables use the first level as the reference cell.

The TEST statement requests SATADJCHI, the Wald chi-square statistic with Satterthwaite correction for the degrees of freedom. This test was chosen because the continuous NHANES 1999-2004 has only 44 denominator degrees of freedom.

Exhibit 15. SAS-Callable SUDAAN Code for REGRESS Procedure

```
PROC REGRESS DATA=one FILETYPE=SAS DESIGN=WR;
  NEST sdmvstra sdmvpsu;
  WEIGHT wtmec6yr;

  SUBPOPX ridageyr ge 20 and riagendr=2 and ridreth2 in(1,2,3) and educ3 in(1,2,3)/
          NAME="Females Aged 20+ yrs, 3 Ethnic Groups";

  CLASS ridreth2 educ3 health3 married;
  REFLEVEL ridreth2=1 educ3=1 health3=1 married=1;
  MODEL bmxbmi = ridreth2 educ3 health3 married age_c age_c_sq;
  TEST satadjchi;

  EFFECTS married educ3 / name="Chunk Test - Married, Educ";
  EFFECTS age_c age_c_sq / name="Chunk Test - Age Lin, Quad";

  EFFECTS ridreth2 = (0 1 -1) / name="NH Black vs. Mex American";
  EFFECTS health3 = (0 1 -1) / name="VG/G vs. Fair/Poor Health";
  EFFECTS educ3 = (0 1 -1) / name="Some College vs. College+";

  EFFECTS health3 = (1 0 -1) / name="Health Linear Trend";
  EFFECTS health3 = (-1 2 -1) / name="Health Deviation from Lin Trend";

  EFFECTS educ3 = (1 0 -1) / name="Educ Linear Trend";
  EFFECTS educ3 = (-1 2 -1) / name="Educ Deviation from Lin Trend";

  CONDMARG ridreth2 health3;
  PREDMARG ridreth2 health3;
  LSMEANS  ridreth2 health3;

  COND_EFF ridreth2 = (-1 1 0) / name="NH Black vs. White";
  COND_EFF ridreth2 = (-1 0 1) / name="Mex American vs. NH White";
  COND_EFF ridreth2 = (0 1 -1) / name="NH Black vs. Mex American";

  SETENV COLSPCE=1 labwidth=25 colwidth=7 decwidth=4;
  PRINT / betas=default t_betafmt=f6.2;

  SETENV TOPMGN=0 COLSPCE=1 LABWIDTH=32 decwidth=2;
  PRINT / tests=default dffmt=f7.0 satadchifmt=f8.2 satadchpfmt=f7.4;

  SETENV TOPMGN=0 COLSPCE=1 LABWIDTH=22 colwidth=6 decwidth=3;
  PRINT / lsmeans=default cond_mrg=default pred_mrg=default predmrgfmt=f9.3
          t_prdmrgfmt=f9.2 condmrgfmt=f11.3 t_cndmrgfmt=f9.2 t_lsmeanfmt=f9.2
          p_lsmeanfmt=f7.4 p_cndmrgfmt=f7.4 p_prdmrgfmt=f7.4;

  SETENV TOPMGN=0 COLSPCE=1 LABWIDTH=28 colwidth=8 decwidth=3;
  PRINT / cnmgcons=default t_pmconfmt=f8.2 p_pmconfmt=f7.4 t_cmconfmt=f8.2
          p_cmconfmt=f7.4;

  rformat health3 health.;
  rformat married yesno.;
  rformat ridreth2 race.;
  rformat educ3 educ.;
  rlabel age_c = "Linear Age (centered)";
  rlabel age_c_sq = "Quadratic Age (centered)";

  RTITLE "Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and
          Age";
  RFOOTNOTE "NHANES 1999-2004" ;
```

Several EFFECTS statements are included. Each statement tests a null hypothesis about specified linear combinations of the population regression coefficients. β is defined as the vector of population regression coefficients, and in this model, is of size (14 x 1) (*i.e.*, 14 rows and 1 column). The vector β includes 10 estimable regression coefficients; four regression coefficients are defined to be 0 because of the parameterization used for the reference cells (one reference cell for each of four independent categorical variables). The order of the regression coefficients in the vector β is determined by the order of the variables on the MODEL statement.

Each EFFECTS statement specifies a contrast matrix C of size (r x 14) (i.e., r rows and 14 columns). The null hypothesis for each EFFECTS statement is that C□ = 0. The degrees of freedom (DF) for the EFFECTS statement (i.e., for the C matrix) is less than or equal to r. Each EFFECTS statement below is labeled with the NAME option to make the printout easier to read.

