Paper SAS060-2014

Making Comparisons Fair: How LS-Means Unify the Analysis of Linear Models

Weijie Cai, SAS Institute Inc.

ABSTRACT

How do you compare group responses when the data are unbalanced or when covariates come into play? Simple averages will not do, but LS-means are just the ticket. Central to postfitting analysis in SAS/STAT® linear modeling procedures, LS-means generalize the simple average for unbalanced data and complicated models. They play a key role both in standard treatment comparisons and Type III tests and in newer techniques such as sliced interaction effects and diffograms. This paper reviews the definition of LS-means, focusing on their interpretation as predicted population marginal means, and it illustrates their broad range of use with numerous examples.

INTRODUCTION

You always hope your data will be balanced, with all combinations of all factors sampled equally often. Balanced data are easy to analyze, because you can simply compare group means to see the treatment effect. You can plot group means or combine them to make higher-level inferences.

When your data are unbalanced, simple averages do not work, because all factors do not have an equal chance to affect the response. That is where LS-means come in. LS-means estimate the averages you would have seen if your data had been balanced; they indicate how a given factor affects the response, all other things being equal. You can use LS-means in all the ways you would use regular means. You can compare LS-means with each other, plot them, or use them to ask more involved questions.

This paper begins with a brief review of LS-means that explains what LS-means are and why you would want to use them. Then it gives an introductory example that shows how to use the LSMEANS statement to perform LS-means comparisons and visualize the results. Next, new LSMEANS statement options and new statements are introduced. Additional examples illustrate how to use these options and statements under various LS-means analysis scenarios.

LS-means were originally a feature of PROC HARVEY (Harvey 1976), a user-written procedure that was developed in the mid-1970s by Walter R. Harvey of Ohio State University. Soon thereafter they were included in PROC GLM. In the 1990s they were added to PROC MIXED, and in the early 2000s to PROC GLIMMIX. With each of these additions, many new features were included, especially for multiple comparisons. Beginning with SAS/STAT 9.22, LS-means are now featured in over a dozen procedures in SAS/STAT and also in SAS/QC® software.

LS-means were originally called “least squares means” (short for “means of least squares predictions”), which is how they were originally computed in the context of general linear models. This shortened form is somewhat misleading in two senses. First, “least squares” should not be construed as modifying “means,” but rather as modifying the predictions over which the means are computed. Second, the extensions of LS-means to mixed models, generalized linear models, and other models are no longer associated with least squares methods. The “LS” comes from the LSMEANS statement, which tells SAS you want to compute them, but the lowercase “means” distinguishes the statistical construct from the SAS syntax.

OVERVIEW OF LEAST SQUARES MEANS

This section explains how LS-means are defined by using a simple two-way interaction model. The nonestimability issue is briefly discussed. The connection between the Type III tests and the joint tests of LS-means differences is also covered in this section.

The basic definition of LS-means is given by Harvey (1975), and it is extensively discussed in Goodnight and Harvey (1978). Basically, LS-means provide an estimate for certain linear combinations of population parameters. The particular linear functions are defined by population marginal means of the corresponding means for balanced data.

The following simple example illustrates the concept of LS-means. In a study of salaries of faculty members selected from different departments at a university, two factors are considered: gender (male and female) and rank (associate and full). Table 1 provides salary means in thousands and sample sizes (shown in parentheses) for each combination of gender and rank.

Table 1 Salary Information for Tenured Professors at a University

Gender	Associate	Full
Male	130 (12)	136 (2)
Female	118 (4)	120 (5)

The mean salary for associate professors is $(130 \times 12 + 118 \times 4)/16 = 127$, and the mean salary for full professors is $(136 \times 2 + 120 \times 5)/7 = 124.6$. The overall mean salary for associate professors is higher than for full professors even though associate professors earn less than full professors in each gender category. The seeming contradiction is caused by the imbalance in the data. The associate professors are predominantly male, and all male professors earn more than their female colleagues in this particular sample. To correct the imbalance, you can compute LS-means for associate and full professors. The LS-means are simply arithmetic means over genders. For associate professors, the LS-mean is $(130 + 118)/2 = 124$. For full professors, the LS-mean is $(136 + 120)/5 = 128$. Thus, the least squares mean of salary for associate professors is lower than for full professors. The LS-means ignore the sample size information in each group and assume a balanced gender distribution in the underlying population. If a balanced design had been available (that is, the sample sizes in all groups had been the same), then the LS-means would be equivalent to the means. Thus, an LS-mean can be thought of as the mean that would be calculated if a balanced design had been obtainable.

Consider a two-way interaction model that has factors A and B. Let Y_{ijk} be the kth observation in the ith row and jth column,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where $i = 1, 2$; $j = 1, 2$; $k = 1, 2, \ldots, n_{ij}$; n_{ij} is the size of each group indexed by (i, j); and ϵ_{ijk} is the random effect that has mean 0. The population mean of each group is thus

$$E(Y_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

The population marginal means of the ith row or jth column are defined as the arithmetic means of population group means for groups of the ith row or jth column:

$$E(Y_{i\cdot}) = \mu + \alpha_i + \bar{\beta}_\cdot + \bar{\gamma}_{i\cdot}$$
$$E(Y_{\cdot j}) = \mu + \bar{\alpha}_\cdot + \beta_j + \bar{\gamma}_{\cdot j}$$

Let $\boldsymbol{\beta}$ be the vector of parameters for the population $[\mu\ \alpha_1\ \alpha_2\ \beta_1\ \beta_2\ \gamma_{11}\ \gamma_{12}\ \gamma_{21}\ \gamma_{22}]'$. Then the population marginal means can be expressed as a linear combination of the population parameters $\mathbf{L}\boldsymbol{\beta}$. For example, the population marginal mean $E(Y_{2\cdot})$ can be represented as $\mathbf{L}\boldsymbol{\beta} = \mu + \alpha_2 + 0.5\beta_1 + 0.5\beta_2 + 0.5\gamma_{12} + 0.5\gamma_{22}$ with $\mathbf{L} = [1\ 0\ 1\ 0.5\ 0.5\ 0\ 0.5\ 0\ 0.5]$. As stated previously, the population marginal mean is defined by a linear combination $\mathbf{L}$. An LS-mean estimates $\mathbf{L}\boldsymbol{\beta}$ by $\mathbf{L}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a least squares estimate for an ordinary linear model, a maximum likelihood estimate for a generalized linear model, and so on.

Sometimes there are missing cells—that is, the responses at certain levels of A and B are not observed for some reason. When there is no information about some population parameters, then any expectation that involves these parameters cannot be estimated. Hence, not all LS-means are estimable. The estimability of LS-means can be formulated in the context of least squares regression and analysis of variance. A linear combination of the parameters $\mathbf{L}\boldsymbol{\beta}$ is estimable if and only if a linear combination of the $\mathbf{Y}$s exists that is an unbiased estimate of $\mathbf{L}\boldsymbol{\beta}$. Because the expectation of the linear combination of $\mathbf{Y}$ is equal to the same linear combination of $\mathbf{X}\boldsymbol{\beta}$ given the design matrix $\mathbf{X}$, $\mathbf{L}\boldsymbol{\beta}$ is estimable if and only if there is a matrix $\mathbf{K}$ such that $\mathbf{L} = \mathbf{KX}$.

For the two-way interaction model, suppose that all cells contain at least one observation. The parameter estimates for $\boldsymbol{\beta}$ by least squares regression are $\hat{\boldsymbol{\beta}} = [\hat{\mu}\ \hat{\alpha}_1\ \hat{\alpha}_2\ \hat{\beta}_1\ \hat{\beta}_2\ \hat{\gamma}_{11}\ \hat{\gamma}_{12}\ \hat{\gamma}_{21}\ \hat{\gamma}_{22}]'$. Then the LS-means for main effects and the interaction effect are represented by $\mathbf{L}\hat{\boldsymbol{\beta}}$ with

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 \\ 1 & 0 & 1 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 1 & 0.5 & 0.5 & 1 & 0 & 0.5 & 0 & 0.5 & 0 \\ 1 & 0.5 & 0.5 & 0 & 1 & 0 & 0.5 & 0 & 0.5 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Suppose the cell that corresponds to the second level of A and B is missing. The parameter estimates vector is thus $\hat{\boldsymbol{\beta}}^* = [\hat{\mu}^*\ \hat{\alpha}_1^*\ \hat{\alpha}_2^*\ \hat{\beta}_1^*\ \hat{\beta}_2^*\ \hat{\gamma}_{11}^*\ \hat{\gamma}_{12}^*\ \hat{\gamma}_{21}^*]'$. There is no parameter estimate for γ_{22} because of the missing cell. Then the LS-means for main effects and the interaction effect are represented by $\mathbf{L}^*\hat{\boldsymbol{\beta}}^*$ with

$$\mathbf{L}^* = \begin{bmatrix} 1 & 1 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 \\ 1 & 0 & 1 & 0.5 & 0.5 & 0 & 0 & 1 \\ 1 & 0.5 & 0.5 & 1 & 0 & 0.5 & 0 & 0.5 \\ 1 & 0.5 & 0.5 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

When the cell is missing, the LS-means that correspond to α_2 and β_2 are nonestimable, because their $\mathbf{L}$ matrices (the second and fourth row of $\mathbf{L}^*$) cannot be formed as linear combinations of the design matrix $\mathbf{X}$.

