Dummy Coding for Dummies

Kathryn Martin, Maternal, Child and Adolescent Health Program,
California Department of Public Health

ABSTRACT

There are a number of ways to incorporate categorical variables into regression analysis in SAS®, one of which is to create dummy variables in the DATA step. Alternatively, parameterization can be automated in PROC LOGISTIC using the CLASS statement. However, it is important to understand SAS default settings and how they affect the interpretation of results. Understanding how variables are coded also allows analysts to make comparisons across different categories of one or more variables using the CONTRAST statement in PROC LOGISITIC. This paper will review aspects of dummy coding in the DATA step and PROC LOGISTIC, SAS default settings, interpretation of regression parameters, and hypothesis testing using the CONTRAST statement.

INTRODUCTION

Parameterization is the process of defining variables in regression models and dummy coding is one way of using categorical variables as predictors in regression modeling. Each observation is assigned a value of 0 or 1 for the categorical variable of interest and these levels are compared with respect to a continuous or categorical outcome using methods like linear or logistic regression. In SAS, dummy coding can be done within the DATA step or in procedures like PROC LOGISTIC using the CLASS statement. The CONTRAST statement, which is available in PROC LOGISTIC, can be used to make comparisons across levels of one or more variables. The goal of this paper is to provide an overview of parameterization in SAS. The basics of making comparisons across levels of a covariate, SAS default settings, interpretation of regression estimates, and hypothesis testing will be reviewed.

SAMPLE DATA

The data set below contains information on drinking during the first trimester of pregnancy (1=Yes, 0 = No) among a hypothetical sample of recent mothers generated by the author for the purposes of this paper. The covariates (i.e. independent variables or predictors), of interest are EDUCATION (1= Less than high school, 2=High school graduate, 3=Some or completed college), AGE ('<35' or '35+'), and INCOME as a percent of the Federal Poverty Level (FPL), which is a continuous variable in the data set. These sample data will be used in the examples in the following sections.

```
data example;
   input agecat $ income education drank @@;

   if agecat = '35+' then age35plus = 1;
     else if agecat = '<35' then age35plus = 0;

   if education = 3 then college = 1;
     else if education in(1,2) then college = 0;

   if education = 2 then highschool = 1;
     else if education in(1,3) then highschool = 0;

   if agecat = '35+' then age35plusE = 1;
     else if agecat = '<35' then age35plusE = -1;

   if education = 3 then collegeE = 1;
     else if education = 2 then collegeE = 0;
     else if education = 1 then collegeE = -1;

   if education = 2 then highschoolE = 1;
     else if education = 3 then highschoolE = 0;
     else if education = 1 then highschoolE = -1;

datalines;
```

```
35+ 201 3 1   <35 205 3 1   35+ 165 3 1   <35 120 . 1
35+ 225 3 1   <35 230 2 1   35+ 80  3 1   <35 164 1 0
35+ 236 3 1   <35 260 1 0   35+ 165 2 1   <35 22  2 0
35+ 224 3 1   <35 320 1 0   .   30  2 0   <35 130 2 1
35+ 264 1 1   <35 230 1 0   35+ 195 2 0   <35 195 1 1
35+ 398 2 1   <35 467 2 1   35+ 105 2 0   <35 45  2 0
.   326 3 1   <35 368 2 0   35+ 164 3 1   <35 50  1 0
35+ 348 3 1   <35 202 2 0   35+ 160 3 0   <35 68  1 0
35+ 301 2 1   <35 267 2 0   35+ 150 2 0   <35 95  1 0
35+ 200 1 1   <35 234 2 1   35+ 159 1 1   <35 84  3 0
35+ 401 1 0   <35 306 1 0   35+ 124 2 0   <35 25  1 0
35+ 465 2 0   <35 340 . 0   35+ 195 3 0   <35 38  1 0
35+ 315 2 0   <35 207 2 0   35+ 45  3 0   <35 75  1 0
35+ 348 2 0   <35 208 3 0   35+ 175 2 0   <35 86  1 0
35+ 267 2 0   <35 230 2 1   35+ 185 . 1   <35 96  2 1
35+ 402 2 1   <35 300 1 1   35+ 160 2 1   <35 134 2 1
35+ 378 3 1   <35 405 1 0   35+ 120 2 1   <35 38  3 1
35+ 267 2 1   <35 200 1 0   35+ 115 2 0   <35 110 3 1
35+ 208 3 1   <35 420 2 0   35+ 126 3 0   <35 .   3 1
35+ 367 3 1   <35 370 2 0   35+ 156 3 0   <35 160 1 0
35+ 349 2 1   <35 .   2 0   35+ 130 2 0   <35 102 1 1
35+ 350 3 0   <35 340 2 0   35+ 138 2 0   <35 64  2 0
35+ 415 2 0   <35 372 2 0   35+ 168 3 0   <35 95  1 0
35+ 420 3 0   <35 245 3 0   35+ 195 3 0   <35 90  1 0
35+ 370 3 1   <35 302 1 0   35+ 104 3 0   <35 32  1 0
35+ 380 1 1   <35 260 1 0   35+ 95  3 0   <35 64  1 0
35+ .   2 1   <35 210 1 0   35+ 185 2 0   <35 95  2 1
35+ 315 3 0   <35 260 2 0   35+ 65  1 0   <35 86  3 0
35+ 360 2 0   <35 280 3 0   35+ 95  1 0   <35 98  1 0
;
```