- EFFECTS #1—"chunk" test of the joint effect of marital status and education; it has 2 df. The null hypothesis is that all regression coefficients for the main effects of marital status and education are equal to 0, conditional on all other variables being in the model.

- EFFECTS #2—"chunk" test of the joint effect of linear and quadratic age, having 2 df; it tests the null hypothesis that the two regression coefficients for linear and quadratic age are equal to 0, conditional on all other variables (except these two) being in the model.

- EFFECTS #3—tests the null hypothesis that the mean BMI among non-Hispanic black women is equal to that of Mexican-American women, conditional on all other variables in the model; it has 1 df.

- EFFECTS #4—tests the null hypothesis that the two regression coefficients for levels 2 and 3 of health status (very good/good vs. fair/poor) are equal to each other, conditional on all other variables in the model; it has 1 df.

- EFFECTS #5—tests the null hypothesis that the two regression coefficients for levels 2 and 3 of education (some college vs. college+) are equal to each other, conditional on all other variables in the model; it has 1 df.

- EFFECTS #6—tests for a linear trend (over three levels) across the health status variable; it has 1 df.

- EFFECTS #7—tests for a deviation from a linear trend across health status; it has 1 df. The two contrasts in statements #6 and #7 are not orthogonal in the weighted data case, but do provide tests of linear and deviation from linear trend, if we make the assumption that the three levels of the qualitative health status variable are equally spaced. These linear contrasts use the orthogonal polynomials for linear or quadratic trend when the classification variable is at three levels. All tests are conditional on all other variables in the model.

- EFFECTS #8—tests for a linear trend (over three levels) of education level, conditional on all other variables in the model; it has 1 df.

- EFFECTS #9—tests for a deviation from a linear trend in education; it has 1 df. The same comments about the contrasts that were made in the previous test for deviation from a linear trend (EFFECTS statement #7) apply here as well.

The CONDMARG statement requests the conditional marginal mean for BMI, with estimated standard error and 95% confidence limits, for the same two categorical variables (race/ethnicity and health status). The CONDMARG is equivalent to LSMEANS and should be used in place of it. LSMEANS remains in REGRESS only for compatibility with earlier releases of SUDAAN. The three COND_EFF statements request that the conditional marginal mean for BMI be compared pairwise for each level of race/ethnicity.

The PREDMARG statement requests the predicted marginal mean for BMI, with estimated standard error, for each level of race/ethnicity and for each level of health status. The first step in calculating the predicted marginal mean for a given level of a categorical variable is to use the estimated regression equation to predict BMI for each observation, setting the value of the specified categorical variable at a given level but using the observation’s values for all other individual covariates. Then, the weighted mean of the predicted BMI values yields the predicted marginal mean.

In the case of linear regression, the conditional marginal mean and the predicted marginal mean are mathematically equivalent, though their variances and confidence limits are different.

The LSMEANS statement requests the estimated least squares mean and standard error for BMI for each level of race/ethnicity and for each level of health status. For a given value or level of race/ethnicity (or of health status), the least squares mean is calculated by using the estimated regression equation with the intercept, the regression coefficient for the given level of the categorical variable specified, and then substituting in the estimated mean for all continuous covariates and the estimated percentage distribution for all other categorical covariates.

Exhibit 16. First Page of REGRESS Results (*.lst file)

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


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method,
Assuming a With Replacement (WR) Design
    Sample Weight: WTMEC6YR
    Stratification Variables(s): SDMVSTRA
    Primary Sampling Unit: SDMVPSU


Number of observations read       :  29402    Weighted count:281175748
Number of observations skipped    :   1724
(WEIGHT variable nonpositive)
Observations in subpopulation     :   6881    Weighted count: 94037229
Observations used in the analysis :   4162    Weighted count: 58735489
Denominator degrees of freedom    :     44


Maximum number of estimable parameters for the model is 10

File ONE contains   87 Clusters
  60 clusters were used to fit the model
Maximum cluster size is 116 records
Minimum cluster size is  25 records