There is a connection between the Type III tests and the joint tests of LS-means differences. The Type III tests examine the significance of each model effect by evaluating partial sums of squares that are associated with the hypotheses $\mathbf{L}\boldsymbol{\beta}$=0. The linear combination $\mathbf{L}$ is constructed in a special way such that (1) other effects that do not contain the effect in question have zero coefficients, and (2) estimable functions of the effect in question are pairwise orthogonal to any effect that contains the effect. For the definition of effect containment and more information about constructing Type III estimable functions, see the *SAS/STAT User's Guide*. On the other hand, the joint tests of LS-means differences examine whether all LS-means of a classification effect are equal. For the two-way interaction model, the Type III hypothesis for effect A gives

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}$$

The linear combination $\mathbf{L}$ corresponds to a specification of the contrast of LS-means for factor A. In general, for a classification effect, the Type III test results are equivalent to the joint tests on LS-means differences, given that the LS-means differences are all estimable. Significant Type III tests suggest further investigation of LS-means comparisons. Significant joint tests of LS-means differences suggest that the effect is significant in the fitted model. In the following example, you can obtain equivalent test results from both the LSMEANS statement and the TEST statement:

```
data a;
   do i = 1 to 1000;
      A = int(5*ranuni(1))+1;
      B = int(5*ranuni(1))+1;
      C = int(5*ranuni(1))+1;
      Y = rannor(1);
      output;
   end;
run;

proc orthoreg data=a;
   class A B C;
   model Y = A|B|C;
   lsmeans A / joint(all);
   test A;
run;
```

Figure 1 shows the test results from both statements.

Figure 1 Joint Test on LS-Means Differences and Type III Test

The ORTHOREG Procedure

Dependent Variable: Y

F Test for Equality of A Least Squares Means				
Num DF	Den DF	F Value	Pr > F
4	875	0.41	0.8015

Type III Tests of Model Effects				
Effect	Num DF	Den DF	F Value	Pr > F
A	4	875	0.41	0.8015

INTRODUCTORY EXAMPLE

This example illustrates how you can use a simple LSMEANS statement to test a hypothesis as opposed to using relatively complex syntax in the ESTIMATE statement. It also shows that the LSMEANS statement uses ODS Graphics to produce default graphics.

Consider a sociological study of Australian Aboriginal and white children (Quine 1975). In the study, the number of days that students were absent from schools was collected from four age groups (final grade in primary schools, and first, second, and third forms in secondary schools) and two cultural groups (Aboriginal and non-Aboriginal). You want to conduct a two-way analysis of the response variable, **Days**, by two classification variables, **Origin** and **Grade**. You can use the GLM procedure to fit a linear model and then save the results for later use in a postfitting analysis. The STORE statement stores the results in a special type of SAS data file called an item store. In this example, that item store is named Ozkidsfit. The following steps read and analyze the data:

```
data Ozkids;
   input Days Origin $ Grade $ @@;
   datalines;
 2 A F0  11 A F0  14 A F0   5 A F0   5 A F0  13 A F0  20 A F0  22 A F0   1 N F3
 6 A F1   6 A F1  15 A F1   7 A F1  14 A F1   6 A F2  32 A F2  53 A F2   9 N F3
57 A F2  14 A F2  16 A F2  16 A F2  17 A F2  40 A F2  43 A F2  46 A F2  22 N F3
12 A F3  15 A F3   8 A F3  23 A F3  23 A F3  28 A F3  34 A F3  36 A F3   3 N F3
38 A F3   3 A F0   5 A F0  11 A F0  24 A F0  45 A F0   5 A F1   6 A F1   5 N F3
 6 A F1   9 A F1  13 A F1  23 A F1  25 A F1  32 A F1  53 A F1  54 A F1  15 N F3
 5 A F1   5 A F1  11 A F1  17 A F1  19 A F1   8 A F2  13 A F2  14 A F2  18 N F3
20 A F2  47 A F2  48 A F2  60 A F2  81 A F2   2 A F2   5 A F3   9 A F3  22 N F3
 7 A F3   0 A F3   2 A F3   3 A F3   5 A F3  10 A F3  14 A F3  21 A F3  37 N F3
36 A F3  40 A F3   6 N F0  17 N F0  67 N F0   0 N F0   0 N F0   2 N F0   2 N F2
 7 N F0  11 N F0  12 N F0   0 N F1   0 N F1   5 N F1   5 N F1   5 N F1   2 N F2
11 N F1  17 N F1   3 N F1   3 N F1  22 N F2  30 N F2  36 N F2   8 N F2   3 N F2
 0 N F2   1 N F2   5 N F2   7 N F2  16 N F2  27 N F2  12 N F3  15 N F3   8 N F2
 0 N F3  30 N F3  10 N F3  14 N F3  27 N F3  41 N F3  69 N F3  25 N F0  10 N F2
10 N F0  11 N F0  20 N F0  33 N F0   5 N F1   7 N F1   0 N F1   1 N F1  12 N F2
 5 N F1   5 N F1   5 N F1   5 N F1   7 N F1  11 N F1  15 N F1   5 N F1   1 N F2
14 N F1   6 N F1   6 N F1   7 N F1  28 N F1   0 N F2   5 N F2  14 N F2   8 N F3
;

proc glm data=Ozkids;
   class Origin Grade;
   model Days = Origin|Grade;
   store Ozkidsfit;
run;
```

Figure 2 displays the Type I and Type III tests of the three effects.

Figure 2 Tests for Effects in a Two-Way Model

The GLM Procedure

Dependent Variable: Days

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Origin	1	2556.742939	2556.742939	12.10	0.0007
Grade	3	2054.774711	684.924904	3.24	0.0239
Origin*Grade	3	3291.796632	1097.265544	5.19	0.0019

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	1	1740.154271	1740.154271	8.24	0.0047
Grade	3	2037.905260	679.301753	3.22	0.0247
Origin*Grade	3	3291.796632	1097.265544	5.19	0.0019

Both main effects and their interaction are significant. Suppose you want to examine the differences between the group of Aboriginal children from the first form in secondary schools and the rest of the groups, adjusting for the other factors. Because the fitted model is saved into an item store, you can use the PLM procedure to perform the postfitting analyses without having to access the original data and refit the model. The following program shows the syntax for using the ESTIMATE statement in PROC PLM to construct the hypotheses:

```
proc plm restore=Ozkidsfit;
   estimate
    'A:F0 vs A:F1'                      Grade [1, 1][-1, 2] Origin*Grade [1, 1 1][-1, 1 2],
    'A:F2 vs A:F1'                      Grade [1, 3][-1, 2] Origin*Grade [1, 1 3][-1, 1 2],
    'A:F3 vs A:F1'                      Grade [1, 4][-1, 2] Origin*Grade [1, 1 4][-1, 1 2],
    'N:F0 vs A:F1' Origin [1, 2][-1, 1] Grade [1, 1][-1, 2] Origin*Grade [1, 2 1][-1, 1 2],
    'N:F1 vs A:F1' Origin [1, 2][-1, 1]                     Origin*Grade [1, 2 2][-1, 1 2],
    'N:F2 vs A:F1' Origin [1, 2][-1, 1] Grade [1, 3][-1, 2] Origin*Grade [1, 2 3][-1, 1 2],
    'N:F3 vs A:F1' Origin [1, 2][-1, 1] Grade [1, 4][-1, 2] Origin*Grade [1, 2 4][-1, 1 2];
run;
```

The $\mathbf{L}$ matrix syntax in the ESTIMATE statement is complicated and error-prone. The syntax in the following LSMEANS statement is much easier:

```
ods graphics on;
proc plm restore=Ozkidsfit;
   lsmeans Origin*Grade/diff=control('A' 'F1');
run;
```

The ODS GRAPHICS ON statement enables ODS Graphics so that PROC PLM will produce graphs.[1] The DIFF option requests LS-means differences. The CONTROL option and its specified levels request that differences be computed with a control group of Aboriginal children from the first form in secondary schools. Figure 3 shows the LS-means differences between the control group and seven other groups. Among the seven other groups, two groups show a significant difference from the control group: Aboriginal children from the second form in secondary schools and non-Aboriginal children from the first form in secondary schools.

[1] Because ODS Graphics is not subsequently disabled, it remains enabled for the rest of this paper.