DUMMY CODING WITHIN THE DATA STEP

Dummy coding is often done within the DATA step, as shown above. Let's review how these variables are used in PROC LOGISITIC to model a categorical outcome and in PROC REG if the outcome is continuous.

EXAMPLE 1

The first dummy variable created is AGE35PLUS, which equals 1 when AGECAT = '35+' and 0 when AGECAT = '<35'. Women with a missing value for AGECAT are also coded as missing for AGE35PLUS.

```
if agecat = '35+' then age35plus = 1;
  else if agecat = '<35' then age35plus = 0;
```

The association between age and drinking during the first trimester of pregnancy is assessed in the logistic regression below. By default SAS models the probability of the response option 0, using 1 as the reference category for the outcome. To obtain the probability that a woman drank during pregnancy, the REF= option can be used to specify that 0 should be used as the reference category.

```
proc logistic data = example;
   model drank (REF='0') = age35plus;
   run;
```

The equation for the logistic regression model above can be written:

$$\ln(p/(1-p)) = \alpha + \beta * \text{age35plus}, \text{ where p is the probability or odds of drinking}$$

The output is below. Notice that SAS is indeed modeling the probability that drank = 1. The equations for the odds of drinking among women 35+ and among women <35, respectively, are:

$$\ln(p/(1-p)) = -0.9649 + 0.8218*1 = -0.1431$$

$$\ln(p/(1-p)) = -0.9649 + 0.8218*0 = -0.9649$$

```
                          Response Profile

                 Ordered                      Total
                   Value         drank     Frequency

                       1             0            72
                       2             1            42

                   Probability modeled is drank=1.

              Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter     DF     Estimate         Error    Chi-Square    Pr > ChiSq

Intercept      1      -0.9649        0.2938       10.7881        0.0010
age35plus      1       0.8218        0.3976        4.2719        0.0387

                         Odds Ratio Estimates

                              Point          95% Wald
           Effect          Estimate      Confidence Limits

           age35plus          2.275         1.043       4.959
```

To compare the odds of drinking among women in the two groups (i.e., calculate an odds ratio) the following equation is used. The interpretation is that women 35+ had 2.275 times the odds of drinking during the first trimester of pregnancy compared with women <35 years old.

$$e^{-0.14} \div e^{-0.96} = e^{-0.14 - (-0.96)} = e^{0.82} = 2.275$$

EXAMPLE 2

Now assume the independent variable has three categories. The number of dummy variables used in a regression should be equal to the number of categories minus one, where the category omitted is the reference group. In the code below, the dummy variable COLLEGE is created, which equals 1 when EDUCATION = 3. The variable HIGHSCHOOL is also created, which equals 1 when EDUCATION = 2. Both dummy variables are set to 0 otherwise and are considered missing if EDUCATION has a missing value.

```
if education = 3 then college = 1;
  else if education in(1,2) then college = 0;

if education = 2 then highschool = 1;
  else if education in(1,3) then highschool = 0;
```

The association between education and drinking during the first trimester of pregnancy is assessed in the logistic regression below.

```
proc logistic data = example;
   model drank (REF='0') = college highschool;
   run;
```

The equation corresponding to the regression model above is:

$$\ln(p/(1-p)) = \alpha + \beta_1 * \text{college} + \beta_2 * \text{highschool}$$