Weighted mean response is 28.365234
Multiple R-Square for the dependent variable BMXBMI: 0.088408
```

REGRESS identified 6,881 individuals in the subpopulation defined as females, aged 20 and older, and either non-Hispanic white or non-Hispanic black or Mexican-American (see ***Exhibit 16***). (This number differs from the number of the subpopulation stated in the DESCRIPT results because REGRESS counts all records in the subpopulation, including those whose weights were zero and those with missing dependent or independent variable values). Of these 6,881 women, 4,162 were used in the linear regression analysis. Thus, 2,719 women did not have full information on all model variables and were excluded from the analysis. This analysis makes the assumption that these subjects are missing at random.

SUDAAN estimated 10 population parameters (regression coefficients) for the model, consistent with the earlier comments about the vector β . SUDAAN used 60 clusters to fit the linear regression model. For the 4,162 women in the analysis, the minimum number of women in a cluster (PSU) was 25, and the maximum number of women was 116. The multiple R-square for the fitted model is 0.088408; this is calculated as the square of the weighted correlation coefficient between the observed BMI and the model-fitted BMI.

Exhibit 17. Regression Coefficient Estimates

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
------------------------------------------------------------------------------------
Independent Variables and                    Lower      Upper
  Effects                                    95%        95%                  P-value
                            Beta             Limit      Limit     T-Test     T-Test
                            Coeff.    SE Beta Beta      Beta      B=0        B=0
------------------------------------------------------------------------------------
Intercept                   26.3921    0.3775   25.6314   27.1529    69.92    0.0000
Linked NH3 Race/Ethnicity
  - Recode
  NH White                   0.0000    0.0000    0.0000    0.0000      .        .
  NH Black                   2.9602    0.3377    2.2796    3.6407     8.77    0.0000
  Mexican American           1.0565    0.4147    0.2208    1.8922     2.55    0.0144
EDUC3
  HS or Less                 0.0000    0.0000    0.0000    0.0000      .        .
  Some College               0.0231    0.2891   -0.5596    0.6058     0.08    0.9368
  College+                  -1.2793    0.3647   -2.0143   -0.5443    -3.51    0.0011
HEALTH3
  Excellent                  0.0000    0.0000    0.0000    0.0000      .        .
  Very Good/Good             2.7074    0.2698    2.1636    3.2513    10.03    0.0000
  Fair/Poor                  4.4618    0.4924    3.4695    5.4541     9.06    0.0000
MARRIED
  1=Yes                      0.0000    0.0000    0.0000    0.0000      .        .
  2=No                       0.4111    0.4271   -0.4498    1.2719     0.96    0.3411
Linear Age (centered)        0.0367    0.0086    0.0194    0.0540     4.27    0.0001
Quadratic Age (centered)    -0.0032    0.0004   -0.0040   -0.0024    -8.38    0.0000
------------------------------------------------------------------------------------
NHANES 1999-2004
```

Exhibit 17 displays estimates of each regression coefficient, with its estimated standard error and 95% confidence interval. The last columns in this table show the results of the tests for the null hypothesis that the population regression coefficient equals 0. The estimated regression vector is of size (14 x 1). Note that four of the 14 regression coefficients are defined to be 0, leaving 10 to be estimated.

The two estimated regression coefficients for race/ethnicity are each significantly different from 0. Both are positive, indicating a higher predicted mean BMI for non-Hispanic black women (by 2.96) and for Mexican-American women (by 1.06), compared to non-Hispanic white women (the reference group), after adjusting for all other variables in the model.

For education level, only the College+ regression coefficient is significantly different from 0. The College+ coefficient is negative and significant, indicating a lower predicted mean BMI for college-educated women compared to women with a high school education or less, after adjusting for all other variables in the model. There is no significant difference between women with some college and those with high school education or less.

The two estimated regression coefficients for health status are each significantly different from 0. Both are positive, indicating a higher predicted mean BMI value for less healthy women, compared to women with excellent health, after adjusting for all other variables in the model.

The estimated regression coefficient for marital status is not significantly different from 0, indicating that married women and unmarried women do not differ on mean BMI, given all other variables in the model.

The estimated regression coefficients for linear age (AGE_C) and for quadratic age (AGE_C_SQ) are each significantly different from 0, given all other variables in the model.

