Chapter 30
The GLM Procedure

Chapter Table of Contents

Chapter 30
The GLM Procedure

Overview

The GLM procedure uses the method of least squares to fit general linear models. Among the statistical methods available in PROC GLM are regression, analysis of variance, analysis of covariance, multivariate analysis of variance, and partial correlation.

PROC GLM analyzes data within the framework of General linear models. PROC GLM handles models relating one or several continuous dependent variables to one or several independent variables. The independent variables may be either *classification* variables, which divide the observations into discrete groups, or *continuous* variables. Thus, the GLM procedure can be used for many different analyses, including

- simple regression
- multiple regression
- analysis of variance (ANOVA), especially for unbalanced data
- analysis of covariance
- response-surface models
- weighted regression
- polynomial regression
- partial correlation
- multivariate analysis of variance (MANOVA)
- repeated measures analysis of variance

PROC GLM Features

The following list summarizes the features in PROC GLM:

- PROC GLM enables you to specify any degree of interaction (crossed effects) and nested effects. It also provides for polynomial, continuous-by-class, and continuous-nesting-class effects.
- Through the concept of estimability, the GLM procedure can provide tests of hypotheses for the effects of a linear model regardless of the number of missing cells or the extent of confounding. PROC GLM displays the Sum of Squares (SS) associated with each hypothesis tested and, upon request, the form of the estimable functions employed in the test. PROC GLM can produce the general form of all estimable functions.

- The REPEATED statement enables you to specify effects in the model that represent repeated measurements on the same experimental unit for the same response, providing both univariate and multivariate tests of hypotheses.
- The RANDOM statement enables you to specify random effects in the model; expected mean squares are produced for each Type I, Type II, Type III, Type IV, and contrast mean square used in the analysis. Upon request, F tests using appropriate mean squares or linear combinations of mean squares as error terms are performed.
- The ESTIMATE statement enables you to specify an $\mathbf{L}$ vector for estimating a linear function of the parameters $\mathbf{L}\beta$.
- The CONTRAST statement enables you to specify a contrast vector or matrix for testing the hypothesis that $\mathbf{L}\beta = 0$. When specified, the contrasts are also incorporated into analyses using the MANOVA and REPEATED statements.
- The MANOVA statement enables you to specify both the hypothesis effects and the error effect to use for a multivariate analysis of variance.
- PROC GLM can create an output data set containing the input dataset in addition to predicted values, residuals, and other diagnostic measures.
- PROC GLM can be used interactively. After specifying and running a model, a variety of statements can be executed without recomputing the model parameters or sums of squares.
- For analysis involving multiple dependent variables but not the MANOVA or REPEATED statements, a missing value in one dependent variable does not eliminate the observation from the analysis for other dependent variables. PROC GLM automatically groups together those variables that have the same pattern of missing values within the data set or within a BY group. This ensures that the analysis for each dependent variable brings into use all possible observations.

PROC GLM Contrasted with Other SAS Procedures

As described previously, PROC GLM can be used for many different analyses and has many special features not available in other SAS procedures. However, for some types of analyses, other procedures are available. As discussed in the "PROC GLM for Unbalanced ANOVA" and "PROC GLM for Quadratic Least Squares Regression" sections (beginning on page 1469), sometimes these other procedures are more efficient than PROC GLM. The following procedures perform some of the same analyses as PROC GLM:

ANOVA performs analysis of variance for balanced designs. The ANOVA procedure is generally more efficient than PROC GLM for these designs.

MIXED fits mixed linear models by incorporating covariance structures in the model fitting process. Its RANDOM and REPEATED statements are similar to those in PROC GLM but offer different functionalities.

NESTED performs analysis of variance and estimates variance components for nested random models. The NESTED procedure is generally more efficient than PROC GLM for these models.

NPAR1WAY performs nonparametric one-way analysis of rank scores. This can also be done using the RANK procedure and PROC GLM.

REG performs simple linear regression. The REG procedure allows several MODEL statements and gives additional regression diagnostics, especially for detection of collinearity. PROC REG also creates plots of model summary statistics and regression diagnostics.

RSREG performs quadratic response-surface regression, and canonical and ridge analysis. The RSREG procedure is generally recommended for data from a response surface experiment.

TTEST compares the means of two groups of observations. Also, tests for equality of variances for the two groups are available. The TTEST procedure is usually more efficient than PROC GLM for this type of data.

VARCOMP estimates variance components for a general linear model.

Getting Started

PROC GLM for Unbalanced ANOVA

Analysis of variance, or ANOVA, typically refers to partitioning the variation in a variable's values into variation between and within several groups or classes of observations. The GLM procedure can perform simple or complicated ANOVA for balanced or unbalanced data.

This example discusses a 2 × 2 ANOVA model. The experimental design is a full factorial, in which each level of one treatment factor occurs at each level of the other treatment factor. The data are shown in a table and then read into a SAS data set.

		A	
		1	2
B	1	12 14	20 18
	2	11 9	17

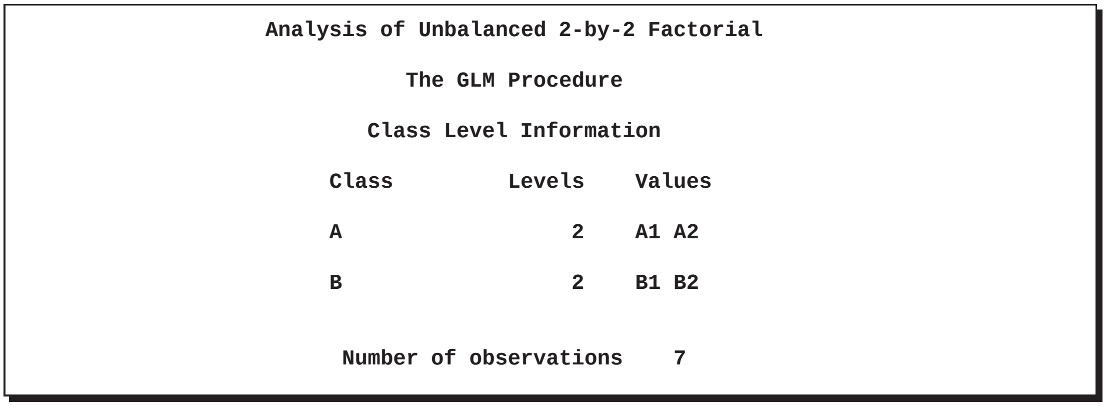
```
title 'Analysis of Unbalanced 2-by-2 Factorial';
data exp;
   input A $ B $ Y @@;
   datalines;
A1 B1 12 A1 B1 14     A1 B2 11 A1 B2 9
A2 B1 20 A2 B1 18     A2 B2 17
;
```

Note that there is only one value for the cell with A=‘A2’ and B=‘B2’. Since one cell contains a different number of values from the other cells in the table, this is an unbalanced design.

The following PROC GLM invocation produces the analysis.

```
proc glm;
   class A B;
   model Y=A B A*B;
run;
```

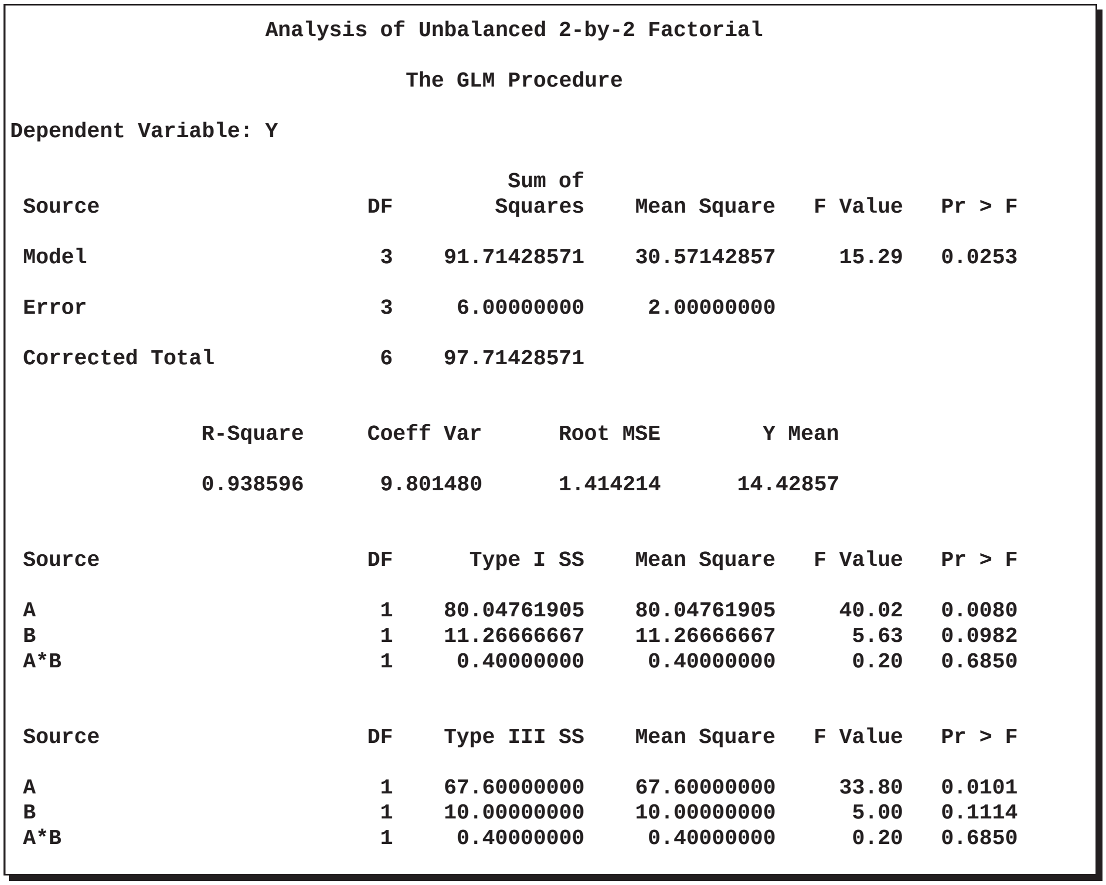
Both treatments are listed in the CLASS statement because they are classification variables. A*B denotes the interaction of the A effect and the B effect. The results are shown in Figure 30.1 and Figure 30.2.



```
                    Analysis of Unbalanced 2-by-2 Factorial

                               The GLM Procedure

                            Class Level Information

                         Class         Levels    Values

                         A                  2    A1 A2

                         B                  2    B1 B2


                          Number of observations    7
```

Figure 30.1. Class Level Information

Figure 30.1 displays information about the classes as well as the number of observations in the data set. Figure 30.2 shows the ANOVA table, simple statistics, and tests of effects.



```
                    Analysis of Unbalanced 2-by-2 Factorial

                               The GLM Procedure

Dependent Variable: Y

                                        Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                        3    91.71428571    30.57142857     15.29   0.0253

 Error                        3     6.00000000     2.00000000

 Corrected Total              6    97.71428571


               R-Square     Coeff Var      Root MSE        Y Mean

               0.938596      9.801480      1.414214      14.42857


 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 A                            1    80.04761905    80.04761905     40.02   0.0080
 B                            1    11.26666667    11.26666667      5.63   0.0982
 A*B                          1     0.40000000     0.40000000      0.20   0.6850


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 A                            1    67.60000000    67.60000000     33.80   0.0101
 B                            1    10.00000000    10.00000000      5.00   0.1114
 A*B                          1     0.40000000     0.40000000      0.20   0.6850
```

Figure 30.2. ANOVA Table and Tests of Effects

The degrees of freedom may be used to check your data. The Model degrees of freedom for a 2×2 factorial design with interaction are $(ab - 1)$, where a is the number of levels of A and b is the number of levels of B; in this case, $(2 \times 2 - 1) = 3$. The Corrected Total degrees of freedom are always one less than the number of observations used in the analysis; in this case, $7 - 1 = 6$.

The overall F test is significant $(F = 15.29, p = 0.0253)$, indicating strong evidence that the means for the four different A×B cells are different. You can further analyze this difference by examining the individual tests for each effect.

Four types of estimable functions of parameters are available for testing hypotheses in PROC GLM. For data with no missing cells, the Type III and Type IV estimable functions are the same and test the same hypotheses that would be tested if the data were balanced. Type I and Type III sums of squares are typically not equal when the data are unbalanced; Type III sums of squares are preferred in testing effects in unbalanced cases because they test a function of the underlying parameters that is independent of the number of observations per treatment combination.

According to a significance level of 5% $(\alpha = 0.05)$, the A*B interaction is not significant $(F = 0.20, p = 0.6850)$. This indicates that the effect of A does not depend on the level of B and vice versa. Therefore, the tests for the individual effects are valid, showing a significant A effect $(F = 33.80, p = 0.0101)$ but no significant B effect $(F = 5.00, p = 0.1114)$.

PROC GLM for Quadratic Least Squares Regression

In polynomial regression, the values of a dependent variable (also called a response variable) are described or predicted in terms of polynomial terms involving one or more independent or explanatory variables. An example of quadratic regression in PROC GLM follows. These data are taken from Draper and Smith (1966, p. 57). Thirteen specimens of 90/10 Cu-Ni alloys are tested in a corrosion-wheel setup in order to examine corrosion. Each specimen has a certain iron content. The wheel is rotated in salt sea water at 30 ft/sec for 60 days. Weight loss is used to quantify the corrosion. The fe variable represents the iron content, and the loss variable denotes the weight loss in milligrams/square decimeter/day in the following DATA step.

```
title 'Regression in PROC GLM';
data iron;
   input fe loss @@;
   datalines;
0.01 127.6   0.48 124.0   0.71 110.8   0.95 103.9
1.19 101.5   0.01 130.1   0.48 122.0   1.44  92.3
0.71 113.1   1.96  83.7   0.01 128.0   1.44  91.4
1.96  86.2
;
```

The GPLOT procedure is used to request a scatter plot of the response variable versus the independent variable.

```
symbol1 c=blue;
proc gplot;
   plot loss*fe / vm=1;
run;
```

The plot in Figure 30.3 displays a strong negative relationship between iron content and corrosion resistance, but it is not clear whether there is curvature in this relationship.

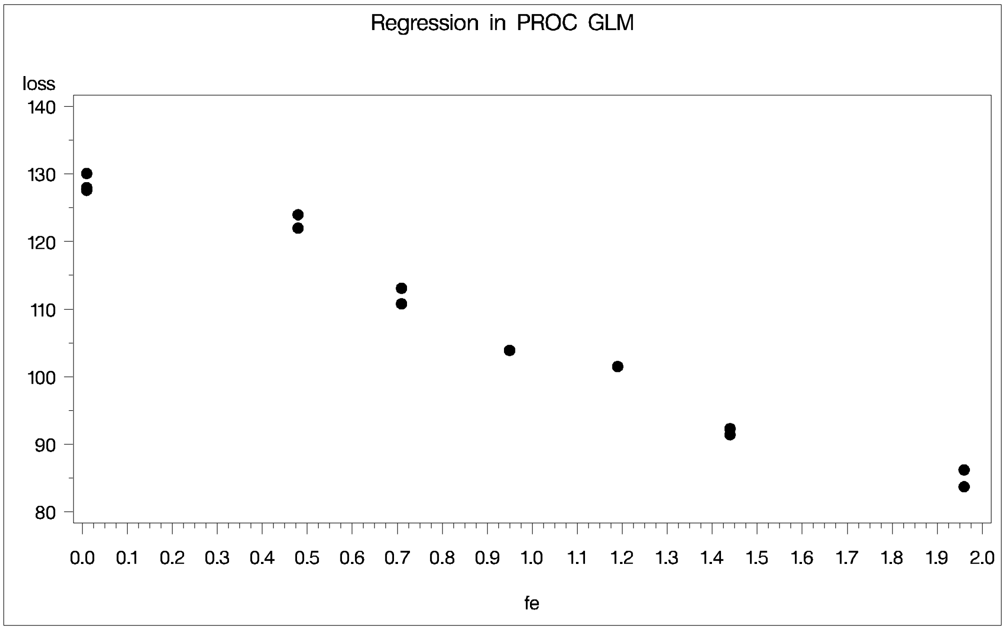


Figure 30.3. Plot of LOSS vs. FE

The following statements fit a quadratic regression model to the data. This enables you to estimate the linear relationship between iron content and corrosion resistance and test for the presence of a quadratic component. The intercept is automatically fit unless the NOINT option is specified.

```
proc glm;
   model loss=fe fe*fe;
run;
```

The CLASS statement is omitted because a regression line is being fitted. Unlike PROC REG, PROC GLM allows polynomial terms in the MODEL statement.

```
                              Regression in PROC GLM

                                The GLM Procedure

                           Number of observations    13
```

Figure 30.4. Class Level Information

The preliminary information in Figure 30.4 informs you that the GLM procedure has been invoked and states the number of observations in the data set. If the model involves classification variables, they are also listed here, along with their levels.

Figure 30.5 shows the overall ANOVA table and some simple statistics. The degrees of freedom can be used to check that the model is correct and that the data have been read correctly. The Model degrees of freedom for a regression is the number of parameters in the model minus 1. You are fitting a model with three parameters in this case,

$$\text{loss} \quad = \quad \beta_0 + \beta_1 \times (\text{fe}) + \beta_2 \times (\text{fe})^2 + \mathit{error}$$

so the degrees of freedom are $3 - 1 = 2$. The Corrected Total degrees of freedom are always one less than the number of observations used in the analysis.

```
                              Regression in PROC GLM

                                The GLM Procedure

Dependent Variable: loss

                                                 Sum of
 Source                        DF               Squares    Mean Square   F Value   Pr > F

 Model                          2           3296.530589    1648.265295    164.68   <.0001

 Error                         10            100.086334      10.008633

 Corrected Total               12           3396.616923


              R-Square     Coeff Var      Root MSE     loss Mean

              0.970534      2.907348      3.163642      108.8154
```

Figure 30.5. ANOVA Table

The R^2 indicates that the model accounts for 97% of the variation in LOSS. The coefficient of variation (C.V.), Root MSE (Mean Square for Error), and mean of the dependent variable are also listed.

The overall F test is significant $(F = 164.68, p < 0.0001)$, indicating that the model as a whole accounts for a significant amount of the variation in LOSS. Thus, it is appropriate to proceed to testing the effects.

Figure 30.6 contains tests of effects and parameter estimates. The latter are displayed by default when the model contains only continuous variables.

```
                              Regression in PROC GLM

                                The GLM Procedure

Dependent Variable: loss

 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 fe                           1    3293.766690    3293.766690    329.09   <.0001
 fe*fe                        1       2.763899       2.763899      0.28   0.6107


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 fe                           1    356.7572421    356.7572421     35.64   0.0001
 fe*fe                        1      2.7638994      2.7638994      0.28   0.6107


                                            Standard
       Parameter         Estimate              Error    t Value    Pr > |t|

       Intercept      130.3199337       1.77096213      73.59      <.0001
       fe             -26.2203900       4.39177557      -5.97      0.0001
       fe*fe            1.1552018       2.19828568       0.53      0.6107
```

Figure 30.6. Tests of Effects and Parameter Estimates

The t tests provided are equivalent to the Type III F tests. The quadratic term is not significant $(F = 0.28, p = 0.6107; t = 0.53, p = 0.6107)$ and thus can be removed from the model; the linear term is significant $(F = 35.64, p = 0.0001; t = -5.97, p = 0.0001)$. This suggests that there is indeed a straight line relationship between loss and fe.

Fitting the model without the quadratic term provides more accurate estimates for β_0 and β_1. PROC GLM allows only one MODEL statement per invocation of the procedure, so the PROC GLM statement must be issued again. The statements used to fit the linear model are

```
proc glm;
   model loss=fe;
run;
```

Figure 30.7 displays the output produced by these statements. The linear term is still significant $(F = 352.27, p < 0.0001)$. The estimated model is now

$$\text{loss} = 129.79 - 24.02 \times \text{fe}$$

```
                             Regression in PROC GLM

                               The GLM Procedure

Dependent Variable: loss

                                       Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F

 Model                       1    3293.766690    3293.766690    352.27   <.0001

 Error                      11     102.850233       9.350021

 Corrected Total            12    3396.616923


               R-Square     Coeff Var      Root MSE     loss Mean

               0.969720      2.810063      3.057780      108.8154


 Source                     DF      Type I SS    Mean Square   F Value   Pr > F

 fe                          1    3293.766690    3293.766690    352.27   <.0001


 Source                     DF    Type III SS    Mean Square   F Value   Pr > F

 fe                          1    3293.766690    3293.766690    352.27   <.0001


                                         Standard
       Parameter         Estimate           Error    t Value    Pr > |t|

       Intercept      129.7865993      1.40273671      92.52      <.0001
       fe             -24.0198934      1.27976715     -18.77      <.0001
```

Figure 30.7. Linear Model Output

Syntax

The following statements are available in PROC GLM.

```
PROC GLM < options > ;
    CLASS variables ;
    MODEL dependents=independents < / options > ;

    ABSORB variables ;
    BY variables ;
    FREQ variable ;
    ID variables ;
    WEIGHT variable ;

    CONTRAST 'label' effect values < . . . effect values > < / options > ;
    ESTIMATE 'label' effect values < . . . effect values > < / options > ;
    LSMEANS effects < / options > ;
    MANOVA < test-options >< / detail-options > ;
    MEANS effects < / options > ;
    OUTPUT < OUT=SAS-data-set >
        keyword=names < . . . keyword=names > < / option > ;
    RANDOM effects < / options > ;
    REPEATED factor-specification < / options > ;
    TEST < H=effects > E=effect < / options > ;
```

Although there are numerous statements and options available in PROC GLM, many applications use only a few of them. Often you can find the features you need by looking at an example or by quickly scanning through this section.

To use PROC GLM, the PROC GLM and MODEL statements are required. You can specify only one MODEL statement (in contrast to the REG procedure, for example, which allows several MODEL statements in the same PROC REG run). If your model contains classification effects, the classification variables must be listed in a CLASS statement, and the CLASS statement must appear before the MODEL statement. In addition, if you use a CONTRAST statement in combination with a MANOVA, RANDOM, REPEATED, or TEST statement, the CONTRAST statement must be entered first in order for the contrast to be included in the MANOVA, RANDOM, REPEATED, or TEST analysis.

The following table summarizes the positional requirements for the statements in the GLM procedure.

Table 30.1. Positional Requirements for PROC GLM Statements

Statement	Must Appear Before the	Must Appear After the
ABSORB	first RUN statement	
BY	first RUN statement	
CLASS	MODEL statement	
CONTRAST	MANOVA, REPEATED, or RANDOM statement	MODEL statement
ESTIMATE		MODEL statement
FREQ	first RUN statement	
ID	first RUN statement	
LSMEANS		MODEL statement
MANOVA		CONTRAST or MODEL statement
MEANS		MODEL statement
MODEL	CONTRAST, ESTIMATE, LSMEANS, or MEANS statement	CLASS statement
OUTPUT		MODEL statement
RANDOM		CONTRAST or MODEL statement
REPEATED		CONTRAST, MODEL, or TEST statement
TEST	MANOVA or REPEATED statement	MODEL statement
WEIGHT	first RUN statement	

The following table summarizes the function of each statement (other than the PROC statement) in the GLM procedure:

Table 30.2. Statements in the GLM Procedure

Statement	Description
ABSORB	absorbs classification effects in a model
BY	specifies variables to define subgroups for the analysis
CLASS	declares classification variables
CONTRAST	constructs and tests linear functions of the parameters
ESTIMATE	estimates linear functions of the parameters
FREQ	specifies a frequency variable
ID	identifies observations on output
LSMEANS	computes least-squares (marginal) means
MANOVA	performs a multivariate analysis of variance
MEANS	computes and optionally compares arithmetic means
MODEL	defines the model to be fit
OUTPUT	requests an output data set containing diagnostics for each observation

Table 30.2. (continued)

Statement	Description
RANDOM	declares certain effects to be random and computes expected mean squares
REPEATED	performs multivariate and univariate repeated measures analysis of variance
TEST	constructs tests using the sums of squares for effects and the error term you specify
WEIGHT	specifies a variable for weighting observations

The rest of this section gives detailed syntax information for each of these statements, beginning with the PROC GLM statement. The remaining statements are covered in alphabetical order.

PROC GLM Statement

PROC GLM < *options* > **;**

The PROC GLM statement starts the GLM procedure. You can specify the following options in the PROC GLM statement:

ALPHA=*p*

specifies the level of significance p for $100(1-p)$% confidence intervals. The value must be between 0 and 1; the default value of $p = 0.05$ results in 95% intervals. This value is used as the default confidence level for limits computed by the following options.

Statement	Options
LSMEANS	CL
MEANS	CLM CLDIFF
MODEL	CLI CLM CLPARM
OUTPUT	UCL= LCL= UCLM= LCLM=

You can override the default in each of these cases by specifying the ALPHA= option for each statement individually.

DATA=*SAS-data-set*

names the SAS data set used by the GLM procedure. By default, PROC GLM uses the most recently created SAS data set.

MANOVA

requests the multivariate mode of eliminating observations with missing values. If any of the dependent variables have missing values, the procedure eliminates that observation from the analysis. The MANOVA option is useful if you use PROC GLM in interactive mode and plan to perform a multivariate analysis.

MULTIPASS

requests that PROC GLM reread the input data set when necessary, instead of writing the necessary values of dependent variables to a utility file. This option decreases disk space usage at the expense of increased execution times, and is useful only in rare situations where disk space is at an absolute premium.

NAMELEN=n

specifies the length of effect names in tables and output data sets to be n characters long, where n is a value between 20 and 200 characters. The default length is 20 characters.

NOPRINT

suppresses the normal display of results. The NOPRINT option is useful when you want only to create one or more output data sets with the procedure. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, "Using the Output Delivery System," for more information.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sorting order for the levels of all classification variables (specified in the CLASS statement). This ordering determines which parameters in the model correspond to each level in the data, so the ORDER= option may be useful when you use CONTRAST or ESTIMATE statements. Note that the ORDER= option applies to the levels for all classification variables. The exception is ORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement in the current PROC GLM run or in the DATA step that created the data set). In this case, the levels are ordered by their internal (numeric) value. Note that this represents a change from previous releases for how class levels are ordered. In releases previous to Version 8, numeric class levels with no explicit format were ordered by their BEST12. formatted values, and in order to revert to the previous ordering you can specify this format explicitly for the affected classification variables. The change was implemented because the former default behavior for ORDER=FORMATTED often resulted in levels not being ordered numerically and usually required the user to intervene with an explicit format or ORDER=INTERNAL to get the more natural ordering. The following table shows how PROC GLM interprets values of the ORDER= option.

Value of ORDER=	**Levels Sorted By**
DATA	order of appearance in the input data set
FORMATTED	external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

By default, ORDER=FORMATTED. For FORMATTED and INTERNAL, the sort order is machine dependent. For more information on sorting order, see the chapter on the SORT procedure in the *SAS Procedures Guide*, and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

OUTSTAT=*SAS-data-set*

names an output data set that contains sums of squares, degrees of freedom, F statistics, and probability levels for each effect in the model, as well as for each CONTRAST that uses the overall residual or error mean square (MSE) as the denominator in constructing the F statistic. If you use the CANONICAL option in the MANOVA statement and do not use an M= specification in the MANOVA statement, the data set also contains results of the canonical analysis. See the section "Output Data Sets" on page 1574 for more information.

ABSORB Statement

ABSORB *variables* **;**

Absorption is a computational technique that provides a large reduction in time and memory requirements for certain types of models. The *variables* are one or more variables in the input data set.

For a main effect variable that does not participate in interactions, you can absorb the effect by naming it in an ABSORB statement. This means that the effect can be adjusted out before the construction and solution of the rest of the model. This is particularly useful when the effect has a large number of levels.

Several variables can be specified, in which case each one is assumed to be nested in the preceding variable in the ABSORB statement.

Note: When you use the ABSORB statement, the data set (or each BY group, if a BY statement appears) must be sorted by the variables in the ABSORB statement. The GLM procedure cannot produce predicted values or least-squares means (LS-means) or create an output data set of diagnostic values if an ABSORB statement is used. If the ABSORB statement is used, it must appear before the first RUN statement or it is ignored.

When you use an ABSORB statement and also use the INT option in the MODEL statement, the procedure ignores the option but computes the uncorrected total sum of squares (SS) instead of the corrected total sums of squares.

See the "Absorption" section on page 1532 for more information.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC GLM to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the GLM procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in base SAS software).

Since sorting the data changes the order in which PROC GLM reads observations, the sorting order for the levels of the classification variables may be affected if you have also specified ORDER=DATA in the PROC GLM statement. This, in turn, affects specifications in CONTRAST and ESTIMATE statements.

If you specify the BY statement, it must appear before the first RUN statement or it is ignored. When you use a BY statement, the interactive features of PROC GLM are disabled.

When both BY and ABSORB statements are used, observations must be sorted first by the variables in the BY statement, and then by the variables in the ABSORB statement.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Contents*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

CLASS Statement

CLASS *variables* **;**

The CLASS statement names the classification variables to be used in the model. Typical class variables are TREATMENT, SEX, RACE, GROUP, and REPLICATION. If you specify the CLASS statement, it must appear before the MODEL statement.

Class levels are determined from up to the first 16 characters of the formatted values of the CLASS variables. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide*, and the discussions for the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

The GLM procedure displays a table summarizing the class variables and their levels, and you can use this to check the ordering of levels and, hence, of the corresponding parameters for main effects. If you need to check the ordering of parameters for interaction effects, use the E option in the MODEL, CONTRAST, ESTIMATE, and LSMEANS statements. See the “Parameterization of PROC GLM Models” section on page 1521 for more information.

CONTRAST Statement

CONTRAST *’label’ effect values* $<$ ... *effect values* $>$ $<$ **/** *options* $>$ **;**

The CONTRAST statement enables you to perform custom hypothesis tests by specifying an $\mathbf{L}$ vector or matrix for testing the univariate hypothesis $\mathbf{L}\beta = 0$ or the multivariate hypothesis $\mathbf{LBM} = 0$. Thus, to use this feature you must be familiar with the details of the model parameterization that PROC GLM uses. For more information, see the “Parameterization of PROC GLM Models” section on page 1521. All of the elements of the $\mathbf{L}$ vector may be given, or if only certain portions of the $\mathbf{L}$ vector are given, the remaining elements are constructed by PROC GLM from the context (in a manner similar to rule 4 discussed in the “Construction of Least-Squares Means” section on page 1555).

There is no limit to the number of CONTRAST statements you can specify, but they must appear after the MODEL statement. In addition, if you use a CONTRAST statement and a MANOVA, REPEATED, or TEST statement, appropriate tests for contrasts are carried out as part of the MANOVA, REPEATED, or TEST analysis. If you use a CONTRAST statement and a RANDOM statement, the expected mean square of the contrast is displayed. As a result of these additional analyses, the CONTRAST statement must appear before the MANOVA, REPEATED, RANDOM, or TEST statement.

In the CONTRAST statement,

label identifies the contrast on the output. A label is required for every contrast specified. Labels must be enclosed in quotes.

effect identifies an effect that appears in the MODEL statement, or the INTERCEPT effect. The INTERCEPT effect can be used when an intercept is fitted in the model. You do not need to include all effects that are in the MODEL statement.

values are constants that are elements of the $\mathbf{L}$ vector associated with the effect.

You can specify the following options in the CONTRAST statement after a slash(/):

E

displays the entire $\mathbf{L}$ vector. This option is useful in confirming the ordering of parameters for specifying $\mathbf{L}$.

E=*effect*

specifies an error term, which must be one of the effects in the model. The procedure uses this effect as the denominator in F tests in univariate analysis. In addition, if you use a MANOVA or REPEATED statement, the procedure uses the effect specified by the E= option as the basis of the $\mathbf{E}$ matrix. By default, the procedure uses the overall residual or error mean square (MSE) as an error term.

ETYPE=*n*

specifies the type (1, 2, 3, or 4, corresponding to Type I, II, III, and IV tests, respectively) of the E= effect. If the E= option is specified and the ETYPE= option is not, the procedure uses the highest type computed in the analysis.

SINGULAR=*number*

checking (GLM) tunes the estimability checking. If ABS($\mathbf{L} - \mathbf{LH}$) $> C \times$*number* for any row in the contrast, then $\mathbf{L}$ is declared nonestimable. $\mathbf{H}$ is the $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ matrix, and C is ABS($\mathbf{L}$) except for rows where $\mathbf{L}$ is zero, and then it is 1. The default value for the SINGULAR= option is 10^{-4}. Values for the SINGULAR= option must be between 0 and 1.

As stated previously, the CONTRAST statement enables you to perform custom hypothesis tests. If the hypothesis is testable in the univariate case, SS($H_0\colon \mathbf{L}\beta = 0$) is computed as

$$(\mathbf{Lb})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{Lb})$$

where $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$. This is the sum of squares displayed on the analysis-of-variance table.

For multivariate testable hypotheses, the usual multivariate tests are performed using

$$\mathbf{H} = \mathbf{M}'(\mathbf{LB})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{LB})\mathbf{M}$$

where $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ and $\mathbf{Y}$ is the matrix of multivariate responses or dependent variables. The degrees of freedom associated with the hypothesis is equal to the row rank of $\mathbf{L}$. The sum of squares computed in this situation are equivalent to the sum of squares computed using an $\mathbf{L}$ matrix with any row deleted that is a linear combination of previous rows.

Multiple-degree-of-freedom hypotheses can be specified by separating the rows of the $\mathbf{L}$ matrix with commas.