Figure 3 LS-Means Difference against the Control Group

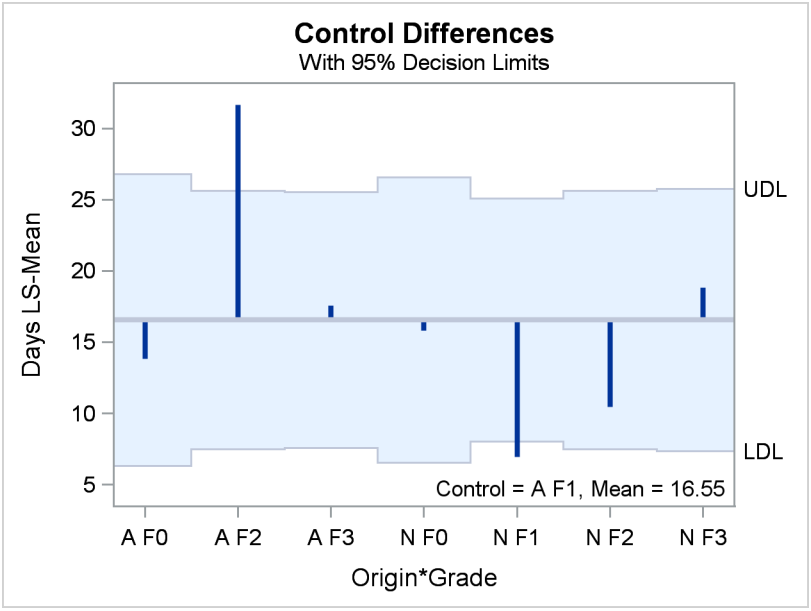
The PLM Procedure

Differences of Origin*Grade Least Squares Means

Origin	Grade	_Origin	_Grade	Estimate	Standard Error	DF	t Value	Pr > |t|
A	F0	A	F1	-2.7038	5.1779	145	-0.52	0.6023
A	F2	A	F1	15.1000	4.5960	145	3.29	0.0013
A	F3	A	F1	1.0214	4.5410	145	0.22	0.8223
N	F0	A	F1	-0.7643	5.0646	145	-0.15	0.8803
N	F1	A	F1	-9.5885	4.3228	145	-2.22	0.0281
N	F2	A	F1	-6.1000	4.5960	145	-1.33	0.1865
N	F3	A	F1	2.2921	4.6561	145	0.49	0.6233

Figure 4 displays the corresponding plot of the LS-means differences. Two vertical line segments exceed the 95% confidence limits of the LS-mean for the control group; they correspond to the two groups that are identified in the LS-means differences.

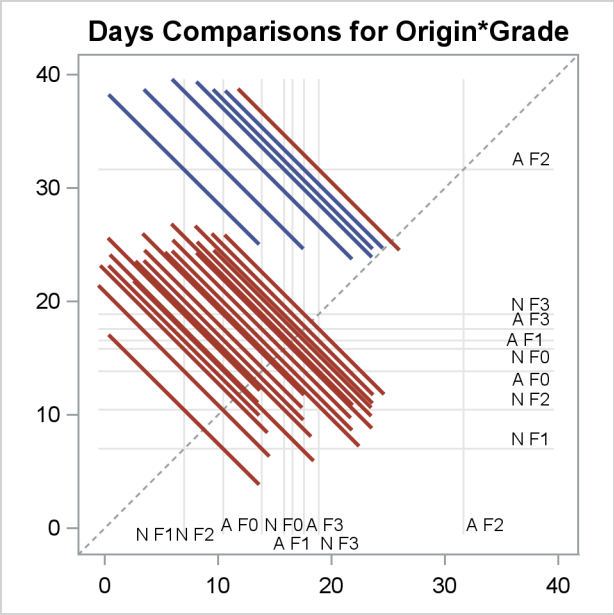
Figure 4 Plot of LS-Means Differences against the Control Group



The LSMEANS statement produces difference plots based on the options that are compatible with the display. The following program requests all pairwise LS-means differences with the multiplicity adjustment; the LSMEANS statement creates the diffogram (Figure 5):

```
proc plm restore=Ozkidsfit;
   lsmeans Origin*Grade/diff adjust=simulate stepdown;
run;
```

Figure 5 Plot of All Pairwise LS-Means Differences



In a diffogram, each line segment corresponds to one pairwise difference between LS-means. A line segment centers at the LS-means in a pair and has both a vertical and a horizontal line that indicate values and levels that correspond to the pair. The length of the line segment reflects the projected width of a confidence interval for the difference. Any line segment that does not cross the 45-degree reference line suggests significant LS-means difference. Notice that the LS-mean of the group of Aboriginal children from the second form in secondary schools is significantly different from any other groups.

USING THE LSMEANS STATEMENT

The syntax for the LSMEANS statement is defined as follows:

LSMEANS < *model-effects* > < / *options* > **;**

By default, LS-means are computed for any effect in the statistical model that involves only classification variables. If you supply the optional model effects, LS-means are computed for the specified effects. You can specify multiple LSMEANS statements, and all LSMEANS statements must appear after the MODEL statement.

In SAS/STAT 9.22, the LSMEANS statement was made available in 11 procedures: GENMOD, GLIMMIX, GLM, LOGISTIC, MIXED, ORTHOREG, PHREG, PLM, SURVEYLOGISTIC, SURVEYPHREG, and SURVEYREG. Starting with SAS/STAT 12.1, the statement is also available in the LIFEREG and PROBIT procedures.

The LSMEANS statement is one of several statements that you can use to analyze LS-means. The new SLICE statement analyzes partitions of higher-order effects that consist of at least two classification variables. It shares all the options found in the LSMEANS statement and also has its own options. The LSMESTIMATE statement is also new. It enables you to perform customized hypothesis testing of linear combinations of LS-means.

The LSMEANS, LSMESTIMATE, and SLICE statements offer the following features:

- incorporation of constructed effects by the EFFECT statement
- customized LS-means and LS-means differences
- multiple comparisons with adjustment
- options for transformations under the context of generalized linear models
- LS-means analysis based on posterior samples from Bayesian models
- ODS statistical graphics of means and means comparisons

INTEGRATION WITH CONSTRUCTED EFFECTS BY THE EFFECT STATEMENT

The EFFECT statement, which appears in many SAS/STAT procedures, became production in SAS 9.3. The EFFECT statement extends the way you can build collections of columns of model effects for design matrices. The collections are called *constructed effects* to distinguish them from the usual model effects that are formed from continuous or classification variables. If a statistical model contains constructed effects such as polynomials or splines, then the LSMEANS, LSMESTIMATE, and SLICE statements incorporate the information and make appropriate computations. Constructed effects can play two roles in a model from which the LS-means computation is possible: When the effects are classification effects, the LS-means can be computed on the effects. When the effects are covariates, LS-means are computed on the classification effects in fitted models. Table 2 shows how LS-means are computed when a model contains constructed effects.

Table 2 LS-Means Computation with Constructed Effects

Type of Constructed Effects	LS-Means for Constructed Effects in Question	LS-Means with Constructed Effects as Covariates
COLLECTION	The effect should be constructed from classification variables only. For each classification variable, the coefficient is 1 at each of its levels and $1/l^*$ at each level of other classification variables.	For each classification variable, the coefficient is $1/l$ at each of its levels. For each continuous variable, the coefficient is its mean.
LAG	At each level of the lag effect, the coefficient is 1.	At each level of the lag effect, the coefficient is $f/(nl)^{**}$.
MULTIMEMBER | MM	At each level of the multimember effect, the coefficient is m^{***}.	At each level of the multimember effect, the coefficient is m/l.
POLYNOMIAL | POLY	Not applicable	The coefficient that corresponds to each column of the polynomial transformations is the mean of that polynomial column.
SPLINE	Not applicable	The coefficient that corresponds to each column of the spline bases is the mean of that basis column.

* l is the number of levels for a classification variable.

** f is the number of total frequencies of all lag effect levels, and n is the total number of observations used for model fitting.

*** m is the total number of variables that are used to construct the effect.

For an example that computes LS-means with a covariate from a constructed effect, see the section "A NONPARAMETRIC MODEL WITH CONSTRUCTED EFFECTS" on page 10.

CUSTOMIZED LS-MEANS AND LS-MEANS DIFFERENCES

The standard LS-means computations have equal coefficients across classification effects. However, you can use the OBSMARGINS option to specify a potentially different weighting scheme for computing the LS-means. The OBSMARGINS option (which can also be written as the OM option) changes the coefficients to be proportional to those found in the original data set or a secondary data set. Furthermore, you can add the BYLEVEL option to modify the observed-margins LS-means. The BYLEVEL option requests separate margins for each level of the LS-means effect in question. For an example that uses the OBSMARGINS and BYLEVEL options to form estimable LS-means comparisons, see the section "NONESTIMABLE LS-MEANS AND LS-MEANS DIFFERENCES" on page 16.

You can add the E option to print the $\mathbf{L}$ matrix coefficients to verify that they are correct. It is possible that the modified LS-means are not estimable when standard means are estimable, or vice versa.

You can use the DIFF option to customize LS-means differences. By default, this option requests all pairwise differences of estimable LS-means. You can customize LS-means differences by using the following values of the DIFF= option:

ANOM requests differences between each LS-mean and the average LS-mean. The average LS-mean is computed as a weighted mean of LS-means, with weights inversely proportional to the diagonal entries of the $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'$ matrix. If there are nonestimable LS-means, the design-based weighted mean is replaced by an equally weighted mean.