Exhibit 18. ANOVA Table

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age

-------------------------------------------------------------------------
Contrast                           Degrees             S_waite    P-value
                                   of        S_waite   Adj        S_waite
                                   Freedom   Adj DF    ChiSq      ChiSq
-------------------------------------------------------------------------
OVERALL MODEL                           10      5.76   26702.73    0.0000
MODEL MINUS INTERCEPT                    9      5.58     174.63    0.0000
INTERCEPT                                .       .          .       .
RIDRETH2                                 2      1.96      79.44    0.0000
EDUC3                                    2      1.80      13.30    0.0010
HEALTH3                                  2      1.67      86.13    0.0000
MARRIED                                  1      1.00       0.93    0.3359
AGE_C                                    1      1.00      18.24    0.0000
AGE_C_SQ                                 1      1.00      70.20    0.0000

Chunk Test - Married, Educ               3      2.45      10.67    0.0080
Chunk Test - Age Lin, Quad               2      1.94      62.59    0.0000
NH Black vs. Mex American                1      1.00      12.79    0.0004
VG/G vs. Fair/Poor Health                1      1.00      21.54    0.0000
Some College vs. College+                1      1.00       9.03    0.0027
Health Linear Trend                      1      1.00      82.11    0.0000
Health Deviation from Lin Trend          1      1.00       4.81    0.0283
Educ Linear Trend                        1      1.00      12.30    0.0005
Educ Deviation from Lin Trend            1      1.00       4.29    0.0384
-------------------------------------------------------------------------
NHANES 1999-2004
```

Exhibit 18 is often called the "analysis of variance" (or ANOVA) table for the linear regression analysis. All tests in the table above, whether default from SUDAAN or user-requested with an EFFECTS statement, use an appropriate contrast matrix *C*. Each *C* matrix specifies a null hypothesis about linear combinations of the components of the β vector, conditional on all other variables in the model.

The first test of the overall model, with 10 degrees of freedom, tests the null hypothesis that all population regression coefficients are equal to 0, (i.e., $\beta = 0$). This null hypothesis is equivalent to stating that the population mean BMI is 0, clearly not of interest in this example. Not surprisingly, the null hypothesis is rejected.

The second test, "model minus intercept," tests the null hypothesis that all regression coefficients in the population are 0, except the intercept. This null hypothesis is rejected, indicating that at least some of the variables in the model are associated with BMI.

The next four tests look at each of the categorical variables in the model, conditional on all other terms being in the model. The test for race/ethnicity (RIDRETH2) tests the null hypothesis that the two estimated regression coefficients for race in the population are both equal to 0. This is equivalent to stating that all three race/ethnicity groups have the same mean BMI. This null hypothesis is rejected, consistent with the results of testing each regression coefficient individually in the preceding table. Similarly, educational level and health status are significantly associated with BMI, and marital status is not.

The next null hypothesis tested is that the regression coefficient for (continuous) linear age is equal to 0. This null hypothesis is rejected. The next test is on the quadratic regression coefficient, and that null hypothesis is rejected. Thus, both the linear and quadratic terms for age are important in the model, conditional on all other variables in the model.

Next, we follow the results of the user requested EFFECTS statements.
The null hypothesis for EFFECTS #1 is rejected. Thus, there is evidence to question the assumption that the regression coefficients for the combined effect of marital status and education are equal to 0, conditional on all other variables (other than these two) being in the model.

EFFECTS #2 tests the null hypothesis that the regression coefficients for linear and quadratic age are both equal to 0, conditional on the remaining variables in the model. The null hypothesis is rejected, as expected. Clearly, age is an important correlate of BMI.

EFFECTS #3 compares the estimated regression coefficients for non-Hispanic black women (2.96) and Mexican-American women (1.06). The null hypothesis is rejected, indicating that the two corresponding regression coefficients in the population are not equal to each other. Thus, non-Hispanic black women have a higher mean BMI than Mexican-American women.

EFFECTS #4 tests the null hypothesis that the two regression coefficients for the lower levels of health are equal to each other. The estimated regression coefficient is 2.71 for women with very good or good health and is 4.46 for women with fair or poor health. The null hypothesis is rejected, indicating that women with fair/poor self-rated health status have a significantly higher mean BMI than women with good or very good health status. In addition, as shown by the *t*-test on individual regression coefficients, women in either of these health status levels have a significantly higher mean BMI compared to women with excellent health. Thus, over the three levels of health status, better self-reported health status is significantly associated with lower BMI.

EFFECTS #5 tests the null hypothesis that the two regression coefficients for the two higher levels of education are equal to each other. The estimated regression coefficient is 0.0231 for women with some college education and -1.28 for women with at least a college degree. The null hypothesis is rejected, indicating that women with at least a college education have a significantly lower mean BMI than those with some college. Thus, over the three levels of education, only the highest level of education significantly reduces BMI.

The null hypothesis for EFFECTS #6 is rejected, indicating a significant linear trend component to health status. This is consistent with the results from DESCRIPT, which showed a decreasing BMI as health status improved. The null hypothesis for EFFECTS #7 is also rejected, indicating a significant deviance from linear trend.

The null hypothesis for EFFECTS #8 is rejected, indicating a significant linear trend component to education. The null hypothesis for EFFECTS #9 is also rejected, indicating a significant deviance from linear trend. This is due to the education threshold effect reported above—only the highest level of education significantly reduced BMI compared to lower educated women.

The predicted and conditional marginals and least squares means within levels of race/ethnicity and health status follow next:

Exhibit 19. Predicted Marginals

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
-----------------------------------------------------------------------------------
Predicted Marginal #1                             Lower     Upper
                         Predicted                95%       95%
                         Marginal         SE      Limit     Limit      T:Marg=0    P-value
-----------------------------------------------------------------------------------
Linked NH3
  Race/Ethnicity -
  Recode
  NH White                  27.930     0.204    27.519    28.341      137.10     0.0000
  NH Black                  30.890     0.314    30.257    31.523       98.40     0.0000
  Mexican American          28.986     0.433    28.114    29.859       66.96     0.0000
HEALTH3
  Excellent                 25.732     0.287    25.155    26.310       89.79     0.0000
  Very Good/Good            28.440     0.204    28.028    28.852      139.16     0.0000
  Fair/Poor                 30.194     0.331    29.527    30.861       91.19     0.0000
-----------------------------------------------------------------------------------
NHANES 1999-2004
```