For example, for the model

```
proc glm;
   class A B;
   model Y=A B;
run;
```

with A at 5 levels and B at 2 levels, the parameter vector is

$$(\mu \;\; \alpha_1 \;\; \alpha_2 \;\; \alpha_3 \;\; \alpha_4 \;\; \alpha_5 \;\; \beta_1 \;\; \beta_2)$$

To test the hypothesis that the pooled A linear and A quadratic effect is zero, you can use the following $\mathbf{L}$ matrix:

$$\mathbf{L} = \begin{bmatrix} 0 & -2 & -1 & 0 & 1 & 2 & 0 & 0 \\ 0 & 2 & -1 & -2 & -1 & 2 & 0 & 0 \end{bmatrix}$$

The corresponding CONTRAST statement is

```
contrast 'A LINEAR & QUADRATIC'
         a -2 -1  0  1  2,
         a  2 -1 -2 -1  2;
```

If the first level of A is a control level and you want a test of control versus others, you can use this statement:

```
contrast 'CONTROL VS OTHERS'  a -1 0.25 0.25 0.25 0.25;
```

See the following discussion of the ESTIMATE statement and the "Specification of ESTIMATE Expressions" section on page 1536 for rules on specification, construction, distribution, and estimability in the CONTRAST statement.

ESTIMATE Statement

ESTIMATE *'label' effect values* $<$ ... *effect values* $>$ $<$ **/** *options* $>$ **;**

The ESTIMATE statement enables you to estimate linear functions of the parameters by multiplying the vector $\mathbf{L}$ by the parameter estimate vector $\mathbf{b}$ resulting in $\mathbf{Lb}$. All of the elements of the $\mathbf{L}$ vector may be given, or, if only certain portions of the $\mathbf{L}$ vector are given, the remaining elements are constructed by PROC GLM from the context (in a manner similar to rule 4 discussed in the “Construction of Least-Squares Means” section on page 1555).

The linear function is checked for estimability. The estimate $\mathbf{Lb}$, where $\mathbf{b} = (\mathbf{X'X})^{-}\mathbf{X'y}$, is displayed along with its associated standard error, $\sqrt{\mathbf{L}(\mathbf{X'X})^{-}\mathbf{L'}s^2}$, and t test. If you specify the CLPARM option in the MODEL statement (see page 1505), confidence limits for the true value are also displayed.

There is no limit to the number of ESTIMATE statements that you can specify, but they must appear after the MODEL statement. In the ESTIMATE statement,

label identifies the estimate on the output. A label is required for every contrast specified. Labels must be enclosed in quotes.

effect identifies an effect that appears in the MODEL statement, or the INTERCEPT effect. The INTERCEPT effect can be used as an effect when an intercept is fitted in the model. You do not need to include all effects that are in the MODEL statement.

values are constants that are the elements of the $\mathbf{L}$ vector associated with the preceding effect. For example,

```
estimate 'A1 VS A2' A  1 -1;
```

forms an estimate that is the difference between the parameters estimated for the first and second levels of the CLASS variable A.

You can specify the following options in the ESTIMATE statement after a slash:

DIVISOR=*number*

specifies a value by which to divide all coefficients so that fractional coefficients can be entered as integer numerators. For example, you can use

```
estimate '1/3(A1+A2) - 2/3A3' a 1 1 -2 / divisor=3;
```

instead of

```
estimate '1/3(A1+A2) - 2/3A3' a 0.33333 0.33333 -0.66667;
```

E

displays the entire $\mathbf{L}$ vector. This option is useful in confirming the ordering of parameters for specifying $\mathbf{L}$.

SINGULAR=*number*

tunes the estimability checking. If ABS$(\mathbf{L} - \mathbf{LH}) > C\times$*number*, then the $\mathbf{L}$ vector is declared nonestimable. $\mathbf{H}$ is the $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ matrix, and C is ABS$(\mathbf{L})$ except for rows where $\mathbf{L}$ is zero, and then it is 1. The default value for the SINGULAR= option is 10^{-4}. Values for the SINGULAR= option must be between 0 and 1.

See also the "Specification of ESTIMATE Expressions" section on page 1536.

FREQ Statement

FREQ *variable* **;**

The FREQ statement names a variable that provides frequencies for each observation in the DATA= data set. Specifically, if *n* is the value of the FREQ variable for a given observation, then that observation is used *n* times.

The analysis produced using a FREQ statement reflects the expanded number of observations. For example, means and total degrees of freedom reflect the expanded number of observations. You can produce the same analysis (without the FREQ statement) by first creating a new data set that contains the expanded number of observations. For example, if the value of the FREQ variable is 5 for the first observation, the first 5 observations in the new data set are identical. Each observation in the old data set is replicated n_i times in the new data set, where n_i is the value of the FREQ variable for that observation.

If the value of the FREQ variable is missing or is less than 1, the observation is not used in the analysis. If the value is not an integer, only the integer portion is used.

If you specify the FREQ statement, it must appear before the first RUN statement or it is ignored.

ID Statement

ID *variables* **;**

When predicted values are requested as a MODEL statement option, values of the variables given in the ID statement are displayed beside each observed, predicted, and residual value for identification. Although there are no restrictions on the length of ID variables, PROC GLM may truncate the number of values listed in order to display them on one line. The GLM procedure displays a maximum of five ID variables.

If you specify the ID statement, it must appear before the first RUN statement or it is ignored.

LSMEANS Statement

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

Least-squares means (LS-means) are computed for each *effect* listed in the LSMEANS statement. You may specify only classification effects in the LSMEANS statement—that is, effects that contain only classification variables. You may also specify options to perform multiple comparisons. In contrast to the MEANS statement, the LSMEANS statement performs multiple comparisons on interactions as well as main effects.

LS-means are *predicted population margins*; that is, they estimate the marginal means over a balanced population. In a sense, LS-means are to unbalanced designs as class and subclass arithmetic means are to balanced designs. Each LS-mean is computed as $\mathbf{L}'\mathbf{b}$ for a certain column vector $\mathbf{L}$, where $\mathbf{b}$ is the vector of parameter estimates—that is, the solution of the normal equations. For further information, see the section “Construction of Least-Squares Means” on page 1555.

Multiple effects can be specified in one LSMEANS statement, or multiple LSMEANS statements can be used, but they must all appear after the MODEL statement. For example,

```
proc glm;
   class A B;
   model Y=A B A*B;
   lsmeans A B A*B;
run;
```

LS-means are displayed for each level of the A, B, and A*B effects.

You can specify the following options in the LSMEANS statement after a slash:

ADJUST=BON
ADJUST=DUNNETT
ADJUST=SCHEFFE
ADJUST=SIDAK
ADJUST=SIMULATE <(*simoptions***)>**
ADJUST=SMM | GT2
ADJUST=TUKEY
ADJUST=T

requests a multiple comparison adjustment for the *p*-values and confidence limits for the differences of LS-means. The ADJUST= option modifies the results of the TDIFF and PDIFF options; thus, if you omit the TDIFF or PDIFF option then the ADJUST= option has no effect. By default, PROC GLM analyzes all pairwise differences unless you specify ADJUST=DUNNETT, in which case PROC GLM analyzes all differences with a control level. The default is ADJUST=T, which really signifies no adjustment for multiple comparisons.

The BON (Bonferroni) and SIDAK adjustments involve correction factors described in the "Multiple Comparisons" section on page 1540 and in Chapter 43, "The MULTTEST Procedure." When you specify ADJUST=TUKEY and your data are unbalanced, PROC GLM uses the approximation described in Kramer (1956) and identifies the adjustment as "Tukey-Kramer" in the results. Similarly, when you specify ADJUST=DUNNETT and the LS-means are correlated, PROC GLM uses the factor-analytic covariance approximation described in Hsu (1992) and identifies the adjustment as "Dunnett-Hsu" in the results. The preceding references also describe the SCHEFFE and SMM adjustments.

The SIMULATE adjustment computes the adjusted *p*-values from the simulated distribution of the maximum or maximum absolute value of a multivariate *t* random vector. The simulation estimates q, the true $(1 - \alpha)$th quantile, where $1 - \alpha$ is the confidence coefficient. The default α is the value of the ALPHA= option in the PROC GLM statement or 0.05 if that option is not specified. You can change this value with the ALPHA= option in the LSMEANS statement.

The number of samples for the SIMULATE adjustment is set so that the tail area for the simulated q is within a certain *accuracy radius* γ of $1 - \alpha$ with an *accuracy confidence* of $100(1 - \epsilon)$%. In equation form,

$$P(|F(\hat{q}) - (1 - \alpha)| \leq \gamma) \quad = \quad 1 - \epsilon$$

where $\hat{q}$ is the simulated q and F is the true distribution function of the maximum; refer to Edwards and Berry (1987) for details. By default, γ = 0.005 and ϵ = 0.01 so that the tail area of $\hat{q}$ is within 0.005 of 0.95 with 99% confidence.

You can specify the following *simoptions* in parentheses after the ADJUST=SIMULATE option.

ACC=*value* specifies the target accuracy radius γ of a $100(1 - \epsilon)$% confidence interval for the true probability content of the estimated $(1 - \alpha)$th quantile. The default value is ACC=0.005. Note that, if you also specify the CVADJUST *simoption*, then the actual accuracy radius will probably be substantially less than this target.

CVADJUST specifies that the quantile should be estimated by the control variate adjustment method of Hsu and Nelson (1998) instead of simply as the quantile of the simulated sample. Specifying the CVADJUST option typically has the effect of significantly reducing the accuracy radius γ of a $100 \times (1 - \epsilon)$% confidence interval for the true probability content of the estimated $(1 - \alpha)$th quantile. The control-variate-adjusted quantile estimate takes roughly twice as long to compute, but it is typically much more accurate than the sample quantile.

EPS=*value* specifies the value ϵ for a $100 \times (1 - \epsilon)$% confidence interval for the true probability content of the estimated $(1 - \alpha)$th quantile. The default value for the accuracy confidence is 99%, corresponding to EPS=0.01.

NSAMP=*n* specifies the sample size for the simulation. By default, n is set based on the values of the target accuracy radius γ and accuracy confidence $100 \times (1 - \epsilon)$true probability content of the estimated $(1 - \alpha)$th quantile. With the default values for γ, ϵ, and α (0.005, 0.01, and 0.05, respectively), NSAMP=12604 by default.

REPORT specifies that a report on the simulation should be displayed, including a listing of the parameters, such as γ, ϵ, and α as well as an analysis of various methods for estimating or approximating the quantile.

SEED=*number* specifies a positive integer less than $2^{31} - 1$. The value of the SEED= option is used to start the pseudo-random number generator for the simulation. The default is a value generated from reading the time of day from the computer's clock.

ALPHA=*p*

specifies the level of significance p for $100(1-p)$% confidence intervals. This option is useful only if you also specify the CL option, and, optionally, the PDIFF option. By default, p is equal to the value of the ALPHA= option in the PROC GLM statement or 0.05 if that option is not specified, This value is used to set the endpoints for confidence intervals for the individual means as well as for differences between means.

AT *variable* = *value*
AT (*variable-list*) = (*value-list*)
AT MEANS

enables you to modify the values of the covariates used in computing LS-means. By default, all covariate effects are set equal to their mean values for computation of standard LS-means. The AT option enables you to set the covariates to whatever values you consider interesting. For more information, see the section "Setting Covariate Values" on page 1556.

BYLEVEL

requests that PROC GLM process the OM data set by each level of the LS-mean effect in question. For more details, see the entry for the OM option in this section.

CL

requests confidence limits for the individual LS-means. If you specify the PDIFF option, confidence limits for differences between means are produced as well. You can control the confidence level with the ALPHA= option. Note that, if you specify an ADJUST= option, the confidence limits for the differences are adjusted for multiple inference but the confidence intervals for individual means are **not** adjusted.

COV

includes variances and covariances of the LS-means in the output data set specified in the OUT= option in the LSMEANS statement. Note that this is the covariance matrix for the LS-means themselves, not the covariance matrix for the differences between the LS-means, which is used in the PDIFF computations. If you omit the OUT= option, the COV option has no effect. When you specify the COV option, you can specify only one effect in the LSMEANS statement.

E

displays the coefficients of the linear functions used to compute the LS-means.

E=*effect*

specifies an effect in the model to use as an error term. The procedure uses the mean square for the *effect* as the error mean square when calculating estimated standard errors (requested with the STDERR option) and probabilities (requested with the STDERR, PDIFF, or TDIFF option). Unless you specify STDERR, PDIFF or TDIFF, the E= option is ignored. By default, if you specify the STDERR, PDIFF, or TDIFF option and do not specify the E= option, the procedure uses the error mean square for calculating standard errors and probabilities.

ETYPE=*n*

specifies the type (1, 2, 3, or 4, corresponding to Type I, II, III, and IV tests, respectively) of the E= effect. If you specify the E= option but not the ETYPE= option, the highest type computed in the analysis is used. If you omit the E= option, the ETYPE= option has no effect.

NOPRINT

suppresses the normal display of results from the LSMEANS statement. This option is useful when an output data set is created with the OUT= option in the LSMEANS statement.

OBSMARGINS
OM

specifies a potentially different weighting scheme for computing LS-means coefficients. The standard LS-means have equal coefficients across classification effects; however, the OM option changes these coefficients to be proportional to those found in the input data set. For more information, see the section "Changing the Weighting Scheme" on page 1557.

The BYLEVEL option modifies the observed-margins LS-means. Instead of computing the margins across the entire data set, the procedure computes separate margins for each level of the LS-mean effect in question. The resulting LS-means are actually equal to raw means in this case. If you specify the BYLEVEL option, it disables the AT option.

OUT=*SAS-data-set*

creates an output data set that contains the values, standard errors, and, optionally, the covariances (see the COV option) of the LS-means. For more information, see the "Output Data Sets" section on page 1574.

PDIFF*<=difftype>*

requests that p-values for differences of the LS-means be produced. The optional *difftype* specifies which differences to display. Possible values for *difftype* are ALL, CONTROL, CONTROLL, and CONTROLU. The ALL value requests all pairwise differences, and it is the default. The CONTROL value requests the differences with a control that, by default, is the first level of each of the specified LS-mean effects.

To specify which levels of the effects are the controls, list the quoted formatted values in parentheses after the keyword CONTROL. For example, if the effects A, B, and C are class variables, each having two levels, '1' and '2', the following LSMEANS statement specifies the '1' '2' level of A*B and the '2' '1' level of B*C as controls:

```
lsmeans A*B B*C / pdiff=control('1' '2', '2' '1');
```

For multiple effect situations such as this one, the ordering of the list is significant, and you should check the output to make sure that the controls are correct.

Two-tailed tests and confidence limits are associated with the CONTROL *difftype*. For one-tailed results, use either the CONTROLL or CONTROLU *difftype*. The CONTROLL *difftype* tests whether the noncontrol levels are significantly less than the control; the lower confidence limits for the control minus the noncontrol levels are considered to be minus infinity. Conversely, the CONTROLU *difftype* tests whether the noncontrol levels are significantly greater than the control; the upper confidence limits for the noncontrol levels minus the control are considered to be infinity.

The default multiple comparisons adjustment for each *difftype* is shown in the following table.

difftype	**Default ADJUST=**
Not specified	T
ALL	TUKEY
CONTROL CONTROLL CONTROLU	DUNNETT

If no *difftype* is specified, the default for the ADJUST= option is T (that is, no adjustment); for PDIFF=ALL, ADJUST=TUKEY is the default; in all other instances, the default value for the ADJUST= option is DUNNETT. If there is a conflict between the PDIFF= and ADJUST= options, the ADJUST= option takes precedence.

For example, in order to compute one-sided confidence limits for differences with a control, adjusted according to Dunnett's procedure, the following statements are equivalent:

```
lsmeans Treatment / pdiff=controll cl;
lsmeans Treatment / pdiff=controll cl adjust=dunnett;
```

SLICE = *fixed-effect*
SLICE = (*fixed-effects***)**

specifies effects within which to test for differences between interaction LS-mean effects. This can produce what are known as tests of simple effects (Winer 1971). For example, suppose that A*B is significant and you want to test for the effect of A within each level of B. The appropriate LSMEANS statement is

```
lsmeans A*B / slice=B;
```

This code tests for the simple main effects of A for B, which are calculated by extracting the appropriate rows from the coefficient matrix for the A*B LS-means and using them to form an *F*-test as performed by the CONTRAST statement.

SINGULAR=*number*

tunes the estimability checking. If ABS$(\mathbf{L} - \mathbf{LH}) > C \times$*number* for any row, then $\mathbf{L}$ is declared nonestimable. $\mathbf{H}$ is the $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ matrix, and C is ABS$(\mathbf{L})$ except for rows where $\mathbf{L}$ is zero, and then it is 1. The default value for the SINGULAR= option is 10^{-4}. Values for the SINGULAR= option must be between 0 and 1.

STDERR

produces the standard error of the LS-means and the probability level for the hypothesis H_0: LS-mean $= 0$.

TDIFF

produces the t values for all hypotheses H_0: LS-mean(i) $=$ LS-mean(j) and the corresponding probabilities.

MANOVA Statement

MANOVA < *test-options* >< / *detail-options* > **;**

If the MODEL statement includes more than one dependent variable, you can perform multivariate analysis of variance with the MANOVA statement. The *test-options* define which effects to test, while the *detail-options* specify how to execute the tests and what results to display.

When a MANOVA statement appears before the first RUN statement, PROC GLM enters a multivariate mode with respect to the handling of missing values; in addition to observations with missing independent variables, observations with *any* missing dependent variables are excluded from the analysis. If you want to use this mode of handling missing values and do not need any multivariate analyses, specify the MANOVA option in the PROC GLM statement.

If you use both the CONTRAST and MANOVA statements, the MANOVA statement must appear after the CONTRAST statement.

Test Options

The following options can be specified in the MANOVA statement as *test-options* in order to define which multivariate tests to perform.

H=*effects* | **INTERCEPT** | **_ALL_**

specifies effects in the preceding model to use as hypothesis matrices. For each $\mathbf{H}$ matrix (the SSCP matrix associated with an effect), the H= specification displays the characteristic roots and vectors of $\mathbf{E}^{-1}\mathbf{H}$ (where $\mathbf{E}$ is the matrix associated with the error effect), Hotelling-Lawley trace, Pillai's trace, Wilks' criterion, and Roy's maximum root criterion with approximate F statistic.

Use the keyword INTERCEPT to produce tests for the intercept. To produce tests for all effects listed in the MODEL statement, use the keyword _ALL_ in place of a list of effects. For background and further details, see the “Multivariate Analysis of Variance” section on page 1558.

E=*effect*

specifies the error effect. If you omit the E= specification, the GLM procedure uses the error SSCP (residual) matrix from the analysis.

M=*equation,. . .,equation* | **(***row-of-matrix,. . .,row-of-matrix***)**

specifies a transformation matrix for the dependent variables listed in the MODEL statement. The equations in the M= specification are of the form

$$\begin{aligned} c_1 \times \textit{dependent-variable} \quad & \pm \quad c_2 \times \textit{dependent-variable} \\ \cdots \quad & \pm \quad c_n \times \textit{dependent-variable} \end{aligned}$$

where the c_i values are coefficients for the various *dependent-variables*. If the value of a given c_i is 1, it can be omitted; in other words $1 \times Y$ is the same as Y. Equations should involve two or more dependent variables. For sample syntax, see the “Examples” section on page 1496.

Alternatively, you can input the transformation matrix directly by entering the elements of the matrix with commas separating the rows and parentheses surrounding the matrix. When this alternate form of input is used, the number of elements in each row must equal the number of dependent variables. Although these combinations actually represent the columns of the $\mathbf{M}$ matrix, they are displayed by rows.

When you include an M= specification, the analysis requested in the MANOVA statement is carried out for the variables defined by the equations in the specification, not the original dependent variables. If you omit the M= option, the analysis is performed for the original dependent variables in the MODEL statement.

If an M= specification is included without either the MNAMES= or PREFIX= option, the variables are labeled MVAR1, MVAR2, and so forth, by default. For further information, see the “Multivariate Analysis of Variance” section on page 1558.

MNAMES=*names*

provides names for the variables defined by the equations in the M= specification. Names in the list correspond to the M= equations or to the rows of the $\mathbf{M}$ matrix (as it is entered).

PREFIX=*name*

is an alternative means of identifying the transformed variables defined by the M= specification. For example, if you specify PREFIX=DIFF, the transformed variables are labeled DIFF1, DIFF2, and so forth.

Detail Options

You can specify the following options in the MANOVA statement after a slash as *detail-options*.

CANONICAL

displays a canonical analysis of the $\mathbf{H}$ and $\mathbf{E}$ matrices (transformed by the $\mathbf{M}$ matrix, if specified) instead of the default display of characteristic roots and vectors.

ETYPE=*n*

specifies the type (1, 2, 3, or 4, corresponding to Type I, II, III, and IV tests, respectively) of the **E** matrix, the SSCP matrix associated with the E= effect. You need this option if you use the E= specification to specify an error effect other than residual error and you want to specify the type of sums of squares used for the effect. If you specify ETYPE=n, the corresponding test must have been performed in the MODEL statement, either by options SSn, En, or the default Type I and Type III tests. By default, the procedure uses an ETYPE= value corresponding to the highest type (largest n) used in the analysis.

HTYPE=*n*

specifies the type (1, 2, 3, or 4, corresponding to Type I, II, III, and IV tests, respectively) of the **H** matrix. See the ETYPE= option for more details.

ORTH

requests that the transformation matrix in the M= specification of the MANOVA statement be orthonormalized by rows before the analysis.

PRINTE

displays the error SSCP matrix $\mathbf{E}$. If the $\mathbf{E}$ matrix is the error SSCP (residual) matrix from the analysis, the partial correlations of the dependent variables given the independent variables are also produced.

For example, the statement

```
   manova / printe;
```

displays the error SSCP matrix and the partial correlation matrix computed from the error SSCP matrix.

PRINTH

displays the hypothesis SSCP matrix $\mathbf{H}$ associated with each effect specified by the H= specification.

SUMMARY

produces analysis-of-variance tables for each dependent variable. When no $\mathbf{M}$ matrix is specified, a table is displayed for each original dependent variable from the MODEL statement; with an $\mathbf{M}$ matrix other than the identity, a table is displayed for each transformed variable defined by the $\mathbf{M}$ matrix.

Examples

The following statements provide several examples of using a MANOVA statement.

```
proc glm;
   class A B;
   model Y1-Y5=A B(A) / nouni;
   manova h=A e=B(A) / printh printe htype=1 etype=1;
   manova h=B(A) / printe;
   manova h=A e=B(A) m=Y1-Y2,Y2-Y3,Y3-Y4,Y4-Y5
          prefix=diff;
   manova h=A e=B(A) m=(1 -1  0  0  0,
                        0  1 -1  0  0,
                        0  0  1 -1  0,
                        0  0  0  1 -1) prefix=diff;
run;
```

Since this MODEL statement requests no options for type of sums of squares, the procedure uses Type I and Type III sums of squares. The first MANOVA statement specifies A as the hypothesis effect and B(A) as the error effect. As a result of the PRINTH option, the procedure displays the hypothesis SSCP matrix associated with the A effect; and, as a result of the PRINTE option, the procedure displays the error SSCP matrix associated with the B(A) effect. The option HTYPE=1 specifies a Type I **H** matrix, and the option ETYPE=1 specifies a Type I **E** matrix.

The second MANOVA statement specifies B(A) as the hypothesis effect. Since no error effect is specified, PROC GLM uses the error SSCP matrix from the analysis as the **E** matrix. The PRINTE option displays this **E** matrix. Since the **E** matrix is the error SSCP matrix from the analysis, the partial correlation matrix computed from this matrix is also produced.

The third MANOVA statement requests the same analysis as the first MANOVA statement, but the analysis is carried out for variables transformed to be successive differences between the original dependent variables. The option PREFIX=DIFF labels the transformed variables as DIFF1, DIFF2, DIFF3, and DIFF4.

Finally, the fourth MANOVA statement has the identical effect as the third, but it uses an alternative form of the M= specification. Instead of specifying a set of equations, the fourth MANOVA statement specifies rows of a matrix of coefficients for the five dependent variables.

As a second example of the use of the M= specification, consider the following:

```
proc glm;
   class group;
   model dose1-dose4=group / nouni;
   manova h = group
          m = -3*dose1 -   dose2 +   dose3 + 3*dose4,
                 dose1 -   dose2 -   dose3 +   dose4,
                -dose1 + 3*dose2 - 3*dose3 +   dose4
          mnames = Linear Quadratic Cubic
          / printe;
run;
```

The M= specification gives a transformation of the dependent variables dose1 through dose4 into orthogonal polynomial components, and the MNAMES= option labels the transformed variables LINEAR, QUADRATIC, and CUBIC, respectively. Since the PRINTE option is specified and the default residual matrix is used as an error term, the partial correlation matrix of the orthogonal polynomial components is also produced.

MEANS Statement

MEANS *effects* < **/** *options* > **;**

Within each group corresponding to each effect specified in the MEANS statement, PROC GLM computes the arithmetic means and standard deviations of all continuous variables in the model (both dependent and independent). You may specify only classification effects in the MEANS statement—that is, effects that contain only classification variables.

Note that the arithmetic means are not adjusted for other effects in the model; for adjusted means, see the "LSMEANS Statement" section on page 1488. If you use a WEIGHT statement, PROC GLM computes weighted means; see the "Weighted Means" section on page 1555.

You may also specify options to perform multiple comparisons. However, the MEANS statement performs multiple comparisons only for main effect means; for multiple comparisons of interaction means, see the "LSMEANS Statement" section on page 1488.

You can use any number of MEANS statements, provided that they appear after the MODEL statement. For example, suppose A and B each have two levels. Then, if you use the following statements

```
proc glm;
   class A B;
   model Y=A B A*B;
   means A B / tukey;
   means A*B;
run;
```

the means, standard deviations, and Tukey's multiple comparisons tests are displayed for each level of the main effects A and B, and just the means and standard deviations are displayed for each of the four combinations of levels for A*B. Since multiple comparisons tests apply only to main effects, the single MEANS statement

```
means A B A*B / tukey;
```

produces the same results.

PROC GLM does not compute means for interaction effects containing continuous variables. Thus, if you have the model

```
class A;
model Y=A X A*X;
```

then the effects X and A*X cannot be used in the MEANS statement. However, if you specify the effect A in the means statement

```
means A;
```

then PROC GLM, by default, displays within-A arithmetic means of both Y and X. Use the DEPONLY option to display means of only the dependent variables.

```
means A / deponly;
```

If you use a WEIGHT statement, PROC GLM computes weighted means and estimates their variance as inversely proportional to the corresponding sum of weights (see the “Weighted Means” section on page 1555). However, note that the statistical interpretation of multiple comparison tests for weighted means is not well understood. See the “Multiple Comparisons” section on page 1540 for formulas. The following table summarizes categories of options available in the MEANS statement.

Task	**Available options**
Modify output	DEPONLY
Perform multiple comparison tests	BON DUNCAN DUNNETT DUNNETTL DUNNETTU GABRIEL GT2 LSD REGWQ SCHEFFE SIDAK SMM SNK T TUKEY WALLER
Specify additional details for multiple comparison tests	ALPHA= CLDIFF CLM E= ETYPE= HTYPE= KRATIO= LINES NOSORT
Test for homogeneity of variances	HOVTEST
Compensate for heterogeneous variances	WELCH

These options are described in the following list.

ALPHA=*p*

specifies the level of significance for comparisons among the means. By default, p is equal to the value of the ALPHA= option in the PROC GLM statement or 0.05 if that option is not specified. You can specify any value greater than 0 and less than 1.

BON

performs Bonferroni t tests of differences between means for all main effect means in the MEANS statement. See the CLDIFF and LINES options for a discussion of how the procedure displays results.

CLDIFF

presents results of the BON, GABRIEL, SCHEFFE, SIDAK, SMM, GT2, T, LSD, and TUKEY options as confidence intervals for all pairwise differences between means, and the results of the DUNNETT, DUNNETTU, and DUNNETTL options as confidence intervals for differences with the control. The CLDIFF option is the default for unequal cell sizes unless the DUNCAN, REGWQ, SNK, or WALLER option is specified.

CLM

presents results of the BON, GABRIEL, SCHEFFE, SIDAK, SMM, T, and LSD options as intervals for the mean of each level of the variables specified in the MEANS statement. For all options except GABRIEL, the intervals are confidence intervals for the true means. For the GABRIEL option, they are *comparison intervals* for comparing means pairwise: in this case, if the intervals corresponding to two means overlap, then the difference between them is insignificant according to Gabriel's method.

DEPONLY

displays only means for the dependent variables. By default, PROC GLM produces means for all continuous variables, including continuous independent variables.

DUNCAN

performs Duncan's multiple range test on all main effect means given in the MEANS statement. See the LINES option for a discussion of how the procedure displays results.

DUNNETT < **(***formatted-control-values***)** >

performs Dunnett's two-tailed t test, testing if any treatments are significantly different from a single control for all main effects means in the MEANS statement.

To specify which level of the effect is the control, enclose the formatted value in quotes in parentheses after the keyword. If more than one effect is specified in the MEANS statement, you can use a list of control values within the parentheses. By default, the first level of the effect is used as the control. For example,

```
means A  / dunnett('CONTROL');
```

where CONTROL is the formatted control value of A. As another example,

```
means A B C / dunnett('CNTLA' 'CNTLB' 'CNTLC');
```

where CNTLA, CNTLB, and CNTLC are the formatted control values for A, B, and C, respectively.

DUNNETTL < **(***formatted-control-value***)** >

performs Dunnett's one-tailed t test, testing if any treatment is significantly less than the control. Control level information is specified as described for the DUNNETT option.

DUNNETTU < **(***formatted-control-value***)** >

performs Dunnett's one-tailed t test, testing if any treatment is significantly greater than the control. Control level information is specified as described for the DUNNETT option.

E=*effect*

specifies the error mean square used in the multiple comparisons. By default, PROC GLM uses the overall residual or error mean square (MS). The effect specified with the E= option must be a term in the model; otherwise, the procedure uses the residual MS.

ETYPE=*n*

specifies the type of mean square for the error effect. When you specify E=*effect*, you may need to indicate which type (1, 2, 3, or 4) of MS is to be used. The n value must be one of the types specified in or implied by the MODEL statement. The default MS type is the highest type used in the analysis.

GABRIEL

performs Gabriel's multiple-comparison procedure on all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

GT2

see the SMM option.

HOVTEST
HOVTEST=BARTLETT
HOVTEST=BF
HOVTEST=LEVENE $<$ **(TYPE= ABS | SQUARE)**$>$
HOVTEST=OBRIEN $<$ **(W=*number*)**$>$

requests a homogeneity of variance test for the groups defined by the MEANS effect. You can optionally specify a particular test; if you do not specify a test, Levene's test (Levene 1960) with TYPE=SQUARE is computed. Note that this option is ignored unless your MODEL statement specifies a simple one-way model.

The HOVTEST=BARTLETT option specifies Bartlett's test (Bartlett 1937), a modification of the normal-theory likelihood ratio test.

The HOVTEST=BF option specifies Brown and Forsythe's variation of Levene's test (Brown and Forsythe 1974).

The HOVTEST=LEVENE option specifies Levene's test (Levene 1960), which is widely considered to be the standard homogeneity of variance test. You can use the TYPE= option in parentheses to specify whether to use the absolute residuals (TYPE=ABS) or the squared residuals (TYPE=SQUARE) in Levene's test. TYPE=SQUARE is the default.

The HOVTEST=OBRIEN option specifies O'Brien's test (O'Brien 1979), which is basically a modification of HOVTEST=LEVENE(TYPE=SQUARE). You can use the W= option in parentheses to tune the variable to match the suspected kurtosis of the underlying distribution. By default, W=0.5, as suggested by O'Brien (1979, 1981).

See the "Homogeneity of Variance in One-Way Models" section on page 1553 for more details on these methods. Example 30.10 on page 1623 illustrates the use of the HOVTEST and WELCH options in the MEANS statement in testing for equal group variances and adjusting for unequal group variances in a one-way ANOVA.

HTYPE=*n*

specifies the MS type for the hypothesis MS. The HTYPE= option is needed only when the WALLER option is specified. The default HTYPE= value is the highest type used in the model.

KRATIO=*value*

specifies the Type 1/Type 2 error seriousness ratio for the Waller-Duncan test. Reasonable values for the KRATIO= option are 50, 100, 500, which roughly correspond for the two-level case to ALPHA levels of 0.1, 0.05, and 0.01, respectively. By default, the procedure uses the value of 100.

LINES

presents results of the BON, DUNCAN, GABRIEL, REGWQ, SCHEFFE, SIDAK, SMM, GT2, SNK, T, LSD, TUKEY, and WALLER options by listing the means in descending order and indicating nonsignificant subsets by line segments beside the corresponding means. The LINES option is appropriate for equal cell sizes, for which it is the default. The LINES option is also the default if the DUNCAN, REGWQ, SNK, or WALLER option is specified, or if there are only two cells of unequal size. The LINES option cannot be used in combination with the DUNNETT, DUNNETTL, or DUNNETTU option. In addition, the procedure has a restriction that no more than 24 overlapping groups of means can exist. If a mean belongs to more than 24 groups, the procedure issues an error message. You can either reduce the number of levels of the variable or use a multiple comparison test that allows the CLDIFF option rather than the LINES option.

Note: If the cell sizes are unequal, the harmonic mean of the cell sizes is used to compute the critical ranges. This approach is reasonable if the cell sizes are not too different, but it can lead to liberal tests if the cell sizes are highly disparate. In this case, you should not use the LINES option for displaying multiple comparisons results; use the TUKEY and CLDIFF options instead.