Based on the output below, the resulting equations for the odds of drinking among college, high school, and less than high school educated women, respectively, are:

$$\ln(p/(1-p)) = -1.3122 + 1.3122*1 + 0.7781*0 = 0$$

$$\ln(p/(1-p)) = -1.3122 + 1.3122*0 + 0.7781*1 = -0.5341$$

$$\ln(p/(1-p)) = -1.3122 + 1.3122*0 + 0.7781*0 = -1.3122$$

```
                 Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter     DF     Estimate        Error    Chi-Square    Pr > ChiSq

Intercept      1      -1.3122       0.4258        9.4962        0.0021
college        1       1.3122       0.5468        5.7593        0.0164
highschool     1       0.7781       0.5240        2.2046        0.1376

                         Odds Ratio Estimates

                        Point          95% Wald
       Effect          Estimate      Confidence Limits

       college           3.714       1.272       10.847
       highschool        2.177       0.780        6.081
```

To compare the odds of drinking among women in college and high school with the odds among women with less than a high school education, the following equations are used, respectively. In other words, when compared with women with a less than high school education, women with a college education had 3.714 times the odds and women who graduated high school had 2.177 times the odds of drinking during the first trimester of pregnancy.

$$e^{0} \div e^{-1.3122} = e^{0-(-1.3122)} = e^{1.3122} = 3.714$$

$$e^{-0.5341} \div e^{-1.3122} = e^{-0.5341-(-1.3122)} = e^{0.7781} = 2.177$$

EXAMPLE 3

Dummy coded variables can also be assessed in relation to a linear outcome using PROC REG. For instance, the code below assesses the relationship between the continuous variable for income as a percent of the Federal Poverty Level (FPL) and the dummy variable for age.

```
proc reg data = example;
   model income = age35plus;
   run;
   quit;
```

The regression equation corresponding to the model above is:

$$y = \alpha + \beta * \text{age35plus}$$

Based on the output below, the resulting equations among women 35+ and among women < 35, respectively, are:

$$y = 187.4821 + 43.2997*1 = 230.7818$$

$$y = 187.4821 + 43.2997*0 = 187.4821$$

```
                         Parameter Estimates

                        Parameter       Standard
Variable      DF         Estimate          Error    t Value    Pr > |t|

Intercept      1         187.4821       15.2557       12.29      <.0001
age35plus      1          43.2997       21.6727        2.00      0.0482
```

In linear regression, the parameter estimate corresponds to the change in the outcome associated with a one unit increase in the independent variable—in this case, a change in AGE35PLUS from 0 to 1. In other words, the average income as a percent of the FPL among women 35+ is about 43 percentage points greater than the average income among women <35.

Note that the intercept in linear regression corresponds to the mean among women in the reference category. In this example, the average income as a percent of the FPL among women <35 is 187.4821.

```
proc means data = example;
   class age35plus;
   var income;
   run;
```

```
                    Analysis Variable : income

                           N
       age35plus         Obs        N            Mean         Std Dev
       ───────────────────────────────────────────────────────────────
               0          58       56        187.4821        117.1177
               1          56       55        230.7818        111.0733
```

EFFECT CODING WITHIN THE DATA STEP

Effect coding is another way to define parameters in regression models. Whereas dummy coding assigns a variable a value of 1 or 0, effect coding assigns a variable a value of 1, 0, or -1, where -1 is the reference category.

EXAMPLE 1

Using effect coding, a variable for age is created where AGE35PLUSE is equal to 1 when AGECAT = '35+' and equal to -1 when AGECAT = '<35'.

```
if agecat = '35+' then age35plusE = 1;
  else if agecat = '<35' then age35plusE = -1;
```

Effect coding is perhaps easiest to understand with a linear outcome. Therefore, first, let's examine the relationship between age and income as a percent of the FPL.

```
proc reg data = example;
   model income = age35plusE;
   run;
   quit;
```

The regression equation corresponding to the model above is:

$$y = \alpha + \beta * \text{age35plusE}$$

Based on the output below, the resulting equations among women 35+ and among women < 35, respectively, are:

$$y = 209.1320 + 21.6498*1 = 230.7818$$

$$y = 209.1320 + 21.6498*-1 = 187.4821$$

```
                         Parameter Estimates

                            Parameter       Standard
Variable          DF         Estimate          Error    t Value    Pr > |t|

Intercept          1         209.1320        10.8363      19.30      <.0001
age35plusE         1          21.6498        10.8363       2.00      0.0482
```

The difference between women in the two groups is still 43.2997, but the SAS output is designed to compare the outcome among women in the non-reference group to the weighted mean of the outcome in the total population. From the prior example, the mean income as a percent of the FPL among women 35+ is 230.7818 and the mean income among women <35 is 187.4821. Adding these values and dividing by 2, produces the weighted mean of 209.1320, which is equivalent to the intercept in the output above.