Exhibit 20. Conditional Marginals

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
-----------------------------------------------------------------------------------
Conditional Marginal
  #1                                              Lower     Upper
                       Conditional                95%       95%
                       Marginal           SE      Limit     Limit      T:Marg=0    P-value
-----------------------------------------------------------------------------------
Linked NH3
  Race/Ethnicity -
  Recode
  NH White                  27.930     0.191    27.545    28.315      146.35     0.0000
  NH Black                  30.890     0.310    30.266    31.514       99.77     0.0000
  Mexican American          28.986     0.430    28.119    29.854       67.34     0.0000
HEALTH3
  Excellent                 25.732     0.283    25.161    26.303       90.83     0.0000
  Very Good/Good            28.440     0.206    28.024    28.855      138.03     0.0000
  Fair/Poor                 30.194     0.340    29.509    30.879       88.81     0.0000
-----------------------------------------------------------------------------------
NHANES 1999-2004
```

Exhibit 21. Least Squares Means

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
--------------------------------------------------------------------------------
LS MEAN #1                                Lower    Upper
                         LS               95%      95%
                         Mean         SE   Limit    Limit     T:Marg=0   P-value
--------------------------------------------------------------------------------
Linked NH3
  Race/Ethnicity -
  Recode
  NH White               27.930    0.191   27.545   28.315      146.35    0.0000
  NH Black               30.890    0.310   30.266   31.514       99.77    0.0000
  Mexican American       28.986    0.430   28.119   29.854       67.34    0.0000
HEALTH3
  Excellent              25.732    0.283   25.161   26.303       90.83    0.0000
  Very Good/Good         28.440    0.206   28.024   28.855      138.03    0.0000
  Fair/Poor              30.194    0.340   29.509   30.879       88.81    0.0000
--------------------------------------------------------------------------------
NHANES 1999-2004
```

The least squares means (***Exhibit 21***) and conditional marginal means (***Exhibit 20***) yield equivalent estimated means, standard errors, and confidence limits. This is expected, since they are mathematically equivalent.

In addition, the predicted marginal mean (***Exhibit 19***) is equal to the conditional marginal mean (***Exhibit 20***), which is only true for the linear regression model. However, the estimated standard errors and confidence limits for the predicted marginal mean still differ from those of the conditional marginal mean. The equality of the conditional and predicted marginal means does not hold for nonlinear models, such as logistic regression.

The marginal means for BMI (whether predicted, conditional, or least squares) are 27.930 for non-Hispanic white adult women, 30.890 for non-Hispanic adult black women, and 28.986 for Mexican-American adult women. The difference between the marginal mean for non-Hispanic black and non-Hispanic white women is (30.89 – 27.93) = 2.96, which is equal to the estimated regression coefficient for the indicator variable for non-Hispanic black. Similarly, the difference between the marginal mean for Mexican-American and non-Hispanic white is (28.986 – 27.930) = 1.056, which is equal to the estimated regression coefficient for Mexican-American women.