LSD

see the T option.

NOSORT

prevents the means from being sorted into descending order when the CLDIFF or CLM option is specified.

REGWQ

performs the Ryan-Einot-Gabriel-Welsch multiple range test on all main effect means in the MEANS statement. See the LINES option for a discussion of how the procedure displays results.

SCHEFFE

performs Scheffé’s multiple-comparison procedure on all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

SIDAK

performs pairwise t tests on differences between means with levels adjusted according to Sidak's inequality for all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

SMM
GT2

performs pairwise comparisons based on the studentized maximum modulus and Sidak's uncorrelated-t inequality, yielding Hochberg's GT2 method when sample sizes are unequal, for all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

SNK

performs the Student-Newman-Keuls multiple range test on all main effect means in the MEANS statement. See the LINES option for discussions of how the procedure displays results.

T
LSD

performs pairwise t tests, equivalent to Fisher's least-significant-difference test in the case of equal cell sizes, for all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

TUKEY

performs Tukey's studentized range test (HSD) on all main effect means in the MEANS statement. (When the group sizes are different, this is the Tukey-Kramer test.) See the CLDIFF and LINES options for discussions of how the procedure displays results.

WALLER

performs the Waller-Duncan k-ratio t test on all main effect means in the MEANS statement. See the KRATIO= and HTYPE= options for information on controlling details of the test, and the LINES option for a discussion of how the procedure displays results.

WELCH

requests Welch's (1951) variance-weighted one-way ANOVA. This alternative to the usual analysis of variance for a one-way model is robust to the assumption of equal within-group variances. This option is ignored unless your MODEL statement specifies a simple one-way model.

Note that using the WELCH option merely produces one additional table consisting of Welch's ANOVA. It does not affect all of the other tests displayed by the GLM procedure, which still require the assumption of equal variance for exact validity.

See the "Homogeneity of Variance in One-Way Models" section on page 1553 for more details on Welch's ANOVA. Example 30.10 on page 1623 illustrates the use of the HOVTEST and WELCH options in the MEANS statement in testing for equal group variances and adjusting for unequal group variances in a one-way ANOVA.

MODEL Statement

MODEL *dependents=independents* < */ options* > **;**

The MODEL statement names the dependent variables and independent effects. The syntax of effects is described in the “Specification of Effects” section on page 1517. If no independent effects are specified, only an intercept term is fit. You can specify only one MODEL statement (in contrast to the REG procedure, for example, which allows several MODEL statements in the same PROC REG run).

The following table summarizes options available in the MODEL statement.

Task	Options
Produce tests for the intercept	INTERCEPT
Omit the intercept parameter from model	NOINT
Produce parameter estimates	SOLUTION
Produce tolerance analysis	TOLERANCE
Suppress univariate tests and output	NOUNI
Display estimable functions	E E1 E2 E3 E4 ALIASING
Control hypothesis tests performed	SS1 SS2 SS3 SS4
Produce confidence intervals	ALPHA= CLI CLM CLPARM
Display predicted and residual values	P
Display intermediate calculations	INVERSE XPX
Tune sensitivity	SINGULAR= ZETA=

These options are described in the following list.

ALIASING

specifies that the estimable functions should be displayed as an *aliasing structure*, for which each row says which linear combination of the parameters is estimated by each estimable function; also, adds a column of the same information to the table of parameter estimates, giving for each parameter the expected value of the estimate associated with that parameter. This option is most useful in fractional factorial experiments that can be analyzed without a CLASS statement.

ALPHA=*p*

specifies the level of significance p for $100(1-p)\%$ confidence intervals. By default, p is equal to the value of the ALPHA= option in the PROC GLM statement, or 0.05 if that option is not specified. You may use values between 0 and 1.

CLI

produces confidence limits for individual predicted values for each observation. The CLI option is ignored if the CLM option is also specified.

CLM

produces confidence limits for a mean predicted value for each observation.

CLPARM

produces confidence limits for the parameter estimates (if the SOLUTION option is also specified) and for the results of all ESTIMATE statements.

E

displays the general form of all estimable functions. This is useful for determining the order of parameters when writing CONTRAST and ESTIMATE statements.

E1

displays the Type I estimable functions for each effect in the model and computes the corresponding sums of squares.

E2

displays the Type II estimable functions for each effect in the model and computes the corresponding sums of squares.

E3

displays the Type III estimable functions for each effect in the model and computes the corresponding sums of squares.

E4

displays the Type IV estimable functions for each effect in the model and computes the corresponding sums of squares.

INTERCEPT
INT

produces the hypothesis tests associated with the intercept as an effect in the model. By default, the procedure includes the intercept in the model but does not display associated tests of hypotheses. Except for producing the uncorrected total sum of squares instead of the corrected total sum of squares, the INT option is ignored when you use an ABSORB statement.

INVERSE
I

displays the augmented inverse (or generalized inverse) $\mathbf{X'X}$ matrix:

$$\begin{bmatrix} (X'X)^- & (X'X)^- X'Y \\ Y'X(X'X)^- & Y'Y - Y'X(X'X)^- X'Y \end{bmatrix}$$

The upper left-hand corner is the generalized inverse of $\mathbf{X'X}$, the upper right-hand corner is the parameter estimates, and the lower right-hand corner is the error sum of squares.

NOINT

omits the intercept parameter from the model.

NOUNI

suppresses the display of univariate statistics. You typically use the NOUNI option with a multivariate or repeated measures analysis of variance when you do not need the standard univariate results. The NOUNI option in a MODEL statement does not affect the univariate output produced by the REPEATED statement.

P

displays observed, predicted, and residual values for each observation that does not contain missing values for independent variables. The Durbin-Watson statistic is also displayed when the P option is specified. The PRESS statistic is also produced if either the CLM or CLI option is specified.

SINGULAR=*number*

tunes the sensitivity of the regression routine to linear dependencies in the design. If a diagonal pivot element is less than $C \times number$ as PROC GLM sweeps the $\mathbf{X'X}$ matrix, the associated design column is declared to be linearly dependent with previous columns, and the associated parameter is zeroed.

The C value adjusts the check to the relative scale of the variable. The C value is equal to the corrected sum of squares for the variable, unless the corrected sum of squares is 0, in which case C is 1. If you specify the NOINT option but not the ABSORB statement, PROC GLM uses the uncorrected sum of squares instead.

The default value of the SINGULAR= option, 10^{-7}, may be too small, but this value is necessary in order to handle the high-degree polynomials used in the literature to compare regression routines.

SOLUTION

produces a solution to the normal equations (parameter estimates). PROC GLM displays a solution by default when your model involves no classification variables, so you need this option only if you want to see the solution for models with classification effects.

SS1

displays the sum of squares associated with Type I estimable functions for each effect. These are also displayed by default.

SS2
displays the sum of squares associated with Type II estimable functions for each effect.

SS3
displays the sum of squares associated with Type III estimable functions for each effect. These are also displayed by default.

SS4
displays the sum of squares associated with Type IV estimable functions for each effect.

TOLERANCE
displays the tolerances used in the SWEEP routine. The tolerances are of the form C/USS or C/CSS, as described in the discussion of the SINGULAR= option. The tolerance value for the intercept is not divided by its uncorrected sum of squares.

XPX
displays the augmented $\mathbf{X}'\mathbf{X}$ crossproducts matrix:

$$\begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix}$$

ZETA=*value*
tunes the sensitivity of the check for estimability for Type III and Type IV functions. Any element in the estimable function basis with an absolute value less than the ZETA= option is set to zero. The default value for the ZETA= option is 10^{-8}.

Although it is possible to generate data for which this absolute check can be defeated, the check suffices in most practical examples. Additional research needs to be performed to make this check relative rather than absolute.

OUTPUT Statement

OUTPUT < **OUT=***SAS-data-set* > *keyword=names*
< . . . *keyword=names* > < */ option* > **;**

The OUTPUT statement creates a new SAS data set that saves diagnostic measures calculated after fitting the model. At least one specification of the form *keyword=names* is required.

All the variables in the original data set are included in the new data set, along with variables created in the OUTPUT statement. These new variables contain the values of a variety of diagnostic measures that are calculated for each observation in the data set. If you want to create a permanent SAS data set, you must specify a two-level name (refer to *SAS Language Reference: Concepts* for more information on permanent SAS data sets).

Details on the specifications in the OUTPUT statement follow.

keyword=names

specifies the statistics to include in the output data set and provides names to the new variables that contain the statistics. Specify a keyword for each desired statistic (see the following list of keywords), an equal sign, and the variable or variables to contain the statistic.

In the output data set, the first variable listed after a keyword in the OUTPUT statement contains that statistic for the first dependent variable listed in the MODEL statement; the second variable contains the statistic for the second dependent variable in the MODEL statement, and so on. The list of variables following the equal sign can be shorter than the list of dependent variables in the MODEL statement. In this case, the procedure creates the new names in order of the dependent variables in the MODEL statement. See the “Examples” section on page 1509.

The keywords allowed and the statistics they represent are as follows:

COOKD Cook’s D influence statistic

COVRATIO standard influence of observation on covariance of parameter estimates

DFFITS standard influence of observation on predicted value

H leverage, $h_i = x_i(\mathbf{X}'\mathbf{X})^{-1}x_i'$

LCL lower bound of a $100(1-p)$% confidence interval for an individual prediction. The p-level is equal to the value of the ALPHA= option in the OUTPUT statement or, if this option is not specified, to the ALPHA= option in the PROC GLM statement. If neither of these options is set then $p = 0.05$ by default, resulting in the lower bound for a 95% confidence interval. The interval also depends on the variance of the error, as well as the variance of the parameter estimates. For the corresponding upper bound, see the UCL keyword.

LCLM lower bound of a $100(1-p)$% confidence interval for the expected value (mean) of the predicted value. The p-level is equal to the value of the ALPHA= option in the OUTPUT statement or, if this option is not specified, to the ALPHA= option in the PROC GLM statement. If neither of these options is set then $p = 0.05$ by default, resulting in the lower bound for a 95% confidence interval. For the corresponding upper bound, see the UCLM keyword.

PREDICTED | P predicted values

PRESS residual for the ith observation that results from dropping it and predicting it on the basis of all other observations. This is the residual divided by $(1-h_i)$ where h_i is the leverage, defined previously.

RESIDUAL | R residuals, calculated as ACTUAL $-$ PREDICTED

RSTUDENT a studentized residual with the current observation deleted

STDI standard error of the individual predicted value

STDP standard error of the mean predicted value

STDR standard error of the residual

STUDENT studentized residuals, the residual divided by its standard error

UCL upper bound of a $100(1-p)$% confidence interval for an individual prediction. The p-level is equal to the value of the ALPHA= option in the OUTPUT statement or, if this option is not specified, to the ALPHA= option in the PROC GLM statement. If neither of these options is set then $p = 0.05$ by default, resulting in the upper bound for a 95% confidence interval. The interval also depends on the variance of the error, as well as the variance of the parameter estimates. For the corresponding lower bound, see the LCL keyword.

UCLM upper bound of a $100(1-p)$% confidence interval for the expected value (mean) of the predicted value. The p-level is equal to the value of the ALPHA= option in the OUTPUT statement or, if this option is not specified, to the ALPHA= option in the PROC GLM statement. If neither of these options is set then $p = 0.05$ by default, resulting in the upper bound for a 95% confidence interval. For the corresponding lower bound, see the LCLM keyword.

OUT=*SAS-data-set*

gives the name of the new data set. By default, the procedure uses the DATAn convention to name the new data set.

The following option is available in the OUTPUT statement and is specified after a slash(/):

ALPHA=*p*

specifies the level of significance p for $100(1-p)$% confidence intervals. By default, p is equal to the value of the ALPHA= option in the PROC GLM statement or 0.05 if that option is not specified. You may use values between 0 and 1.

See Chapter 3, “Introduction to Regression Procedures,” and the “Influence Diagnostics” section in Chapter 55, “The REG Procedure,” for details on the calculation of these statistics.

Examples

The following statements show the syntax for creating an output data set with a single dependent variable.

```
proc glm;
   class a b;
   model y=a b a*b;
   output out=new p=yhat r=resid stdr=eresid;
run;
```

These statements create an output data set named new. In addition to all the variables from the original data set, new contains the variable yhat, with values that are predicted values of the dependent variable y; the variable resid, with values that are the residual values of y; and the variable eresid, with values that are the standard errors of the residuals.

The following statements show a situation with five dependent variables.

```
proc glm;
   by group;
   class a;
   model y1-y5=a x(a);
   output out=pout predicted=py1-py5;
run;
```

Data set pout contains five new variables, py1 through py5. The values of py1 are the predicted values of y1; the values of py2 are the predicted values of y2; and so on.

For more information on the data set produced by the OUTPUT statement, see the section "Output Data Sets" on page 1574.

RANDOM Statement

RANDOM *effects* $<$ **/** *options* $>$ **;**

When some model effects are random (that is, assumed to be sampled from a normal population of effects), you can specify these effects in the RANDOM statement in order to compute the expected values of mean squares for various model effects and contrasts and, optionally, to perform random effects analysis of variance tests. You can use as many RANDOM statements as you want, provided that they appear after the MODEL statement. If you use a CONTRAST statement with a RANDOM statement and you want to obtain the expected mean squares for the contrast hypothesis, you must enter the CONTRAST statement before the RANDOM statement.

Note: PROC GLM uses only the information pertaining to expected mean squares when you specify the TEST option in the RANDOM statement and, even then, only in the extra F tests produced by the RANDOM statement. Other features in the GLM procedure—including the results of the LSMEANS and ESTIMATE statements—assume that all effects are fixed, so that all tests and estimability checks for these statements are based on a fixed effects model, even when you use a RANDOM statement. Therefore, you should use the MIXED procedure to compute tests involving these features that take the random effects into account; see the section "PROC GLM versus PROC MIXED for Random Effects Analysis" on page 1567 and Chapter 41, "The MIXED Procedure," for more information.

When you use the RANDOM statement, by default the GLM procedure produces the Type III expected mean squares for model effects and for contrasts specified before the RANDOM statement in the program code. In order to obtain expected values for other types of mean squares, you need to specify which types of mean squares are of interest in the MODEL statement. See the section "Computing Type I, II, and IV Expected Mean Squares" on page 1570 for more information.

The list of effects in the RANDOM statement should contain one or more of the pure classification effects specified in the MODEL statement (that is, main effects, crossed effects, or nested effects involving only class variables). The coefficients corresponding to each effect specified are assumed to be normally and independently distributed with common variance. Levels in different effects are assumed to be independent.

You can specify the following options in the RANDOM statement after a slash:

Q

displays all quadratic forms in the fixed effects that appear in the expected mean squares. For some designs, large mixed-level factorials, for example, the Q option may generate a substantial amount of output.

TEST

performs hypothesis tests for each effect specified in the model, using appropriate error terms as determined by the expected mean squares.

Caution: PROC GLM does not automatically declare interactions to be random when the effects in the interaction are declared random. For example,

```
   random a b / test;
```

does not produce the same expected mean squares or tests as

```
   random a b a*b / test;
```

To ensure correct tests, you need to list all random interactions and random main effects in the RANDOM statement.

See the section “Random Effects Analysis” on page 1567 for more information on the calculation of expected mean squares and tests and on the similarities and differences between the GLM and MIXED procedures. See Chapter 4, “Introduction to Analysis-of-Variance Procedures,” and Chapter 41, “The MIXED Procedure,” for more information on random effects.

REPEATED Statement

REPEATED *factor-specification* < */ options* > **;**

When values of the dependent variables in the MODEL statement represent repeated measurements on the same experimental unit, the REPEATED statement enables you to test hypotheses about the measurement factors (often called *within-subject factors*) as well as the interactions of within-subject factors with independent variables in the MODEL statement (often called *between-subject factors*). The REPEATED statement provides multivariate and univariate tests as well as hypothesis tests for a variety of single-degree-of-freedom contrasts. There is no limit to the number of within-subject factors that can be specified.

The REPEATED statement is typically used for handling repeated measures designs with one repeated response variable. Usually, the variables on the left-hand side of the equation in the MODEL statement represent one repeated response variable. This does not mean that only one factor can be listed in the REPEATED statement. For example, one repeated response variable (hemoglobin count) might be measured 12 times (implying variables Y1 to Y12 on the left-hand side of the equal sign in the MODEL statement), with the associated within-subject factors treatment and time (implying two factors listed in the REPEATED statement). See the "Examples" section on page 1514 for an example of how PROC GLM handles this case. Designs with two or more repeated response variables can, however, be handled with the IDENTITY transformation; see page 1513 for more information, and Example 30.9 on page 1618 for an example of analyzing a doubly-multivariate repeated measures design.

When a REPEATED statement appears, the GLM procedure enters a multivariate mode of handling missing values. If any values for variables corresponding to each combination of the within-subject factors are missing, the observation is excluded from the analysis.

If you use a CONTRAST or TEST statement with a REPEATED statement, you must enter the CONTRAST or TEST statement before the REPEATED statement.

The simplest form of the REPEATED statement requires only a *factor-name*. With two repeated factors, you must specify the *factor-name* and number of levels (*levels*) for each factor. Optionally, you can specify the actual values for the levels (*level-values*), a *transformation* that defines single-degree-of freedom contrasts, and *options* for additional analyses and output. When you specify more than one within-subject factor, the *factor-names* (and associated level and transformation information) must be separated by a comma in the REPEATED statement. These terms are described in the following section, "Syntax Details."

Syntax Details

You can specify the following terms in the REPEATED statement.

factor-specification

The *factor-specification* for the REPEATED statement can include any number of individual factor specifications, separated by commas, of the following form:

factor-name levels < **(***level-values***)** > < *transformation* >

where

factor-name names a factor to be associated with the dependent variables. The name should not be the same as any variable name that already exists in the data set being analyzed and should conform to the usual conventions of SAS variable names.

When specifying more than one factor, list the dependent variables in the MODEL statement so that the within-subject factors defined in the REPEATED statement are nested; that is, the first factor defined in the REPEATED statement should be the one with values that change least frequently.

levels gives the number of levels associated with the factor being defined. When there is only one within-subject factor, the number of levels is equal to the number of dependent variables. In this case, *levels* is optional. When more than one within-subject factor is defined, however, *levels* is required, and the product of the number of levels of all the factors must equal the number of dependent variables in the MODEL statement.

(*level-values*) gives values that correspond to levels of a repeated-measures factor. These values are used to label output and as spacings for constructing orthogonal polynomial contrasts if you specify a POLYNOMIAL transformation. The number of values specified must correspond to the number of levels for that factor in the REPEATED statement. Enclose the *level-values* in parentheses.

The following *transformation* keywords define single-degree-of-freedom contrasts for factors specified in the REPEATED statement. Since the number of contrasts generated is always one less than the number of levels of the factor, you have some control over which contrast is omitted from the analysis by which transformation you select. The only exception is the IDENTITY transformation; this transformation is not composed of contrasts and has the same degrees of freedom as the factor has levels. By default, the procedure uses the CONTRAST transformation.

CONTRAST < **(***ordinal-reference-level* **)** > generates contrasts between levels of the factor and a reference level. By default, the procedure uses the last level as the reference level; you can optionally specify a reference level in parentheses after the keyword CONTRAST. The reference level corresponds to the ordinal value of the level rather than the level value specified. For example, to generate contrasts between the first level of a factor and the other levels, use

```
contrast(1)
```

HELMERT generates contrasts between each level of the factor and the mean of subsequent levels.

IDENTITY generates an identity transformation corresponding to the associated factor. This transformation is *not* composed of contrasts; it has n degrees of freedom for an n-level factor, instead of $n - 1$. This can be used for doubly-multivariate repeated measures.

MEAN < **(***ordinal-reference-level* **)** > generates contrasts between levels of the factor and the mean of all other levels of the factor. Specifying a reference level eliminates the contrast between that level and the mean. Without a reference level, the contrast involving the last level is omitted. See the CONTRAST transformation for an example.

POLYNOMIAL generates orthogonal polynomial contrasts. Level values, if provided, are used as spacings in the construction of the polynomials; otherwise, equal spacing is assumed.

PROFILE generates contrasts between adjacent levels of the factor.

You can specify the following options in the REPEATED statement after a slash.

CANONICAL
performs a canonical analysis of the $\mathbf{H}$ and $\mathbf{E}$ matrices corresponding to the transformed variables specified in the REPEATED statement.

HTYPE=*n*
specifies the type of the $\mathbf{H}$ matrix used in the multivariate tests and the type of sums of squares used in the univariate tests. See the HTYPE= option in the specifications for the MANOVA statement for further details.

MEAN
generates the overall arithmetic means of the within-subject variables.

NOM
displays only the results of the univariate analyses.

NOU
displays only the results of the multivariate analyses.

PRINTE
displays the $\mathbf{E}$ matrix for each combination of within-subject factors, as well as partial correlation matrices for both the original dependent variables and the variables defined by the transformations specified in the REPEATED statement. In addition, the PRINTE option provides sphericity tests for each set of transformed variables. If the requested transformations are not orthogonal, the PRINTE option also provides a sphericity test for a set of orthogonal contrasts.

PRINTH
displays the $\mathbf{H}$ (SSCP) matrix associated with each multivariate test.

PRINTM
displays the transformation matrices that define the contrasts in the analysis. PROC GLM always displays the $\mathbf{M}$ matrix so that the transformed variables are defined by the rows, not the columns, of the displayed $\mathbf{M}$ matrix. In other words, PROC GLM actually displays $\mathbf{M}'$.

PRINTRV
displays the characteristic roots and vectors for each multivariate test.

SUMMARY
produces analysis-of-variance tables for each contrast defined by the within-subject factors. Along with tests for the effects of the independent variables specified in the MODEL statement, a term labeled MEAN tests the hypothesis that the overall mean of the contrast is zero.

Examples

When specifying more than one factor, list the dependent variables in the MODEL statement so that the within-subject factors defined in the REPEATED statement are nested; that is, the first factor defined in the REPEATED statement should be the one with values that change least frequently. For example, assume that three treatments are administered at each of four times, for a total of twelve dependent variables on each experimental unit. If the variables are listed in the MODEL statement as Y1 through Y12, then the following REPEATED statement

```
proc glm;
    classes group;
    model Y1-Y12=group / nouni;
    repeated trt 3, time 4;
run;
```

implies the following structure:

	Dependent Variables											
	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
Value of trt	1	1	1	1	2	2	2	2	3	3	3	3
Value of time	1	2	3	4	1	2	3	4	1	2	3	4

The REPEATED statement always produces a table like the preceding one. For more information, see the section "Repeated Measures Analysis of Variance" on page 1560.

TEST Statement

TEST < **H=***effects* > **E=***effect* < **/** *options* > **;**

Although an F value is computed for all sums of squares in the analysis using the residual MS as an error term, you may request additional F tests using other effects as error terms. You need a TEST statement when a nonstandard error structure (as in a split-plot design) exists. Note, however, that this may not be appropriate if the design is unbalanced, since in most unbalanced designs with nonstandard error structures, mean squares are not necessarily independent with equal expectations under the null hypothesis.

Caution: The GLM procedure does not check any of the assumptions underlying the F statistic. When you specify a TEST statement, you assume sole responsibility for the validity of the F statistic produced. To help validate a test, you can use the RANDOM statement and inspect the expected mean squares, or you can use the TEST option of the RANDOM statement.

You may use as many TEST statements as you want, provided that they appear after the MODEL statement.

You can specify the following terms in the TEST statement.

H=*effects* specifies which effects in the preceding model are to be used as hypothesis (numerator) effects.

E=*effect* specifies one, and only one, effect to use as the error (denominator) term. The E= specification is required.

By default, the sum of squares type for all hypothesis sum of squares and error sum of squares is the highest type computed in the model. If the hypothesis type or error type is to be another type that was computed in the model, you should specify one or both of the following options after a slash.

ETYPE=*n*

specifies the type of sum of squares to use for the error term. The type must be a type computed in the model (n=1, 2, 3, or 4).

HTYPE=*n*

specifies the type of sum of squares to use for the hypothesis. The type must be a type computed in the model (n=1, 2, 3, or 4).

This example illustrates the TEST statement with a split-plot model:

```
proc glm;
   class a b c;
   model y=a  b(a) c a*c b*c(a);
   test h=a e=b(a)/ htype=1 etype=1;
   test h=c a*c e=b*c(a) / htype=1 etype=1;
run;
```

WEIGHT Statement

WEIGHT *variable* **;**

When a WEIGHT statement is used, a weighted residual sum of squares

$$\sum_i w_i(y_i - \hat{y}_i)^2$$

is minimized, where w_i is the value of the variable specified in the WEIGHT statement, y_i is the observed value of the response variable, and $\hat{y}_i$ is the predicted value of the response variable.

If you specify the WEIGHT statement, it must appear before the first RUN statement or it is ignored.

An observation is used in the analysis only if the value of the WEIGHT statement variable is nonmissing and greater than zero.

The WEIGHT statement has no effect on degrees of freedom or number of observations, but it is used by the MEANS statement when calculating means and performing multiple comparison tests (as described in the “MEANS Statement” section beginning on page 1497). The normal equations used when a WEIGHT statement is present are

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{Y}$$

where $\mathbf{W}$ is a diagonal matrix consisting of the values of the variable specified in the WEIGHT statement.

If the weights for the observations are proportional to the reciprocals of the error variances, then the weighted least-squares estimates are best linear unbiased estimators (BLUE).

Details

Statistical Assumptions for Using PROC GLM

The basic statistical assumption underlying the least-squares approach to general linear modeling is that the observed values of each dependent variable can be written as the sum of two parts: a fixed component $x'\beta$, which is a linear function of the independent coefficients, and a random noise, or error, component ϵ:

$$y \quad = \quad x'\beta + \epsilon$$

The independent coefficients x are constructed from the model effects as described in the "Parameterization of PROC GLM Models" section on page 1521. Further, the errors for different observations are assumed to be uncorrelated with identical variances. Thus, this model can be written

$$E(Y) \quad = \quad X\beta, \qquad \text{Var}(Y) \quad = \quad \sigma^2 I$$

where Y is the vector of dependent variable values, X is the matrix of independent coefficients, I is the identity matrix, and σ^2 is the common variance for the errors. For multiple dependent variables, the model is similar except that the errors for different dependent variables within the same observation are not assumed to be uncorrelated. This yields a multivariate linear model of the form

$$E(Y) \quad = \quad XB, \qquad \text{Var}(\text{vec}(Y)) \quad = \quad \Sigma \otimes I$$

where Y and B are now matrices, with one column for each dependent variable, vec(Y) strings Y out by rows, and $\otimes$ indicates the Kronecker matrix product.

Under the assumptions thus far discussed, the least-squares approach provides estimates of the linear parameters that are unbiased and have minimum variance among linear estimators. Under the further assumption that the errors have a normal (or Gaussian) distribution, the least-squares estimates are the maximum likelihood estimates and their distribution is known. All of the significance levels ("p values") and confidence limits calculated by the GLM procedure require this assumption of normality in order to be exactly valid, although they are good approximations in many other cases.

Specification of Effects

Each term in a model, called an *effect*, is a variable or combination of variables. Effects are specified with a special notation using variable names and operators. There are two kinds of variables: *classification* (or *class*) *variables* and *continuous variables*. There are two primary operators: *crossing* and *nesting*. A third operator, the *bar operator*, is used to simplify effect specification.

In an analysis-of-variance model, independent variables must be variables that identify classification levels. In the SAS System, these are called *class variables* and are declared in the CLASS statement. (They can also be called *categorical*, *qualitative*, *discrete*, or *nominal variables*.) Class variables can be either *numeric* or *character*. The values of a class variable are called *levels*. For example, the class variable Sex has the levels “male” and “female.”

In a model, an independent variable that is not declared in the CLASS statement is assumed to be continuous. Continuous variables, which must be numeric, are used for response variables and covariates. For example, the heights and weights of subjects are continuous variables.

Types of Effects

There are seven different types of effects used in the GLM procedure. In the following list, assume that A, B, C, D, and E are class variables and that X1, X2, and Y are continuous variables:

- Regressor effects are specified by writing continuous variables by themselves: X1 X2.
- Polynomial effects are specified by joining two or more continuous variables with asterisks: X1*X1 X1*X2.
- Main effects are specified by writing class variables by themselves: A B C.
- Crossed effects (interactions) are specified by joining class variables with asterisks: A*B B*C A*B*C.
- Nested effects are specified by following a main effect or crossed effect with a class variable or list of class variables enclosed in parentheses. The main effect or crossed effect is nested within the effects listed in parentheses:

  B(A) C(B*A) D*E(C*B*A) .

  In this example, B(A) is read “B nested within A.”
- Continuous-by-class effects are written by joining continuous variables and class variables with asterisks: X1*A.
- Continuous-nesting-class effects consist of continuous variables followed by a class variable interaction enclosed in parentheses: X1(A) X1*X2(A*B).

One example of the general form of an effect involving several variables is

X1*X2*A*B*C(D*E)

This example contains crossed continuous terms by crossed classification terms nested within more than one class variable. The continuous list comes first, followed by the crossed list, followed by the nesting list in parentheses. Note that asterisks can appear within the nested list but not immediately before the left parenthesis. For details on how the design matrix and parameters are defined with respect to the effects specified in this section, see the section "Parameterization of PROC GLM Models" on page 1521.

The MODEL statement and several other statements use these effects. Some examples of MODEL statements using various kinds of effects are shown in the following table; a, b, and c represent class variables, and y, y1, y2, x, and z represent continuous variables.

Specification	Kind of Model
model y=x;	simple regression
model y=x z;	multiple regression
model y=x x*x;	polynomial regression
model y1 y2=x z;	multivariate regression
model y=a;	one-way ANOVA
model y=a b c;	main effects model
model y=a b a*b;	factorial model (with interaction)
model y=a b(a) c(b a);	nested model
model y1 y2=a b;	multivariate analysis of variance (MANOVA)
model y=a x;	analysis-of-covariance model
model y=a x(a);	separate-slopes model
model y=a x x*a;	homogeneity-of-slopes model

The Bar Operator

You can shorten the specification of a large factorial model using the bar operator. For example, two ways of writing the model for a full three-way factorial model are

```
proc glm;                 and          proc glm;
   class A B C;                           class A B C;
   model Y=A B C A*B                      model Y=A|B|C;
         A*C B*C A*B*C;                run;
run;
```

When the bar (|) is used, the right- and left-hand sides become effects, and the cross of them becomes an effect. Multiple bars are permitted. The expressions are expanded from left to right, using rules 2–4 given in Searle (1971, p. 390).

- Multiple bars are evaluated left to right. For instance, A|B|C is evaluated as follows.

$$\begin{aligned} \text{A | B | C} &\rightarrow \{ \text{A | B} \} \text{ | C} \\ &\rightarrow \{ \text{A B A*B} \} \text{ | C} \\ &\rightarrow \text{A B A*B A*C B*C A*B*C} \end{aligned}$$

- Crossed and nested groups of variables are combined. For example, A(B) | C(D) generates A*C(B D), among other terms.
- Duplicate variables are removed. For example, A(C) | B(C) generates A*B(C C), among other terms, and the extra C is removed.
- Effects are discarded if a variable occurs on both the crossed and nested parts of an effect. For instance, A(B) | B(D E) generates A*B(B D E), but this effect is eliminated immediately.

You can also specify the maximum number of variables involved in any effect that results from bar evaluation by specifying that maximum number, preceded by an @ sign, at the end of the bar effect. For example, the specification A | B | C@2 would result in only those effects that contain 2 or fewer variables: in this case, A B A*B C A*C and B*C.

The following table gives more examples of using the bar and at operators.

A | C(B)	is equivalent to	A C(B) A*C(B)
A(B) | C(B)	is equivalent to	A(B) C(B) A*C(B)
A(B) | B(D E)	is equivalent to	A(B) B(D E)
A | B(A) | C	is equivalent to	A B(A) C A*C B*C(A)
A | B(A) | C@2	is equivalent to	A B(A) C A*C
A | B | C | D@2	is equivalent to	A B A*B C A*C B*C D A*D B*D C*D
A*B(C*D)	is equivalent to	A*B(C D)

Using PROC GLM Interactively

You can use the GLM procedure interactively. After you specify a model with a MODEL statement and run PROC GLM with a RUN statement, you can execute a variety of statements without reinvoking PROC GLM.

The “Syntax” section (page 1477) describes which statements can be used interactively. These interactive statements can be executed singly or in groups by following the single statement or group of statements with a RUN statement. Note that the MODEL statement cannot be repeated; PROC GLM allows only one MODEL statement.

If you use PROC GLM interactively, you can end the GLM procedure with a DATA step, another PROC step, an ENDSAS statement, or a QUIT statement.

When you are using PROC GLM interactively, additional RUN statements do not end the procedure but tell PROC GLM to execute additional statements.

When you specify a WHERE statement with PROC GLM, it should appear before the first RUN statement. The WHERE statement enables you to select only certain observations for analysis without using a subsetting DATA step. For example, the statement `where group ne 5` omits observations with GROUP=5 from the analysis. Refer to *SAS Language Reference: Dictionary* for details on this statement.

When you specify a BY statement with PROC GLM, interactive processing is not possible; that is, once the first RUN statement is encountered, processing proceeds for each BY group in the data set, and no further statements are accepted by the procedure.

Interactivity is also disabled when there are different patterns of missing values among the dependent variables. For details, see the "Missing Values" section on page 1571.

Parameterization of PROC GLM Models

The GLM procedure constructs a linear model according to the specifications in the MODEL statement. Each effect generates one or more columns in a design matrix $\mathbf{X}$. This section shows precisely how $\mathbf{X}$ is built.