EXAMPLE 2
If the independent variable has more than two categories and effect coding is used, all non-reference categories are assigned a value of 1 for their category and 0 otherwise. The reference category is always assigned a value of -1. For instance:

```
if education = 3 then collegeE = 1;
  else if education = 2 then collegeE = 0;
  else if education = 1 then collegeE = -1;

if education = 2 then highschoolE = 1;
  else if education = 3 then highschoolE = 0;
  else if education = 1 then highschoolE = -1;
```

The SAS code for the regression is:

```
proc reg data = example;
   model income = collegeE highschoolE;
   run;
   quit;
```

The equation corresponding to the model above is:

$$y = \alpha + \beta_1\text{*collegeE} + \beta_2\text{*highschoolE}$$

Based on the output below, the resulting equations among women with college, high school, and less than high school educations, respectively, are:

$$y = 205.5051 + 2.8889*1 + 27.0404*0 = 208.3940$$

$$y = 205.5051 + 2.8889*0 + 27.0404*1 = 232.5455$$

$$y = 205.5051 + 2.8889*{-1} + 27.0404*{-1} = 175.5758$$

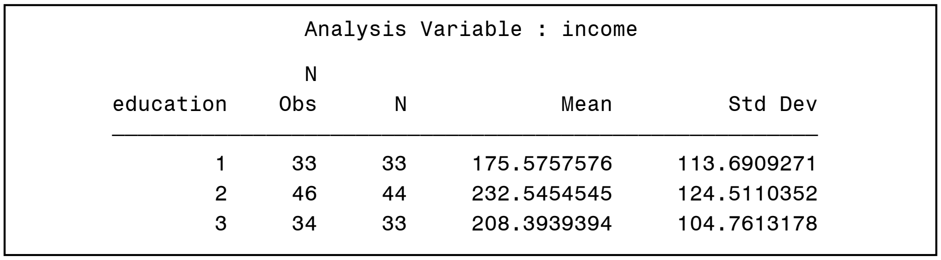
```
                        Parameter Estimates

                              Parameter       Standard
Variable          DF           Estimate          Error    t Value    Pr > |t|

Intercept          1           205.5051        11.1297      18.46      <.0001
collegeE           1             2.8889        16.0935       0.18      0.8579
highschoolE        1            27.0404        15.0073       1.80      0.0744
```

Again, the intercept corresponds to the weighted mean income in the total population. The parameter estimates correspond to the difference in income among college- and high school-educated women and the mean income in the total population. The mean income as a percent of the FPL is 208.3939 among college-educated women, 232.5455 among high-school educated women, and 175.5758 among women with less than a high school education. Adding these values together and dividing by three produces the weighted group mean of 205.5051.

```
proc means data = example;
   class education;
   var income;
   run;
```



```
                 Analysis Variable : income

                     N
     education     Obs       N            Mean          Std Dev
     ______________________________________________________________

             1      33      33     175.5757576      113.6909271
             2      46      44     232.5454545      124.5110352
             3      34      33     208.3939394      104.7613178
```

EXAMPLE 3
When effect coding is used in logistic regression, the parameter estimate can be interpreted as the change in the log odds of the outcome associated with a one unit increase in the independent variable. However, because the variable is coded 1 or -1, a one unit increase is not meaningful. Therefore, the parameter estimate and the odds ratio are not directly interpretable. Let's look at the example below.

```
proc logistic data = example;
   model drank (REF='0') = age35plusE;
   run;
```