CONTROL requests differences by using a control, which is the first valid level of the specified effect by default. You can instead specify a quoted, formatted value of the classification variable as a control.

CONTROLL tests whether the noncontrol levels are significantly smaller than the control.

CONTROLU tests whether the noncontrol levels are significantly larger than the control.

Sometimes testing hypotheses on LS-means involves linear functions other than simple differences. For example, you might want to test whether the LS-mean for the first level is significantly less than the average of the LS-means for the second and third levels for a classification effect. You can use the LSMESTIMATE statement to perform the analysis.

MULTIPLE COMPARISONS WITH ADJUSTMENT

You usually use the LSMEANS statement to compute and display standard differences between LS-means. If you need to consider more than one difference, you need to adjust for multiplicity, because making a larger number of comparisons increases the chance of finding differences that appear to be significant when they are actually not. You can use the ADJUST= option to adjust individual test p-values to control the probability of making erroneous inferences. For more information about various adjustment methods, see Westfall et al. (1999). For an example where Bonferroni adjustment is applied for comparing LS-means, see the section "LS-MEANS WITH GENERALIZED LINEAR MODELS" on page 12.

TRANSFORMATIONS OF LS-MEANS AND LS-MEANS DIFFERENCES

Like the structure of general linear models that are fit by ordinary least squares, the structure of the design matrix $\mathbf{X}$ for models that involve nonlinear optimization (such as generalized linear models and proportional hazards models) informs the analysis about levels of existing classification variables. LS-means are meaningful quantities in such models because they correspond to averaged predictions at the scale of the linear predictor that is formed as $\mathbf{X}\boldsymbol{\beta}$. For example, in a binomial model that uses the logit link, the LS-means are predicted population margins of the log odds, and the LS-means differences are predicted population margins of log odds ratios. For an example, see the section "LS-MEANS WITH GENERALIZED LINEAR MODELS" on page 12. You can transform the LS-means to the data scale by using the ILINK option, and you can transform the LS-means differences to odds ratios by using the ODDSRATIO option. In proportional hazards models, the LS-means differences are predicted population margins of log hazard ratios, and you can produce estimates of hazard ratios by using the EXP option. Note that the EXP, ILINK, and ODDSRATIO options produce nonlinear transformations of the LS-means differences, not differences of nonlinear transformations of LS-means.

LS-MEANS ANALYSIS BASED ON BAYESIAN POSTERIOR SAMPLES

Some procedures (GENMOD, LIFEREG, and PHREG) fit Bayesian models via the BAYES statement. When you perform postfitting analysis on Bayesian models, you use posterior samples of covariates to construct test statistics. You can use the LSMEANS, LSMESTIMATE, and SLICE statements to hypothesize and visualize LS-means of the posterior samples. For an example, see the section "LS-MEANS ANALYSIS OF A BAYESIAN PROPORTIONAL ODDS MODEL" on page 14.

ODS STATISTICAL GRAPHICS OF LS-MEANS AND DIFFERENCES

The LSMEANS statement includes a PLOTS= option for depicting LS-means and their differences. When you specify this option, the LSMEANS statement produces appropriate ODS statistical graphics, depending on the analysis that is performed on the LS-means. These graphics could be ANOM plots, control plots, or

diffograms for LS-means comparisons; marginal or interaction means plots; or box plots or histograms for posterior samples of LS-means or LS-means differences for Bayesian analysis. You can also request specific types of plots in the PLOTS= option. The section "INTRODUCTORY EXAMPLE" on page 4 illustrates how to produce graphics.

EXAMPLES OF USING THE LSMEANS STATEMENT

The following examples illustrate different ways you can use the LSMEANS and associated statements. You can perform the LS-means analyses by adding the LSMEANS statement after the MODEL statement, or you can get the same results by applying the LSMEANS statement in PROC PLM to a saved item store from a model fit.

A NONPARAMETRIC MODEL WITH CONSTRUCTED EFFECTS

This example illustrates how you can analyze LS-means of spline functions that model nonlinear dependencies. Consider an agronomic study of the effect of different amounts of a fertilizer on the growth of two flower species. A researcher applies the fertilizer to each of 100 plants of the two species and records their growth at harvest. The following DATA step reads the data:

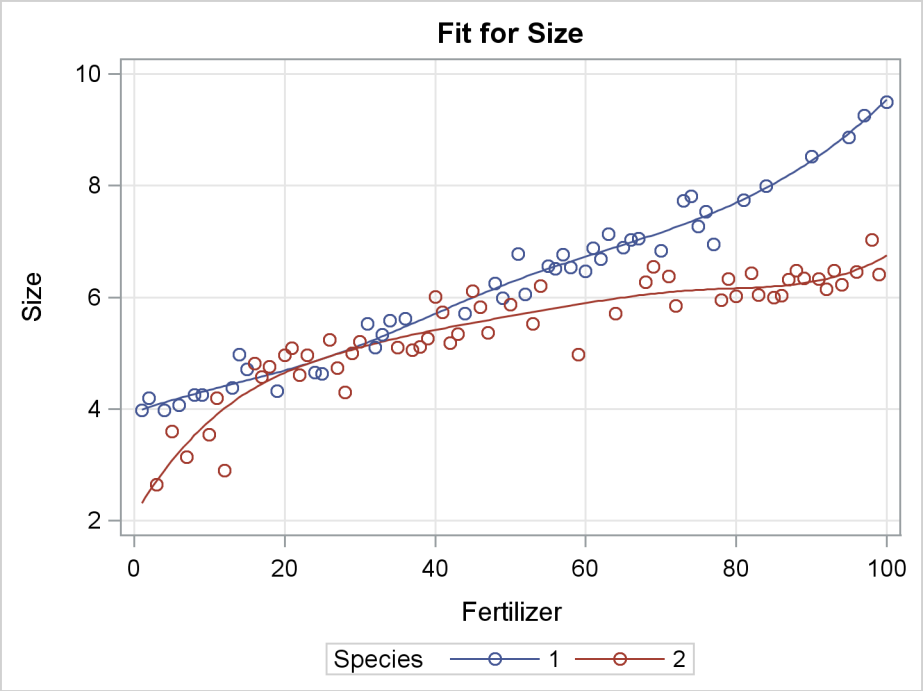
```
data Flowers;
   input Species Size @@;
   Fertilizer = _n_;
   datalines;
1 3.980   1 4.199   2 2.640   1 3.974   2 3.603   1 4.065   2 3.139   1 4.251
1 4.253   2 3.540   2 4.195   2 2.892   1 4.379   1 4.971   1 4.712   2 4.811
2 4.574   2 4.755   1 4.316   2 4.961   2 5.088   2 4.607   2 4.959   1 4.653
1 4.629   2 5.237   2 4.734   2 4.299   2 5.002   2 5.201   1 5.520   1 5.105
1 5.329   1 5.580   2 5.098   1 5.613   2 5.052   2 5.108   2 5.257   2 6.005
2 5.726   2 5.179   2 5.338   1 5.707   2 6.105   2 5.828   2 5.368   1 6.252
1 5.984   2 5.867   1 6.771   1 6.052   2 5.522   2 6.200   1 6.562   1 6.517
1 6.769   1 6.534   2 4.969   1 6.460   1 6.873   1 6.678   1 7.135   2 5.705
1 6.893   1 7.023   1 7.050   2 6.273   2 6.549   1 6.836   2 6.375   2 5.841
1 7.727   1 7.806   1 7.269   1 7.533   1 6.948   2 5.954   2 6.326   2 6.017
1 7.744   2 6.431   2 6.040   1 7.995   2 5.996   2 6.028   2 6.321   2 6.479
2 6.337   1 8.516   2 6.326   2 6.144   2 6.474   2 6.221   1 8.867   2 6.453
1 9.253   2 7.024   2 6.403   1 9.498
;
```

The researcher believes that the growth rate for each species is a nonlinear function of the fertilizer amount. Accordingly, she uses the EFFECT statement to fit separate spline functions of **Fertilizer** for each **Species**, which fits an ANCOVA model. The following statements save the fit results in an item store named FlowerModel.

```
proc orthoreg data=Flowers;
   class Species;
   effect SmoothF = spline(Fertilizer);
   model Size = Species|SmoothF;
   effectplot / obs;
   store FlowerModel;
run;
```

In the fit plot produced by the EFFECTPLOT statement in Figure 6, you can observe the nonlinear dependence of the variable **Size** on the variable **Fertilizer**, especially for the second flower species. There are no major differences in predicted species growth from fertilizer amount 20 to 40. When the fertilizer amount is greater than 60, the predicted growth for the first flower species is larger than that of the second one.

Figure 6 Predicted Values by Group



The researcher decides to compare the LS-means of **Species** at two specific fertilizer amounts (30 and 80) to see how much the LS-means differ and whether the differences between the two species are significant. She uses the AT option in the LSMEANS statement in PROC PLM to perform the postfitting analysis:

```
proc plm restore=FlowerModel;
   lsmeans Species / diff at Fertilizer=30;
   lsmeans Species / diff at Fertilizer=80;
run;
```

Figure 7 confirms the observation in the fit plot (Figure 6): there is no significant difference between the two species at fertilizer amount 30, but there is a significant difference at fertilizer amount 80.