Intercept

All models include a column of 1s by default to estimate an intercept parameter μ. You can use the NOINT option to suppress the intercept.

Regression Effects

Regression effects (covariates) have the values of the variables copied into the design matrix directly. Polynomial terms are multiplied out and then installed in $\mathbf{X}$.

Main Effects

If a class variable has m levels, PROC GLM generates m columns in the design matrix for its main effect. Each column is an indicator variable for one of the levels of the class variable. The default order of the columns is the sort order of the values of their levels; this order can be controlled with the ORDER= option in the PROC GLM statement, as shown in the following table.

Data		Design Matrix					
			A		B		
A	B	μ	A1	A2	B1	B2	B3
1	1	1	1	0	1	0	0
1	2	1	1	0	0	1	0
1	3	1	1	0	0	0	1
2	1	1	0	1	1	0	0
2	2	1	0	1	0	1	0
2	3	1	0	1	0	0	1

There are more columns for these effects than there are degrees of freedom for them; in other words, PROC GLM is using an over-parameterized model.

Crossed Effects

First, PROC GLM reorders the terms to correspond to the order of the variables in the CLASS statement; thus, B*A becomes A*B if A precedes B in the CLASS statement. Then, PROC GLM generates columns for all combinations of levels that occur in the data. The order of the columns is such that the rightmost variables in the cross index faster than the leftmost variables. No columns are generated corresponding to combinations of levels that do not occur in the data.

Data		Design Matrix													
			A		B			A*B							
A	B	μ	A1	A2	B1	B2	B3	A1B1	A1B2	A1B3	A2B1	A2B2	A2B3
1	1	1	1	0	1	0	0	1	0	0	0	0	0
1	2	1	1	0	0	1	0	0	1	0	0	0	0
1	3	1	1	0	0	0	1	0	0	1	0	0	0
2	1	1	0	1	1	0	0	0	0	0	1	0	0
2	2	1	0	1	0	1	0	0	0	0	0	1	0
2	3	1	0	1	0	0	1	0	0	0	0	0	1

In this matrix, main-effects columns are not linearly independent of crossed-effect columns; in fact, the column space for the crossed effects contains the space of the main effect.

Nested Effects

Nested effects are generated in the same manner as crossed effects. Hence, the design columns generated by the following statements are the same (but the ordering of the columns is different):

`model y=a b(a);` (B nested within A)

`model y=a a*b;` (omitted main effect for B)

The nesting operator in PROC GLM is more a notational convenience than an operation distinct from crossing. Nested effects are characterized by the property that the nested variables never appear as main effects. The order of the variables within nesting parentheses is made to correspond to the order of these variables in the CLASS statement. The order of the columns is such that variables outside the parentheses index faster than those inside the parentheses, and the rightmost nested variables index faster than the leftmost variables.

Data		Design Matrix								
			A		B(A)					
A	B	μ	A1	A2	B1A1	B2A1	B3A1	B1A2	B2A2	B3A2
1	1	1	1	0	1	0	0	0	0	0
1	2	1	1	0	0	1	0	0	0	0
1	3	1	1	0	0	0	1	0	0	0
2	1	1	0	1	0	0	0	1	0	0
2	2	1	0	1	0	0	0	0	1	0
2	3	1	0	1	0	0	0	0	0	1

Continuous-Nesting-Class Effects

When a continuous variable nests with a class variable, the design columns are constructed by multiplying the continuous values into the design columns for the class effect.

Data		Design Matrix				
			A		X(A)	
X	A	μ	A1	A2	X(A1)	X(A2)
21	1	1	1	0	21	0
24	1	1	1	0	24	0
22	1	1	1	0	22	0
28	2	1	0	1	0	28
19	2	1	0	1	0	19
23	2	1	0	1	0	23

This model estimates a separate slope for X within each level of A.

Continuous-by-Class Effects

Continuous-by-class effects generate the same design columns as continuous-nesting-class effects. The two models differ by the presence of the continuous variable as a regressor by itself, in addition to being a contributor to X*A.

Data		Design Matrix					
				A		X*A	
X	A	μ	X	A1	A2	X*A1	X*A2
21	1	1	21	1	0	21	0
24	1	1	24	1	0	24	0
22	1	1	22	1	0	22	0
28	2	1	28	0	1	0	28
19	2	1	19	0	1	0	19
23	2	1	23	0	1	0	23

Continuous-by-class effects are used to test the homogeneity of slopes. If the continuous-by-class effect is nonsignificant, the effect can be removed so that the response with respect to X is the same for all levels of the class variables.

General Effects

An example that combines all the effects is

X1*X2*A*B*C(D E)

The continuous list comes first, followed by the crossed list, followed by the nested list in parentheses.

The sequencing of parameters is important to learn if you use the CONTRAST or ESTIMATE statement to compute or test some linear function of the parameter estimates.

Effects may be retitled by PROC GLM to correspond to ordering rules. For example, B*A(E D) may be retitled A*B(D E) to satisfy the following:

- Class variables that occur outside parentheses (crossed effects) are sorted in the order in which they appear in the CLASS statement.
- Variables within parentheses (nested effects) are sorted in the order in which they appear in a CLASS statement.

The sequencing of the parameters generated by an effect can be described by which variables have their levels indexed faster:

- Variables in the crossed part index faster than variables in the nested list.
- Within a crossed or nested list, variables to the right index faster than variables to the left.

For example, suppose a model includes four effects—A, B, C, and D—each having two levels, 1 and 2. If the CLASS statement is

```
class A B C D;
```

then the order of the parameters for the effect B*A(C D), which is retitled A*B(C D), is as follows.

$A_1B_1C_1D_1$
$A_1B_2C_1D_1$
$A_2B_1C_1D_1$
$A_2B_2C_1D_1$
$A_1B_1C_1D_2$
$A_1B_2C_1D_2$
$A_2B_1C_1D_2$
$A_2B_2C_1D_2$
$A_1B_1C_2D_1$
$A_1B_2C_2D_1$
$A_2B_1C_2D_1$
$A_2B_2C_2D_1$

$A_1B_1C_2D_2$
$A_1B_2C_2D_2$
$A_2B_1C_2D_2$
$A_2B_2C_2D_2$

Note that first the crossed effects B and A are sorted in the order in which they appear in the CLASS statement so that A precedes B in the parameter list. Then, for each combination of the nested effects in turn, combinations of A and B appear. The B effect changes fastest because it is rightmost in the (renamed) cross list. Then A changes next fastest. The D effect changes next fastest, and C is the slowest since it is leftmost in the nested list.

When numeric class variables are used, their levels are sorted by their character format, which may not correspond to their numeric sort sequence. Therefore, it is advisable to include a format for numeric class variables or to use the ORDER=INTERNAL option in the PROC GLM statement to ensure that levels are sorted by their internal values.

Degrees of Freedom

For models with classification (categorical) effects, there are more design columns constructed than there are degrees of freedom for the effect. Thus, there are linear dependencies among the columns. In this event, the parameters are not jointly estimable; there is an infinite number of least-squares solutions. The GLM procedure uses a generalized (g2) inverse to obtain values for the estimates; see the “Computational Method” section on page 1574 for more details. The solution values are not produced unless the SOLUTION option is specified in the MODEL statement. The solution has the characteristic that estimates are zero whenever the design column for that parameter is a linear combination of previous columns. (Strictly termed, the solution values should not be called estimates, since the parameters may not be formally estimable.) With this full parameterization, hypothesis tests are constructed to test linear functions of the parameters that are estimable.

Other procedures (such as the CATMOD procedure) reparameterize models to full rank using certain restrictions on the parameters. PROC GLM does not reparameterize, making the hypotheses that are commonly tested more understandable. See Goodnight (1978) for additional reasons for not reparameterizing.

PROC GLM does not actually construct the entire design matrix $\mathbf{X}$; rather, a row x_i of $\mathbf{X}$ is constructed for each observation in the data set and used to accumulate the crossproduct matrix $\mathbf{X}'\mathbf{X} = \sum_i x_i' x_i$.

Hypothesis Testing in PROC GLM

See Chapter 12, “The Four Types of Estimable Functions,” for a complete discussion of the four standard types of hypothesis tests.

Example

To illustrate the four types of tests and the principles upon which they are based, consider a two-way design with interaction based on the following data:

		B	
		1	2
	1	23.5 23.7	28.7
A	2	8.9	5.6 8.9
	3	10.3 12.5	13.6 14.6

Invoke PROC GLM and specify all the estimable functions options to examine what the GLM procedure can test. The following statements are followed by the summary ANOVA table. See Figure 30.8.

```
data example;
   input a b y @@;
   datalines;
1 1 23.5  1 1 23.7  1 2 28.7  2 1  8.9  2 2  5.6
2 2  8.9  3 1 10.3  3 1 12.5  3 2 13.6  3 2 14.6
;

proc glm;
   class a b;
   model y=a b a*b / e e1 e2 e3 e4;
run;
```

```
                               The GLM Procedure

Dependent Variable: y

                                      Sum of
 Source                     DF       Squares    Mean Square   F Value   Pr > F

 Model                       5    520.4760000    104.0952000     49.66   0.0011

 Error                       4      8.3850000      2.0962500

 Corrected Total             9    528.8610000


               R-Square     Coeff Var      Root MSE        y Mean

               0.984145      9.633022      1.447843      15.03000
```

Figure 30.8. Summary ANOVA Table from PROC GLM

The following sections show the general form of estimable functions and discuss the four standard tests, their properties, and abbreviated output for the two-way crossed example.

Estimability

Figure 30.9 is the general form of estimable functions for the example. In order to be testable, a hypothesis must be able to fit within the framework displayed here.

```
                               The GLM Procedure

                      General Form of Estimable Functions

                      Effect            Coefficients

                      Intercept         L1

                      a         1       L2
                      a         2       L3
                      a         3       L1-L2-L3

                      b         1       L5
                      b         2       L1-L5

                      a*b       1 1     L7
                      a*b       1 2     L2-L7
                      a*b       2 1     L9
                      a*b       2 2     L3-L9
                      a*b       3 1     L5-L7-L9
                      a*b       3 2     L1-L2-L3-L5+L7+L9
```

Figure 30.9. General Form of Estimable Functions

If a hypothesis is estimable, the Ls in the preceding scheme can be set to values that match the hypothesis. All the standard tests in PROC GLM can be shown in the preceding format, with some of the Ls zeroed and some set to functions of other Ls.

The following sections show how many of the hypotheses can be tested by comparing the model sum-of-squares regression from one model to a submodel. The notation used is

$$\mathrm{SS}(B\ \mathit{effects}|A\ \mathit{effects}) = \mathrm{SS}(B\ \mathit{effects}, A\ \mathit{effects}) - \mathrm{SS}(A\ \mathit{effects})$$

where SS(*A effects*) denotes the regression model sum of squares for the model consisting of *A effects*. This notation is equivalent to the reduction notation defined by Searle (1971) and summarized in Chapter 12, “The Four Types of Estimable Functions.”

Type I Tests

Type I sums of squares (SS), also called *sequential sums of squares*, are the incremental improvement in error sums of squares as each effect is added to the model. They can be computed by fitting the model in steps and recording the difference in error sum of squares at each step.

Source	Type I SS
A	$\mathrm{SS}(A \mid \mu)$
B	$\mathrm{SS}(B \mid \mu, A)$
$A * B$	$\mathrm{SS}(A * B\ \mid \mu, A, B)$

Type I sums of squares are displayed by default because they are easy to obtain and can be used in various hand calculations to produce sum of squares values for a series of different models. Nelder (1994) and others have argued that Type I and II sums are essentially the only appropriate ones for testing ANOVA effects; however, refer also to the discussion of Nelder’s article, especially Rodriguez, Tobias, and Wolfinger (1995) and Searle (1995).

The Type I hypotheses have these properties:

- Type I sum of squares for all effects add up to the model sum of squares. None of the other sum of squares types have this property, except in special cases.
- Type I hypotheses can be derived from rows of the Forward-Dolittle transformation of $\mathbf{X}'\mathbf{X}$ (a transformation that reduces $\mathbf{X}'\mathbf{X}$ to an upper triangular matrix by row operations).
- Type I sum of squares are statistically independent of each other under the usual assumption that the true residual errors are independent and identically normally distributed (see page 1517).
- Type I hypotheses depend on the order in which effects are specified in the MODEL statement.

- Type I hypotheses are uncontaminated by parameters corresponding to effects that precede the effect being tested; however, the hypotheses usually involve parameters for effects following the tested effect in the model. For example, in the model

  ```
  Y=A B;
  ```

  the Type I hypothesis for B does not involve A parameters, but the Type I hypothesis for A does involve B parameters.
- Type I hypotheses are functions of the cell counts for unbalanced data; the hypotheses are not usually the same hypotheses that are tested if the data are balanced.
- Type I sums of squares are useful for polynomial models where you want to know the contribution of a term as though it had been made orthogonal to preceding effects. Thus, in polynomial models, Type I sums of squares correspond to tests of the orthogonal polynomial effects.

The Type I estimable functions and associated tests for the example are shown in Figure 30.10. (This combines tables from several pages of output.)

```
                               The GLM Procedure

                           Type I Estimable Functions

                           ----------------Coefficients----------------
         Effect            a                        b             a*b

         Intercept         0                        0             0

         a         1       L2                       0             0
         a         2       L3                       0             0
         a         3       -L2-L3                   0             0

         b         1       0.1667*L2-0.1667*L3      L5            0
         b         2       -0.1667*L2+0.1667*L3     -L5           0

         a*b       1 1     0.6667*L2                0.2857*L5     L7
         a*b       1 2     0.3333*L2                -0.2857*L5    -L7
         a*b       2 1     0.3333*L3                0.2857*L5     L9
         a*b       2 2     0.6667*L3                -0.2857*L5    -L9
         a*b       3 1     -0.5*L2-0.5*L3           0.4286*L5     -L7-L9
         a*b       3 2     -0.5*L2-0.5*L3           -0.4286*L5    L7+L9
```

```
                               The GLM Procedure

Dependent Variable: y

 Source                      DF       Type I SS     Mean Square    F Value    Pr > F

 a                            2     494.0310000     247.0155000     117.84    0.0003
 b                            1      10.7142857      10.7142857       5.11    0.0866
 a*b                          2      15.7307143       7.8653571       3.75    0.1209
```

Figure 30.10. Type I Estimable Functions and Associated Tests

Type II Tests

The Type II tests can also be calculated by comparing the error sums of squares (SS) for subset models. The Type II SS are the reduction in error SS due to adding the term after all other terms have been added to the model except terms that contain the effect being tested. An effect is contained in another effect if it can be derived by deleting variables from the latter effect. For example, A and B are both contained in A*B. For this model

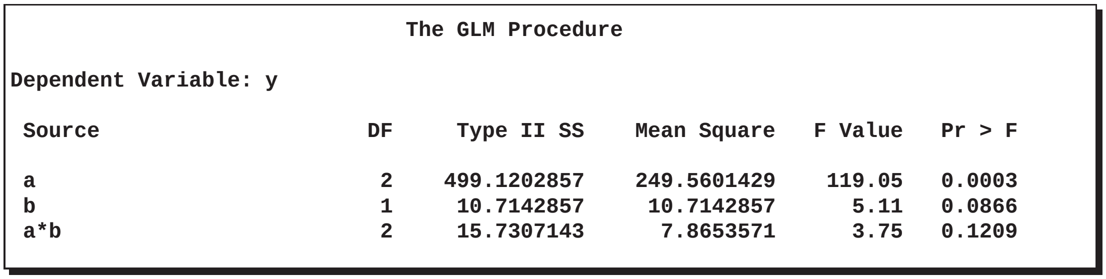
Source	Type II SS
A	SS$(A \mid \mu, B)$
B	SS$(B \mid \mu, A)$
$A * B$	SS$(A * B \mid \mu, A, B)$

Type II SS have these properties:

- Type II SS do not necessarily sum to the model SS.
- The hypothesis for an effect does not involve parameters of other effects except for containing effects (which it must involve to be estimable).
- Type II SS are invariant to the ordering of effects in the model.
- For unbalanced designs, Type II hypotheses for effects that are contained in other effects are not usually the same hypotheses that are tested if the data are balanced. The hypotheses are generally functions of the cell counts.

The Type II estimable functions and associated tests for the example are shown in Figure 30.11. (Again, this combines tables from several pages of output.)

```
                               The GLM Procedure

                          Type II Estimable Functions

                           ----------------Coefficients----------------
         Effect            a                        b             a*b

         Intercept         0                        0             0

         a         1       L2                       0             0
         a         2       L3                       0             0
         a         3       -L2-L3                   0             0

         b         1       0                        L5            0
         b         2       0                        -L5           0

         a*b       1 1     0.619*L2+0.0476*L3       0.2857*L5     L7
         a*b       1 2     0.381*L2-0.0476*L3       -0.2857*L5    -L7
         a*b       2 1     -0.0476*L2+0.381*L3      0.2857*L5     L9
         a*b       2 2     0.0476*L2+0.619*L3       -0.2857*L5    -L9
         a*b       3 1     -0.5714*L2-0.4286*L3     0.4286*L5     -L7-L9
         a*b       3 2     -0.4286*L2-0.5714*L3     -0.4286*L5    L7+L9
```



```
                               The GLM Procedure

Dependent Variable: y

 Source                      DF     Type II SS    Mean Square   F Value   Pr > F

 a                            2    499.1202857    249.5601429    119.05   0.0003
 b                            1     10.7142857     10.7142857      5.11   0.0866
 a*b                          2     15.7307143      7.8653571      3.75   0.1209
```

Figure 30.11. Type II Estimable Functions and Associated Tests

Type III and Type IV Tests

Type III and Type IV sums of squares (SS), sometimes referred to as *partial sums of squares*, are considered by many to be the most desirable; see Searle (1987, Section 4.6). These SS cannot, in general, be computed by comparing model SS from several models using PROC GLM's parameterization. However, they can sometimes be computed by reduction for methods that reparameterize to full rank, when such a reparameterization effectively imposes Type III linear constraints on the parameters. In PROC GLM, they are computed by constructing a hypothesis matrix $\mathbf{L}$ and then computing the SS associated with the hypothesis $\mathbf{L}\beta = 0$. As long as there are no missing cells in the design, Type III and Type IV SS are the same.

These are properties of Type III and Type IV SS:

- The hypothesis for an effect does not involve parameters of other effects except for containing effects (which it must involve to be estimable).
- The hypotheses to be tested are invariant to the ordering of effects in the model.
- The hypotheses are the same hypotheses that are tested if there are no missing cells. They are not functions of cell counts.
- The SS do not generally add up to the model SS and, in some cases, can exceed the model SS.

The SS are constructed from the general form of estimable functions. Type III and Type IV tests are different only if the design has missing cells. In this case, the Type III tests have an orthogonality property, while the Type IV tests have a balancing property. These properties are discussed in Chapter 12, "The Four Types of Estimable Functions." For this example, since the data contains observations for all pairs of levels of A and B, Type IV tests are identical to the Type III tests that are shown in Figure 30.12. (This combines tables from several pages of output.)

```
                               The GLM Procedure

                          Type III Estimable Functions

                              -------------Coefficients------------
            Effect            a                 b             a*b

            Intercept         0                 0             0

            a         1       L2                0             0
            a         2       L3                0             0
            a         3       -L2-L3            0             0

            b         1       0                 L5            0
            b         2       0                 -L5           0

            a*b       1 1     0.5*L2            0.3333*L5     L7
            a*b       1 2     0.5*L2            -0.3333*L5    -L7
            a*b       2 1     0.5*L3            0.3333*L5     L9
            a*b       2 2     0.5*L3            -0.3333*L5    -L9
            a*b       3 1     -0.5*L2-0.5*L3    0.3333*L5     -L7-L9
            a*b       3 2     -0.5*L2-0.5*L3    -0.3333*L5    L7+L9
```

```
                               The GLM Procedure

Dependent Variable: y

 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 a                            2    479.1078571    239.5539286    114.28   0.0003
 b                            1      9.4556250      9.4556250      4.51   0.1009
 a*b                          2     15.7307143      7.8653571      3.75   0.1209
```

Figure 30.12. Type III Estimable Functions and Associated Tests

Absorption

Absorption is a computational technique used to reduce computing resource needs in certain cases. The classic use of absorption occurs when a blocking factor with a large number of levels is a term in the model.

For example, the statements

```
proc glm;
   absorb herd;
   class a b;
   model y=a b a*b;
run;
```

are equivalent to

```
proc glm;
   class herd a b;
   model y=herd a b a*b;
run;
```

The exception to the previous statements is that the Type II, Type III, or Type IV SS for HERD are not computed when HERD is absorbed.

The algorithm for absorbing variables is similar to the one used by the NESTED procedure for computing a nested analysis of variance. As each new row of $[X|Y]$ (corresponding to the nonabsorbed independent effects and the dependent variables) is constructed, it is adjusted for the absorbed effects in a Type I fashion. The efficiency of the absorption technique is due to the fact that this adjustment can be done in one pass of the data and without solving any linear equations, assuming that the data have been sorted by the absorbed variables.

Several effects can be absorbed at one time. For example, these statements

```
proc glm;
   absorb herd cow;
   class a b;
   model y=a b a*b;
run;
```

are equivalent to

```
proc glm;
   class herd cow a b;
   model y=herd cow(herd) a b a*b;
run;
```

When you use absorption, the size of the $\mathbf{X}'\mathbf{X}$ matrix is a function only of the effects in the MODEL statement. The effects being absorbed do not contribute to the size of the $\mathbf{X}'\mathbf{X}$ matrix.

For the preceding example, a and b can be absorbed:

```
proc glm;
   absorb a b;
   class herd cow;
   model y=herd cow(herd);
run;
```

Although the sources of variation in the results are listed as

```
a b(a) herd cow(herd)
```

all types of estimable functions for herd and cow(herd) are free of a, b, and a*b parameters.

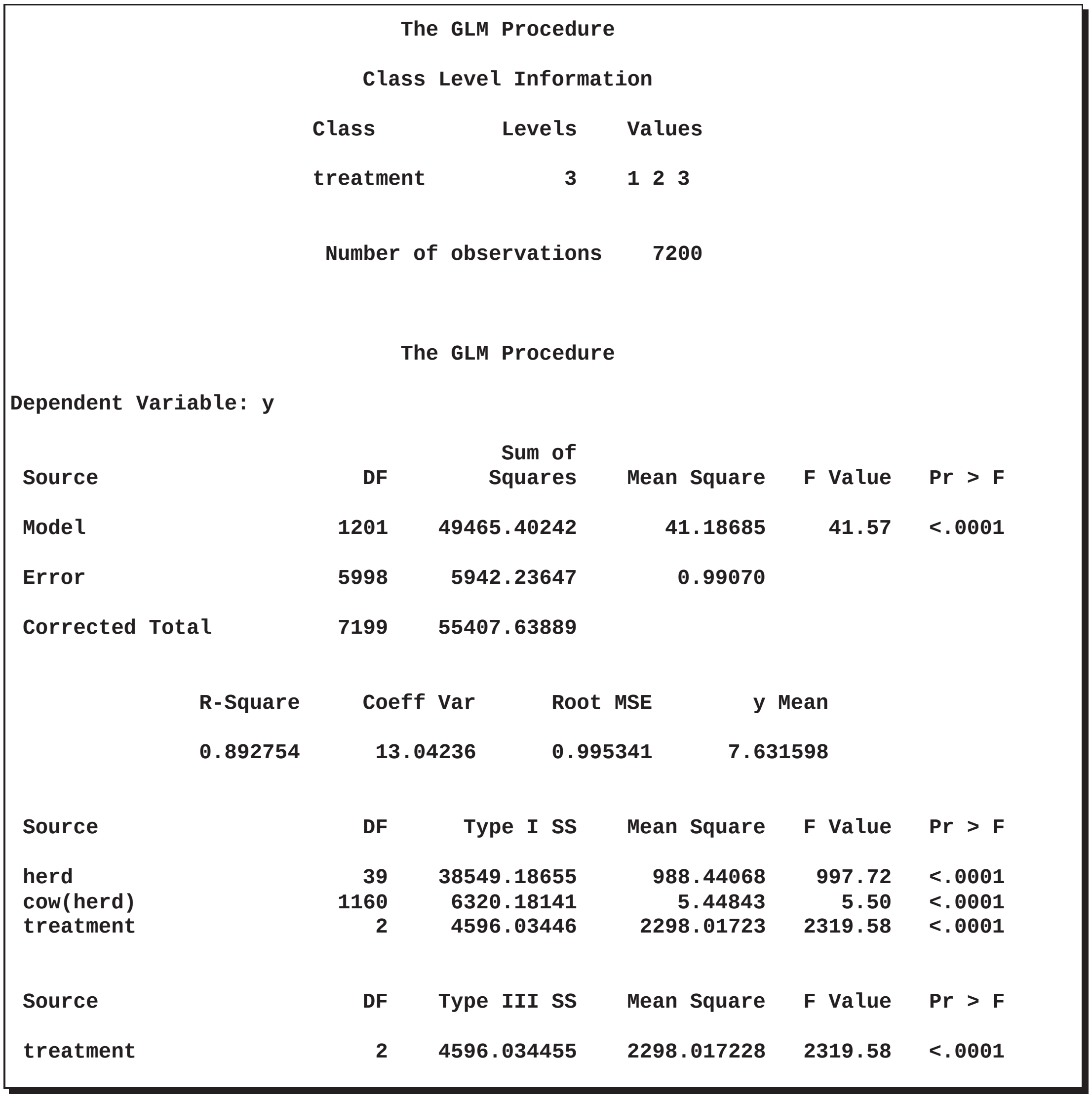
To illustrate the savings in computing using the ABSORB statement, PROC GLM is run on generated data with 1147 degrees of freedom in the model with the following statements:

```
data a;
   do herd=1 to 40;
      do cow=1 to 30;
         do treatment=1 to 3;
            do rep=1 to 2;
               y = herd/5 + cow/10 + treatment + rannor(1);
               output;
            end;
         end;
      end;
   end;

proc glm;
   class herd cow treatment;
   model y=herd cow(herd) treatment;
run;
```

This analysis would have required over 6 megabytes of memory for the $\mathbf{X}'\mathbf{X}$ matrix had PROC GLM solved it directly. However, in the following statements, the GLM procedure needs only a 4×4 matrix for the intercept and treatment because the other effects are absorbed.

```
proc glm;
   absorb herd cow;
   class treatment;
   model y = treatment;
run;
```

These statements produce the results shown in Figure 30.13.



```
                               The GLM Procedure

                            Class Level Information

                        Class          Levels    Values

                        treatment           3    1 2 3


                         Number of observations    7200



                               The GLM Procedure

Dependent Variable: y

                                            Sum of
 Source                      DF          Squares    Mean Square   F Value   Pr > F

 Model                     1201    49465.40242       41.18685     41.57   <.0001

 Error                     5998     5942.23647        0.99070

 Corrected Total           7199    55407.63889


               R-Square     Coeff Var      Root MSE        y Mean

               0.892754      13.04236      0.995341      7.631598


 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 herd                        39    38549.18655      988.44068    997.72   <.0001
 cow(herd)                 1160     6320.18141        5.44843      5.50   <.0001
 treatment                    2     4596.03446     2298.01723   2319.58   <.0001


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 treatment                    2    4596.034455    2298.017228   2319.58   <.0001
```

Figure 30.13. Absorption of Effects

Specification of ESTIMATE Expressions

Consider the model

$$E(Y) \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 .$$

The corresponding MODEL statement for PROC GLM is

```
model y=x1 x2 x3;
```

To estimate the difference between the parameters for x_1 and x_2,

$$\beta_1 - \beta_2 \quad = \quad (\begin{array}{cccc} 0 & 1 & -1 & 0 \end{array})\boldsymbol{\beta}, \text{ where } \boldsymbol{\beta} = (\begin{array}{cccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 \end{array})'$$

you can use the following ESTIMATE statement:

```
estimate 'B1-B2'  x1 1  x2 -1;
```

To predict y at $x_1 = 1$, $x_2 = 0$, and $x_3 = -2$, you can estimate

$$\beta_0 + \beta_1 - 2\beta_3 \quad = \quad (\begin{array}{cccc} 1 & 1 & 0 & -2 \end{array})\boldsymbol{\beta}$$

with the following ESTIMATE statement:

```
estimate 'B0+B1-2B3' intercept 1 x1 1 x3 -2;
```

Now consider models involving class variables such as

```
model y=A B A*B;
```

with the associated parameters:

$$(\begin{array}{cccccccccccc} \mu & \alpha_1 & \alpha_2 & \alpha_3 & \beta_1 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{21} & \gamma_{22} & \gamma_{31} & \gamma_{32} \end{array})$$

The LS-mean for the first level of A is $\mathbf{L}\beta$, where

$$\mathbf{L} = (\begin{array}{ccccccccccccccc} 1 & | & 1 & 0 & 0 & | & 0.5 & 0.5 & | & 0.5 & 0.5 & 0 & 0 & 0 & 0 \end{array})$$

You can estimate this with the following ESTIMATE statement:

```
estimate 'LS-mean(A1)' intercept 1 A 1 B 0.5 0.5 A*B 0.5 0.5;
```

Note in this statement that only one element of **L** is specified following the A effect, even though A has three levels. Whenever the list of constants following an effect name is shorter than the effect's number of levels, zeros are used as the remaining constants. (If the list of constants is longer than the number of levels for the effect, the extra constants are ignored, and a warning message is displayed.)

To estimate the A linear effect in the preceding model, assuming equally spaced levels for A, you can use the following **L**:

$$\mathbf{L} = (\ 0 \ \mid \ -1 \quad 0 \quad 1 \ \mid \ 0 \quad 0 \ \mid \ -0.5 \quad -0.5 \quad 0 \quad 0 \quad 0.5 \quad 0.5\ )$$

The ESTIMATE statement for this **L** is written as

```
estimate 'A Linear' A -1 0 1;
```

If you do not specify the elements of **L** for an effect that contains a specified effect, then the elements of the specified effect are equally distributed over the corresponding levels of the higher-order effect. In addition, if you specify the intercept in an ESTIMATE or CONTRAST statement, it is distributed over all classification effects that are not contained by any other specified effect. The distribution of lower-order coefficients to higher-order effect coefficients follows the same general rules as in the LSMEANS statement, and it is similar to that used to construct Type IV tests. In the previous example, the -1 associated with α_1 is divided by the number n_{1j} of γ_{1j} parameters; then each γ_{1j} coefficient is set to $-1/n_{1j}$. The 1 associated with α_3 is distributed among the γ_{3j} parameters in a similar fashion. In the event that an unspecified effect contains several specified effects, only that specified effect with the most factors in common with the unspecified effect is used for distribution of coefficients to the higher-order effect.

Numerous syntactical expressions for the ESTIMATE statement were considered, including many that involved specifying the effect and level information associated with each coefficient. For models involving higher-level effects, the requirement of specifying level information can lead to very bulky specifications. Consequently, the simpler form of the ESTIMATE statement described earlier was implemented. The syntax of this ESTIMATE statement puts a burden on you to know a priori the order of the parameter list associated with each effect. You can use the ORDER= option in the PROC GLM statement to ensure that the levels of the classification effects are sorted appropriately.

Note: If you use the ESTIMATE statement with unspecified effects, use the E option to make sure that the actual **L** constructed by the preceding rules is the one you intended.

A Check for Estimability

Each $\mathbf{L}$ is checked for estimability using the relationship: $\mathbf{L} = \mathbf{LH}$ where $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$. The $\mathbf{L}$ vector is declared nonestimable, if for any i

$$\text{ABS}(\mathbf{L}_i - (\mathbf{LH})_i > \begin{cases} \epsilon & \text{if } \mathbf{L}_i = 0 \text{ or} \\ \epsilon \times \text{ABS}(\mathbf{L}_i) & \text{otherwise} \end{cases}$$

where $\epsilon = 10^{-4}$ by default; you can change this with the SINGULAR= option. Continued fractions (like 1/3) should be specified to at least six decimal places, or the DIVISOR parameter should be used.

Comparing Groups

An important task in analyzing data with classification effects is to estimate the typical response for each level of a given effect; often, you also want to compare these estimates to determine which levels are equivalent in terms of the response. You can perform this task in two ways with the GLM procedure: with direct, arithmetic group means; and with so-called *least-squares means* (LS-means).

Means Versus LS-Means

Computing and comparing arithmetic means—either simple or weighted within-group averages of the input data—is a familiar and well-studied statistical process. This is the right approach to summarizing and comparing groups for one-way and balanced designs. However, in unbalanced designs with more than one effect, the arithmetic mean for a group may not accurately reflect the “typical” response for that group, since it does not take other effects into account.

