Paper 194-29

Head of the CLASS: Impress your colleagues with a superior understanding of the CLASS statement in PROC LOGISTIC

Michelle L. Pritchard and David J. Pasta

Ovation Research Group, San Francisco, CA

ABSTRACT

Data analysts and statistical programmers in a variety of disciplines use PROC LOGISTIC to fit logistic and proportional odds models to binary and ordinal responses. These models frequently contain categorical explanatory variables, and programmers must use the CLASS statement to identify which variables are categorical predictors. SAS Version 8 allows analysts more power and flexibility when handling categorized effects, specifically by allowing you to specify the method of parameterization and the reference category. But, be warned, this new set of tools should be used with caution! The default parameterization may not furnish what you would expect if you are used to the default provided by the CLASS statement in PROC GLM. The selected parameterization method has a profound effect on how CONTRAST statements are specified and in the interpretation of the parameter estimates. Also, the syntax is new and does take some getting used to. This paper provides a series of examples to allow programmers to become comfortable using the CLASS statement to specify and test desired hypotheses in PROC LOGISTIC.

INTRODUCTION

Beginning in SAS Version 8 (SAS version 8.02 for Windows 2000 will be used throughout this paper), the LOGISTIC procedure enables programmers to specify a full-rank parameterization (with the choice of effect, reference, polynomial, or orthogonal polynomial coding), or a less than full-rank parameterization (as in the GLM procedure). This is a big step forward from the days of doing your own coding of dummy variables, like you would for regression.

Let's begin by stepping through the new syntax.

The GLM parameterization is only available as a global option (i.e., for all variables in the class statement), but the full rank parameterizations can be specified either globally or for individual variables. **Global parameterization** is specified by the PARAM= <*EFFECT GLM ORTHPOLY POLYNOMIAL REFERENCE*> option after the forward slash (/) in the CLASS statement, as follows:

```
proc logistic data = temp01;
  class <classvar1> <classvar2>/param = glm;
  model <response> = <classvar1> <classvar2>;
  title3"Example 1: PARAM = GLM Global Option";
run;
quit;

proc logistic data = temp01;
  class <classvar1> <classvar2>/param = ref;
  model <response> = <classvar1> <classvar2>;
  title3"Example 2: PARAM = REF Global Option";
run;
quit;
```

Alternatively, **parameterization for individual variables** can be specified in parentheses immediately following the variable in the CLASS statement, as follows:

```
proc logistic data = temp01;
  class <classvar1> (param = ref) <classvar2> (param = effect);
  model <response> = <classvar1> <classvar2>;
  title3"Example 3: Individual Variable Parameterization Options";
run;
quit;
```

As an extra bit of flexibility, the REF= option in the CLASS statement determines the **reference level** for the EFFECT and REFERENCE coding. As with the PARAM= option, the REF= option can be coded globally or separately for each classification variable. To program the **reference level for each variable individually**, use the REF = " " option in parentheses immediately following the variable in the CLASS statement, or use the keyword FIRST or LAST (FIRST designates the first ordered category as the reference and LAST designates the last ordered category as the reference). Note that unless you use FIRST or LAST, the reference level should be placed in either single or double quotation marks, as follows:

```
proc logistic data = temp01;
  class <classvar1> (param = ref ref = “<refcat>“) <classvar2> (param = effect ref = last);
  model <response> = <classvar1> <classvar2>;
  title3"Example 4: Individual Variable Parameterization Options with Individual Specification of
        Reference Category";
run;
quit;
```

Alternatively, the **reference category can be applied to all variables** in the CLASS statement by using the keyword FIRST or LAST in the REF= option after the forward slash (/) in the CLASS statement.

```
proc logistic data = temp01;
  class <classvar1> (param = ref) <classvar2> (param = effect)/ref = first;
  model <response> = <classvar1> <classvar2>;
  title3"Example 5: Individual Variable Parameterization Options with Global Specification of Reference
        Category";
run;
quit;
```

Noteworthy notes:

1. If you neglect to specify a coding method and/or reference category, then the **default parameterization** method is PARAM=EFFECT and the **default reference category** is the last ordered category (REF=LAST).
2. Individual parameterization trumps the global option, unless the global option is GLM.
3. REF= is only valid when PARAM=EFFECT or PARAM=REF.
4. If a format has been applied to the classification variables, then either the **formatted value** (or FIRST or LAST) must be used when specifying the reference category. If the keyword FIRST or LAST is used in this situation, then it pertains to the first or last ordered formatted category.
5. We recommend limiting the classification variables to no more than one parameterization method, unless there are specific multiple hypotheses that you want to test **and** you are confident in what you are doing **and** you won't be showing the output to anyone else **and** you are sure you'll remember later why you did what you did **and** … you get the idea.