Figure 7 LS-Means at Two Fertilizer Amounts

The PLM Procedure

Species Least Squares Means

Species	Fertilizer	Estimate	Standard Error	DF	t Value	Pr > |t|
1	30.00	5.1394	0.1245	86	41.29	<.0001
2	30.00	5.1026	0.1030	86	49.52	<.0001

Differences of Species Least Squares Means

Species	_Species	Fertilizer	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2	30.00	0.03686	0.1616	86	0.23	0.8201

Species Least Squares Means

Species	Fertilizer	Estimate	Standard Error	DF	t Value	Pr > |t|
1	80.00	7.6941	0.1173	86	65.62	<.0001
2	80.00	6.1616	0.09387	86	65.64	<.0001

Differences of Species Least Squares Means

Species	_Species	Fertilizer	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2	80.00	1.5325	0.1502	86	10.20	<.0001

The researcher also wants to test how effective different amounts of the fertilizer are on the average growth across species. Specific questions include: (1) At what level does the fertilizer begin to affect growth? (2) Does the fertilizer yield an average growth larger than 6 at amount 40? The fit plot in Figure 6 suggests that the fertilizer begins to affect the growth of both species approximately at amount 10. She uses the following statement to carry out these two tests:

```
proc plm restore=FlowerModel;
   lsmestimate Species 'impact at 10' 1 1 divisor=2 / at Fertilizer=10;
   lsmestimate Species 'impact at 40' 1 1 divisor=2 / at Fertilizer=40 testvalue=6 upper;
run;
```

Figure 8 shows that there is minimal evidence that the fertilizer affects average species growth at amount 10. There is no evidence that the fertilizer could yield average growth larger than 6 with amount 40. These results warrant further investigation, probably at a higher fertilizer amount.

Figure 8 Average LS-Means at Two Fertilizer Amounts

The PLM Procedure

Least Squares Means Estimate

Effect	Label	Fertilizer	Estimate	Standard Error	DF	t Value	Tails	Pr > |t|
Species	impact at 10	10.00	4.0636	0.09029	86	45.01	Both	<.0001

Least Squares Means Estimate

Effect	Label	Fertilizer	Estimate	Standard Error	DF	Test Value t	Value	Tails	Pr > t
Species	impact at 40	40.00	5.5609	0.06528	86	6.000	-6.73	Upper	1.0000

The fitted model in this example contains a constructed spline effect on fertilizer amount. The saved item store FlowerModel contains all the necessary information about the spline construction and all associated parameters for spline bases. When the researcher specifies the AT option for **Fertilizer**, the LSMEANS and LSMESTIMATE statements recognize that the variable forms the constructed effect **SmoothF** in the fitted model, and they form design values for spline bases based on the specified AT= values in order to compute LS-means and differences. You can perform equivalent analyses by using the ESTIMATE statement, but it is more difficult. The TESTVALUE= and UPPER options provide convenient ways to customize hypotheses.

LS-MEANS WITH GENERALIZED LINEAR MODELS

This example illustrates how you can use the LSMEANS statement to compute odds ratios with confidence limits adjusted for multiplicity of an interaction effect in a logistic model. Consider a study of analgesic effects of treatments of elderly patients who have neuralgia. Two treatments and a placebo are compared. The response variable is whether the patient reported pain or not. Researchers recorded the age and gender of 60 patients along with the duration of complaint before the treatment began. The following DATA step creates the data set Neuralgia:

```
Data Neuralgia;
   input Treatment $ Sex $ Age Duration Pain $ @@;
   datalines;
P F 68  1 No   B M 74 16 No   P F 67 30 No   P M 66 26 Yes  B M 70 22 No
B F 67 28 No   B F 77 16 No   A F 71 12 No   B F 72 50 No   A M 65 15 No
B F 76  9 Yes  A M 71 17 Yes  A F 63 27 No   A F 69 18 Yes  P F 67  1 Yes
B F 66 12 No   A M 62 42 No   P F 64  1 Yes  A F 64 17 No   A M 67 10 No
P M 74  4 No   A F 72 25 No   P M 70  1 Yes  B M 66 19 No   P F 72 11 Yes
B M 59 29 No   A F 64 30 No   A M 70 28 No   A M 69  1 No   A F 74  1 No
B F 78  1 No   P M 83  1 Yes  B F 69 42 No   B M 75 30 Yes  B M 80 21 Yes
P M 77 29 Yes  P F 79 20 Yes  A M 70 12 No   A F 69 12 No   A F 69  3 No
B F 65 14 No   B M 70  1 No   B M 67 23 No   A M 76 25 Yes  B F 65  7 No
P M 78 12 Yes  B M 77  1 Yes  B F 69 24 No   P M 66  4 Yes  P F 68 27 Yes
P F 65 29 No   P M 60 26 Yes  A M 78 15 Yes  B M 75 21 Yes  P M 68 11 Yes
A F 67 11 No   P F 72 27 No   P F 70 13 Yes  A M 75  6 Yes  P M 67 17 Yes
;
```

There are five variables in the data set. **Treatment** is a categorical variable that has three levels: A and B are two test treatments, and P represents the placebo treatment. **Sex** is a categorical variable for gender. **Age** is the age of each patient. **Duration** is the duration of the complaint (in months) before the treatment began. **Pain** is the binary response variable that indicates whether there was pain.

You can use the following program to fit a logistic model that consists of three explanatory variables (**Treatment**, **Sex**, and **Age**) and save the model in an item store named NeuralgiaModel:

```
proc logistic data=Neuralgia;
   class Treatment Sex / param=glm;
   model Pain = Treatment|Sex Age;
   store NeuralgiaModel;
run;
```

You can use the LSMEANS statement to compute the odds ratios for the **Treatment** variable for female patients, adjusting for their ages as follows:

```
proc plm restore=NeuralgiaModel;
   lsmeans Treatment*Sex/diff oddsratio adjust=bon cl;
run;
```

The results of the preceding program include all pairwise difference comparisons at all levels of both **Treatment** and **Sex**. You need to filter out unwanted results. The Bonferroni adjustment in multiple comparison takes the extra number of comparisons into account. However, if you are interested only in a certain level of the interaction effect, you can use the SLICE statement. The following program computes odds ratios for **Treatment** for female patients:

```
proc plm restore=NeuralgiaModel;
   slice Treatment*Sex/sliceby(sex='F') diff oddsratio adjust=bon cl;
run;
```

The Bonferroni adjustment is used in the multiple comparisons between LS-means. The confidence limits for both LS-means differences and odds ratios are also reported. The chi-square test in Figure 9 rejects the hypothesis that the LS-means for **Treatment*Sex** are all equal for female patients. Then individual tests at each level of **Treatment** are performed.

Figure 9 Chi-Square Test for **Treatment*Sex**

The PLM Procedure

Chi-Square Test for Treatment*Sex Least Squares Means Slice

Slice	Num DF	Chi-Square	Pr > ChiSq
Sex F	2	8.22	0.0164

Figure 10 lists the partial results that the SLICE statement produces.

Figure 10 Odds Ratios and (Adjusted) Confidence Limits

Simple Differences of Treatment*Sex Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

Slice	Treatment	_Treatment	Estimate	Standard Error	z Value	Pr > |z|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Sex F	A	B	-0.9224	1.6311	-0.57	0.5717	1.0000	0.05	-4.1193	2.2744	-4.8272	2.9824
Sex F	A	P	2.8269	1.3207	2.14	0.0323	0.0970	0.05	0.2384	5.4154	-0.3348	5.9886
Sex F	B	P	3.7493	1.4933	2.51	0.0120	0.0361	0.05	0.8225	6.6761	0.1744	7.3243

Simple Differences of Treatment*Sex Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

Slice	Treatment	_Treatment	Odds Ratio	Lower Confidence Limit for Odds Ratio	Upper Confidence Limit for Odds Ratio	Adj Lower Odds Ratio	Adj Upper Odds Ratio
Sex F	A	B	0.398	0.016	9.722	0.008	19.734
Sex F	A	P	16.892	1.269	224.838	0.715	398.848
Sex F	B	P	42.492	2.276	793.254	1.190	>999.999

According to the fitted logistic model, the odds that a female patient does not report pain when she receives treatment B are significantly greater than the odds that she does not report pain when she receives the placebo. There is slight evidence that the odds of reporting pain with treatment A are greater than with the placebo, after you adjust for multiplicity.