For example, consider the following analysis of an unbalanced two-way design:

```
data twoway;
   input Treatment Block y @@;
   datalines;
1 1 17   1 1 28   1 1 19   1 1 21   1 1 19
1 2 43   1 2 30   1 2 39   1 2 44   1 2 44
1 3 16
2 1 21   2 1 21   2 1 24   2 1 25
2 2 39   2 2 45   2 2 42   2 2 47
2 3 19   2 3 22   2 3 16
3 1 22   3 1 30   3 1 33   3 1 31
3 2 46
3 3 26   3 3 31   3 3 26   3 3 33   3 3 29   3 3 25
;

title1 "Unbalanced Two-way Design";
ods select ModelANOVA Means LSMeans;
proc glm data=twoway;
   class Treatment Block;
   model y = Treatment|Block;
```

```
      means Treatment;
      lsmeans Treatment;
   run;
   ods select all;
```

The ANOVA results are shown in Figure 30.14.

```
                           Unbalanced Two-way Design

                               The GLM Procedure

Dependent Variable: y

 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 Treatment                    2       8.060606       4.030303      0.24   0.7888
 Block                        2    2621.864124    1310.932062     77.95   <.0001
 Treatment*Block              4      32.684361       8.171090      0.49   0.7460


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 Treatment                    2     266.130682     133.065341      7.91   0.0023
 Block                        2    1883.729465     941.864732     56.00   <.0001
 Treatment*Block              4      32.684361       8.171090      0.49   0.7460
```

Figure 30.14. ANOVA Results for Unbalanced Two-Way Design

```
                           Unbalanced Two-way Design

                               The GLM Procedure

               Level of             --------------y-------------
               Treatment      N             Mean         Std Dev

               1             11       29.0909091      11.5104695
               2             11       29.1818182      11.5569735
               3             11       30.1818182       6.3058414
```

Figure 30.15. Treatment Means for Unbalanced Two-Way Design

```
                           Unbalanced Two-way Design

                               The GLM Procedure
                              Least Squares Means

                           Treatment        y LSMEAN

                           1              25.6000000
                           2              28.3333333
                           3              34.4444444
```

Figure 30.16. Treatment LS-means for Unbalanced Two-Way Design

No matter how you look at it, this data exhibits a strong effect due to the blocks (F-test $p < 0.0001$) and no significant interaction between treatments and blocks (F-test $p > 0.7$). But the lack of balance affects how the treatment effect is interpreted: in a main-effects-only model, there are no significant differences between the treatment means themselves (Type I F-test $p > 0.7$), but there are highly significant differences between the treatment means corrected for the block effects (Type III F-test $p < 0.01$).

LS-means are, in effect, within-group means appropriately adjusted for the other effects in the model. More precisely, they estimate the marginal means for a balanced population (as opposed to the unbalanced design). For this reason, they are also called *estimated population marginal means* by Searle, Speed, and Milliken (1980). In the same way that the Type I F-test assesses differences between the arithmetic treatment means (when the treatment effect comes first in the model), the Type III F-test assesses differences between the LS-means. Accordingly, for the unbalanced two-way design, the discrepancy between the Type I and Type III tests is reflected in the arithmetic treatment means and treatment LS-means, as shown in Figure 30.15 and Figure 30.16. See the section "Construction of Least-Squares Means" on page 1555 for more on LS-means.

Note that, while the arithmetic means are always uncorrelated (under the usual assumptions for analysis of variance; see page 1517), the LS-means may not be. This fact complicates the problem of multiple comparisons for LS-means; see the following section.

Multiple Comparisons

When comparing more than two means, an ANOVA F-test tells you whether the means are significantly different from each other, but it does not tell you which means differ from which other means. Multiple comparison procedures (MCPs), also called *mean separation tests*, give you more detailed information about the differences among the means. The goal in multiple comparisons is to compare the average effects of three or more "treatments" (for example, drugs, groups of subjects) to decide which treatments are better, which ones are worse, and by how much, while controlling the probability of making an incorrect decision. A variety of multiple comparison methods are available with the MEANS and LSMEANS statement in the GLM procedure.

The following classification is due to Hsu (1996). Multiple comparison procedures can be categorized in two ways: by the comparisons they make and by the strength of inference they provide. With respect to which comparisons are made, the GLM procedure offers two types:

- comparisons between all pairs of means
- comparisons between a control and all other means

The strength of inference says what can be inferred about the structure of the means when a test is significant; it is related to what type of error rate the MCP controls. MCPs available in the GLM procedure provide one of the following types of inference, in order from weakest to strongest.

- Individual: differences between means, unadjusted for multiplicity
- Inhomogeneity: means are different
- Inequalities: which means are different
- Intervals: simultaneous confidence intervals for mean differences

Methods that control only individual error rates are not true MCPs at all. Methods that yield the strongest level of inference, simultaneous confidence intervals, are usually preferred, since they enable you not only to say which means are different but also to put confidence bounds on *how much* they differ, making it easier to assess the practical significance of a difference. They are also less likely to lead nonstatisticians to the invalid conclusion that nonsignificantly different sample means imply equal population means. Interval MCPs are available for both arithmetic means and LS-means via the MEANS and LSMEANS statements, respectively.[*]

Table 30.3 and Table 30.4 display MCPs available in PROC GLM for all pairwise comparisons and comparisons with a control, respectively, along with associated strength of inference and the syntax (when applicable) for both the MEANS and the LSMEANS statements.

Table 30.3. Multiple Comparisons Procedures for All Pairwise Comparison

	Strength of	Syntax	
Method	Inference	MEANS	LSMEANS
Student's t	Individual	T	PDIFF ADJUST=T
Duncan	Individual	DUNCAN	
Student-Newman-Keuls	Inhomogeneity	SNK	
REGWQ	Inequalities	REGWQ	
Tukey-Kramer	Intervals	TUKEY	PDIFF ADJUST=TUKEY
Bonferroni	Intervals	BON	PDIFF ADJUST=BON
Sidak	Intervals	SIDAK	PDIFF ADJUST=SIDAK
Scheffé	Intervals	SCHEFFE	PDIFF ADJUST=SCHEFFE
SMM	Intervals	SMM	PDIFF ADJUST=SMM
Gabriel	Intervals	GABRIEL	
Simulation	Intervals		PDIFF ADJUST=SIMULATE

Table 30.4. Multiple Comparisons Procedures for Comparisons with a Control

	Strength of	Syntax	
Method	Inference	MEANS	LSMEANS
Student's t	Individual		PDIFF=CONTROL ADJUST=T
Dunnett	Intervals	DUNNETT	PDIFF=CONTROL ADJUST=DUNNETT
Bonferroni	Intervals		PDIFF=CONTROL ADJUST=BON
Sidak	Intervals		PDIFF=CONTROL ADJUST=SIDAK
Scheffé	Intervals		PDIFF=CONTROL ADJUST=SCHEFFE
SMM	Intervals		PDIFF=CONTROL ADJUST=SMM
Simulation	Intervals		PDIFF=CONTROL ADJUST=SIMULATE

[*] The Duncan-Waller method does not fit into the preceding scheme, since it is based on the Bayes risk rather than any particular error rate.

Note: One-sided Dunnett's tests are also available from the MEANS statement with the DUNNETTL and DUNNETTU options and from the LSMEANS statement with PDIFF=CONTROLL and PDIFF=CONTROLU.

Details of these multiple comparison methods are given in the following sections.

Pairwise Comparisons

All the methods discussed in this section depend on the standardized pairwise differences $t_{ij} = (\bar{y}_i - \bar{y}_j)/\hat{\sigma}_{ij}$, where

- i and j are the indices of two groups
- $\bar{y}_i$ and $\bar{y}_j$ are the means or LS-means for groups i and j
- $\hat{\sigma}_{ij}$ is the square-root of the estimated variance of $\bar{y}_i - \bar{y}_j$. For simple arithmetic means, $\hat{\sigma}^2_{ij} = s^2(1/n_i + 1/n_j)$, where n_i and n_j are the sizes of groups i and j, respectively, and s^2 is the mean square for error, with ν degrees of freedom. For weighted arithmetic means, $\hat{\sigma}^2_{ij} = s^2(1/w_i + 1/w_j)$, where w_i and w_j are the sums of the weights in groups i and j, respectively. Finally, for LS-means defined by the linear combinations $l_i'b$ and $l_j'b$ of the parameter estimates, $\hat{\sigma}^2_{ij} = s^2 l_i'(\mathbf{X}'\mathbf{X})^- l_j$.

Furthermore, all of the methods are discussed in terms of significance tests of the form

$$|t_{ij}| \quad \geq \quad c(\alpha)$$

where $c(\alpha)$ is some constant depending on the significance level. Such tests can be inverted to form confidence intervals of the form

$$(\bar{y}_i - \bar{y}_j) - \hat{\sigma}_{ij} c(\alpha) \quad \leq \quad \mu_i - \mu_j \quad \leq \quad (\bar{y}_i - \bar{y}_j) + \hat{\sigma}_{ij} c(\alpha)$$

The simplest approach to multiple comparisons is to do a t test on every pair of means (the T option in the MEANS statement, ADJUST=T in the LSMEANS statement). For the ith and jth means, you can reject the null hypothesis that the population means are equal if

$$|t_{ij}| \quad \geq \quad t(\alpha; \nu)$$

where α is the significance level, ν is the number of error degrees of freedom, and $t(\alpha; \nu)$ is the two-tailed critical value from a Student's t distribution. If the cell sizes are all equal to, say, n, the preceding formula can be rearranged to give

$$|\bar{y}_i - \bar{y}_j| \quad \geq \quad t(\alpha; \nu) s \sqrt{\frac{2}{n}}$$

the value of the right-hand side being Fisher's least significant difference (LSD).

There is a problem with repeated t tests, however. Suppose there are ten means and each t test is performed at the 0.05 level. There are 10(10-1)/2=45 pairs of means to compare, each with a 0.05 probability of a type 1 error (a false rejection of the null hypothesis). The chance of making at least one type 1 error is much higher than 0.05. It is difficult to calculate the exact probability, but you can derive a pessimistic approximation by assuming that the comparisons are independent, giving an upper bound to the probability of making at least one type 1 error (the experimentwise error rate) of

$$1 - (1 - 0.05)^{45} \quad = \quad 0.90$$

The actual probability is somewhat less than 0.90, but as the number of means increases, the chance of making at least one type 1 error approaches 1.

If you decide to control the individual type 1 error rates for each comparison, you are controlling the individual or comparisonwise error rate. On the other hand, if you want to control the overall type 1 error rate for all the comparisons, you are controlling the experimentwise error rate. It is up to you to decide whether to control the comparisonwise error rate or the experimentwise error rate, but there are many situations in which the experimentwise error rate should be held to a small value. Statistical methods for comparing three or more means while controlling the probability of making at least one type 1 error are called *multiple comparisons procedures*.

It has been suggested that the experimentwise error rate can be held to the α level by performing the overall ANOVA F-test at the α level and making further comparisons only if the F-test is significant, as in Fisher's protected LSD. This assertion is false if there are more than three means (Einot and Gabriel 1975). Consider again the situation with ten means. Suppose that one population mean differs from the others by such a sufficiently large amount that the power (probability of correctly rejecting the null hypothesis) of the F-test is near 1 but that all the other population means are equal to each other. There will be $9(9-1)/2 = 36$ t tests of true null hypotheses, with an upper limit of 0.84 on the probability of at least one type 1 error. Thus, you must distinguish between the experimentwise error rate under the complete null hypothesis, in which all population means are equal, and the experimentwise error rate under a partial null hypothesis, in which some means are equal but others differ. The following abbreviations are used in the discussion:

CER	comparisonwise error rate
EERC	experimentwise error rate under the complete null hypothesis
MEER	maximum experimentwise error rate under any complete or partial null hypothesis

These error rates are associated with the different strengths of inference discussed on page 1541: individual tests control the CER; tests for inhomogeneity of means control the EERC; tests that yield confidence inequalities or confidence intervals control the MEER. A preliminary F-test controls the EERC but not the MEER.

You can control the MEER at the α level by setting the CER to a sufficiently small value. The Bonferroni inequality (Miller 1981) has been widely used for this purpose.

If

$$\text{CER} = \frac{\alpha}{c}$$

where c is the total number of comparisons, then the MEER is less than α. Bonferroni t tests (the BON option in the MEANS statement, ADJUST=BON in the LSMEANS statement) with MEER $< \alpha$ declare two means to be significantly different if

$$|t_{ij}| \geq t(\epsilon; \nu)$$

where

$$\epsilon = \frac{2\alpha}{k(k-1)}$$

for comparison of k means.

Sidak (1967) has provided a tighter bound, showing that

$$\text{CER} = 1 - (1-\alpha)^{1/c}$$

also ensures that MEER $\leq \alpha$ for any set of c comparisons. A Sidak t test (Games 1977), provided by the SIDAK option, is thus given by

$$|t_{ij}| \geq t(\epsilon; \nu)$$

where

$$\epsilon = 1 - (1-\alpha)^{\frac{2}{k(k-1)}}$$

for comparison of k means.

You can use the Bonferroni additive inequality and the Sidak multiplicative inequality to control the MEER for any set of contrasts or other hypothesis tests, not just pairwise comparisons. The Bonferroni inequality can provide simultaneous inferences in any statistical application requiring tests of more than one hypothesis. Other methods discussed in this section for pairwise comparisons can also be adapted for general contrasts (Miller 1981).

Scheffé (1953, 1959) proposes another method to control the MEER for any set of contrasts or other linear hypotheses in the analysis of linear models, including pairwise comparisons, obtained with the SCHEFFE option. Two means are declared significantly different if

$$|t_{ij}| \geq \sqrt{(k-1)F(\alpha; k-1, \nu)}$$

where $F(\alpha; k-1, \nu)$ is the α-level critical value of an F distribution with $k-1$ numerator degrees of freedom and ν denominator degrees of freedom.

Scheffé's test is compatible with the overall ANOVA F-test in that Scheffé's method never declares a contrast significant if the overall F-test is nonsignificant. Most other multiple comparison methods can find significant contrasts when the overall F-test is nonsignificant and, therefore, suffer a loss of power when used with a preliminary F-test.

Scheffé's method may be more powerful than the Bonferroni or Sidak methods if the number of comparisons is large relative to the number of means. For pairwise comparisons, Sidak t tests are generally more powerful.

Tukey (1952, 1953) proposes a test designed specifically for pairwise comparisons based on the studentized range, sometimes called the "honestly significant difference test," that controls the MEER when the sample sizes are equal. Tukey (1953) and Kramer (1956) independently propose a modification for unequal cell sizes. The Tukey or Tukey-Kramer method is provided by the TUKEY option in the MEANS statement and the ADJUST=TUKEY option in the LSMEANS statement. This method has fared extremely well in Monte Carlo studies (Dunnett 1980). In addition, Hayter (1984) gives a proof that the Tukey-Kramer procedure controls the MEER for means comparisons, and Hayter (1989) describes the extent to which the Tukey-Kramer procedure has been proven to control the MEER for LS-means comparisons. The Tukey-Kramer method is more powerful than the Bonferroni, Sidak, or Scheffé methods for pairwise comparisons. Two means are considered significantly different by the Tukey-Kramer criterion if

$$|t_{ij}| \quad \geq \quad q(\alpha; k, \nu)$$

where $q(\alpha; k, \nu)$ is the α-level critical value of a studentized range distribution of k independent normal random variables with ν degrees of freedom.

Hochberg (1974) devised a method (the GT2 or SMM option) similar to Tukey's, but it uses the studentized maximum modulus instead of the studentized range and employs Sidak's (1967) uncorrelated t inequality. It is proven to hold the MEER at a level not exceeding α with unequal sample sizes. It is generally less powerful than the Tukey-Kramer method and always less powerful than Tukey's test for equal cell sizes. Two means are declared significantly different if

$$|t_{ij}| \quad \geq \quad m(\alpha; c, \nu)$$

where $m(\alpha; c, \nu)$ is the α-level critical value of the studentized maximum modulus distribution of c independent normal random variables with ν degrees of freedom and $c = k(k-1)/2$.

Gabriel (1978) proposes another method (the GABRIEL option) based on the studentized maximum modulus. This method is applicable only to arithmetic means. It rejects if

$$\frac{|\bar{y}_i - \bar{y}_j|}{s\left(\frac{1}{\sqrt{2n_i}} + \frac{1}{\sqrt{2n_j}}\right)} \quad \geq \quad m(\alpha; k, \nu)$$

For equal cell sizes, Gabriel's test is equivalent to Hochberg's GT2 method. For unequal cell sizes, Gabriel's method is more powerful than GT2 but may become liberal with highly disparate cell sizes (refer also to Dunnett 1980). Gabriel's test is the only method for unequal sample sizes that lends itself to a graphical representation as intervals around the means. Assuming $\bar{y}_i > \bar{y}_j$, you can rewrite the preceding inequality as

$$\bar{y}_i - m(\alpha; k, \nu)\frac{s}{\sqrt{2n_i}} \quad \geq \quad \bar{y}_j + m(\alpha; k, \nu)\frac{s}{\sqrt{2n_j}}$$

The expression on the left does not depend on j, nor does the expression on the right depend on i. Hence, you can form what Gabriel calls an (l, u)-interval around each sample mean and declare two means to be significantly different if their (l, u)-intervals do not overlap. See Hsu (1996, section 5.2.1.1) for a discussion of other methods of graphically representing all pair-wise comparisons.

Comparing All Treatments to a Control

One special case of means comparison is that in which the only comparisons that need to be tested are between a set of new treatments and a single control. In this case, you can achieve better power by using a method that is restricted to test only comparisons to the single control mean. Dunnett (1955) proposes a test for this situation that declares a mean significantly different from the control if

$$|t_{i0}| \quad \geq \quad d(\alpha; k, \nu, \rho_1, \ldots, \rho_{k-1})$$

where $\bar{y}_0$ is the control mean and $d(\alpha; k, \nu, \rho_1, \ldots, \rho_{k-1})$ is the critical value of the "many-to-one t statistic" (Miller 1981; Krishnaiah and Armitage 1966) for k means to be compared to a control, with ν error degrees of freedom and correlations $\rho_1, \ldots, \rho_{k-1}$, $\rho_i = n_i/(n_0 + n_i)$. The correlation terms arise because each of the treatment means is being compared to the same control. Dunnett's test holds the MEER to a level not exceeding the stated α.

Approximate and Simulation-based Methods

Both Tukey's and Dunnett's tests are based on the same general quantile calculation:

$$q^t(\alpha, \nu, R) \quad = \quad \{q \ni P(\max(|t_1|, \ldots, |t_n|) > q) = \alpha\}$$

where the t_i have a joint multivariate t distribution with ν degrees of freedom and correlation matrix R. In general, evaluating $q^t(\alpha, \nu, R)$ requires repeated numerical calculation of an $(n + 1)$-fold integral. This is usually intractable, but the problem reduces to a feasible 2-fold integral when R has a certain symmetry in the case of Tukey's test, and a *factor analytic structure* (cf. Hsu 1992) in the case of Dunnett's test. The R matrix has the required symmetry for exact computation of Tukey's test if the t_is are studentized differences between

- $k(k-1)/2$ pairs of k uncorrelated means with equal variances—that is, equal sample sizes

- $k(k-1)/2$ pairs of k LS-means from a *variance-balanced* design (for example, a balanced incomplete block design)

Refer to Hsu (1992, 1996) for more information. The R matrix has the factor analytic structure for exact computation of Dunnett's test if the t_is are studentized differences between

- $k-1$ means and a control mean, all uncorrelated. (Dunnett's one-sided methods depend on a similar probability calculation, without the absolute values.) Note that it is not required that the variances of the means (that is, the sample sizes) be equal.
- $k-1$ LS-means and a control LS-mean from either a *variance-balanced* design, or a design in which the other factors are *orthogonal* to the treatment factor (for example, a randomized block design with proportional cell frequencies).

However, other important situations that do **not** result in a correlation matrix R that has the structure for exact computation include

- all pairwise differences with unequal sample sizes
- differences between LS-means in many unbalanced designs

In these situations, exact calculation of $q^t(\alpha, \nu, R)$ is intractable in general. Most of the preceding methods can be viewed as using various approximations for $q^t(\alpha, \nu, R)$. When the sample sizes are unequal, the Tukey-Kramer test is equivalent to another approximation. For comparisons with a control when the correlation R does not have a factor analytic structure, Hsu (1992) suggests approximating R with a matrix R^* that does have such a structure and correspondingly approximating $q^t(\alpha, \nu, R)$ with $q^t(\alpha, \nu, R^*)$. When you request Dunnett's test for LS-means (the PDIFF=CONTROL and ADJUST=DUNNETT options), the GLM procedure automatically uses Hsu's approximation when appropriate.

Finally, Edwards and Berry (1987) suggest calculating $q^t(\alpha, \nu, R)$ by simulation. Multivariate t vectors are sampled from a distribution with the appropriate ν and R parameters, and Edwards and Berry (1987) suggest estimating $q^t(\alpha, \nu, R)$ by $\hat{q}$, the α percentile of the observed values of $\max(|t_1|, \ldots, |t_n|)$. Sufficient samples are generated for the true $P(\max(|t_1|, \ldots, |t_n|) > \hat{q})$ to be within a certain accuracy radius γ of α with accuracy confidence $100(1-\epsilon)$. You can approximate $q^t(\alpha, \nu, R)$ by simulation for comparisons between LS-means by specifying ADJUST=SIM (with either PDIFF=ALL or PDIFF=CONTROL). By default, $\gamma = 0.005$ and $\epsilon = 0.01$, so that the tail area of $\hat{q}$ is within 0.005 of α with 99% confidence. You can use the ACC= and EPS= options with ADJUST=SIM to reset γ and ϵ, or you can use the NSAMP= option to set the sample size directly. You can also control the random number sequence with the SEED= option.

Hsu and Nelson (1998) suggest a more accurate simulation method for estimating $q^t(\alpha, \nu, R)$, using a control variate adjustment technique. The same independent, standardized normal variates that are used to generate multivariate t vectors from a distribution with the appropriate ν and R parameters are also used to generate multivariate t vectors from a distribution for which the exact value of $q^t(\alpha, \nu, R)$ is known. $\max(|t_1|, \ldots, |t_n|)$ for the second sample is used as a control variate for adjusting the quantile estimate based on the first sample; refer to Hsu and Nelson (1998) for more details. The control variate adjustment has the drawback that it takes somewhat longer than the crude technique of Edwards and Berry (1987), but it typically yields an estimate that is many times more accurate. In most cases, if you are using ADJUST=SIM, then you should specify ADJUST=SIM(CVADJUST). You can also specify ADJUST=SIM(CVADJUST REPORT) to display a summary of the simulation that includes, among other things, the actual accuracy radius γ, which should be substantially smaller than the target accuracy radius (0.005 by default).

Multiple-Stage Tests

You can use all of the methods discussed so far to obtain simultaneous confidence intervals (Miller 1981). By sacrificing the facility for simultaneous estimation, you can obtain simultaneous tests with greater power using multiple-stage tests (MSTs). MSTs come in both step-up and step-down varieties (Welsch 1977). The step-down methods, which have been more widely used, are available in SAS/STAT software.

Step-down MSTs first test the homogeneity of all of the means at a level γ_k. If the test results in a rejection, then each subset of $k - 1$ means is tested at level γ_{k-1}; otherwise, the procedure stops. In general, if the hypothesis of homogeneity of a set of p means is rejected at the γ_p level, then each subset of $p - 1$ means is tested at the γ_{p-1} level; otherwise, the set of p means is considered not to differ significantly and none of its subsets are tested. The many varieties of MSTs that have been proposed differ in the levels γ_p and the statistics on which the subset tests are based. Clearly, the EERC of a step-down MST is not greater than γ_k, and the CER is not greater than γ_2, but the MEER is a complicated function of γ_p, $p = 2, \ldots, k$.

With unequal cell sizes, PROC GLM uses the harmonic mean of the cell sizes as the common sample size. However, since the resulting operating characteristics can be undesirable, MSTs are recommended only for the balanced case. When the sample sizes are equal and if the range statistic is used, you can arrange the means in ascending or descending order and test only contiguous subsets. But if you specify the F statistic, this shortcut cannot be taken. For this reason, only range-based MSTs are implemented. It is common practice to report the results of an MST by writing the means in such an order and drawing lines parallel to the list of means spanning the homogeneous subsets. This form of presentation is also convenient for pairwise comparisons with equal cell sizes.

The best known MSTs are the Duncan (the DUNCAN option) and Student-Newman-Keuls (the SNK option) methods (Miller 1981). Both use the studentized range statistic and, hence, are called *multiple range tests*. Duncan's method is often called the "new" multiple range test despite the fact that it is one of the oldest MSTs in current use.

The Duncan and SNK methods differ in the γ_p values used. For Duncan's method, they are

$$\gamma_p \quad = \quad 1-(1-\alpha)^{p-1}$$

whereas the SNK method uses

$$\gamma_p \quad = \quad \alpha$$

Duncan's method controls the CER at the α level. Its operating characteristics appear similar to those of Fisher's unprotected LSD or repeated t tests at level α (Petrinovich and Hardyck 1969). Since repeated t tests are easier to compute, easier to explain, and applicable to unequal sample sizes, Duncan's method is not recommended. Several published studies (for example, Carmer and Swanson 1973) have claimed that Duncan's method is superior to Tukey's because of greater power without considering that the greater power of Duncan's method is due to its higher type 1 error rate (Einot and Gabriel 1975).

The SNK method holds the EERC to the α level but does not control the MEER (Einot and Gabriel 1975). Consider ten population means that occur in five pairs such that means within a pair are equal, but there are large differences between pairs. If you make the usual sampling assumptions and also assume that the sample sizes are very large, all subset homogeneity hypotheses for three or more means are rejected. The SNK method then comes down to five independent tests, one for each pair, each at the α level. Letting α be 0.05, the probability of at least one false rejection is

$$1-(1-0.05)^5 \quad = \quad 0.23$$

As the number of means increases, the MEER approaches 1. Therefore, the SNK method cannot be recommended.

A variety of MSTs that control the MEER have been proposed, but these methods are not as well known as those of Duncan and SNK. An approach developed by Ryan (1959, 1960), Einot and Gabriel (1975), and Welsch (1977) sets

$$\gamma_p \quad = \quad \begin{cases} 1-(1-\alpha)^{p/k} & \text{for } p < k-1 \\ \alpha & \text{for } p \geq k-1 \end{cases}$$

You can use range statistics, leading to what is called the REGWQ method after the authors' initials. If you assume that the sample means have been arranged in descending order from $\bar{y}_1$ through $\bar{y}_k$, the homogeneity of means $\bar{y}_i, \ldots, \bar{y}_j, i < j$, is rejected by REGWQ if

$$\bar{y}_i - \bar{y}_j \quad \geq \quad q(\gamma_p; p, \nu)\frac{s}{\sqrt{n}}$$

where $p = j - i + 1$ and the summations are over $u = i, \ldots, j$ (Einot and Gabriel 1975). To ensure that the MEER is controlled, the current implementation checks whether $q(\gamma_p; p, \nu)$ is monotonically increasing in p. If not, then a set of critical values that are increasing in p is substituted instead.

REGWQ appears to be the most powerful step-down MST in the current literature (for example, Ramsey 1978). Use of a preliminary F-test decreases the power of all the other multiple comparison methods discussed previously except for Scheffé's test.

Bayesian Approach

Waller and Duncan (1969) and Duncan (1975) take an approach to multiple comparisons that differs from all the methods previously discussed in minimizing the Bayes risk under additive loss rather than controlling type 1 error rates. For each pair of population means μ_i and μ_j, null (H_0^{ij}) and alternative (H_a^{ij}) hypotheses are defined:

$$\begin{aligned} H_0^{ij}: &\quad \mu_i - \mu_j \le 0 \\ H_a^{ij}: &\quad \mu_i - \mu_j > 0 \end{aligned}$$

For any i, j pair, let d_0 indicate a decision in favor of H_0^{ij} and d_a indicate a decision in favor of H_a^{ij}, and let $\delta = \mu_i - \mu_j$. The loss function for the decision on the i, j pair is

$$\begin{aligned} L(d_0 \mid \delta) &= \begin{cases} 0 & \text{if } \delta \le 0 \\ \delta & \text{if } \delta > 0 \end{cases} \\ L(d_a \mid \delta) &= \begin{cases} -k\delta & \text{if } \delta \le 0 \\ 0 & \text{if } \delta > 0 \end{cases} \end{aligned}$$

where k represents a constant that you specify rather than the number of means. The loss for the joint decision involving all pairs of means is the sum of the losses for each individual decision. The population means are assumed to have a normal prior distribution with unknown variance, the logarithm of the variance of the means having a uniform prior distribution. For the i, j pair, the null hypothesis is rejected if

$$\bar{y}_i - \bar{y}_j \quad \ge \quad t_B s \sqrt{\frac{2}{n}}$$

where t_B is the Bayesian t value (Waller and Kemp 1976) depending on k, the F statistic for the one-way ANOVA, and the degrees of freedom for F. The value of t_B is a decreasing function of F, so the Waller-Duncan test (specified by the WALLER option) becomes more liberal as F increases.

Recommendations

In summary, if you are interested in several individual comparisons and are not concerned about the effects of multiple inferences, you can use repeated t tests or Fisher's unprotected LSD. If you are interested in all pairwise comparisons or all comparisons with a control, you should use Tukey's or Dunnett's test, respectively, in order to make the strongest possible inferences. If you have weaker inferential requirements and, in particular, if you don't want confidence intervals for the mean differences, you should use the REGWQ method. Finally, if you agree with the Bayesian approach and Waller and Duncan's assumptions, you should use the Waller-Duncan test.

Interpretation of Multiple Comparisons

When you interpret multiple comparisons, remember that failure to reject the hypothesis that two or more means are equal should not lead you to conclude that the population means are, in fact, equal. Failure to reject the null hypothesis implies only that the difference between population means, if any, is not large enough to be detected with the given sample size. A related point is that nonsignificance is nontransitive: that is, given three sample means, the largest and smallest may be significantly different from each other, while neither is significantly different from the middle one. Nontransitive results of this type occur frequently in multiple comparisons.

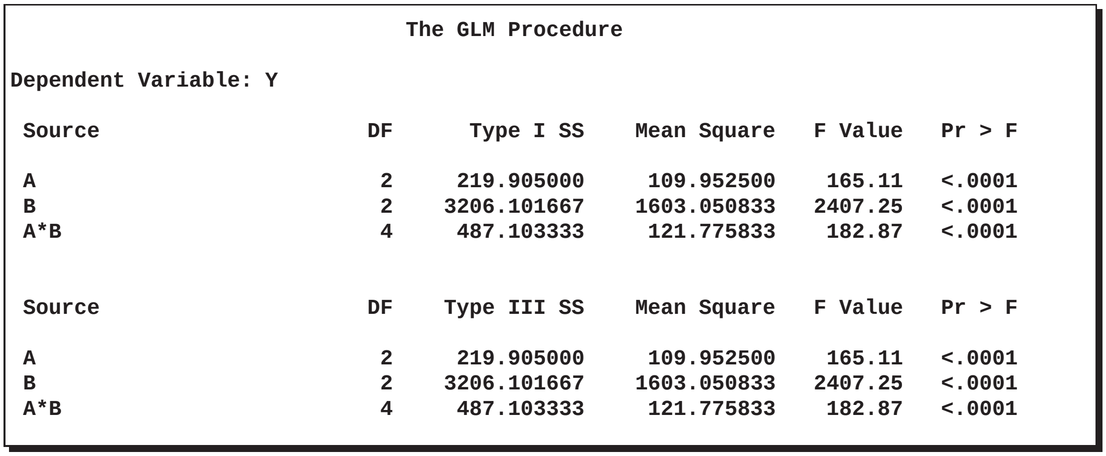
Multiple comparisons can also lead to counter-intuitive results when the cell sizes are unequal. Consider four cells labeled A, B, C, and D, with sample means in the order A>B>C>D. If A and D each have two observations, and B and C each have 10,000 observations, then the difference between B and C may be significant, while the difference between A and D is not.

Simple Effects

Suppose you use the following statements to fit a full factorial model to a two-way design:

```
data twoway;
   input A B Y @@;
   datalines;
1 1 10.6   1 1 11.0   1 1 10.6   1 1 11.3
1 2 -0.2   1 2  1.3   1 2 -0.2   1 2  0.2
1 3  0.1   1 3  0.4   1 3 -0.4   1 3  1.0
2 1 19.7   2 1 19.3   2 1 18.5   2 1 20.4
2 2 -0.2   2 2  0.5   2 2  0.8   2 2 -0.4
2 3 -0.9   2 3 -0.1   2 3 -0.2   2 3 -1.7
3 1 29.7   3 1 29.6   3 1 29.0   3 1 30.2
3 2  1.5   3 2  0.2   3 2 -1.5   3 2  1.3
3 3  0.2   3 3  0.4   3 3 -0.4   3 3 -2.2
;
proc glm data=twoway;
   class A B;
   model Y = A B A*B;
run;
```

Partial results for the analysis of variance are shown in Figure 30.17. The Type I and Type III results are the same because this is a balanced design.



```
                              The GLM Procedure

Dependent Variable: Y

 Source                      DF       Type I SS     Mean Square    F Value    Pr > F

 A                            2      219.905000      109.952500     165.11    <.0001
 B                            2     3206.101667     1603.050833    2407.25    <.0001
 A*B                          4      487.103333      121.775833     182.87    <.0001


 Source                      DF     Type III SS     Mean Square    F Value    Pr > F

 A                            2      219.905000      109.952500     165.11    <.0001
 B                            2     3206.101667     1603.050833    2407.25    <.0001
 A*B                          4      487.103333      121.775833     182.87    <.0001
```

Figure 30.17. Two-way Design with Significant Interaction

The interaction A*B is significant, indicating that the effect of A depends on the level of B. In some cases, you may be interested in looking at the differences between predicted values across A for different levels of B. Winer (1971) calls this the *simple effects* of A. You can compute simple effects with the LSMEAN statement by specifying the SLICE= option. In this case, since the GLM procedure is interactive, you can compute the simple effects of A by submitting the following statements after the preceding statements.

```
lsmeans A*B / slice=B;
run;
```

The results are shown Figure 30.18. Note that A has a significant effect for B=1 but not for B=2 and B=3.

```
                          The GLM Procedure
                         Least Squares Means

                        A    B        Y LSMEAN

                        1    1      10.8750000
                        1    2       0.2750000
                        1    3       0.2750000
                        2    1      19.4750000
                        2    2       0.1750000
                        2    3      -0.7250000
                        3    1      29.6250000
                        3    2       0.3750000
                        3    3      -0.5000000



                          The GLM Procedure
                         Least Squares Means

                     A*B Effect Sliced by B for Y

                                Sum of
   B        DF         Squares     Mean Square    F Value    Pr > F

   1         2      704.726667      352.363333     529.13    <.0001
   2         2        0.080000        0.040000       0.06    0.9418
   3         2        2.201667        1.100833       1.65    0.2103
```

Figure 30.18. Interaction LS-means and Simple Effects

Homogeneity of Variance in One-Way Models

One of the usual assumptions for the GLM procedure is that the underlying errors are all uncorrelated with homogeneous variances (see page 1517). You can test this assumption in PROC GLM by using the HOVTEST option in the MEANS statement, requesting a *homogeneity of variance* test. This section discusses the computational details behind these tests. Note that the GLM procedure allows homogeneity of variance testing for simple one-way models only. Homogeneity of variance testing for more complex models is a subject of current research.