DATA

Now that we are familiar with the new PARAM= and REF= syntax, let's consider a model with a dichotomous response and three classification variables. The variables involved are distributed as follows:

```
                                     Cumulative    Cumulative
response    Frequency     Percent     Frequency      Percent
----------------------------------------------------------------------------------------------------------
       0         329       59.82           329        59.82
       1         221       40.18           550       100.00
```

```
                                      Cumulative    Cumulative
   female    Frequency     Percent     Frequency      Percent
----------------------------------------------------------------------------------------------------------
0. Male           259       47.09           259        47.09
1. Female         291       52.91           550       100.00
```

```
                                       Cumulative    Cumulative
    smoker    Frequency     Percent     Frequency      Percent
----------------------------------------------------------------------------------------------------------
1. Never           261       47.45           261        47.45
2. Past            214       38.91           475        86.36
3. Current          75       13.64           550       100.00
```

```
                                     Cumulative    Cumulative
  agecat    Frequency     Percent     Frequency      Percent
----------------------------------------------------------------------------------------------------------
1. 18-34         210       38.18           210        38.18
2. 35-49         167       30.36           377        68.55
3. 50-64         117       21.27           494        89.82
4. 65+            56       10.18           550       100.00
```

```
female    smoker    agecat    response    Frequency     Percent
-----------------------------------------------------------------------------------------------------------
     0         1         1           0         34         6.18
     0         1         1           1         21         3.82
     0         1         2           0         16         2.91
     0         1         2           1         10         1.82
     0         1         3           0         15         2.73
     0         1         3           1          9         1.64
     0         1         4           0          7         1.27
     0         1         4           1          3         0.55
     0         2         1           0         26         4.73
     0         2         1           1         19         3.45
     0         2         2           0         27         4.91
     0         2         2           1         17         3.09
     0         2         3           0         10         1.82
     0         2         3           1          5         0.91
     0         2         4           0          1         0.18
     0         2         4           1          1         0.18
     0         3         1           0          3         0.55
     0         3         1           1          2         0.36
     0         3         2           0          6         1.09
     0         3         2           1          6         1.09
     0         3         3           0          6         1.09
     0         3         3           1          6         1.09
     0         3         4           0          5         0.91
     0         3         4           1          4         0.73
     1         1         1           0         28         5.09
     1         1         1           1         18         3.27
     1         1         2           0         22         4.00
     1         1         2           1         20         3.64
     1         1         3           0         25         4.55
     1         1         3           1         17         3.09
     1         1         4           0          8         1.45
     1         1         4           1          8         1.45
     1         2         1           0         26         4.73
     1         2         1           1         20         3.64
     1         2         2           0         25         4.55
     1         2         2           1         12         2.18
     1         2         3           0         15         2.73
     1         2         3           1          7         1.27
     1         2         4           0          2         0.36
     1         2         4           1          1         0.18
     1         3         1           0          7         1.27
     1         3         1           1          6         1.09
     1         3         2           0          5         0.91
     1         3         2           1          1         0.18
     1         3         3           0          1         0.18
     1         3         3           1          1         0.18
     1         3         4           0          9         1.64
     1         3         4           1          7         1.27
```

MODEL A: DEFAULT PARAMETERIZATION (EFFECT)

```
proc logistic data = temp01  descending;
  class female smoker agecat ;
  model response = female smoker agecat;
  title3"Model A: Logistic regression with three categorical predictors and default options PARAM=EFFECT
       and REF=LAST";
run;
quit;
```

In Model A, the method of parameterization is not specified, so the default EFFECT parameterization will be used. (Also, by default the last ordered category will be used as the reference category.) The columns of the design matrix are created based on the EFFECT coding scheme, as follows: For each variable, *c*-1 columns comprise the design matrix, where *c* is the number of levels of the classification variable. For the nonreference levels, the columns represent group membership (1=member, 0 =non-member), and for the reference level, the row is populated with a –1. In our Model A, the reference level is the last ordered category, so the design matrix is:

```
                                      Class Level Information

                   Design Variables

Class      Value      1      2      3

female     0          1
           1         -1

smoker     1          1      0
           2          0      1
           3         -1     -1

agecat     1          1      0      0
           2          0      1      0
           3          0      0      1
           4         -1     -1     -1
```