LS-MEANS ANALYSIS OF A BAYESIAN PROPORTIONAL ODDS MODEL

This example illustrates how you can perform an LS-means analysis of a Bayesian proportional odds model that is fit by PROC LIFEREG. The following DATA step creates the SAS data set Larynx from the larynx cancer data in Klein and Moeschberger (2003). The variable **Time** is the logarithm of the intervals (in years) between first treatment and either death or the end of the study. The variable **Age** records each patient's age at the time of diagnosis. The variable **Year** records the year of diagnosis. The variable **Stage** records the stage of the patient's cancer.

```
data Larynx;
   input Stage Time Age Year Death @@;
   label Time='log(Time)';
   datalines;
1 0.6 77 76 1  1 1.3 53 71 1  1 2.4 45 71 1  1 2.5 57 78 0  1 3.2 58 74 1  3 9.3 69 71 0
1 3.2 51 77 0  1 3.3 76 74 1  1 3.3 63 77 0  1 3.5 43 71 1  1 3.5 60 73 1  3 10.1 51 71 0
1 4.0 52 71 1  1 4.0 63 76 1  1 4.3 86 74 1  1 4.5 48 76 0  1 4.5 68 76 0  4 0.1 65 72 1
1 5.3 81 72 1  1 5.5 70 75 0  1 5.9 58 75 0  1 5.9 47 75 0  1 6.0 75 73 1  4 0.3 71 76 1
1 6.1 77 75 0  1 6.2 64 75 0  1 6.4 77 72 1  1 6.5 67 70 1  1 6.5 79 74 0  4 0.4 76 77 1
1 6.7 61 74 0  1 7.0 66 74 0  1 7.4 68 71 1  1 7.4 73 73 0  1 8.1 56 73 0  4 0.8 65 76 1
1 8.1 73 73 0  1 9.6 58 71 0  1 10.7 68 70 0 2 0.2 86 74 1  2 1.8 64 77 1  4 0.8 78 77 1
2 2.0 63 75 1  2 2.2 71 78 0  2 2.6 67 78 0  2 3.3 51 77 0  2 3.6 70 77 1  4 1.0 41 77 1
2 3.6 72 77 0  2 4.0 81 71 1  2 4.3 47 76 0  2 4.3 64 76 0  2 5.0 66 76 0  4 1.5 68 73 1
2 6.2 74 72 1  2 7.0 62 73 1  2 7.5 50 73 0  2 7.6 53 73 0  2 9.3 61 71 0  4 2.0 69 76 1
3 0.3 49 72 1  3 0.3 71 76 1  3 0.5 57 74 1  3 0.7 79 77 1  3 0.8 82 74 1  4 2.3 62 71 1
3 1.0 49 76 1  3 1.3 60 76 1  3 1.6 64 72 1  3 1.8 74 71 1  3 1.9 72 74 1  4 2.9 74 78 0
3 1.9 53 74 1  3 3.2 54 75 1  3 3.5 81 74 1  3 3.7 52 77 0  3 4.5 66 76 0  4 3.6 71 75 1
3 4.8 54 76 0  3 4.8 63 76 0  3 5.0 59 73 1  3 5.0 49 76 0  3 5.1 69 76 0  4 3.8 84 74 1
3 6.3 70 72 1  3 6.4 65 72 1  3 6.5 65 74 0  3 7.8 68 72 1  3 8.0 78 73 0  4 4.3 48 76 0
;
```

The study investigates the effects of a patient's age and cancer stage on survival. The following statements sort the data by the variable **Stage** in descending order and then fit a Bayesian proportional odds model by using the LIFEREG procedure with the BAYES statement. A log-logistic distribution is assumed for a patient's survival time. The fitted model is saved in an item store named LifeModel.

```
proc sort data=Larynx;
   by DESCENDING Stage;
run;

proc lifereg data=Larynx order=data;
   class Stage;
   model Time*Death(0) = Age Stage / dist = llogistic;
   bayes seed=100 nmc=500 nbi=500 diagnostic=none;
   store LifeModel;
run;
```

Suppose you want to test whether the odds of survival at cancer stages 4, 3, and 2 are different from those at stage 1. You can use the LSMESTIMATE statement in PROC PLM to perform the postfitting analysis, as shown in the following program. The PERCENTILES= option requests the 5th and 95th percentiles of the LS-means and differences from the posterior sample. The EXP option requests exponentiation of the LS-means and differences. The transformed values correspond to odds of survival under the context of the proportional odds model.

```
proc plm restore=LifeModel percentiles=(5,95);
   lsmestimate Stage '4 vs 1' 1 0 0 -1,
                     '3 vs 1' 0 1 0 -1,
                     '2 vs 1' 0 0 1 -1 / cl exp
                      plots=boxplot(orient=horizontal);
run;
```

Figure 11 shows that the odds of survival at stage 2 are not much different from those at stage 1, whereas the odds of survival at stage 3 and 4 are significantly lower than those at stage 1.

Figure 11 Custom LS-Means Tests

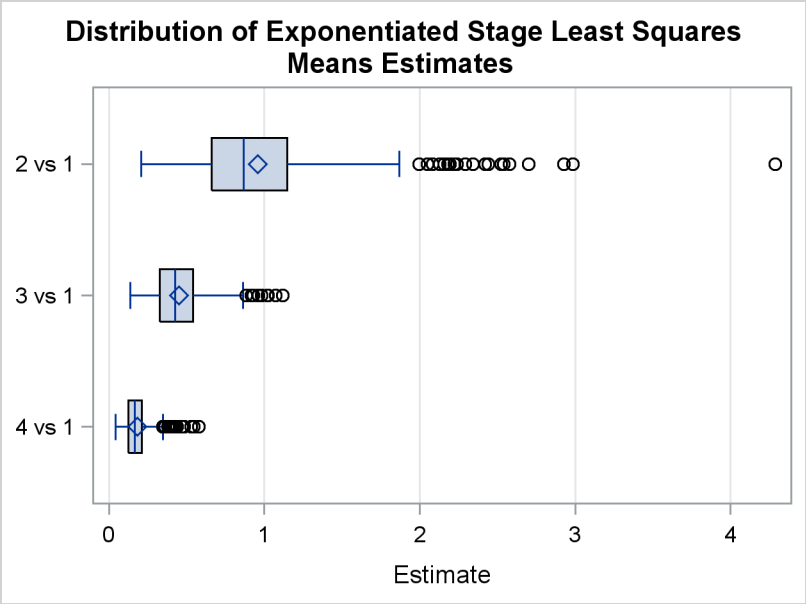
The PLM Procedure

Sample Least Squares Means Estimates

Effect Label	N	Estimate	Standard Deviation	Percentiles 5th	Percentiles 95th	Alpha	Lower HPD	Upper HPD	Exponentiated	Standard Deviation of Exponentiated	Percentiles for Exponentiated 5th	Percentiles for Exponentiated 95th	Lower HPD of Exponentiated	Upper HPD of Exponentiated
Stage 4 vs 1	500	-1.8001	0.4321	-2.5007	-1.0549	0.05	-2.6279	-0.8897	0.1815	0.082975	0.0820	0.3482	0.05786	0.3605
Stage 3 vs 1	500	-0.8694	0.3727	-1.5109	-0.2743	0.05	-1.6033	-0.2031	0.4488	0.168055	0.2207	0.7601	0.1748	0.7795
Stage 2 vs 1	500	-0.1410	0.4413	-0.8678	0.5695	0.05	-1.1401	0.6252	0.9585	0.462376	0.4199	1.7675	0.3198	1.8686

The box plot in Figure 12 shows the distribution of the LS-means differences that are computed from the posterior samples at the exponential scale.

Figure 12 Box Plot of LS-Means Differences



NONESTIMABLE LS-MEANS AND LS-MEANS DIFFERENCES

This example describes the nonestimability problem in LS-means and how you can use the OM option for different weighting schemes that enable you to construct estimable LS-means. In a study of wheat yield in a designed experiment, a researcher considers two factors: soil type (variable **Soil**) and fertilizer type (variable **Fertilizer**). The response variable is the wheat yield (variable **Yield**). The following DATA step creates the data set Yield.

```
data Yield;
  input Soil $ Fertilizer $ Yield @@;
  datalines;
S1 B 4 S1 B 7 S1 C 0 S1 E 9 S1 E 2 S1 B 5 S1 A 5
S1 D 5 S2 E 7 S2 E 2 S2 E 6 S2 A 5 S2 D 5
;
```

The researcher first uses the FREQ procedure to inspect the crosstabulation:

```
proc freq data=Yield;
   table Soil*Fertilizer/norow;
run;
```

Figure 13 shows that this is an unbalanced design, because frequencies of soil types and fertilizer types vary across cells. Also, there are two missing cells: wheat yield with fertilizer B and C at soil type S2.