Bartlett (1937) proposes a test for equal variances that is a modification of the normal-theory likelihood ratio test (the HOVTEST=BARTLETT option). While Bartlett's test has accurate Type I error rates and optimal power when the underlying distribution of the data is normal, it can be very inaccurate if that distribution is even slightly nonnormal (Box 1953). Therefore, Bartlett's test is not recommended for routine use.

An approach that leads to tests that are much more robust to the underlying distribution is to transform the original values of the dependent variable to derive a *dispersion variable* and then to perform analysis of variance on this variable. The significance level for the test of homogeneity of variance is the *p*-value for the ANOVA *F*-test on the dispersion variable. All of the homogeneity of variance tests available in PROC GLM except Bartlett's use this approach.

Levene's test (Levene 1960) is widely considered to be the standard homogeneity of variance test (the HOVTEST=LEVENE option). Levene's test is of the dispersion-variable-ANOVA form discussed previously, where the dispersion variable is either

$$\begin{aligned} z_{ij}^2 &= (y_{ij} - \bar{y}_i)^2 && \text{(TYPE=SQUARE, the default)} \\ z_{ij} &= |y_{ij} - \bar{y}_i| && \text{(TYPE=ABS)} \end{aligned}$$

O'Brien (1979) proposes a test (HOVTEST=OBRIEN) that is basically a modification of Levene's z_{ij}^2, using the dispersion variable

$$z_{ij}^W = \frac{(W + n_i - 2)n_i(y_{ij} - \bar{y}_i)^2 - W(n_i - 1)\sigma_i^2}{(n_i - 1)(n_i - 2)}$$

where n_i is the size of the i^{th} group and σ_i^2 is its sample variance. You can use the W= option in parentheses to tune O'Brien's z_{ij}^W dispersion variable to match the suspected kurtosis of the underlying distribution. The choice of the value of the W= option is rarely critical. By default, W=0.5, as suggested by O'Brien (1979, 1981).

Finally, Brown and Forsythe (1974) suggest using the absolute deviations from the group *medians*:

$$z_{ij}^{\text{BF}} = |y_{ij} - m_i|$$

where m_i is the median of the i^{th} group. You can use the HOVTEST=BF option to specify this test.

Simulation results (Conover, Johnson, and Johnson 1981; Olejnik and Algina 1987) show that, while all of these ANOVA-based tests are reasonably robust to the underlying distribution, the Brown-Forsythe test seems best at providing power to detect variance differences while protecting the Type I error probability. However, since the within-group medians are required for the Brown-Forsythe test, it can be resource intensive if there are very many groups or if some groups are very large.

If one of these tests rejects the assumption of homogeneity of variance, you should use Welch's ANOVA instead of the usual ANOVA to test for differences between group means. However, this conclusion holds only if you use one of the robust homogeneity of variance tests (that is, not for HOVTEST=BARTLETT); even then, any homogeneity of variance test has too little power to be relied upon always to detect when Welch's ANOVA is appropriate. Unless the group variances are extremely different or the number of groups is large, the usual ANOVA test is relatively robust when the groups are all about the same size. As Box (1953) notes, "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!"

Example 30.10 on page 1623 illustrates the use of the HOVTEST and WELCH options in the MEANS statement in testing for equal group variances and adjusting for unequal group variances in a one-way ANOVA.

Weighted Means

In previous releases, if you specified a WEIGHT statement and one or more of the multiple comparisons options, PROC GLM estimated the variance of the difference between weighted group means for group i and j as

$$MSE \times \left(\frac{1}{n_i} + \frac{1}{n_j} \right)$$

where MSE is the (weighted) mean square for error and n_i is the size of group i. This variance is involved in all of the multiple comparison methods. Beginning with Release 6.12, the variance of the difference between weighted group means for group i and j is computed as

$$MSE \times \left(\frac{1}{w_i} + \frac{1}{w_j} \right)$$

where w_i is the sum of the weights for the observations in group i.

Construction of Least-Squares Means

To construct a least-squares mean (LS-mean) for a given level of a given effect, construct a row vector L according to the following rules and use it in an ESTIMATE statement to compute the value of the LS-mean:

1. Set all L_i corresponding to covariates (continuous variables) to their mean value.
2. Consider effects contained by the given effect. Set the L_i corresponding to levels associated with the given level equal to 1. Set all other L_i in these effects equal to 0. (See Chapter 12, "The Four Types of Estimable Functions," for a definition of *containing*.)
3. Consider the given effect. Set the L_i corresponding to the given level equal to 1. Set the L_i corresponding to other levels equal to 0.
4. Consider the effects that contain the given effect. If these effects are not nested within the given effect, then set the L_i corresponding to the given level to $1/k$, where k is the number of such columns. If these effects are nested within the given effect, then set the L_i corresponding to the given level to $1/(k_1 k_2)$, where k_1 is the number of nested levels within this combination of nested effects, and k_2 is the number of such combinations. For L_i corresponding to other levels, use 0.
5. Consider the other effects not yet considered. If there are no nested factors, then set all L_i corresponding to this effect to $1/j$, where j is the number of levels in the effect. If there are nested factors, then set all L_i corresponding to this effect to $1/(j_1 j_2)$, where j_1 is the number of nested levels within a given combination of nested effects and j_2 is the number of such combinations.

The consequence of these rules is that the sum of the Xs within any classification effect is 1. This set of Xs forms a linear combination of the parameters that is checked for estimability before it is evaluated.

For example, consider the following model:

```
proc glm;
   class A B C;
   model Y=A B A*B C Z;
   lsmeans A B A*B C;
run;
```

Assume A has 3 levels, B has 2 levels, and C has 2 levels, and assume that every combination of levels of A and B exists in the data. Assume also that Z is a continuous variable with an average of 12.5. Then the least-squares means are computed by the following linear combinations of the parameter estimates:

		A			B		A*B						C		
	μ	1	2	3	1	2	11	12	21	22	31	32	1	2	Z
LSM( )	1	1/3	1/3	1/3	1/2	1/2	1/6	1/6	1/6	1/6	1/6	1/6	1/2	1/2	12.5
LSM(A1)	1	1	0	0	1/2	1/2	1/2	1/2	0	0	0	0	1/2	1/2	12.5
LSM(A2)	1	0	1	0	1/2	1/2	0	0	1/2	1/2	0	0	1/2	1/2	12.5
LSM(A3)	1	0	0	1	1/2	1/2	0	0	0	0	1/2	1/2	1/2	1/2	12.5
LSM(B1)	1	1/3	1/3	1/3	1	0	1/3	0	1/3	0	1/3	0	1/2	1/2	12.5
LSM(B2)	1	1/3	1/3	1/3	0	1	0	1/3	0	1/3	0	1/3	1/2	1/2	12.5
LSM(AB11)	1	1	0	0	1	0	1	0	0	0	0	0	1/2	1/2	12.5
LSM(AB12)	1	1	0	0	0	1	0	1	0	0	0	0	1/2	1/2	12.5
LSM(AB21)	1	0	1	0	1	0	0	0	1	0	0	0	1/2	1/2	12.5
LSM(AB22)	1	0	1	0	0	1	0	0	0	1	0	0	1/2	1/2	12.5
LSM(AB31)	1	0	0	1	1	0	0	0	0	0	1	0	1/2	1/2	12.5
LSM(AB32)	1	0	0	1	0	1	0	0	0	0	0	1	1/2	1/2	12.5
LSM(C1)	1	1/3	1/3	1/3	1/2	1/2	1/6	1/6	1/6	1/6	1/6	1/6	1	0	12.5
LSM(C2)	1	1/3	1/3	1/3	1/2	1/2	1/6	1/6	1/6	1/6	1/6	1/6	0	1	12.5

Setting Covariate Values

By default, all covariate effects are set equal to their mean values for computation of standard LS-means. The AT option in the LSMEANS statement enables you to set the covariates to whatever values you consider interesting.

If there is an effect containing two or more covariates, the AT option sets the effect equal to the product of the individual means rather than the mean of the product (as with standard LS-means calculations). The AT MEANS option leaves covariates equal to their mean values (as with standard LS-means) and incorporates this adjustment to cross products of covariates.

As an example, the following is a model with a classification variable A and two continuous variables, x1 and x2:

```
class A;
model y = A x1 x2 x1*x2;
```

The coefficients for the continuous effects with various AT specifications are shown in the following table.

Syntax	x1	x2	x1*x2
`lsmeans A;`	$\overline{x_1}$	$\overline{x_2}$	$\overline{x_1 x_2}$
`lsmeans A / at means;`	$\overline{x_1}$	$\overline{x_2}$	$\overline{x_1} \cdot \overline{x_2}$
`lsmeans A / at x1=1.2;`	1.2	$\overline{x_2}$	$1.2 \cdot \overline{x_2}$
`lsmeans A / at (x1 x2)=(1.2 0.3);`	1.2	0.3	$1.2 \cdot 0.3$

For the first two LSMEANS statements, the A LS-mean coefficient for x1 is $\overline{x_1}$ (the mean of x1) and for x2 is $\overline{x_2}$ (the mean of x2). However, for the first LSMEANS statement, the coefficient for x1*x2 is $\overline{x_1 x_2}$, but for the second LSMEANS statement the coefficient is $\overline{x_1} \cdot \overline{x_2}$. The third LSMEANS statement sets the coefficient for x1 equal to 1.2 and leaves that for x2 at $\overline{x_2}$, and the final LSMEANS statement sets these values to 1.2 and 0.3, respectively.

If you specify a WEIGHT variable, then weighted means are used for the covariate values. Also, observations with missing dependent variables are included in computing the covariate means, unless these observations form a missing cell. You can use the E option in conjunction with the AT option to check that the modified LS-means coefficients are the ones you desire.

The AT option is disabled if you specify the BYLEVEL option, in which case the coefficients for the covariates are set equal to their means within each level of the LS-mean effect in question.

Changing the Weighting Scheme

The standard LS-means have equal coefficients across classification effects; however, the OM option in the LSMEANS statement changes these coefficients to be proportional to those found in the input data set. This adjustment is reasonable when you want your inferences to apply to a population that is not necessarily balanced but has the margins observed in the original data set.

In computing the observed margins, PROC GLM uses all observations for which there are no missing independent variables, including those for which there are missing dependent variables. Also, if there is a WEIGHT variable, PROC GLM uses weighted margins to construct the LS-means coefficients. If the analysis data set is balanced or if you specify a simple one-way model, the LS-means will be unchanged by the OM option.

The BYLEVEL option modifies the observed-margins LS-means. Instead of computing the margins across the entire data set, PROC GLM computes separate margins for each level of the LS-mean effect in question. The resulting LS-means are actually equal to raw means in this case. The BYLEVEL option disables the AT option if it is specified.

Note that the MIXED procedure implements a more versatile form of the OM option, enabling you to specifying an alternative data set over which to compute observed margins. If you use the BYLEVEL option, too, then this data set is effectively the "population" over which the population marginal means are computed. See Chapter 41, "The MIXED Procedure," for more information.

You may want to use the E option in conjunction with either the OM or BYLEVEL option to check that the modified LS-means coefficients are the ones you desire. It is possible that the modified LS-means are not estimable when the standard ones are, or vice versa.

Multivariate Analysis of Variance

If you fit several dependent variables to the same effects, you may want to make tests jointly involving parameters of several dependent variables. Suppose you have p dependent variables, k parameters for each dependent variable, and n observations. The models can be collected into one equation:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{Y}$ is $n \times p$, $\mathbf{X}$ is $n \times k$, $\boldsymbol{\beta}$ is $k \times p$, and ϵ is $n \times p$. Each of the p models can be estimated and tested separately. However, you may also want to consider the joint distribution and test the p models simultaneously.

For multivariate tests, you need to make some assumptions about the errors. With p dependent variables, there are $n \times p$ errors that are independent across observations but not across dependent variables. Assume

$$\text{vec}(\epsilon) \sim N(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma})$$

where vec(ϵ) strings ϵ out by rows, $\otimes$ denotes Kronecker product multiplication, and $\mathbf{\Sigma}$ is $p \times p$. $\mathbf{\Sigma}$ can be estimated by

$$\mathbf{S} = \frac{\mathbf{e}'\mathbf{e}}{n-r} = \frac{(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})}{n-r}$$

where $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$, r is the rank of the $\mathbf{X}$ matrix, and $\mathbf{e}$ is the matrix of residuals.

If $\mathbf{S}$ is scaled to unit diagonals, the values in $\mathbf{S}$ are called *partial correlations of the Ys adjusting for the Xs*. This matrix can be displayed by PROC GLM if PRINTE is specified as a MANOVA option.

The multivariate general linear hypothesis is written

$$\mathbf{L}\beta\mathbf{M} = 0$$

You can form hypotheses for linear combinations across columns, as well as across rows of $\boldsymbol{\beta}$.

The MANOVA statement of the GLM procedure tests special cases where $\mathbf{L}$ corresponds to Type I, Type II, Type III, or Type IV tests, and $\mathbf{M}$ is the $p \times p$ identity matrix. These tests are joint tests that the given type of hypothesis holds for all dependent variables in the model, and they are often sufficient to test all hypotheses of interest.

Finally, when these special cases are not appropriate, you can specify your own $\mathbf{L}$ and $\mathbf{M}$ matrices by using the CONTRAST statement before the MANOVA statement and the M= specification in the MANOVA statement, respectively. Another alternative is to use a REPEATED statement, which automatically generates a variety of $\mathbf{M}$ matrices useful in repeated measures analysis of variance. See the "REPEATED Statement" section on page 1511 and the "Repeated Measures Analysis of Variance" section on page 1560 for more information.

One useful way to think of a MANOVA analysis with an $\mathbf{M}$ matrix other than the identity is as an analysis of a set of transformed variables defined by the columns of the $\mathbf{M}$ matrix. You should note, however, that PROC GLM always displays the $\mathbf{M}$ matrix in such a way that the transformed variables are defined by the rows, not the columns, of the displayed $\mathbf{M}$ matrix.

All multivariate tests carried out by the GLM procedure first construct the matrices $\mathbf{H}$ and $\mathbf{E}$ corresponding to the numerator and denominator, respectively, of a univariate F-test.

$$\begin{aligned}\mathbf{H} &= \mathbf{M}'(\mathbf{Lb})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{Lb})\mathbf{M} \\ \mathbf{E} &= \mathbf{M}'(\mathbf{Y}'\mathbf{Y} - \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b})\mathbf{M}\end{aligned}$$

The diagonal elements of $\mathbf{H}$ and $\mathbf{E}$ correspond to the hypothesis and error SS for univariate tests. When the $\mathbf{M}$ matrix is the identity matrix (the default), these tests are for the original dependent variables on the left-hand side of the MODEL statement. When an $\mathbf{M}$ matrix other than the identity is specified, the tests are for transformed variables defined by the columns of the $\mathbf{M}$ matrix. These tests can be studied by requesting the SUMMARY option, which produces univariate analyses for each original or transformed variable.

Four multivariate test statistics, all functions of the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ (or $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$), are constructed:

- Wilks' lambda = det($\mathbf{E}$)/det($\mathbf{H}+\mathbf{E}$)
- Pillai's trace = trace($\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$)
- Hotelling-Lawley trace = trace($\mathbf{E}^{-1}\mathbf{H}$)
- Roy's maximum root = λ, largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$

All four are reported with F approximations. For further details on these four statistics, see the "Multivariate Tests" section in Chapter 3, "Introduction to Regression Procedures."

Repeated Measures Analysis of Variance

When several measurements are taken on the same experimental unit (person, plant, machine, and so on), the measurements tend to be correlated with each other. When the measurements represent qualitatively different things, such as weight, length, and width, this correlation is best taken into account by use of multivariate methods, such as multivariate analysis of variance. When the measurements can be thought of as responses to levels of an experimental factor of interest, such as time, treatment, or dose, the correlation can be taken into account by performing a repeated measures analysis of variance.

PROC GLM provides both univariate and multivariate tests for repeated measures for one response. For an overall reference on univariate repeated measures, refer to Winer (1971). The multivariate approach is covered in Cole and Grizzle (1966). For a discussion of the relative merits of the two approaches, see LaTour and Miniard (1983).

Another approach to analysis of repeated measures is via general mixed models. This approach can handle balanced as well as unbalanced or missing within-subject data, and it offers more options for modeling the within-subject covariance. The main drawback of the mixed models approach is that it generally requires iteration and, thus, may be less computationally efficient. For further details on this approach, see Chapter 41, “The MIXED Procedure,” and Wolfinger and Chang (1995).

Organization of Data for Repeated Measures Analysis

In order to deal efficiently with the correlation of repeated measures, the GLM procedure uses the multivariate method of specifying the model, even if only a univariate analysis is desired. In some cases, data may already be entered in the univariate mode, with each repeated measure listed as a separate observation along with a variable that represents the experimental unit (subject) on which measurement is taken. Consider the following data set old:

```
SUBJ   GROUP    TIME       Y
   1       1       1      15
   1       1       2      19
   1       1       3      25
   2       1       1      21
   2       1       2      18
   2       1       3      17
   1       2       1      14
   1       2       2      12
   1       2       3      16
   2       2       1      11
   2       2       2      20
                .
                .
                .
  10       3       1      14
  10       3       2      18
  10       3       3      16
```

There are three observations for each subject, corresponding to measurements taken at times 1, 2, and 3. These data could be analyzed using the following statements:

```
proc glm data=old;
   class group subj time;
   model y=group subj(group) time group*time;
   test h=group e=subj(group);
run;
```

However, this analysis assumes subjects' measurements are uncorrelated across time. A repeated measures analysis does not make this assumption. It uses a data set new:

```
GROUP        Y1     Y2     Y3
    1        15     19     25
    1        21     18     17
    2        14     12     16
    2        11     20     21
                 .
                 .
                 .
    3        14     18     16
```

In the data set new, the three measurements for a subject are all in one observation. For example, the measurements for subject 1 for times 1, 2, and 3 are 15, 19, and 25. For these data, the statements for a repeated measures analysis (assuming default options) are

```
proc glm data=new;
   class group;
   model y1-y3=group / nouni;
   repeated time;
run;
```

To convert the univariate form of repeated measures data to the multivariate form, you can use a program like the following:

```
proc sort data=old;
   by group subj;
run;

data new(keep=y1-y3 group);
   array yy(3)  y1-y3;
   do time=1 to 3;
      set old;
      by group subj;
      yy(time)=y;
      if last.subj then return;
   end;
run;
```

Alternatively, you could use PROC TRANSPOSE to achieve the same results with a program like this one:

```
proc sort data=old;
   by group subj;
run;

proc transpose out=new(rename=(_1=y1 _2=y2 _3=y3));
   by group subj;
   id time;
run;
```

Refer to the discussions in *SAS Language Reference: Concepts* for more information on rearrangement of data sets.

Hypothesis Testing in Repeated Measures Analysis

In repeated measures analysis of variance, the effects of interest are

- between-subject effects (such as GROUP in the previous example)
- within-subject effects (such as TIME in the previous example)
- interactions between the two types of effects (such as GROUP*TIME in the previous example)

Repeated measures analyses are distinguished from MANOVA because of interest in testing hypotheses about the within-subject effects and the within-subject-by-between-subject interactions.

For tests that involve only between-subjects effects, both the multivariate and univariate approaches give rise to the same tests. These tests are provided for all effects in the MODEL statement, as well as for any CONTRASTs specified. The ANOVA table for these tests is labeled “Tests of Hypotheses for Between Subjects Effects” on the PROC GLM results. These tests are constructed by first adding together the dependent variables in the model. Then an analysis of variance is performed on the sum divided by the square root of the number of dependent variables. For example, the statements

```
model y1-y3=group;
repeated time;
```

give a one-way analysis of variance using $(Y1 + Y2 + Y3)/\sqrt{3}$ as the dependent variable for performing tests of hypothesis on the between-subject effect GROUP. Tests for between-subject effects are equivalent to tests of the hypothesis $\mathbf{L}\beta\mathbf{M} = 0$, where $\mathbf{M}$ is simply a vector of 1s.

For within-subject effects and for within-subject-by-between-subject interaction effects, the univariate and multivariate approaches yield different tests. These tests are provided for the within-subject effects and for the interactions between these effects and the other effects in the MODEL statement, as well as for any CONTRASTs specified. The univariate tests are displayed in a table labeled “Univariate Tests of

Hypotheses for Within Subject Effects." Results for multivariate tests are displayed in a table labeled "Repeated Measures Analysis of Variance."

The multivariate tests provided for within-subjects effects and interactions involving these effects are Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's maximum root. For further details on these four statistics, see the "Multivariate Tests" section in Chapter 3, "Introduction to Regression Procedures." As an example, the statements

```
   model y1-y3=group;
   repeated time;
```

produce multivariate tests for the within-subject effect TIME and the interaction TIME*GROUP.

The multivariate tests for within-subject effects are produced by testing the hypothesis $\mathbf{L}\beta\mathbf{M} = 0$, where the $\mathbf{L}$ matrix is the usual matrix corresponding to Type I, Type II, Type III, or Type IV hypotheses tests, and the $\mathbf{M}$ matrix is one of several matrices depending on the transformation that you specify in the REPEATED statement. The only assumption required for valid tests is that the dependent variables in the model have a multivariate normal distribution with a common covariance matrix across the between-subject effects.

The univariate tests for within-subject effects and interactions involving these effects require some assumptions for the probabilities provided by the ordinary F-tests to be correct. Specifically, these tests require certain patterns of covariance matrices, known as Type H covariances (Huynh and Feldt 1970). Data with these patterns in the covariance matrices are said to satisfy the Huynh-Feldt condition. You can test this assumption (and the Huynh-Feldt condition) by applying a sphericity test (Anderson 1958) to any set of variables defined by an orthogonal contrast transformation. Such a set of variables is known as a set of orthogonal components. When you use the PRINTE option in the REPEATED statement, this sphericity test is applied both to the transformed variables defined by the REPEATED statement and to a set of orthogonal components if the specified transformation is not orthogonal. It is the test applied to the orthogonal components that is important in determining whether your data have Type H covariance structure. When there are only two levels of the within-subject effect, there is only one transformed variable, and a sphericity test is not needed. The sphericity test is labeled "Test for Sphericity" on the output.

If your data satisfy the preceding assumptions, use the usual F-tests to test univariate hypotheses for the within-subject effects and associated interactions.

If your data do not satisfy the assumption of Type H covariance, an adjustment to numerator and denominator degrees of freedom can be used. Two such adjustments, based on a degrees of freedom adjustment factor known as ϵ (epsilon) (Box 1954), are provided in PROC GLM. Both adjustments estimate ϵ and then multiply the numerator and denominator degrees of freedom by this estimate before determining significance levels for the F-tests. Significance levels associated with the adjusted tests are labeled "Adj Pr > F" on the output. The first adjustment, initially proposed for use in data analysis by Greenhouse and Geisser (1959), is labeled "Greenhouse-

Geisser Epsilon” and represents the maximum-likelihood estimate of Box’s ϵ factor. Significance levels associated with adjusted F-tests are labeled “G-G” on the output. Huynh and Feldt (1976) have shown that the G-G estimate tends to be biased downward (that is, too conservative), especially for small samples, and they have proposed an alternative estimator that is constructed using unbiased estimators of the numerator and denominator of Box’s ϵ. Huynh and Feldt’s estimator is labeled “Huynh-Feldt Epsilon” on the PROC GLM output, and the significance levels associated with adjusted F-tests are labeled “H-F.” Although ϵ must be in the range of 0 to 1, the H-F estimator can be outside this range. When the H-F estimator is greater than 1, a value of 1 is used in all calculations for probabilities, and the H-F probabilities are not adjusted. In summary, if your data do not meet the assumptions, use adjusted F-tests. However, when you strongly suspect that your data may not have Type H covariance, all these univariate tests should be interpreted cautiously. In such cases, you should consider using the multivariate tests instead.

The univariate sums of squares for hypotheses involving within-subject effects can be easily calculated from the $\mathbf{H}$ and $\mathbf{E}$ matrices corresponding to the multivariate tests described in the “Multivariate Analysis of Variance” section on page 1558. If the $\mathbf{M}$ matrix is orthogonal, the univariate sums of squares is calculated as the trace (sum of diagonal elements) of the appropriate $\mathbf{H}$ matrix; if it is not orthogonal, PROC GLM calculates the trace of the $\mathbf{H}$ matrix that results from an orthogonal $\mathbf{M}$ matrix transformation. The appropriate error term for the univariate F-tests is constructed in a similar way from the error SSCP matrix and is labeled Error(*factorname*), where *factorname* indicates the $\mathbf{M}$ matrix that is used in the transformation.

When the design specifies more than one repeated measures factor, PROC GLM computes the $\mathbf{M}$ matrix for a given effect as the direct (Kronecker) product of the $\mathbf{M}$ matrices defined by the REPEATED statement if the factor is involved in the effect or as a vector of 1s if the factor is not involved. The test for the main effect of a repeated-measures factor is constructed using an $\mathbf{L}$ matrix that corresponds to a test that the mean of the observation is zero. Thus, the main effect test for repeated measures is a test that the means of the variables defined by the $\mathbf{M}$ matrix are all equal to zero, while interactions involving repeated-measures effects are tests that the between-subjects factors involved in the interaction have no effect on the means of the transformed variables defined by the $\mathbf{M}$ matrix. In addition, you can specify other $\mathbf{L}$ matrices to test hypotheses of interest by using the CONTRAST statement, since hypotheses defined by CONTRAST statements are also tested in the REPEATED analysis. To see which combinations of the original variables the transformed variables represent, you can specify the PRINTM option in the REPEATED statement. This option displays the transpose of $\mathbf{M}$, which is labeled as M in the PROC GLM results. The tests produced are the same for any choice of transformation $(\mathbf{M})$ matrix specified in the REPEATED statement; however, depending on the nature of the repeated measurements being studied, a particular choice of transformation matrix, coupled with the CANONICAL or SUMMARY option, can provide additional insight into the data being studied.

Transformations Used in Repeated Measures Analysis of Variance

As mentioned in the specifications of the REPEATED statement, several different $\mathbf{M}$ matrices can be generated automatically, based on the transformation that you specify in the REPEATED statement. Remember that both the univariate and multivariate tests that PROC GLM performs are unaffected by the choice of transformation; the choice of transformation is important only when you are trying to study the nature of a repeated measures effect, particularly with the CANONICAL and SUMMARY options. If one of these matrices does not meet your needs for a particular analysis, you may want to use the M= option in the MANOVA statement to perform the tests of interest.

The following sections describe the transformations available in the REPEATED statement, provide an example of the $\mathbf{M}$ matrix that is produced, and give guidelines for the use of the transformation. As in the PROC GLM output, the displayed matrix is labeled M. This is the $\mathbf{M}'$ matrix.

CONTRAST Transformation

This is the default transformation used by the REPEATED statement. It is useful when one level of the repeated measures effect can be thought of as a control level against which the others are compared. For example, if five drugs are administered to each of several animals and the first drug is a control or placebo, the statements

```
proc glm;
   model d1-d5= / nouni;
   repeated drug 5 contrast(1) / summary printm;
run;
```

produce the following $\mathbf{M}$ matrix:

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

When you examine the analysis of variance tables produced by the SUMMARY option, you can tell which of the drugs differed significantly from the placebo.

POLYNOMIAL Transformation

This transformation is useful when the levels of the repeated measure represent quantitative values of a treatment, such as dose or time. If the levels are unequally spaced, *level values* can be specified in parentheses after the number of levels in the REPEATED statement. For example, if five levels of a drug corresponding to 1, 2, 5, 10 and 20 milligrams are administered to different treatment groups, represented by the variable group, the statements

```
proc glm;
   class group;
   model r1-r5=group / nouni;
   repeated dose 5 (1 2 5 10 20) polynomial / summary printm;
run;
```

produce the following $\mathbf{M}$ matrix.

$$\mathbf{M} = \begin{bmatrix} -0.4250 & -0.3606 & -0.1674 & 0.1545 & 0.7984 \\ 0.4349 & 0.2073 & -0.3252 & -0.7116 & 0.3946 \\ -0.4331 & 0.1366 & 0.7253 & -0.5108 & 0.0821 \\ 0.4926 & -0.7800 & 0.3743 & -0.0936 & 0.0066 \end{bmatrix}$$

The SUMMARY option in this example provides univariate ANOVAs for the variables defined by the rows of this $\mathbf{M}$ matrix. In this case, they represent the linear, quadratic, cubic, and quartic trends for dose and are labeled dose$_1$, dose$_2$, dose$_3$, and dose$_4$, respectively.

HELMERT Transformation

Since the Helmert transformation compares a level of a repeated measure to the mean of subsequent levels, it is useful when interest lies in the point at which responses cease to change. For example, if four levels of a repeated measures factor represent responses to treatments administered over time to males and females, the statements

```
proc glm;
   class sex;
   model resp1-resp4=sex / nouni;
   repeated trtmnt 4 helmert / canon printm;
run;
```

produce the following $\mathbf{M}$ matrix:

$$\mathbf{M} = \begin{bmatrix} 1 & -0.33333 & -0.33333 & -0.33333 \\ 0 & 1 & -0.50000 & -0.50000 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

MEAN Transformation

This transformation can be useful in the same types of situations in which the CONTRAST transformation is useful. If you substitute the following statement for the REPEATED statement shown in the “CONTRAST Transformation” section,

```
repeated drug 5 mean / printm;
```

the following $\mathbf{M}$ matrix is produced:

$$\mathbf{M} = \begin{bmatrix} 1 & -0.25 & -0.25 & -0.25 & -0.25 \\ -0.25 & 1 & -0.25 & -0.25 & -0.25 \\ -0.25 & -0.25 & 1 & -0.25 & -0.25 \\ -0.25 & -0.25 & -0.25 & 1 & -0.25 \end{bmatrix}$$

As with the CONTRAST transformation, if you want to omit a level other than the last, you can specify it in parentheses after the keyword MEAN in the REPEATED statement.

PROFILE Transformation

When a repeated measure represents a series of factors administered over time, but a polynomial response is unreasonable, a profile transformation may prove useful. As an example, consider a training program in which four different methods are employed to teach students at several different schools. The repeated measure is the score on tests administered after each of the methods is completed. The statements

```
proc glm;
   class school;
   model t1-t4=school / nouni;
   repeated method 4 profile / summary nom printm;
run;
```

produce the following $\mathbf{M}$ matrix:

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

To determine the point at which an improvement in test scores takes place, you can examine the analyses of variance for the transformed variables representing the differences between adjacent tests. These analyses are requested by the SUMMARY option in the REPEATED statement, and the variables are labeled METHOD.1, METHOD.2, and METHOD.3.

Random Effects Analysis

When some model effects are random (that is, assumed to be sampled from a normal population of effects), you can specify these effects in the RANDOM statement in order to compute the expected values of mean squares for various model effects and contrasts and, optionally, to perform random effects analysis of variance tests.

PROC GLM versus PROC MIXED for Random Effects Analysis

Other SAS procedures that can be used to analyze models with random effects include the MIXED and VARCOMP procedures. Note that, for these procedures, the random effects specification is an integral part of the model, affecting how both random and fixed effects are fit; for PROC GLM, the random effects are treated in a *post hoc* fashion after the complete fixed effect model is fit. This distinction affects other features in the GLM procedure, such as the results of the LSMEANS and ESTIMATE statements. These features assume that all effects are fixed, so that all tests and estimability checks for these statements are based on a fixed effects model, even when you use a RANDOM statement. Standard errors for estimates and LS-means based on the fixed effects model may be significantly smaller than those based on a true random effects model; in fact, some functions that are estimable under a true random effects model may not even be estimable under the fixed effects model. Therefore, you should use the MIXED procedure to compute tests involving these features that take the random effects into account; see Chapter 41, “The MIXED Procedure,” for more information.

Note that, for balanced data, the test statistics computed when you specify the TEST option on the RANDOM statement have an exact F distribution only when the design is balanced; for unbalanced designs, the p values for the F-tests are approximate. For balanced data, the values obtained by PROC GLM and PROC MIXED agree; for unbalanced data, they usually do not.

Computation of Expected Mean Squares for Random Effects

The RANDOM statement in PROC GLM declares one or more effects in the model to be random rather than fixed. By default, PROC GLM displays the coefficients of the expected mean squares for all terms in the model. In addition, when you specify the TEST option in the RANDOM statement, the procedure determines what tests are appropriate and provides F ratios and probabilities for these tests.