Exhibit 22 compares the adjusted or marginal mean BMI (conditional, predicted, or least squares) with the unadjusted mean BMI (from the DESCRIPT results). Even after adjusting for other variables in the linear regression model, the overall pattern is the same.

Exhibit 22. Adjusted and Unadjusted Mean BMI, by Race/Ethnicity, Women Aged 20+ Years

Race/Ethnicity	Adjusted Mean BMI	Unadjusted Mean BMI
Non-Hispanic white	27.930	27.821
Non-Hispanic black	30.890	31.385
Mexican American	28.986	29.390

The marginal means for BMI (whether predicted, conditional or least squares) are 25.732 for women in excellent health, 28.440 for women in very good or good health, and 30.194 for women in fair or poor health. As for race/ethnicity, the differences in the adjusted mean BMIs for women in excellent health compared to the other two health status levels is equal to the estimated regression coefficients for health status.

The null hypothesis of the user-requested COND_EFF statement is that the conditional marginal mean BMI (i.e., adjusted for all other variables in the model) is the same for non-Hispanic white vs. black women. The results displayed in ***Exhibit 23*** indicate that the null hypothesis is rejected. Note that the estimated contrast and its estimated standard error are equal to the estimated regression coefficient for non-Hispanic black women and its estimated standard error, respectively (see ***Exhibit 17***). Note also that the *t*-statistic and the *p*-value (directly above) are equivalent to the same quantities in the output of estimated regression coefficients. These equalities hold only for linear regression; they do not hold for nonlinear models such as logistic regression.

Exhibit 23. Contrasted Conditional Marginals (Black vs. White)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
-----------------------------------------------------------------------
Contrasted Conditional
  Marginal #1                  CONDMARG
                               Contrast         SE     T-Stat   P-value
-----------------------------------------------------------------------
NH Black vs. White                2.960      0.338       8.77    0.0000
-----------------------------------------------------------------------
NHANES 1999-2004
```

Exhibit 24. Contrasted Conditional Marginals (Mex Amer vs. White)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
-----------------------------------------------------------------------
Contrasted Conditional
  Marginal #2                  CONDMARG
                               Contrast         SE     T-Stat   P-value
-----------------------------------------------------------------------
Mex American vs. NH White         1.056      0.415       2.55    0.0144
-----------------------------------------------------------------------
NHANES 1999-2004
```

The null hypothesis of equality of the conditional marginal mean BMI for Mexican-American women and non-Hispanic white women is rejected (***Exhibit 24***). The results are identical to those of estimated regression coefficients (***Exhibit 17***), because we are fitting a linear model.

Exhibit 25. Contrasted Conditional Marginals (Black vs. Mex Amer)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMXBMI: Body Mass Index (kg/m**2)
For Subpopulation: Females Aged 20+ yrs, 3 Ethnic Groups

Linear Reg of BMI on Race/Eth, Educ, Health Status, Marital Status, and Age
-----------------------------------------------------------------------
Contrasted Conditional
  Marginal #3                CONDMARG
                             Contrast         SE     T-Stat   P-value
-----------------------------------------------------------------------
NH Black vs. Mex American       1.904      0.532       3.58    0.0009
-----------------------------------------------------------------------
NHANES 1999-2004
```

The conditional marginal mean BMI differs significantly for non-Hispanic black women vs. Mexican-American women (***Exhibit 25***). This is not one of the default comparisons contained in the regression coefficient table, but EFFECTS #3 in the ANOVA table (***Exhibit 18***) makes this comparison. Note that the square of the *t*-statistic (3.58) directly above equals the chi-square statistic (12.8) calculated for EFFECTS #3 in the ANOVA table. This is true only for the linear model.

A summary statement of the fitted model is as follows. The covariate marital status is not significantly related to mean BMI, given all other variables in the model. Conditional on all other variables in the model, lower mean BMI is significantly associated with higher education and with better self-reported health status. Conditional on all other variables in the model, age is significantly related to mean BMI in a quadratic relationship. The effect of race/ethnicity on mean BMI, the main question of the analysis, is significant. Controlling on all other variables in the model, non-Hispanic white women have a significantly lower mean BMI than non-Hispanic black women (by 2.96 units) and Mexican-American women (by 1.06 units), and Mexican-American women have a significantly lower mean BMI than non-Hispanic black women (by 1.90 units).