Figure 13 Two-Way Frequencies

The FREQ Procedure

Frequency Percent Col Pct	Table of Soil by Fertilizer Fertilizer					
Soil	A	B	C	D	E	Total
S1	1	3	1	1	2	8
	7.69	23.08	7.69	7.69	15.38	61.54
	50.00	100.00	100.00	50.00	40.00	
S2	1	0	0	1	3	5
	7.69	0.00	0.00	7.69	23.08	38.46
	50.00	0.00	0.00	50.00	60.00	
Total	2	3	1	2	5	13
	15.38	23.08	7.69	15.38	38.46	100.00

The researcher then decides to fit a two-way interaction model and compute LS-means and differences for the variable **Fertilizer**. For his analysis, he chooses PROC MIXED from the many SAS/STAT procedures that could perform the model fitting and postfitting analysis:

```
proc mixed data=Yield;
   class Soil Fertilizer;
   model Yield = Soil|Fertilizer;
   lsmeans fertilizer / diff;
run;
```

Figure 14 displays fertilizer LS-means and differences. Among the LS-means at all levels of **Fertilizer**, the LS-means for fertilizer B and C are nonestimable because they have missing cells for the two levels at soil type S2. The differences between other estimable LS-means and these two nonestimable LS-means are nonestimable. However, the difference between these two nonestimable LS-means themselves is actually estimable.

Figure 14 LS-Means and Differences

The Mixed Procedure

Least Squares Means

Effect	Fertilizer	Estimate	Standard Error	DF	t Value	Pr > |t|
Fertilizer	A	5.0000	2.0777	5	2.41	0.0611
Fertilizer	B	Non-est	.	.	.	.
Fertilizer	C	Non-est	.	.	.	.
Fertilizer	D	5.0000	2.0777	5	2.41	0.0611
Fertilizer	E	5.2500	1.3411	5	3.91	0.0112

Differences of Least Squares Means

Effect	Fertilizer	_Fertilizer	Estimate	Standard Error	DF	t Value	Pr > |t|
Fertilizer	A	B	Non-est	.	.	.	.
Fertilizer	A	C	Non-est	.	.	.	.
Fertilizer	A	D	-111E-18	2.9383	5	-0.00	1.0000
Fertilizer	A	E	-0.2500	2.4729	5	-0.10	0.9234
Fertilizer	B	C	5.3333	3.3928	5	1.57	0.1768
Fertilizer	B	D	Non-est	.	.	.	.
Fertilizer	B	E	Non-est	.	.	.	.
Fertilizer	C	D	Non-est	.	.	.	.
Fertilizer	C	E	Non-est	.	.	.	.
Fertilizer	D	E	-0.2500	2.4729	5	-0.10	0.9234

The $\mathbf{L}$ matrix for the fertilizer LS-means is

$$\begin{bmatrix} 1 & 0.5 & 0.5 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 0.5 & 0.5 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 0.5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 \\ 1 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \end{bmatrix}$$

The $\mathbf{L}$ coefficients that correspond to the difference between fertilizer B and C at soil type S2 are

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

These coefficients correspond to the difference between fertilizer B and C at soil type S1 because the LSMEANS statement uses averages from nonmissing cells, which yield coefficient 1 at soil type S1 instead of equal weights at both soil types. By taking the LS-means difference, the analysis compares fertilizers at soil S1, and the difference is estimable.

The default weighting scheme for data that contain missing cells causes nonestimable LS-means. The researcher then investigates a different set of LS-means and differences by adding the OM option to the LSMEANS statement, as follows:

```
proc mixed data=Yield;
   class Soil Fertilizer;
   model Yield = Soil|Fertilizer;
   lsmeans Fertilizer / diff om;
run;
```

Figure 15 shows that none of the LS-means are estimable and only two LS-means differences are estimable: one is between fertilizer A and D, and the other is between fertilizer B and C.

Figure 15 LS-Means and Differences with the OM Option

The Mixed Procedure

Least Squares Means

Effect	Fertilizer	Margins	Estimate	Standard Error	DF	t Value	Pr > |t|
Fertilizer	A	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	B	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	C	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	D	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	E	WORK.YIELD	Non-est	.	.	.	.

Differences of Least Squares Means

Effect	Fertilizer	_Fertilizer	Margins	Estimate	Standard Error	DF	t Value	Pr > |t|
Fertilizer	A	B	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	A	C	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	A	D	WORK.YIELD	-111E-18	2.9383	5	-0.00	1.0000
Fertilizer	A	E	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	B	C	WORK.YIELD	5.3333	3.3928	5	1.57	0.1768
Fertilizer	B	D	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	B	E	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	C	D	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	C	E	WORK.YIELD	Non-est	.	.	.	.
Fertilizer	D	E	WORK.YIELD	Non-est	.	.	.	.

The $\mathbf{L}$ matrix for the fertilizer LS-means is

$$\begin{bmatrix} 1 & 0.6154 & 0.3846 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 0.6154 & 0.3846 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.6154 & 0.3846 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.6154 & 0.3846 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 \\ 1 & 0.6154 & 0.3846 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.4 & 0 & 0 & 0.6 \end{bmatrix}$$

The OM option requests coefficients that are proportional to the cell frequencies. As shown in Figure 13, the percentage is 61.54% for soil type S1 and 38.46% for soil type S2. These percentages determine the coefficients for **Soil** levels. For levels of **Soil*Fertilizer**, the coefficients are computed from column percentages that are reported in Figure 13. None of the LS-means are estimable, because their $\mathbf{L}$ matrices cannot be represented as linear combinations of $\mathbf{X}$. However, the difference between fertilizer B and C is estimable, because it is actually the difference at soil type S1. The corresponding $\mathbf{L}$ coefficients for the difference between fertilizer A and D are

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0.5 & 0 & -1 & 0 & 0 & 0.5 & 0 & 0 \end{bmatrix}$$

These coefficients can be represented as a linear combination of $\mathbf{X}$.

The researcher performs one more step to obtain another set of results for LS-means and differences by adding a BYLEVEL option to the LSMEANS statement, as in the following statements. The BYLEVEL option requests observed margins at each level of **Fertilizer** in order to scale the means. So the resulting LS-means are actually equal to raw means.

```
proc mixed data=Yield;
   class Soil Fertilizer;
   model Yield = Soil|Fertilizer;
   lsmeans Fertilizer / diff om bylevel;
run;
```

This time, all the LS-means and their differences are estimable, as shown in Figure 16.

Figure 16 LS-Means and Differences with the OM and BYLEVEL Options

The Mixed Procedure

Least Squares Means

Effect	Fertilizer	Margins	By Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Fertilizer	A	WORK.YIELD	Yes	5.0000	2.0777	5	2.41	0.0611
Fertilizer	B	WORK.YIELD	Yes	5.3333	1.6964	5	3.14	0.0256
Fertilizer	C	WORK.YIELD	Yes	8.88E-16	2.9383	5	0.00	1.0000
Fertilizer	D	WORK.YIELD	Yes	5.0000	2.0777	5	2.41	0.0611
Fertilizer	E	WORK.YIELD	Yes	5.2000	1.3140	5	3.96	0.0108

Differences of Least Squares Means

Effect	Fertilizer	_Fertilizer	Margins	By Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Fertilizer	A	B	WORK.YIELD	Yes	-0.3333	2.6822	5	-0.12	0.9059
Fertilizer	A	C	WORK.YIELD	Yes	5.0000	3.5986	5	1.39	0.2234
Fertilizer	A	D	WORK.YIELD	Yes	-111E-18	2.9383	5	-0.00	1.0000
Fertilizer	A	E	WORK.YIELD	Yes	-0.2000	2.4583	5	-0.08	0.9383
Fertilizer	B	C	WORK.YIELD	Yes	5.3333	3.3928	5	1.57	0.1768
Fertilizer	B	D	WORK.YIELD	Yes	0.3333	2.6822	5	0.12	0.9059
Fertilizer	B	E	WORK.YIELD	Yes	0.1333	2.1458	5	0.06	0.9529
Fertilizer	C	D	WORK.YIELD	Yes	-5.0000	3.5986	5	-1.39	0.2234
Fertilizer	C	E	WORK.YIELD	Yes	-5.2000	3.2187	5	-1.62	0.1671
Fertilizer	D	E	WORK.YIELD	Yes	-0.2000	2.4583	5	-0.08	0.9383

The $\mathbf{L}$ matrix for the fertilizer LS-means is

$$\begin{bmatrix} 1 & 0.5 & 0.5 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 0.5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 \\ 1 & 0.4 & 0.6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.4 & 0 & 0 & 0.6 \end{bmatrix}$$

The coefficients for **Soil** are different between fertilizer levels, because the BYLEVEL option requests that the weights for averaging be computed from column percentages individually at each fertilizer level instead of using the global column percentage. The coefficients of levels of **Soil*Fertilizer** are computed the same way as when only the OM option was used. The LS-means and their differences based on the $\mathbf{L}$ matrix are all estimable.

By definition, LS-means are averages of cell means that have equal weights. Sometimes prior knowledge or analysis suggests alternative weighting schemes. Both the OM option and the BYLEVEL option provide ways to modify LS-means to approximate underlying population characteristics.