The expected mean squares are computed as follows. Consider the model

$$Y = X_0\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 + \cdots + X_k\boldsymbol{\beta}_k + \epsilon$$

where $\boldsymbol{\beta}_0$ represents the fixed effects and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \epsilon$ represent the random effects. Random effects are assumed to be normally and independently distributed. For any $\mathbf{L}$ in the row space of $\mathbf{X} = (X_0 \mid X_1 \mid X_2 \mid \cdots \mid X_k)$, the expected value of the sum of squares for $\mathbf{L}\beta$ is

$$E(\mathrm{SS}_L) = \boldsymbol{\beta}_0'\mathbf{C}_0'\mathbf{C}_0\boldsymbol{\beta}_0 + \mathrm{SSQ}(\mathbf{C}_1)\sigma_1^2 + \mathrm{SSQ}(\mathbf{C}_2)\sigma_2^2 + \cdots + \mathrm{SSQ}(\mathbf{C}_k)\sigma_k^2 + \mathrm{rank}(\mathbf{L})\sigma_\epsilon^2$$

where $\mathbf{C}$ is of the same dimensions as $\mathbf{L}$ and is partitioned as the $\mathbf{X}$ matrix. In other words,

$$\mathbf{C} = (\mathbf{C}_0 \mid \mathbf{C}_1 \mid \cdots \mid \mathbf{C}_k)$$

Furthermore, $\mathbf{C} = \mathbf{ML}$, where $\mathbf{M}$ is the inverse of the lower triangular Cholesky decomposition matrix of $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'$. SSQ($\mathbf{A}$) is defined as tr($\mathbf{A}'\mathbf{A}$).

For the model in the following MODEL statement

```
model Y=A B(A) C A*C;
random B(A);
```

with B(A) declared as random, the expected mean square of each effect is displayed as

$$\mathrm{Var(Error)} + \textit{constant} \times \mathrm{Var}(\mathsf{B(A)}) + Q(\mathsf{A}, \mathsf{C}, \mathsf{A*C})$$

If any fixed effects appear in the expected mean square of an effect, the letter Q followed by the list of fixed effects in the expected value is displayed. The actual numeric values of the quadratic form ($\mathbf{Q}$ matrix) can be displayed using the Q option.

To determine appropriate means squares for testing the effects in the model, the TEST option in the RANDOM statement performs the following.

1. First, it forms a matrix of coefficients of the expected mean squares of those effects that were declared to be random.
2. Next, for each effect in the model, it determines the combination of these expected mean squares that produce an expectation that includes all the terms in the expected mean square of the effect of interest except the one corresponding to the effect of interest. For example, if the expected mean square of an effect A*B is

   $$\text{Var(Error)} + 3 \times \text{Var}(\mathsf{A}) + \text{Var}(\mathsf{A} * \mathsf{B})$$

   PROC GLM determines the combination of other expected mean squares in the model that has expectation

   $$\text{Var(Error)} + 3 \times \text{Var}(\mathsf{A})$$

3. If the preceding criterion is met by the expected mean square of a single effect in the model (as is often the case in balanced designs), the F test is formed directly. In this case, the mean square of the effect of interest is used as the numerator, the mean square of the single effect with an expected mean square that satisfies the criterion is used as the denominator, and the degrees of freedom for the test are simply the usual model degrees of freedom.
4. When more than one mean square must be combined to achieve the appropriate expectation, an approximation is employed to determine the appropriate degrees of freedom (Satterthwaite 1946). When effects other than the effect of interest are listed after the Q in the output, tests of hypotheses involving the effect of interest are not valid unless all other fixed effects involved in it are assumed to be zero. When tests such as these are performed by using the TEST option in the RANDOM statement, a note is displayed reminding you that further assumptions are necessary for the validity of these tests. Remember that although the tests are not valid unless these assumptions are made, this does not provide a basis for these assumptions to be true. The particulars of a given experiment must be examined to determine whether the assumption is reasonable.

Refer to Goodnight and Speed (1978), Milliken and Johnson (1984, Chapters 22 and 23), and Hocking (1985) for further theoretical discussion.

Sum-to-Zero Assumptions

The formulation and parameterization of the expected mean squares for random effects in mixed models is an ongoing item of controversy in the statistical literature. Confusion arises over whether or not to assume that terms involving fixed effects sum to zero. Cornfield and Tukey (1956), Winer (1971), and others assume that they do sum to zero; Searle (1971), Hocking (1973), and others (including PROC GLM) do not. The assumption usually makes no difference for balanced data, but with unbalanced designs it can yield different expected mean squares for certain terms, and, hence, different F and p values.

For arguments in favor of not assuming that terms involving fixed effects sum to zero, see Section 9.7 of Searle (1971) and Sections 1 and 4 of McLean, Sanders, and Stroup (1991). Other references are Hartley and Searle (1969) and Searle, Casella, McCulloch (1992).

Computing Type I, II, and IV Expected Mean Squares

When you use the RANDOM statement, by default the GLM procedure produces the Type III expected mean squares for model effects and for contrasts specified before the RANDOM statement. In order to obtain expected values for other types of mean squares, you need to specify which types of mean squares are of interest in the MODEL statement. For example, in order to obtain the Type IV expected mean squares for effects in the RANDOM and CONTRAST statements, specify the SS4 option in the MODEL statement. If you want both Type III and Type IV expected mean squares, specify both the SS3 and SS4 options in the MODEL statement. Since the estimable function basis is not automatically calculated for Type I and Type II SS, the E1 (for Type I) or E2 (for Type II) option must be specified in the MODEL statement in order for the RANDOM statement to produce the expected mean squares for the Type I or Type II sums of squares. Note that it is important to list the fixed effects first in the MODEL statement when requesting the Type I expected mean squares.

For example, suppose you have a two-way design with factors A and B in which the main effect for B and the interaction are random. In order to compute the Type III expected mean squares (in addition to the fixed-effect analysis), you can use the following statements:

```
proc glm;
   class A B;
   model Y = A B A*B;
   random B A*B;
run;
```

If you use the SS4 option in the MODEL statement,

```
proc glm;
   class A B;
   model Y = A B A*B / ss4;
   random B A*B;
run;
```

then only the Type IV expected mean squares are computed (as well as the Type IV fixed-effect tests). For the Type I expected mean squares, you can use the following statements:

```
proc glm;
   class A B;
   model Y = A B A*B / e1;
   random B A*B;
run;
```

For each of these cases, in order to perform random effect analysis of variance tests for each effect specified in the model, you need to specify the TEST option in the RANDOM statement, as follows:

```
proc glm;
   class A B;
   model Y = A B A*B;
   random B A*B / test;
run;
```

The GLM procedure automatically determines the appropriate error term for each test, based on the expected mean squares.

Missing Values

For an analysis involving one dependent variable, PROC GLM uses an observation if values are nonmissing for that dependent variable and all the class variables.

For an analysis involving multiple dependent variables without the MANOVA or REPEATED statement, or without the MANOVA option in the PROC GLM statement, a missing value in one dependent variable does not eliminate the observation from the analysis of other nonmissing dependent variables. On the other hand, for an analysis with the MANOVA or REPEATED statement, or with the MANOVA option in the PROC GLM statement, PROC GLM uses an observation if values are nonmissing for all dependent variables and all the variables used in independent effects.

During processing, the GLM procedure groups the dependent variables by their pattern of missing values across observations so that sums and cross products can be collected in the most efficient manner.

If your data have different patterns of missing values among the dependent variables, interactivity is disabled. This can occur when some of the variables in your data set have missing values and

- you do not use the MANOVA option in the PROC GLM statement
- you do not use a MANOVA or REPEATED statement before the first RUN statement

Note that the REG procedure handles missing values differently in this case; see Chapter 55, "The REG Procedure," for more information.

Computational Resources

Memory

For large problems, most of the memory resources are required for holding the $\mathbf{X}'\mathbf{X}$ matrix of the sums and cross products. The section "Parameterization of PROC GLM Models" on page 1521 describes how columns of the $\mathbf{X}$ matrix are allocated for various types of effects. For each level that occurs in the data for a combination of class variables in a given effect, a row and column for $\mathbf{X}'\mathbf{X}$ is needed.

The following example illustrates the calculation. Suppose A has 20 levels, B has 4 levels, and C has 3 levels. Then consider the model

```
proc glm;
  class A B C;
  model Y1 Y2 Y3=A B A*B C A*C B*C A*B*C X1 X2;
run;
```

The $\mathbf{X}'\mathbf{X}$ matrix (bordered by $\mathbf{X}'\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Y}$) can have as many as 425 rows and columns:

1	for the intercept term
20	for A
4	for B
80	for A*B
3	for C
60	for A*C
12	for B*C
240	for A*B*C
2	for X1 and X2 (continuous variables)
3	for Y1, Y2, and Y3 (dependent variables)

The matrix has 425 rows and columns only if all combinations of levels occur for each effect in the model. For m rows and columns, $8m^2$ bytes are needed for cross products. In this case, $8 \cdot 425^2 = 1,445,000$ bytes, or about $1,445,000/1024 = 1411K$.

The required memory grows as the square of the number of columns of $\mathbf{X}$; most of the memory is for the A*B*C interaction. Without A*B*C, you have 185 columns and need 268K for $\mathbf{X}'\mathbf{X}$. Without either A*B*C or A*B, you need 86K. If A is recoded to have ten levels, then the full model has only 220 columns and requires 378K.

The second time that a large amount of memory is needed is when Type III, Type IV, or contrast sums of squares are being calculated. This memory requirement is a function of the number of degrees of freedom of the model being analyzed and the maximum degrees of freedom for any single source. Let Rank equal the sum of the model degrees of freedom, MaxDF be the maximum number of degrees of freedom for any single source, and N_y be the number of dependent variables in the model. Then the memory requirement in bytes is

$$\begin{aligned} \left(8 \times \left(\frac{\text{Rank } \times (\text{Rank } + 1)}{2}\right)\right) \quad & + \quad (N_y \times \text{Rank }) \\ & + \quad \left(\frac{\text{MaxDF} \times (\text{MaxDF} + 1)}{2}\right) \\ & + \quad (N_y \times \text{MaxDF}) \end{aligned}$$

Unfortunately, these quantities are not available when the $\mathbf{X}'\mathbf{X}$ matrix is being constructed, so PROC GLM may occasionally request additional memory even after you have increased the memory allocation available to the program.

If you have a large model that exceeds the memory capacity of your computer, these are your options:

- Eliminate terms, especially high-level interactions.
- Reduce the number of levels for variables with many levels.
- Use the ABSORB statement for parts of the model that are large.
- Use the REPEATED statement for repeated measures variables.
- Use PROC ANOVA or PROC REG rather than PROC GLM, if your design allows.

CPU Time

For large problems, two operations consume a lot of CPU time: the collection of sums and cross products and the solution of the normal equations.

The time required for collecting sums and cross products is difficult to calculate because it is a complicated function of the model. For a model with m columns and n rows (observations) in $\mathbf{X}$, the worst case occurs if all columns are continuous variables, involving $nm^2/2$ multiplications and additions. If the columns are levels of a classification, then only m sums may be needed, but a significant amount of time may be spent in look-up operations. Solving the normal equations requires time for approximately $m^3/2$ multiplications and additions.

Suppose you know that Type IV sums of squares are appropriate for the model you are analyzing (for example, if your design has no missing cells). You can specify the SS4 option in your MODEL statement, which saves CPU time by requesting the Type IV sums of squares instead of the more computationally burdensome Type III sums of squares. This proves especially useful if you have a factor in your model that has many levels and is involved in several interactions.

Computational Method

Let $\mathbf{X}$ represent the $n \times p$ design matrix and $\mathbf{Y}$ the $n \times 1$ vector of dependent variables. (See the section "Parameterization of PROC GLM Models" on page 1521 for information on how $\mathbf{X}$ is formed from your model specification.)

The normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$ are solved using a modified sweep routine that produces a generalized (g2) inverse $(\mathbf{X}'\mathbf{X})^-$ and a solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ (Pringle and Raynor 1971).

For each effect in the model, a matrix $\mathbf{L}$ is computed such that the rows of $\mathbf{L}$ are estimable. Tests of the hypothesis $\mathbf{L}\beta = 0$ are then made by first computing

$$\mathrm{SS}(\mathbf{L}\beta = 0) = (\mathbf{Lb})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^-\mathbf{L}')^{-1}(\mathbf{Lb})$$

and then computing the associated F value using the mean squared error.

Output Data Sets

OUT= Data Set Created by the OUTPUT Statement

The OUTPUT statement produces an output data set that contains the following:

- all original data from the SAS data set input to PROC GLM
- the new variables corresponding to the diagnostic measures specified with statistics keywords in the OUTPUT statement (PREDICTED=, RESIDUAL=, and so on).

With multiple dependent variables, a name can be specified for any of the diagnostic measures for each of the dependent variables in the order in which they occur in the MODEL statement.

For example, suppose that the input data set A contains the variables y1, y2, y3, x1, and x2. Then you can use the following statements:

```
proc glm data=A;
   model y1 y2 y3=x1;
   output out=out p=y1hat y2hat y3hat
                  r=y1resid lclm=y1lcl uclm=y1ucl;
run;
```

The output data set out contains y1, y2, y3, x1, x2, y1hat, y2hat, y3hat, y1resid, y1lcl, and y1ucl. The variable x2 is output even though it is not used by PROC GLM. Although predicted values are generated for all three dependent variables, residuals are output for only the first dependent variable.

When any independent variable in the analysis (including all class variables) is missing for an observation, then all new variables that correspond to diagnostic measures are missing for the observation in the output data set.

When a dependent variable in the analysis is missing for an observation, then some new variables that correspond to diagnostic measures are missing for the observation in the output data set, and some are still available. Specifically, in this case, the new variables that correspond to COOKD, COVRATIO, DFFITS, PRESS, R, RSTUDENT, STDR, and STUDENT are missing in the output data set. The variables corresponding to H, LCL, LCLM, P, STDI, STDP, UCL, and UCLM are not missing.

OUT= Data Set Created by the LSMEANS Statement

The OUT= option in the LSMEANS statement produces an output data set that contains

- the unformatted values of each classification variable specified in any effect in the LSMEANS statement
- a new variable, LSMEAN, which contains the LS-mean for the specified levels of the classification variables
- a new variable, STDERR, which contains the standard error of the LS-mean

The variances and covariances among the LS-means are also output when the COV option is specified along with the OUT= option. In this case, only one effect can be specified in the LSMEANS statement, and the following variables are included in the output data set:

- new variables, COV1, COV2, . . . , COVn, where n is the number of levels of the effect specified in the LSMEANS statement. These variables contain the covariances of each LS-mean with each other LS-mean.
- a new variable, NUMBER, which provides an index for each observation to identify the covariances that correspond to that observation. The covariances for the observation with NUMBER equal to n can be found in the variable COVn.

OUTSTAT= Data Set

The OUTSTAT= option in the PROC GLM statement produces an output data set that contains

- the BY variables, if any
- _TYPE_, a new character variable. _TYPE_ may take the values ‘SS1’, ‘SS2’, ‘SS3’, ‘SS4’, or ‘CONTRAST’, corresponding to the various types of sums of squares generated, or the values ‘CANCORR’, ‘STRUCTUR’, or ‘SCORE’, if a canonical analysis is performed through the MANOVA statement and no M= matrix is specified.
- _SOURCE_, a new character variable. For each observation in the data set, _SOURCE_ contains the name of the model effect or contrast label from which the corresponding statistics are generated.
- _NAME_, a new character variable. For each observation in the data set, _NAME_ contains the name of one of the dependent variables in the model or, in the case of canonical statistics, the name of one of the canonical variables (CAN1, CAN2, and so forth).

- four new numeric variables: SS, DF, F, and PROB, containing sums of squares, degrees of freedom, F values, and probabilities, respectively, for each model or contrast sum of squares generated in the analysis. For observations resulting from canonical analyses, these variables have missing values.
- if there is more than one dependent variable, then variables with the same names as the dependent variables represent
  - for _TYPE_=SS1, SS2, SS3, SS4, or CONTRAST, the crossproducts of the hypothesis matrices
  - for _TYPE_=CANCORR, canonical correlations for each variable
  - for _TYPE_=STRUCTUR, coefficients of the total structure matrix
  - for _TYPE_=SCORE, raw canonical score coefficients

The output data set can be used to perform special hypothesis tests (for example, with the IML procedure in SAS/IML software), to reformat output, to produce canonical variates (through the SCORE procedure), or to rotate structure matrices (through the FACTOR procedure).

Displayed Output

The GLM procedure produces the following output by default:

- The overall analysis-of-variance table breaks down the Total Sum of Squares for the dependent variable into the portion attributed to the Model and the portion attributed to Error.
- The Mean Square term is the Sum of Squares divided by the degrees of freedom (DF).
- The Mean Square for Error is an estimate of σ^2, the variance of the true errors.
- The F Value is the ratio produced by dividing the Mean Square for the Model by the Mean Square for Error. It tests how well the model as a whole (adjusted for the mean) accounts for the dependent variable's behavior. An F-test is a joint test to determine that all parameters except the intercept are zero.
- A small significance probability, Pr > F, indicates that some linear function of the parameters is significantly different from zero.
- R-Square, R^2, measures how much variation in the dependent variable can be accounted for by the model. R^2, which can range from 0 to 1, is the ratio of the sum of squares for the model divided by the sum of squares for the corrected total. In general, the larger the value of R^2, the better the model's fit.
- Coef Var, the coefficient of variation, which describes the amount of variation in the population, is 100 times the standard deviation estimate of the dependent variable, Root MSE (Mean Square for Error), divided by the Mean. The coefficient of variation is often a preferred measure because it is unitless.
- Root MSE estimates the standard deviation of the dependent variable (or equivalently, the error term) and equals the square root of the Mean Square for Error.
- Mean is the sample mean of the dependent variable.

These tests are used primarily in analysis-of-variance applications:

- The Type I SS (sum of squares) measures incremental sums of squares for the model as each variable is added.
- The Type III SS is the sum of squares for a balanced test of each effect, adjusted for every other effect.

These items are used primarily in regression applications:

- The Estimates for the model Parameters (the intercept and the coefficients)
- t Value is the Student's t value for testing the null hypothesis that the parameter (if it is estimable) equals zero.
- The significance level, Pr > |t|, is the probability of getting a larger value of t if the parameter is truly equal to zero. A very small value for this probability leads to the conclusion that the independent variable contributes significantly to the model.
- The Standard Error is the square root of the estimated variance of the estimate of the true value of the parameter.

Other portions of output are discussed in the following examples.

ODS Table Names

PROC GLM assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 30.5. ODS Tables Produced in PROC GLM

ODS Table Name	Description	Statement / Option
Aliasing	Type 1,2,3,4 aliasing structure	MODEL / (E1 E2 E3 or E4) and ALIASING
AltErrContrasts	ANOVA table for contrasts with alternative error	CONTRAST / E=
AltErrTests	ANOVA table for tests with alternative error	TEST / E=
Bartlett	Bartlett's homogeneity of variance test	MEANS / HOVTEST=BARTLETT
CLDiffs	Multiple comparisons of pairwise differences	MEANS / CLDIFF or DUNNETT or (Unequal cells and not LINES)
CLDiffsInfo	Information for multiple comparisons of pairwise differences	MEANS / CLDIFF or DUNNETT or (Unequal cells and not LINES)
CLMeans	Multiple comparisons of means with confidence/comparison interval	MEANS / CLM

Table 30.5. (continued)

ODS Table Name	Description	Statement / Option
CLMeansInfo	Information for multiple comparison of means with confidence/comparison interval	MEANS / CLM
CanAnalysis	Canonical analysis	(MANOVA or REPEATED) / CANONICAL
CanCoefficients	Canonical coefficients	(MANOVA or REPEATED) / CANONICAL
CanStructure	Canonical structure	(MANOVA or REPEATED) / CANONICAL
CharStruct	Characteristic roots and vectors	(MANOVA / not CANONICAL) or (REPEATED / PRINTRV)
ClassLevels	Classification variable levels	CLASS statement
ContrastCoef	**L** matrix for contrast	CONTRAST / EST
Contrasts	ANOVA table for contrasts	CONTRAST statement
DependentInfo	Simultaneously analyzed dependent variables	default when there are multiple dependent variables with different patterns of missing values
Diff	PDiff matrix of Least-Squares Means	LSMEANS / PDIFF
Epsilons	Greenhouse-Geisser and Huynh-Feldt epsilons	REPEATED statement
ErrorSSCP	Error SSCP matrix	(MANOVA or REPEATED) / PRINTE
EstFunc	Type 1,2,3,4 estimable functions	MODEL / (E1 E2 E3 or E4)
Estimates	Estimate statement results	ESTIMATE statement
ExpectedMeanSquares	Expected mean squares	RANDOM statement
FitStatistics	R-Square, C.V., Root MSE, and dependent mean	default
GAliasing	General form of aliasing structure	MODEL / E and ALIASING
GEstFunc	General form of estimable functions	MODEL / E
HOVFTest	Homogeneity of variance ANOVA	MEANS / HOVTEST
HypothesisSSCP	Hypothesis SSCP matrix	(MANOVA or REPEATED) / PRINTH
Inv	inv(**X'X**) matrix	MODEL / INVERSE
LSMeanCL	Confidence interval for LS-means	LSMEANS / CL
LSMeanCoef	Coefficients of Least-Squares Means	LSMEANS / E
LSMeanDiffCL	Confidence interval for LS-mean differences	LSMEANS / PDIFF and CL
LSMeans	Least-Squares means	LSMEANS statement

Table 30.5. (continued)

ODS Table Name	Description	Statement / Option
MANOVATransform	Multivariate transformation matrix	MANOVA / M=
MCLines	Multiple comparisons LINES output	MEANS / LINES or ((DUNCAN or WALLER or SNK or REGWQ) and not (CLDIFF or CLM)) or (Equal cells and not CLDIFF)
MCLinesInfo	Information for multiple comparison LINES output	MEANS / LINES or ((DUNCAN or WALLER or SNK or REGWQ) and not (CLDIFF or CLM)) or (Equal cells and not CLDIFF)
MCLinesRange	Ranges for multiple range MC tests	MEANS / LINES or ((DUNCAN or WALLER or SNK or REGWQ) and not (CLDIFF or CLM)) or (Equal cells and not CLDIFF)
MTests	Multivariate tests	MANOVA statement
MatrixRepresentation	**X** matrix element representation	as needed for other options
Means	Group means	MEANS statement
ModelANOVA	ANOVA for model terms	default
NObs	Number of observations	default
OverallANOVA	Over-all ANOVA	default
ParameterEstimates	Estimated linear model coefficients	MODEL / SOLUTION
PartialCorr	Partial correlation matrix	(MANOVA or REPEATED) / PRINTE
PredictedInfo	Predicted values info	MODEL / PREDICTED or CLM or CLI
PredictedValues	Predicted values	MODEL / PREDICTED or CLM or CLI
QForm	Quadratic form for expected mean squares	RANDOM / Q
RandomModelANOVA	Random effect tests	RANDOM / TEST
RepeatedLevelInfo	Correspondence between dependents and repeated measures levels	REPEATED statement
RepeatedTransform	Repeated Measures Transformation Matrix	REPEATED / PRINTM
SimDetails	Details of difference quantile simulation	LSMEANS / ADJUST=SIMULATE(REPORT)
SimResults	Evaluation of difference quantile simulation	LSMEANS / ADJUST=SIMULATE(REPORT)
SlicedANOVA	Sliced effect ANOVA table	LSMEANS / SLICE
Sphericity	Sphericity tests	REPEATED / PRINTE
Tests	Summary ANOVA for specified MANOVA H= effects	MANOVA / H= SUMMARY

Table 30.5. (continued)

ODS Table Name	Description	Statement / Option
Tolerances	**X'X** Tolerances	MODEL / TOLERANCE
Welch	Welch's ANOVA	MEANS / WELCH
XpX	**X'X** matrix	MODEL / XPX

Examples

Example 30.1. Balanced Data from Randomized Complete Block with Means Comparisons and Contrasts

The following example* analyzes an experiment to investigate how snapdragons grow in various soils. To eliminate the effect of local fertility variations, the experiment is run in blocks, with each soil type sampled in each block. Since these data are balanced, the Type I and Type III SS are the same and are equal to the traditional ANOVA SS.

First, the standard analysis is shown followed by an analysis that uses the SOLUTION option and includes MEANS and CONTRAST statements. The ORDER=DATA option in the second PROC GLM statement is used so that the ordering of coefficients in the CONTRAST statement can correspond to the ordering in the input data. The SOLUTION option requests a display of the parameter estimates, which are only produced by default if there are no CLASS variables. A MEANS statement is used to request a table of the means with two multiple comparison procedures requested. In experiments with focused treatment questions, CONTRAST statements are preferable to general means comparison methods. The following statements produce Output 30.1.1 through Output 30.1.5:

```
title 'Balanced Data from Randomized Complete Block';
data plants;
   input Type $ @;
   do Block = 1 to 3;
      input StemLength @;
      output;
      end;
   datalines;
Clarion  32.7 32.3 31.5
Clinton  32.1 29.7 29.1
Knox     35.7 35.9 33.1
O'Neill  36.0 34.2 31.2
Compost  31.8 28.0 29.2
Wabash   38.2 37.8 31.9
Webster  32.5 31.1 29.7
;

proc glm;
   class Block Type;
   model StemLength = Block Type;
run;
```

*reported by Stenstrom (1940)

```
proc glm order=data;
   class Block Type;
   model StemLength = Block Type / solution;

/*------------------------------------clrn-cltn-knox-onel-cpst-wbsh-wstr */
contrast ’Compost vs. others’  Type   -1   -1   -1   -1    6   -1   -1;
contrast ’River soils vs. non’ Type   -1   -1   -1   -1    0    5   -1,
                               Type   -1    4   -1   -1    0    0   -1;
contrast ’Glacial vs. drift’   Type   -1    0    1    1    0    0   -1;
contrast ’Clarion vs. Webster’ Type   -1    0    0    0    0    0    1;
contrast ’’Knox vs. O’Neill’’  Type    0    0    1   -1    0    0    0;
run;

    means Type / waller regwq;
run;
```

Output 30.1.1. Standard Analysis for Randomized Complete Block

```
                  Balanced Data from Randomized Complete Block

                              The GLM Procedure

                           Class Level Information

Class         Levels    Values

Block              3    1 2 3

Type               7    Clarion Clinton Compost Knox O’Neill Wabash Webster


                        Number of observations    21
                  Balanced Data from Randomized Complete Block

                               The GLM Procedure

Dependent Variable: StemLength

                                         Sum of
 Source                      DF         Squares    Mean Square   F Value   Pr > F

 Model                        8     142.1885714     17.7735714     10.80   0.0002

 Error                       12      19.7428571      1.6452381

 Corrected Total             20     161.9314286


            R-Square     Coeff Var      Root MSE    StemLength Mean

            0.878079      3.939745      1.282668           32.55714


 Source                      DF       Type I SS    Mean Square   F Value   Pr > F

 Block                        2      39.0371429     19.5185714     11.86   0.0014
 Type                         6     103.1514286     17.1919048     10.45   0.0004


 Source                      DF     Type III SS    Mean Square   F Value   Pr > F

 Block                        2      39.0371429     19.5185714     11.86   0.0014
 Type                         6     103.1514286     17.1919048     10.45   0.0004
```

This analysis shows that the stem length is significantly different for the different soil types. In addition, there are significant differences in stem length between the three blocks in the experiment.

Output 30.1.2. Standard Analysis Again

```
                  Balanced Data from Randomized Complete Block

                               The GLM Procedure

                            Class Level Information

  Class         Levels    Values

  Block              3    1 2 3

  Type               7    Clarion Clinton Compost Knox O'Neill Wabash Webster


                          Number of observations    21
```

The GLM procedure is invoked again, this time with the ORDER=DATA option. This enables you to write accurate contrast statements more easily because you know the order SAS is using for the levels of the variable Type. The standard analysis is displayed again.

Output 30.1.3. Contrasts and Solutions

```
                  Balanced Data from Randomized Complete Block

                               The GLM Procedure

Dependent Variable: StemLength

 Contrast                  DF    Contrast SS    Mean Square   F Value   Pr > F

 Compost vs. others         1    29.24198413    29.24198413     17.77   0.0012
 River soils vs. non        2    48.24694444    24.12347222     14.66   0.0006
 Glacial vs. drift          1    22.14083333    22.14083333     13.46   0.0032
 Clarion vs. Webster        1     1.70666667     1.70666667      1.04   0.3285
 Knox vs. O'Neill           1     1.81500000     1.81500000      1.10   0.3143


                                                  Standard
  Parameter                 Estimate                 Error    t Value    Pr > |t|

  Intercept              29.35714286 B      0.83970354      34.96      <.0001
  Block     1             3.32857143 B      0.68561507       4.85      0.0004
  Block     2             1.90000000 B      0.68561507       2.77      0.0169
  Block     3             0.00000000 B       .                .         .
  Type      Clarion       1.06666667 B      1.04729432       1.02      0.3285
  Type      Clinton      -0.80000000 B      1.04729432      -0.76      0.4597
  Type      Knox          3.80000000 B      1.04729432       3.63      0.0035
  Type      O'Neill       2.70000000 B      1.04729432       2.58      0.0242
  Type      Compost      -1.43333333 B      1.04729432      -1.37      0.1962
  Type      Wabash        4.86666667 B      1.04729432       4.65      0.0006
  Type      Webster       0.00000000 B       .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

Output 30.1.3 shows the tests for contrasts that you specified as well as the estimated parameters. The contrast label, degrees of freedom, sum of squares, Mean Square, F Value, and Pr > F are shown for each contrast requested. In this example, the contrast results show that at the 5% significance level,

- the stem length of plants grown in compost soil is significantly different from the average stem length of plants grown in other soils
- the stem length of plants grown in river soils is significantly different from the average stem length of those grown in nonriver soils
- the average stem length of plants grown in glacial soils (Clarion and Webster) is significantly different from the average stem length of those grown in drift soils (Knox and O'Neill)
- stem lengths for Clarion and Webster are not significantly different
- stem lengths for Knox and O'Neill are not significantly different

In addition to the estimates for the parameters of the model, the results of t tests about the parameters are also displayed. The 'B' following the parameter estimates indicates that the estimates are biased and do not represent a unique solution to the normal equations.

Output 30.1.4. Waller-Duncan tests

```
                  Balanced Data from Randomized Complete Block

                               The GLM Procedure

                  Waller-Duncan K-ratio t Test for StemLength

NOTE: This test minimizes the Bayes risk under additive loss and certain other
                                  assumptions.


                    Kratio                                100
                    Error Degrees of Freedom               12
                    Error Mean Square                1.645238
                    F Value                             10.45
                    Critical Value of t               2.12034
                    Minimum Significant Difference   2.2206


           Means with the same letter are not significantly different.


            Waller Grouping          Mean      N    Type

                          A        35.967      3    Wabash
                          A
                          A        34.900      3    Knox
                          A
                     B    A        33.800      3    O’Neill
                     B
                     B    C        32.167      3    Clarion
                          C
                     D    C        31.100      3    Webster
                     D    C
                     D    C        30.300      3    Clinton
                     D
                     D             29.667      3    Compost
```

Output 30.1.5. Ryan-Einot-Gabriel-Welsch Multiple Range Test

```
                Balanced Data from Randomized Complete Block

                             The GLM Procedure

         Ryan-Einot-Gabriel-Welsch Multiple Range Test for StemLength

        NOTE: This test controls the Type I experimentwise error rate.


                     Alpha                            0.05
                     Error Degrees of Freedom           12
                     Error Mean Square            1.645238


Number of Means          2         3         4         5         6         7
Critical Range  2.9876649  3.283833 3.4396257 3.5402242 3.5402242 3.6634222


         Means with the same letter are not significantly different.