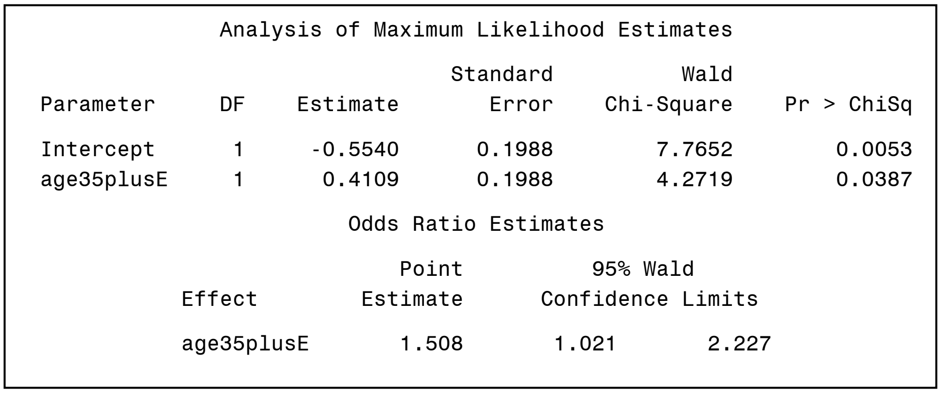
The regression equation for this model is:

$$\ln(p/(1-p)) = \alpha + \beta * \text{age35plusE}$$

Based on the output below, the resulting equations for the odds of drinking among women 35+ and women < 35, respectively, are:

$$\ln(p/(1-p)) = -0.5540 + 0.4109*1 = -0.1431$$

$$\ln(p/(1-p)) = -0.5540 + 0.4109*-1 = 0.9649$$



```
                  Analysis of Maximum Likelihood Estimates

                                         Standard          Wald
  Parameter     DF     Estimate         Error     Chi-Square     Pr > ChiSq

  Intercept      1      -0.5540        0.1988        7.7652         0.0053
  age35plusE     1       0.4109        0.1988        4.2719         0.0387

                            Odds Ratio Estimates

                               Point          95% Wald
             Effect         Estimate      Confidence Limits

             age35plusE        1.508         1.021         2.227
```

Again, the odds ratio and parameter estimate are not directly interpretable because they correspond to a one unit increase in the independent variable (i.e. from 0 to 1 or from -1 to 0). To calculate an odds ratio comparing women 35+ with women <35 the following equation is used:

$$e^{-0.1431} \div e^{-0.9649} = e^{-0.1431 - (-0.9649)} = e^{0.8218} = 2.275$$

THE CLASS STATEMENT

Dummy and effect coding can be automated in PROC LOGISTIC by listing the independent variable(s) in the CLASS statement. You can use the REF= option to specify the reference group. The reference category for both categorical and numeric variables is placed in quotes. If a format is applied to the variable, the reference category specified should be the format value instead of the data value.

```
proc logistic data = example;
   class agecat (REF='<35');
   model drank (REF='0') = agecat;
   run;
```

Notice under CLASS LEVEL INFORMATION in the output below that SAS parameterizes the model using effect coding. This is the default setting. Therefore, the parameter estimate for AGECAT is equal to the parameter estimate in the example above on effect coding (0.4109). However, the correct odds ratio comparing women 35+ with women < 35 is computed.

```
                        Class Level Information
                                          Design
                    Class      Value     Variables

                    agecat     35+              1
                               <35             -1

                Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
Parameter        DF     Estimate        Error    Chi-Square Pr > ChiSq

Intercept         1      -0.5540       0.1988        7.7652     0.0053
agecat    35+     1       0.4109       0.1988        4.2719     0.0387

                       Odds Ratio Estimates

                               Point          95% Wald
       Effect               Estimate      Confidence Limits

       agecat 35+ vs <35       2.275        1.043       4.959
```

To request dummy coding, the PARAM=REF option can used within the CLASS statement.

```
proc logistic data = example;
   class agecat (REF='<35')/param=ref;
   model drank (REF='0') = agecat;
   run;
```

Notice that the odds ratio in the output below is the same, but now the parameter estimates differ.

```
                      Class Level Information
                                         Design
                   Class      Value     Variables

                   agecat     35+              1
                              <35              0

               Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
Parameter        DF     Estimate        Error    Chi-Square Pr > ChiSq

Intercept         1      -0.9649       0.2938       10.7881     0.0010
agecat    35+     1       0.8218       0.3976        4.2719     0.0387

                        Odds Ratio Estimates

                              Point          95% Wald
      Effect               Estimate      Confidence Limits

      agecat 35+ vs <35       2.275        1.043       4.959
```