Using EFFECT coding, the beta estimates are estimating the difference in the effect of each nonreference level compared to the average effect over all levels.

```
                             Analysis of Maximum Likelihood Estimates

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

Intercept        1      -0.3705       0.1057        12.2880        0.0005
female     0     1      -0.0170       0.0877         0.0374        0.8466
smoker     1     1      -0.0120       0.1253         0.0091        0.9240
smoker     2     1      -0.1098       0.1353         0.6578        0.4174
agecat     1     1       0.0470       0.1431         0.1079        0.7425
agecat     2     1      -0.0108       0.1524         0.0051        0.9433
agecat     3     1      -0.0753       0.1677         0.2017        0.6534
```

Therefore, the odds of having response = 1 for never vs. current smokers is: exp(-0.0120)/exp[-(-0.012 – 0.1098)] = 0.875.

```
                                      Odds Ratio Estimates

                     Point          95% Wald
Effect            Estimate      Confidence Limits

female 0 vs 1        0.967        0.685       1.363
smoker 1 vs 3        0.875        0.511       1.499
smoker 2 vs 3        0.793        0.451       1.397
agecat 1 vs 4        1.008        0.536       1.897
agecat 2 vs 4        0.951        0.498       1.818
agecat 3 vs 4        0.892        0.456       1.745
```

MODEL B: GLM PARAMETERIZATION

```
proc logistic data = temp01  descending;
  class female smoker agecat/param = glm ;
  model response = female smoker agecat;
  title3"Model B: Logistic regression with three categorical predictors and PARAM = GLM option";
run;
quit;
```

In Model B, the GLM parameterization has been specified. With GLM parameterization, the programmer does not have control over the reference category using the REF= option; instead, SAS includes variables as long as they are linearly independent from those that went before. This has the effect of making the last ordered category the omitted (reference) category. The columns of the design matrix are created based on the GLM coding scheme, as follows: For each variable, *c* columns comprise the design matrix, where *c* is the number of levels of the classification variable. For all levels, the columns represent group membership (1=member, 0 =non-member). The design matrix for Model B is:

```
                                         Class Level Information

                         Design Variables

Class      Value      1      2      3      4

female     0          1      0
           1          0      1

smoker     1          1      0      0
           2          0      1      0
           3          0      0      1

agecat     1          1      0      0      0
           2          0      1      0      0
           3          0      0      1      0
           4          0      0      0      1
```

Using GLM coding, the beta estimates are estimating the difference in the effect of each nonreference level compared to the reference (last) level.

```
                                 Analysis of Maximum Likelihood Estimates

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

intercept       1     -0.1927      0.3163        0.3712        0.5423
female    0     1     -0.0339      0.1754        0.0374        0.8466
female    1     0           0           .         .              .
smoker    1     1     -0.1337      0.2747        0.2368        0.6266
smoker    2     1     -0.2315      0.2887        0.6431        0.4226
smoker    3     0           0           .         .              .
agecat    1     1     0.00790      0.3226        0.0006        0.9805
agecat    2     1     -0.0500      0.3304        0.0229        0.8798
agecat    3     1     -0.1144      0.3424        0.1117        0.7382
agecat    4     0           0           .         .              .
```

Therefore, the odds of having response = 1 for never vs. current smokers is: exp(-0.1337) = 0.875.

```
                                          Odds Ratio Estimates

                    Point          95% Wald
Effect           Estimate      Confidence Limits

female 0 vs 1       0.967       0.685       1.363
smoker 1 vs 3       0.875       0.511       1.499
smoker 2 vs 3       0.793       0.451       1.397
agecat 1 vs 4       1.008       0.536       1.897
agecat 2 vs 4       0.951       0.498       1.818
agecat 3 vs 4       0.892       0.456       1.745
```

One of the reasons this parameterization may be desirable is that it corresponds to the behavior of GLM and MIXED. That parameterization is familiar to long-time users of those PROCs. One disadvantage is that in order to control which category is omitted (treated as the reference category), it is necessary to order the categories so the category to be omitted is last.