LS-MEANS IN A RANDOMIZED BLOCK DESIGN

This example shows how you can perform the LS-means analysis in a randomized block design and make adjusted multiple comparisons of the LS-means. Gotway and Stroup (1997) analyze data from an agronomic field trial. In these data, researchers study 16 varieties of wheat to determine their resistance to infestation by the Hessian fly. They form the randomized block design by assigning the varieties on an 8×8 grid. Each 4×4 quadrant of the grid contains 16 entries, which all together constitute a block. The researchers count the number of damaged plants and the total number of plants growing in each entry; they identify them by the block number and the entry number within each block, respectively. Two auxiliary variables indicate the coordinates of an entry in the grid. The following DATA step creates the data set HessianFly:

```
data HessianFly;
  label Y = 'No. of damaged plants'
        n = 'No. of plants';
  input Block Entry Lat Lng N Y @@;
  datalines;
1 14 1 1  8 2    1 16 1 2  9 1    3  7 1 5  7 7    3 13 1 6  7 0    2 12 7 3  9 2
1  7 1 3 13 9    1  6 1 4  9 9    3  8 1 7 13 3    3 14 1 8  9 0    2 16 7 4  9 0
1 13 2 1  9 2    1 15 2 2 14 7    3  4 2 5 15 11   3 10 2 6  9 7    4  3 7 7  9 9
1  8 2 3  8 6    1  5 2 4 11 8    3  3 2 7 15 11   3  9 2 8 13 5    4 10 7 8  6 6
1 11 3 1 12 7    1 12 3 2 11 8    3  6 3 5 16 9    3  1 3 6  8 8    2  9 8 1 14 9
1  2 3 3 10 8    1  3 3 4 12 5    3 15 3 7  7 0    3 12 3 8 12 8    2  1 8 2 13 12
1 10 4 1  9 7    1  9 4 2 15 8    3 11 4 5  8 1    3 16 4 6 15 1    4  2 8 5 12 8
1  4 4 3 19 6    1  1 4 4  8 7    3  5 4 7 12 7    3  2 4 8 16 12   4 11 8 6  9 7
2 15 5 1 15 6    2  3 5 2 11 9    4  9 5 5 15 8    4  4 5 6 10 6    2  8 8 3 12 3
2 10 5 3 12 5    2  2 5 4  9 9    4 12 5 7 13 5    4  1 5 8 15 9    2  4 8 4 14 7
2 11 6 1 20 10   2  7 6 2 10 8    4 15 6 5 17 6    4  6 6 6  8 2    4  5 8 7 11 10
2 14 6 3 12 4    2  6 6 4 10 7    4 14 6 7 12 5    4  7 6 8 15 8    4 16 8 8 15 7
2  5 7 1  8 8    2 13 7 2  6 0    4 13 7 5 13 2    4  8 7 6 13 9
;
```

The researchers consider a standard generalized linear model for independent binomial counts by assuming that infestations are independent among entries and that all plants within each entry have the same propensity for infestation. They use the following GLIMMIX procedure statements to fit the model and save it in an item store named HessianFlyModel:

```
proc glimmix data=HessianFly;
   class Block Entry;
   model y/n = Block Entry / dist=binomial link=logit;
   store HessianFlyModel;
run;
```

The "Class Level Information" table in Figure 17 lists the levels of the variable that is specified in the CLASS statement and shows the ordering of the levels.

Figure 17 Class Level Information

The GLIMMIX Procedure

Class Level Information	
Class	**Levels** **Values**
Block	4 1 2 3 4
Entry	16 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

The **Block** variable has four levels, and the **Entry** variable has 16 levels. The researchers want to compute the LS-means for **Entry**. They use the following PROC PLM statements to conduct the postfitting analysis:

```
proc plm restore=HessianFlyModel;
   lsmeans Entry/ilink adj=tukey lines;
run;
```

The ILINK option in the LSMEANS statement reports the LS-means on the scale of the mean, which are predicted probabilities. The ADJ=TUKEY option requests a Tukey-Kramer-type multiple comparison adjustment for the p-values for the differences of LS-means. This option implies the DIFF option. The LINES option presents the comparison results in a table that lists the LS-means in descending order and indicates nonsignificant subsets by line segments beside the corresponding means.

Figure 18 lists LS-means for **Entry**, the corresponding predicted probabilities, and their standard errors. The LS-means for many **Entry** levels are significantly different from 0. This means the infestation probabilities vary at different levels of **Entry**.

Given 16 levels of **Entry**, there are 120 pairs of LS-means differences. You can either check each pairwise difference in the LS-means differences table (not displayed here because of its large size) or use the LINES table to get a general sense of the grouping structures among **Entry** levels.

Figure 18 LS-Means

The PLM Procedure

Entry Least Squares Means

Entry	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error of Mean
1	1.4864	0.3921	45	3.79	0.0004	0.8155	0.05899
2	1.3453	0.3585	45	3.75	0.0005	0.7934	0.05877
3	0.9963	0.3278	45	3.04	0.0039	0.7303	0.06457
4	0.07592	0.2643	45	0.29	0.7753	0.5190	0.06599
5	1.3139	0.3775	45	3.48	0.0011	0.7882	0.06302
6	0.5758	0.3180	45	1.81	0.0768	0.6401	0.07325
7	0.8608	0.3302	45	2.61	0.0123	0.7028	0.06896
8	-0.1639	0.2975	45	-0.55	0.5843	0.4591	0.07387
9	0.09605	0.2662	45	0.36	0.7200	0.5240	0.06641
10	0.8413	0.3635	45	2.31	0.0253	0.6987	0.07651
11	0.03126	0.2883	45	0.11	0.9141	0.5078	0.07205
12	0.04234	0.2996	45	0.14	0.8882	0.5106	0.07486
13	-2.0941	0.5330	45	-3.93	0.0003	0.1097	0.05204
14	-1.0185	0.3538	45	-2.88	0.0061	0.2653	0.06897
15	-0.6303	0.2883	45	-2.19	0.0340	0.3475	0.06536
16	-1.4645	0.3713	45	-3.94	0.0003	0.1878	0.05663

Figure 19 displays results of comparisons between all pairs of LS-means by listing the means in descending order and indicating nonsignificant subsets by line segments.

Figure 19 Tukey-Kramer Grouping for LS-Means

Tukey-Kramer Grouping for Entry Least Squares Means (Alpha=0.05)
LS-means with the same letter are not significantly different.

Entry	Estimate					
1	1.4864				A	
					A	
2	1.3453				A	
					A	
5	1.3139				A	
					A	
3	0.9963				A	
					A	
7	0.8608		B		A	
			B		A	
10	0.8413		B		A	
			B		A	
6	0.5758		B		A	C
			B		A	C
9	0.09605		B	D	A	C
			B	D	A	C
4	0.07592		B	D	A	C
			B	D	A	C
12	0.04234	E	B	D	A	C
		E	B	D	A	C
11	0.03126	E	B	D	A	C
		E	B	D	A	C
8	-0.1639	E	B	D	A	C
		E	B	D		C
15	-0.6303	E	B	D		C
		E		D		C
14	-1.0185	E		D		C
		E		D		
16	-1.4645	E		D		
		E				
13	-2.0941	E				

Figure 19 identifies five subsets within which LS-means are not significantly different from each other. Aside from the LS-means, wihch are listed in descending order, five vertical character strings are connected by character symbols that indicate nonsignificant subsets. For example, subset A contains the following levels: 1, 2, 5, 3, 7, 10, 6, 9, 4, 12, 11, and 8. The pairwise differences among these levels in subset A are not significant at the 0.05 nominal level.

SUMMARY

The LSMEANS statement can perform many postfitting tasks. Examples in this paper show that the LSMEANS statement provides a tool for constructing hypotheses, customizing comparisons, producing appropriate and informative graphics, and so on. The LSMEANS statement greatly simplifies programs for analyzing LS-means that could done instead by using the ESTIMATE and CONTRAST statements. The LSMEANS statement along with its rich set of options is available in many SAS/STAT procedures that fit a broad range of models, from general linear models to proportional hazards models. If you want to perform analyses on predicted population margins, consider using the LSMEANS statement.

REFERENCES

Goodnight, J. H. and Harvey, W. R. (1978), *Least-Squares Means in the Fixed-Effects General Linear Models*, Technical Report R-103, SAS Institute Inc., Cary, NC.

Gotway, C. A. and Stroup, W. W. (1997), "A Generalized Linear Model Approach to Spatial Data and Prediction," *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 157–187.

Harvey, W. R. (1975), *Least-Squares Analysis of Data with Unequal Subclass Numbers*, Technical Report ARS H-4, U.S. Department of Agriculture, Agriculture Research Service.

Harvey, W. R. (1976), "Use of the HARVEY Procedure," in *Proceedings of the First Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Klein, J. P. and Moeschberger, M. L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd Edition, New York: Springer-Verlag.

Quine, S. (1975), *Achievement Orientation of Aboriginal and White Adolescents*, Ph.D. diss., Australian National University, Canberra.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., and Hochberg, Y. (1999), *Multiple Comparisons and Multiple Tests Using the SAS System*, Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The author is grateful to Randy Tobias, Phil Gibbs, and Warren Kuhfeld for their valuable discussion and assistance in the preparation of this paper. The author also thanks Anne Baxter and Ed Huddleston for editorial assistance.

CONTACT INFORMATION

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

Weijie Cai
SAS Institute Inc.
600 Research Drive
Cary, NC 27513
Web: http://support.sas.com/statistics