THE CONTRAST STATEMENT

The CONTRAST statement in PROC LOGISITIC allows the user to make comparisons across levels of one or more covariates. This is particularly useful when interaction terms are included in the model. An interaction term allows the association between an independent variable and an outcome to differ across levels of another covariate. The code below examines the relationship between age 35+, college education, and drinking during the first trimester of pregnancy. Here we allow for a different odds ratio associated with age across education levels. We also allow for a different odds ratio associated with education across age groups.

```
proc logistic data = example;
   model drank (REF='0') = age35plus college age35plus*college;
   run;
```

The regression equation corresponding to the model above is:

$$\ln(p/(1-p)) = \alpha + \beta_1\text{*age35plus} + \beta_2\text{*college} + \beta_3\text{*age35plus*college}$$

Based on the output below, we can calculate regression equations for women <35 who have a college education, and women <35 without a college education, respectively.

$$\ln(p/(1-p)) = -1.1856 + 0.8602*0 + 0.9625*1 – 0.6371*0 = -0.2231$$

$$\ln(p/(1-p)) = -1.1856 + 0.8602*0 + 0.9625*0 – 0.6371*0 = -1.1856$$

```
                Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept       1     -1.1856      0.3445       11.8438        0.0006
age35plus       1      0.8602      0.5012        2.9460        0.0861
college         1      0.9625      0.7541        1.6290        0.2018
age35plus*      1     -0.6371      0.9316        0.4677        0.4941
college
```

To compare the odds of drinking among women in the two groups the following equation is used:

$$e^{-0.2231} \div e^{-1.1856} = e^{-0.2231 – (-1.1856)} = e^{0.9625} = 2.618$$

Using the CONTRAST statement, these types of calculations can be automated in SAS. All we need is a basic understanding of matrix algebra. Take the matrices below for women <35 with and without a college education. The values for the dummy variables for women in each group are indicated. To compare women in the two groups, we take the difference of the matrices.

$$\begin{array}{lccccc} & \begin{array}{c}\text{Women <35}\\ \text{with a college}\\ \text{education}\end{array} & & \begin{array}{c}\text{Women <35}\\ \text{with no college}\\ \text{education}\end{array} & & \begin{array}{c}\text{Difference}\\ \text{between}\\ \text{groups}\end{array} \\ \begin{array}{l}\text{AGE35PLUS}\\ \text{COLLEGE}\\ \text{AGE35PLUS*COLLEGE}\end{array} & \begin{bmatrix}0\\1\\0\end{bmatrix} & — & \begin{bmatrix}0\\0\\0\end{bmatrix} & = & \begin{bmatrix}0\\1\\0\end{bmatrix}\end{array}$$

The matrix of the difference between groups is specified in the CONTRAST statement. A title or description is given in quotes. Then each parameter in the model is specified, directly followed by the corresponding value from the matrix for the difference between groups. The option ESTIMATE=EXP tells SAS to calculate an odds ratio from the parameter estimate.

```
proc logistic data = example;
   model drank (REF='0') = age35plus college age35plus*college;
   contrast "<35 college vs. <35 no college"
      age35plus 0 college 1 age35plus*college 0/estimate=exp;
   run;
```

Notice the odds ratio produced is exactly what we calculated above. Other comparisons can be made using these methods (e.g. between women 35+ with a college education and women 35+ without a college education).

```
                        Contrast Rows Estimation and Testing Results

                                                        Standard
Contrast                        Type Row  Estimate        Error  Alpha   Confidence Limits

<35 college vs. <35 no college  EXP   1     2.6182       1.9744   0.05    0.5972   11.4789
```

CONCLUSION

This paper has provided a basic overview of parameterization in SAS. Dummy- and effect-coded variables can be created in the DATA step. Alternatively, the CLASS statement can be used to define regression parameters and reference categories. The CONTRAST statement was also used to make comparisons across levels of one or more covariates. Hopefully this overview will provide both novice and experienced SAS users with a better understanding of defining parameters in regression modeling so that they can interpret results correctly and make full use of the statistical procedures offered by SAS.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kathryn Martin
Maternal, Child and Adolescent Health Program
California Department of Public Health
1615 Capitol Avenue, MS 8304, PO Box 997420
Sacramento, CA 95899-7420
E-mail: Katie.Martin@cdph.ca.gov