MODEL C: REFERENCE PARAMETERIZATION

The REFERENCE parameterization is similar to GLM except you actually omit one of the categories (so the resulting design matrix is full rank) and you have explicit control over which category is to be omitted; it need not be the last category.

```
proc logistic data = temp01  descending;
  class female smoker agecat/param = ref;
  model response = female smoker agecat;
  title3"Model C: Logistic regression with three categorical predictors and PARAM = REF option";
run;
quit;
```

In Model C, the REFERENCE parameterization has been specified. (Also, by default the last ordered category will be used as the reference category.) The columns of the design matrix are created based on the REFERENCE coding scheme, as follows: For each variable, *c*-1 columns comprise the design matrix, where *c* is the number of levels of the classification variable. For the nonreference levels, the columns represent group membership (1=member, 0 =non-member), and for the reference level, the row is populated with a 0. In our Model C, the reference level is the last ordered category, so the design matrix is:

```
                                              Class Level Information

                     Design Variables

Class      Value      1      2      3

female     0          1
           1          0

smoker     1          1      0
           2          0      1
           3          0      0

agecat     1          1      0      0
           2          0      1      0
           3          0      0      1
           4          0      0      0
```

Using REFERENCE coding, the beta estimates are estimating the difference in the effect of each nonreference level compared to the effect of the reference level.

```
                                Analysis of Maximum Likelihood Estimates

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

Intercept       1      -0.1927        0.3163        0.3712        0.5423
female    0     1      -0.0339        0.1754        0.0374        0.8466
smoker    1     1      -0.1337        0.2747        0.2368        0.6266
smoker    2     1      -0.2315        0.2887        0.6431        0.4226
agecat    1     1      0.00790        0.3226        0.0006        0.9805
agecat    2     1      -0.0500        0.3304        0.0229        0.8798
agecat    3     1      -0.1144        0.3424        0.1117        0.7382
```

Therefore, as with the GLM parameterization, the odds of having response = 1 for never vs. current smokers is: exp(-0.1337) = 0.875.

```
                                           Odds Ratio Estimates

                    Point          95% Wald
Effect           Estimate      Confidence Limits

female 0 vs 1       0.967       0.685       1.363
smoker 1 vs 3       0.875       0.511       1.499
smoker 2 vs 3       0.793       0.451       1.397
agecat 1 vs 4       1.008       0.536       1.897
agecat 2 vs 4       0.951       0.498       1.818
agecat 3 vs 4       0.892       0.456       1.745
```

MODEL D: REFERENCE PARAMETERIZATION WITH REFERENCE CATEGORY SPECIFICATION

In this model we use the reference parameterization but specify REF=FIRST. This changes the design matrix and the odds ratios are inverted for binary variables.

```
proc logistic data = temp01  descending;
  class female smoker agecat/ param = ref ref = first ;
  model response = female smoker agecat;
  title3"Model D: Logistic regression with three categorical predictors and PARAM=REF and REF=first";
run;
quit;
```

```
                                         Class Level Information

                     Design Variables

Class      Value      1      2      3

female     0          0
           1          1

smoker     1          0      0
           2          1      0
           3          0      1

agecat     1          0      0      0
           2          1      0      0
           3          0      1      0
           4          0      0      1
```

```
                                Analysis of Maximum Likelihood Estimates

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

Intercept       1      -0.3524      0.1915       3.3869        0.0657
female    1     1       0.0339      0.1754       0.0374        0.8466
smoker    2     1      -0.0978      0.1920       0.2597        0.6103
smoker    3     1       0.1337      0.2747       0.2368        0.6266
agecat    2     1      -0.0579      0.2121       0.0745        0.7850
agecat    3     1      -0.1223      0.2376       0.2650        0.6067
agecat    4     1     -0.00790      0.3226       0.0006        0.9805
```

Therefore, the odds of having response = 1 for current vs. never smokers is: exp(0.1337) = 1.143.

```
                                          Odds Ratio Estimates

                    Point          95% Wald
Effect           Estimate      Confidence Limits

female 1 vs 0       1.035       0.734       1.459
smoker 2 vs 1       0.907       0.622       1.321
smoker 3 vs 1       1.143       0.667       1.958
agecat 2 vs 1       0.944       0.623       1.430
agecat 3 vs 1       0.885       0.555       1.410
agecat 4 vs 1       0.992       0.527       1.867
```

MODEL E: DEFAULT WITH INTERACTIONS

One thing that analysts new to the CLASS statement sometimes find mysterious is that when the specify interactions sometimes SAS changes the order of the variables. They specified female*smoker, for example, but in the output they see smoker*female. What is SAS doing? It turns out it's paying attention to the order of the variables in the CLASS statement.

```
proc logistic data = temp01  descending;
  class female smoker agecat;
  model response = female smoker agecat female*smoker female*agecat smoker*agecat;
  title3"Model E1: Logistic regression with three categorical predictors plus two-way interactions,
        default options";
run;
quit;
```

```
proc logistic data = temp01  descending;
  class smoker agecat female;
  model response = female smoker agecat female*smoker female*agecat smoker*agecat;
  title3"Model E2: Logistic regression with three categorical predictors plus two-way interactions,
        default options";
run;
quit;
```

MODEL F: PARAM=GLM WITH INTERACTIONS

```
proc logistic data = temp01  descending;
  class female smoker agecat/param = glm;
  model response = female smoker agecat female*smoker female*agecat smoker*agecat;
  title3"Model F: Logistic regression with three categorical predictors plus two-way interactions, with
       PARAM = GLM option";
run;
```

HYPOTHESIS TESTING WITH CONTRAST STATEMENTS

The CONTRAST statement in PROC LOGISTIC works much like the corresponding statement in other procedures such as GLM and MIXED. It allows you to specify one or more linear combinations of the parameters to test against zero. You can test each of the linear combinations against zero or specify that several linear combinations be tested against zero simultaneously (a multiple-degrees-of-freedom test). In order to correctly specify the CONTRAST statement, you need to pay close attention to the parameterization. For example, consider the AGECAT variable as parameterized in Model A, the default EFFECT coding. To test group 1 against group 2, the contrast statement would be

```
  contrast '1 vs. 2' agecat 1 –1 0 / e;
```

The 0 at the end is not necessary but is a good reminder that there are three parameters for the AGECAT variable. The E option is extremely useful: it requests that the linear combination be displayed, giving you a chance to make sure the coefficients are lined up with the proper parameters. We strongly recomment using the E option when testing new constrast statements.

Testing that groups 1, 2, and 3 are simultaneously equal requires a two degree of freedom test:

```
  contrast '1 = 2 = 3' agecat 1 –1 0 , agecat 1 0 -1 / e;
```

or, equivalently,

```
  contrast '1 = 2 = 3' agecat 1 –1 0 , agecat 0 1 -1 / e;
```

or

```
  contrast '1 = 2 = 3' agecat 1 0 –1 , agecat 0 1 -1 / e;
```

all of which test the same hypothesis, that the first three AGECAT groups are equal.

Because the fourth AGECAT group is coded as the negation of the first three, contrasts involving that group look rather different. If you label the parameters for the first three groups A1, A2, and A3, then the parameter for the fourth group is –A1-A2-A3. Thus if you want to compare group 1 to group 4, you want to test (A1)-(-A1-A2-A3)=0, or 2*A1+A2+A3=0:

```
  contrast '1 vs. 4' agecat 2 1 1 / e;
```

It is clear how useful it is to label each contrast – it would be easy to forget what this contrast is designed to test!

The construction of CONTRAST statements can be very tricky in the presence of many interactions. To be sure you are testing the hypothesis you want to test, use the E option and write down the resulting equation separately. It may be worth seeking advice from a statistician for especially complex models.

CONCLUSION

The different model parameterizations allowed in PROC LOGISTIC are a powerful tool for specifying models and testing hypotheses. With that power comes responsibility – responsibility to be sure the results you get are what you intended. To make sure that happens, remember to:

1. pay attention to the order of the levels of the class variables
2. pay attention to the order of the variables in the CLASS statement
3. pay attention to the design matrix
4. understand the beta estimates and hypothesis tests
5. understand the syntax and defaults, and what options are and are not available for each method of parameterization
6. be aware of how formatted data are treated

CONTACT INFORMATION

The authors welcome questions and comments. Please direct inquiries to:

Michelle L. Pritchard
Ovation Research Group
120 Howard St., Ste. 600
San Francisco, CA 94105
mpritchard@ovation.org
(310) 821-9145

David J. Pasta
Ovation Research Group
120 Howard St., Ste. 600
San Francisco, CA 94105
dpasta@ovation.org
(415) 371-2111