             REGWQ Grouping          Mean      N    Type

                     A             35.967      3    Wabash
                     A
                B    A             34.900      3    Knox
                B    A
                B    A    C        33.800      3    O'Neill
                B         C
                B    D    C        32.167      3    Clarion
                     D    C
                     D    C        31.100      3    Webster
                     D
                     D             30.300      3    Clinton
                     D
                     D             29.667      3    Compost
```

The final two pages of output (Output 30.1.4 and Output 30.1.5) present results of the Waller-Duncan and REGWQ multiple comparison procedures. For each test, notes and information pertinent to the test are given on the output. The Type means are arranged from highest to lowest. Means with the same letter are not significantly different. For this example, while some pairs of means are significantly different, there are no clear equivalence classes among the different soils.

Example 30.2. Regression with Mileage Data

A car is tested for gas mileage at various speeds to determine at what speed the car achieves the greatest gas mileage. A quadratic model is fit to the experimental data. The following statements produce Output 30.2.1 through Output 30.2.4:

```
title 'Gasoline Mileage Experiment';
data mileage;
   input mph mpg @@;
   datalines;
20 15.4
30 20.2
40 25.7
50 26.2  50 26.6  50 27.4
55   .
60 24.8
;

proc glm;
   model mpg=mph mph*mph / p clm;
   output out=pp p=mpgpred r=resid;

axis1 minor=none major=(number=5);
axis2 minor=none major=(number=8);
symbol1 c=black i=none   v=plus;
symbol2 c=black i=spline v=none;
proc gplot data=pp;
   plot mpg*mph=1 mpgpred*mph=2 / overlay haxis=axis1
        vaxis=axis2;
run;
```

Output 30.2.1. Standard Regression Analysis Output from PROC GLM

```
                          Gasoline Mileage Experiment

                               The GLM Procedure

                          Number of observations    8

NOTE: Due to missing values, only 7 observations can be used in this analysis.
```

```
                          Gasoline Mileage Experiment

                               The GLM Procedure

Dependent Variable: mpg

                                             Sum of
 Source                      DF         Squares    Mean Square   F Value   Pr > F

 Model                        2     111.8086183     55.9043091     77.96   0.0006

 Error                        4       2.8685246      0.7171311

 Corrected Total              6     114.6771429


               R-Square     Coeff Var      Root MSE      mpg Mean

               0.974986      3.564553      0.846836      23.75714


 Source                      DF       Type I SS    Mean Square   F Value   Pr > F

 mph                          1     85.64464286    85.64464286    119.43   0.0004
 mph*mph                      1     26.16397541    26.16397541     36.48   0.0038


 Source                      DF     Type III SS    Mean Square   F Value   Pr > F

 mph                          1     41.01171219    41.01171219     57.19   0.0016
 mph*mph                      1     26.16397541    26.16397541     36.48   0.0038


                                               Standard
       Parameter         Estimate             Error    t Value    Pr > |t|

       Intercept     -5.985245902       3.18522249      -1.88      0.1334
       mph            1.305245902       0.17259876       7.56      0.0016
       mph*mph       -0.013098361       0.00216852      -6.04      0.0038
```

The overall F statistic is significant. The tests of mph and mph*mph in the Type I sums of squares show that both the linear and quadratic terms in the regression model are significant. The model fits well, with an R^2 of 0.97. The table of parameter estimates indicates that the estimated regression equation is

$$\text{mpg} \quad = \quad -5.9852 + 1.3052 \times \text{mph} - 0.0131 \times \text{mph}^2$$

Output 30.2.2. Results of Requesting the P and CLM Options

```
                          Gasoline Mileage Experiment

                               The GLM Procedure

     Observation             Observed          Predicted           Residual

               1          15.40000000        14.88032787         0.51967213
               2          20.20000000        21.38360656        -1.18360656
               3          25.70000000        25.26721311         0.43278689
               4          26.20000000        26.53114754        -0.33114754
               5          26.60000000        26.53114754         0.06885246
               6          27.40000000        26.53114754         0.86885246
               7 *          .                26.18073770          .
               8          24.80000000        25.17540984        -0.37540984

                                    95% Confidence Limits for
                Observation            Mean Predicted Value

                          1          12.69701317     17.06364257
                          2          20.01727192     22.74994119
                          3          23.87460041     26.65982582
                          4          25.44573423     27.61656085
                          5          25.44573423     27.61656085
                          6          25.44573423     27.61656085
                          7 *        24.88679308     27.47468233
                          8          23.05954977     27.29126990


* Observation was not used in this analysis
```

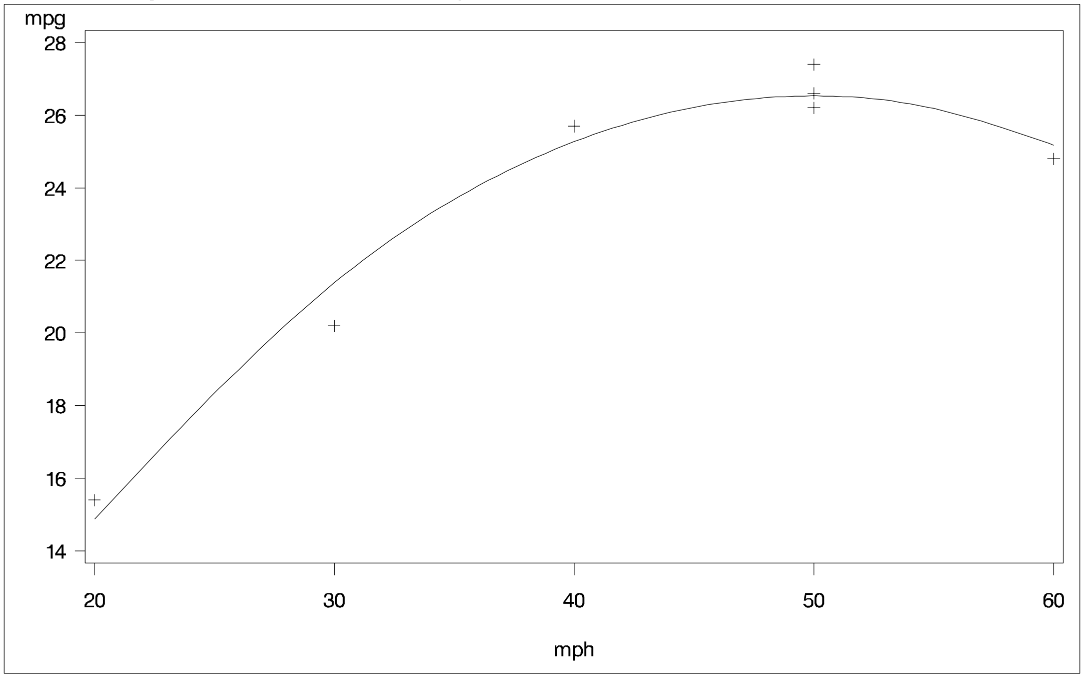
The P and CLM options in the MODEL statement produce the table shown in Output 30.2.2. For each observation, the observed, predicted, and residual values are shown. In addition, the 95% confidence limits for a mean predicted value are shown for each observation. Note that the observation with a missing value for mph is not used in the analysis, but predicted and confidence limit values are shown.

Output 30.2.3. Additional Results of Requesting the P and CLM Options

```
                          Gasoline Mileage Experiment

                               The GLM Procedure

              Sum of Residuals                         -0.00000000
              Sum of Squared Residuals                  2.86852459
              Sum of Squared Residuals - Error SS      -0.00000000
              PRESS Statistic                          23.18107335
              First Order Autocorrelation              -0.54376613
              Durbin-Watson D                           2.94425592
```

The final portion of output gives some additional information on the residuals. The Press statistic gives the sum of squares of predicted residual errors, as described in Chapter 3, “Introduction to Regression Procedures.” The First Order Autocorrelation and the Durbin-Watson *D* statistic, which measures first-order autocorrelation, are also given.

Output 30.2.4. Plot of Mileage Data



Output 30.2.4 shows the actual and predicted values for the data. The quadratic relationship between mpg and mph is evident.

Example 30.3. Unbalanced ANOVA for Two-Way Design with Interaction

This example uses data from Kutner (1974, p. 98) to illustrate a two-way analysis of variance. The original data source is Afifi and Azen (1972, p. 166). These statements produce Output 30.3.1.

```
/*---------------------------------------------------------*/
/* Note: Kutner’s 24 for drug 2, disease 1 changed to 34.  */
/*---------------------------------------------------------*/
title ’Unbalanced Two-Way Analysis of Variance’;
data a;
   input drug disease @;
   do i=1 to 6;
      input y @;
      output;
   end;
   datalines;
1 1 42 44 36 13 19 22
1 2 33  . 26  . 33 21
1 3 31 -3  . 25 25 24
2 1 28  . 23 34 42 13
2 2  . 34 33 31  . 36
2 3  3 26 28 32  4 16
3 1  .  .  1 29  . 19
3 2  . 11  9  7  1 -6
3 3 21  1  .  9  3  .
4 1 24  .  9 22 -2 15
```

```
4 2 27 12 12 -5 16 15
4 3 22  7 25  5 12  .
;

proc glm;
   class drug disease;
   model y=drug disease drug*disease / ss1 ss2 ss3 ss4;
run;
```

Output 30.3.1. Unbalanced ANOVA for Two-Way Design with Interaction

```
                    Unbalanced Two-Way Analysis of Variance

                               The GLM Procedure

                            Class Level Information

                        Class         Levels    Values

                        drug               4    1 2 3 4

                        disease            3    1 2 3


                          Number of observations    72

NOTE: Due to missing values, only 58 observations can be used in this analysis.
```

```
                    Unbalanced Two-Way Analysis of Variance

                               The GLM Procedure

Dependent Variable: y

                                          Sum of
 Source                      DF          Squares    Mean Square   F Value   Pr > F

 Model                       11      4259.338506     387.212591      3.51   0.0013

 Error                       46      5080.816667     110.452536

 Corrected Total             57      9340.155172


               R-Square     Coeff Var      Root MSE        y Mean

               0.456024      55.66750      10.50964      18.87931


 Source                      DF        Type I SS    Mean Square   F Value   Pr > F

 drug                         3      3133.238506    1044.412835      9.46   <.0001
 disease                      2       418.833741     209.416870      1.90   0.1617
 drug*disease                 6       707.266259     117.877710      1.07   0.3958


 Source                      DF       Type II SS    Mean Square   F Value   Pr > F

 drug                         3      3063.432863    1021.144288      9.25   <.0001
 disease                      2       418.833741     209.416870      1.90   0.1617
 drug*disease                 6       707.266259     117.877710      1.07   0.3958


 Source                      DF      Type III SS    Mean Square   F Value   Pr > F

 drug                         3      2997.471860     999.157287      9.05   <.0001
 disease                      2       415.873046     207.936523      1.88   0.1637
 drug*disease                 6       707.266259     117.877710      1.07   0.3958


 Source                      DF       Type IV SS    Mean Square   F Value   Pr > F

 drug                         3      2997.471860     999.157287      9.05   <.0001
 disease                      2       415.873046     207.936523      1.88   0.1637
 drug*disease                 6       707.266259     117.877710      1.07   0.3958
```

Note the differences between the four types of sums of squares. The Type I sum of squares for drug essentially tests for differences between the expected values of the arithmetic mean response for different drugs, unadjusted for the effect of disease. By contrast, the Type II sum of squares for drug measure the differences between arithmetic means for each drug after adjusting for disease. The Type III sum of squares measures the differences between predicted drug means over a balanced drug×disease population—that is, between the LS-means for drug. Finally, the Type IV sum of squares is the same as the Type III sum of squares in this case, since there is data for every drug-by-disease combination.

No matter which sum of squares you prefer to use, this analysis shows a significant difference among the four drugs, while the disease effect and the drug-by-disease interaction are not significant. As the previous discussion indicates, Type III sums of squares correspond to differences between LS-means, so you can follow up the Type III tests with a multiple comparisons analysis of the drug LS-means. Since the GLM procedure is interactive, you can accomplish this by submitting the following statements after the previous ones that performed the ANOVA.

```
   lsmeans drug / pdiff=all adjust=tukey;
run;
```

Both the LS-means themselves and a matrix of adjusted p-values for pairwise differences between them are displayed; see Output 30.3.2.

Output 30.3.2. LS-Means for Unbalanced ANOVA

```
                    Unbalanced Two-Way Analysis of Variance

                               The GLM Procedure
                              Least Squares Means
               Adjustment for Multiple Comparisons: Tukey-Kramer

                                                     LSMEAN
                         drug        y LSMEAN        Number

                         1         25.9944444             1
                         2         26.5555556             2
                         3          9.7444444             3
                         4         13.5444444             4
```

```
                    Unbalanced Two-Way Analysis of Variance

                               The GLM Procedure
                              Least Squares Means
               Adjustment for Multiple Comparisons: Tukey-Kramer

                       Least Squares Means for effect drug
                       Pr > |t| for H0: LSMean(i)=LSMean(j)

                              Dependent Variable: y

          i/j              1             2             3             4

             1                      0.9989        0.0016        0.0107
             2        0.9989                      0.0011        0.0071
             3        0.0016        0.0011                      0.7870
             4        0.0107        0.0071        0.7870
```

The multiple comparisons analysis shows that drugs 1 and 2 have very similar effects, and that drugs 3 and 4 are also insignificantly different from each other. Evidently, the main contribution to the significant drug effect is the difference between the 1/2 pair and the 3/4 pair.

Example 30.4. Analysis of Covariance

Analysis of covariance combines some of the features of both regression and analysis of variance. Typically, a continuous variable (the covariate) is introduced into the model of an analysis-of-variance experiment.

Data in the following example are selected from a larger experiment on the use of drugs in the treatment of leprosy (Snedecor and Cochran 1967, p. 422).

Variables in the study are

Drug	- two antibiotics (A and D) and a control (F)
PreTreatment	- a pre-treatment score of leprosy bacilli
PostTreatment	- a post-treatment score of leprosy bacilli

Ten patients are selected for each treatment (Drug), and six sites on each patient are measured for leprosy bacilli.

The covariate (a pretreatment score) is included in the model for increased precision in determining the effect of drug treatments on the posttreatment count of bacilli.

The following code creates the data set, performs a parallel-slopes analysis of covariance with PROC GLM, and computes Drug LS-means. These statements produce Output 30.4.1.

```
data drugtest;
   input Drug $ PreTreatment PostTreatment @@;
   datalines;
A 11  6   A  8  0   A  5  2   A 14  8   A 19 11
A  6  4   A 10 13   A  6  1   A 11  8   A  3  0
D  6  0   D  6  2   D  7  3   D  8  1   D 18 18
D  8  4   D 19 14   D  8  9   D  5  1   D 15  9
F 16 13   F 13 10   F 11 18   F  9  5   F 21 23
F 16 12   F 12  5   F 12 16   F  7  1   F 12 20
;

proc glm;
   class Drug;
   model PostTreatment = Drug PreTreatment / solution;
   lsmeans Drug / stderr pdiff cov out=adjmeans;
run;

proc print data=adjmeans;
run;
```

Output 30.4.1. Overall Analysis of Variance

```
                              The GLM Procedure

                           Class Level Information

                        Class         Levels    Values

                        Drug               3    A D F


                         Number of observations    30
```

```
                              The GLM Procedure

Dependent Variable: PostTreatment

                                             Sum of
 Source                        DF          Squares    Mean Square   F Value   Pr > F

 Model                          3     871.497403     290.499134     18.10   <.0001

 Error                         26     417.202597      16.046254

 Corrected Total               29    1288.700000


           R-Square     Coeff Var      Root MSE    PostTreatment Mean

           0.676261      50.70604      4.005778              7.900000
```

This model assumes that the slopes relating posttreatment scores to pretreatment scores are parallel for all drugs. You can check this assumption by including the class-by-covariate interaction, Drug*PreTreatment, in the model and examining the ANOVA test for the significance of this effect. This extra test is omitted in this example, but it is insignificant, justifying the equal-slopes assumption.

In Output 30.4.2, the Type I SS for Drug (293.6) gives the between-drug sums of squares that are obtained for the analysis-of-variance model PostTreatment=Drug. This measures the difference between arithmetic means of posttreatment scores for different drugs, disregarding the covariate. The Type III SS for Drug (68.5537) gives the Drug sum of squares adjusted for the covariate. This measures the differences between Drug LS-means, controlling for the covariate. The Type I test is highly significant ($p = 0.001$), but the Type III test is not. This indicates that, while there is a statistically significant difference between the arithmetic drug means, this difference is reduced to below the level of background noise when you take the pretreatment scores into account. From the table of parameter estimates, you can derive the least-squares predictive formula model for estimating posttreatment score based on pretreatment score and drug.

$$\text{post} \;=\; \begin{cases} (-0.435 + -3.446) & + \;\; 0.987 \cdot \text{pre}, \quad \text{if Drug=A} \\ (-0.435 + -3.337) & + \;\; 0.987 \cdot \text{pre}, \quad \text{if Drug=D} \\ \;\; -0.435 & + \;\; 0.987 \cdot \text{pre}, \quad \text{if Drug=F} \end{cases}$$

Output 30.4.2. Tests and Parameter Estimates

```
                               The GLM Procedure

Dependent Variable: PostTreatment

 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 Drug                         2    293.6000000    146.8000000      9.15   0.0010
 PreTreatment                 1    577.8974030    577.8974030     36.01   <.0001


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 Drug                         2     68.5537106     34.2768553      2.14   0.1384
 PreTreatment                 1    577.8974030    577.8974030     36.01   <.0001


                                            Standard
    Parameter              Estimate            Error    t Value    Pr > |t|

    Intercept          -0.434671164 B      2.47135356      -0.18      0.8617
    Drug         A     -3.446138280 B      1.88678065      -1.83      0.0793
    Drug         D     -3.337166948 B      1.85386642      -1.80      0.0835
    Drug         F      0.000000000 B       .                .         .
    PreTreatment        0.987183811        0.16449757       6.00      <.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

Output 30.4.3 displays the LS-means, which are, in a sense, the means adjusted for the covariate. The STDERR option in the LSMEANS statement causes the standard error of the LS-means and the probability of getting a larger t value under the hypothesis $H_0\colon$LS-mean $= 0$ to be included in this table as well. Specifying the PDIFF option causes all probability values for the hypothesis $H_0\colon$LS-mean$(i) =$ LS-mean(j) to be displayed, where the indexes i and j are numbered treatment levels.

Output 30.4.3. LS-means

```
                               The GLM Procedure
                              Least Squares Means

                          Post
                     Treatment        Standard                  LSMEAN
          Drug          LSMEAN           Error    Pr > |t|      Number

          A          6.7149635       1.2884943      <.0001           1
          D          6.8239348       1.2724690      <.0001           2
          F         10.1611017       1.3159234      <.0001           3


                      Least Squares Means for effect Drug
                      Pr > |t| for H0: LSMean(i)=LSMean(j)

                       Dependent Variable: PostTreatment

                 i/j              1             2             3

                    1                      0.9521        0.0793
                    2        0.9521                      0.0835
                    3        0.0793        0.0835


NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

The OUT= and COV options in the LSMEANS statement create a data set of the estimates, their standard errors, and the variances and covariances of the LS-means, which is displayed in Output 30.4.4

Output 30.4.4. LS-means Output Data Set

```
Obs     _NAME_      Drug   LSMEAN   STDERR  NUMBER    COV1      COV2      COV3

 1   PostTreatment   A     6.7150  1.28849     1     1.66022   0.02844  -0.08403
 2   PostTreatment   D     6.8239  1.27247     2     0.02844   1.61918  -0.04299
 3   PostTreatment   F    10.1611  1.31592     3    -0.08403  -0.04299   1.73165
```

Example 30.5. Three-Way Analysis of Variance with Contrasts

This example uses data from Cochran and Cox (1957, p. 176) to illustrate the analysis of a three-way factorial design with replication, including the use of the CONTRAST statement with interactions, the OUTSTAT= data set, and the SLICE= option in the LSMEANS statement.

The object of the study is to determine the effects of electric current on denervated muscle. The variables are

Rep — the replicate number, 1 or 2

Time — the length of time the current is applied to the muscle, ranging from 1 to 4

Current — the level of electric current applied, ranging from 1 to 4

Number — the number of treatments per day, ranging from 1 to 3

MuscleWeight — the weight of the denervated muscle

The following code produces Output 30.5.1 through Output 30.5.4.

```
data muscles;
   do Rep=1 to 2;
      do Time=1 to 4;
         do Current=1 to 4;
            do Number=1 to 3;
               input MuscleWeight @@;
               output;
            end;
         end;
      end;
   end;
   datalines;
72 74 69 61 61 65 62 65 70 85 76 61
67 52 62 60 55 59 64 65 64 67 72 60
57 66 72 72 43 43 63 66 72 56 75 92
57 56 78 60 63 58 61 79 68 73 86 71
46 74 58 60 64 52 71 64 71 53 65 66
44 58 54 57 55 51 62 61 79 60 78 82
53 50 61 56 57 56 56 56 71 56 58 69
46 55 64 56 55 57 64 66 62 59 58 88
;

proc glm outstat=summary;
   class Rep Current Time Number;
   model MuscleWeight = Rep Current|Time|Number;
   contrast ’Time in Current 3’
      Time 1 0 0 -1 Current*Time 0 0 0 0 0 0 0 0 1 0 0 -1,
      Time 0 1 0 -1 Current*Time 0 0 0 0 0 0 0 0 0 1 0 -1,
      Time 0 0 1 -1 Current*Time 0 0 0 0 0 0 0 0 0 0 1 -1;
   contrast ’Current 1 versus 2’ Current 1 -1;
   lsmeans Current*Time / slice=Current;
run;

proc print data=summary;
run;
```

The first CONTRAST statement examines the effects of Time within level 3 of Current. This is also called the *simple effect* of Time within Current*Time. Note that, since there are three degrees of freedom, it is necessary to specify three rows in the CONTRAST statement, separated by commas. Since the parameterization that PROC GLM uses is determined in part by the ordering of the variables in the CLASS statement, Current is specified before Time so that the Time parameters are nested within the Current*Time parameters; thus, the Current*Time contrast coefficients in each row are simply the Time coefficients of that row within the appropriate level of Current.

The second CONTRAST statement isolates a single degree of freedom effect corresponding to the difference between the first two levels of Current. You can use such a contrast in a large experiment where certain preplanned comparisons are important, but you want to take advantage of the additional error degrees of freedom available when all levels of the factors are considered.

The LSMEANS statement with the SLICE= option is an alternative way to test for the simple effect of Time within Current*Time. In addition to listing the LS-means for each current strength and length of time, it gives a table of F-tests for differences between the LS-means across Time within each Current level. In some cases, this can be a way to disentangle a complex interaction.

Output 30.5.1. Overall Analysis

```
                               The GLM Procedure

                            Class Level Information

                        Class         Levels    Values

                        Rep                2    1 2

                        Current            4    1 2 3 4

                        Time               4    1 2 3 4

                        Number             3    1 2 3


                          Number of observations    96



                               The GLM Procedure

Dependent Variable: MuscleWeight

                                       Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F

 Model                      48    5782.916667     120.477431      1.77   0.0261

 Error                      47    3199.489583      68.074246

 Corrected Total            95    8982.406250


           R-Square     Coeff Var      Root MSE    MuscleWeight Mean

           0.643805      13.05105      8.250712             63.21875
```

The output, shown in Output 30.5.2 and Output 30.5.3, indicates that the main effects for Rep, Current, and Number are significant (with p-values of 0.0045, <0.0001, and 0.0461, respectively), but Time is not significant, indicating that, in general, it doesn't matter how long the current is applied. None of the interaction terms are significant, nor are the contrasts significant. Notice that the row in the sliced ANOVA table corresponding to level 3 of current matches the "Time in Current 3" contrast.

Output 30.5.2. Individual Effects and Contrasts

```
                               The GLM Procedure

Dependent Variable: MuscleWeight

 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 Rep                          1     605.010417     605.010417      8.89   0.0045
 Current                      3    2145.447917     715.149306     10.51   <.0001
 Time                         3     223.114583      74.371528      1.09   0.3616
 Current*Time                 9     298.677083      33.186343      0.49   0.8756
 Number                       2     447.437500     223.718750      3.29   0.0461
 Current*Number               6     644.395833     107.399306      1.58   0.1747
 Time*Number                  6     367.979167      61.329861      0.90   0.5023
 Current*Time*Number         18    1050.854167      58.380787      0.86   0.6276


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 Rep                          1     605.010417     605.010417      8.89   0.0045
 Current                      3    2145.447917     715.149306     10.51   <.0001
 Time                         3     223.114583      74.371528      1.09   0.3616
 Current*Time                 9     298.677083      33.186343      0.49   0.8756
 Number                       2     447.437500     223.718750      3.29   0.0461
 Current*Number               6     644.395833     107.399306      1.58   0.1747
 Time*Number                  6     367.979167      61.329861      0.90   0.5023
 Current*Time*Number         18    1050.854167      58.380787      0.86   0.6276


 Contrast                    DF    Contrast SS    Mean Square   F Value   Pr > F

 Time in Current 3            3    34.83333333    11.61111111      0.17   0.9157
 Current 1 versus 2           1    99.18750000    99.18750000      1.46   0.2334
```

Output 30.5.3. Simple Effects of Time

```
                               The GLM Procedure
                              Least Squares Means

             Current*Time Effect Sliced by Current for MuscleWeight

                                 Sum of
     Current        DF          Squares     Mean Square    F Value    Pr > F

     1               3       271.458333       90.486111       1.33    0.2761
     2               3       120.666667       40.222222       0.59    0.6241
     3               3        34.833333       11.611111       0.17    0.9157
     4               3        94.833333       31.611111       0.46    0.7085
```

The SS, F statistics, and p-values can be stored in an OUTSTAT= data set, as shown in Output 30.5.4.

Output 30.5.4. Contents of the OUTSTAT= Data Set

```
Obs     _NAME_     _SOURCE_             _TYPE_    DF     SS       F       PROB

  1  MuscleWeight  ERROR                ERROR     47  3199.49    .       .
  2  MuscleWeight  Rep                  SS1        1   605.01   8.8875  0.00454
  3  MuscleWeight  Current              SS1        3  2145.45  10.5054  0.00002
  4  MuscleWeight  Time                 SS1        3   223.11   1.0925  0.36159
  5  MuscleWeight  Current*Time         SS1        9   298.68   0.4875  0.87562
  6  MuscleWeight  Number               SS1        2   447.44   3.2864  0.04614
  7  MuscleWeight  Current*Number       SS1        6   644.40   1.5777  0.17468
  8  MuscleWeight  Time*Number          SS1        6   367.98   0.9009  0.50231
  9  MuscleWeight  Current*Time*Number  SS1       18  1050.85   0.8576  0.62757
 10  MuscleWeight  Rep                  SS3        1   605.01   8.8875  0.00454
 11  MuscleWeight  Current              SS3        3  2145.45  10.5054  0.00002
 12  MuscleWeight  Time                 SS3        3   223.11   1.0925  0.36159
 13  MuscleWeight  Current*Time         SS3        9   298.68   0.4875  0.87562
 14  MuscleWeight  Number               SS3        2   447.44   3.2864  0.04614
 15  MuscleWeight  Current*Number       SS3        6   644.40   1.5777  0.17468
 16  MuscleWeight  Time*Number          SS3        6   367.98   0.9009  0.50231
 17  MuscleWeight  Current*Time*Number  SS3       18  1050.85   0.8576  0.62757
 18  MuscleWeight  Time in Current 3    CONTRAST   3    34.83   0.1706  0.91574
 19  MuscleWeight  Current 1 versus 2   CONTRAST   1    99.19   1.4570  0.23344
```

Example 30.6. Multivariate Analysis of Variance

The following example employs multivariate analysis of variance (MANOVA) to measure differences in the chemical characteristics of ancient pottery found at four kiln sites in Great Britain. The data are from Tubb, Parker, and Nickless (1980), as reported in Hand et al. (1994).

For each of 26 samples of pottery, the percentages of oxides of five metals are measured. The following statements create the data set and invoke the GLM procedure to perform a one-way MANOVA. Additionally, it is of interest to know whether the pottery from one site in Wales (Llanederyn) differs from the samples from other sites; a CONTRAST statement is used to test this hypothesis.

```
data pottery;
   title1 "Romano-British Pottery";
   input Site $12. Al Fe Mg Ca Na;
   datalines;
Llanederyn   14.4 7.00 4.30 0.15 0.51
Llanederyn   13.8 7.08 3.43 0.12 0.17
Llanederyn   14.6 7.09 3.88 0.13 0.20
Llanederyn   11.5 6.37 5.64 0.16 0.14
Llanederyn   13.8 7.06 5.34 0.20 0.20
Llanederyn   10.9 6.26 3.47 0.17 0.22
Llanederyn   10.1 4.26 4.26 0.20 0.18
Llanederyn   11.6 5.78 5.91 0.18 0.16
Llanederyn   11.1 5.49 4.52 0.29 0.30
Llanederyn   13.4 6.92 7.23 0.28 0.20
Llanederyn   12.4 6.13 5.69 0.22 0.54
Llanederyn   13.1 6.64 5.51 0.31 0.24
Llanederyn   12.7 6.69 4.45 0.20 0.22
Llanederyn   12.5 6.44 3.94 0.22 0.23
Caldicot     11.8 5.44 3.94 0.30 0.04
Caldicot     11.6 5.39 3.77 0.29 0.06
IslandThorns 18.3 1.28 0.67 0.03 0.03
IslandThorns 15.8 2.39 0.63 0.01 0.04
IslandThorns 18.0 1.50 0.67 0.01 0.06
IslandThorns 18.0 1.88 0.68 0.01 0.04
IslandThorns 20.8 1.51 0.72 0.07 0.10
AshleyRails  17.7 1.12 0.56 0.06 0.06
AshleyRails  18.3 1.14 0.67 0.06 0.05
AshleyRails  16.7 0.92 0.53 0.01 0.05
AshleyRails  14.8 2.74 0.67 0.03 0.05
AshleyRails  19.1 1.64 0.60 0.10 0.03
;
proc glm data=pottery;
   class Site;
   model Al Fe Mg Ca Na = Site;
   contrast ’Llanederyn vs. the rest’ Site 1 1 1 -3;
   manova h=_all_ / printe printh;
run;
```

After the summary information, displayed in Output 30.6.1, PROC GLM produces the univariate analyses for each of the dependent variables, as shown in Output 30.6.2. These analyses show that sites are significantly different for all oxides individually. You can suppress these univariate analyses by specifying the NOUNI option in the MODEL statement.

Output 30.6.1. Summary Information on Groups

```
                        Romano-British Pottery

                          The GLM Procedure

                       Class Level Information

Class         Levels    Values

Site               4    AshleyRails Caldicot IslandThorns Llanederyn


                     Number of observations    26
```

Output 30.6.2. Univariate Analysis of Variance for Each Dependent

```
                             Romano-British Pottery

                               The GLM Procedure

Dependent Variable: Al

                                       Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F

 Model                       3    175.6103187     58.5367729     26.67   <.0001

 Error                      22     48.2881429      2.1949156

 Corrected Total            25    223.8984615


               R-Square     Coeff Var      Root MSE       Al Mean

               0.784330      10.22284      1.481525      14.49231


 Source                     DF      Type I SS    Mean Square   F Value   Pr > F

 Site                        3    175.6103187     58.5367729     26.67   <.0001


 Source                     DF    Type III SS    Mean Square   F Value   Pr > F

 Site                        3    175.6103187     58.5367729     26.67   <.0001


  Contrast                     DF   Contrast SS   Mean Square  F Value  Pr > F

  Llanederyn vs. the rest       1   58.58336640   58.58336640    26.69  <.0001
```

```
                             Romano-British Pottery

                               The GLM Procedure

Dependent Variable: Fe

                                           Sum of
 Source                      DF          Squares    Mean Square   F Value   Pr > F

 Model                        3      134.2216158     44.7405386     89.88   <.0001

 Error                       22       10.9508457      0.4977657

 Corrected Total             25      145.1724615


               R-Square     Coeff Var      Root MSE       Fe Mean

               0.924567      15.79171      0.705525      4.467692


 Source                      DF        Type I SS    Mean Square   F Value   Pr > F

 Site                         3      134.2216158     44.7405386     89.88   <.0001


 Source                      DF      Type III SS    Mean Square   F Value   Pr > F

 Site                         3      134.2216158     44.7405386     89.88   <.0001


  Contrast                       DF    Contrast SS    Mean Square  F Value  Pr > F

  Llanederyn vs. the rest         1    71.15144132    71.15144132   142.94  <.0001
```

```
                             Romano-British Pottery

                               The GLM Procedure

Dependent Variable: Mg

                                       Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F

 Model                       3    103.3505270     34.4501757     49.12   <.0001

 Error                      22     15.4296114      0.7013460

 Corrected Total            25    118.7801385


               R-Square     Coeff Var      Root MSE       Mg Mean

               0.870099      26.65777      0.837464      3.141538


 Source                     DF      Type I SS    Mean Square   F Value   Pr > F

 Site                        3    103.3505270     34.4501757     49.12   <.0001


 Source                     DF    Type III SS    Mean Square   F Value   Pr > F

 Site                        3    103.3505270     34.4501757     49.12   <.0001


  Contrast                     DF   Contrast SS   Mean Square  F Value  Pr > F

  Llanederyn vs. the rest       1   56.59349339   56.59349339    80.69  <.0001
```

```
                             Romano-British Pottery

                               The GLM Procedure

Dependent Variable: Ca

                                        Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                        3     0.20470275     0.06823425     29.16   <.0001

 Error                       22     0.05148571     0.00234026

 Corrected Total             25     0.25618846


               R-Square     Coeff Var      Root MSE       Ca Mean

               0.799032      33.01265      0.048376      0.146538


 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 Site                         3     0.20470275     0.06823425     29.16   <.0001


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 Site                         3     0.20470275     0.06823425     29.16   <.0001


  Contrast                      DF    Contrast SS   Mean Square  F Value  Pr > F

  Llanederyn vs. the rest        1     0.03531688    0.03531688    15.09  0.0008
```

```
                             Romano-British Pottery

                               The GLM Procedure

Dependent Variable: Na

                                           Sum of
 Source                     DF          Squares    Mean Square   F Value   Pr > F

 Model                       3       0.25824560     0.08608187      9.50   0.0003

 Error                      22       0.19929286     0.00905877

 Corrected Total            25       0.45753846


               R-Square     Coeff Var      Root MSE       Na Mean

               0.564424      60.06350      0.095178      0.158462


 Source                     DF        Type I SS    Mean Square   F Value   Pr > F

 Site                        3       0.25824560     0.08608187      9.50   0.0003


 Source                     DF      Type III SS    Mean Square   F Value   Pr > F

 Site                        3       0.25824560     0.08608187      9.50   0.0003


  Contrast                     DF    Contrast SS    Mean Square  F Value  Pr > F

  Llanederyn vs. the rest       1     0.23344446     0.23344446    25.77  <.0001
```

The PRINTE option in the MANOVA statement displays the elements of the error matrix, also called the Error Sums of Squares and Crossproducts matrix. See Output 30.6.3. The diagonal elements of this matrix are the error sums of squares from the corresponding univariate analyses.

The PRINTE option also displays the partial correlation matrix associated with the E matrix. In this example, none of the oxides are very strongly correlated; the strongest correlation ($r = 0.488$) is between magnesium oxide and calcium oxide.

Output 30.6.3. Error SSCP Matrix and Partial Correlations

```
                              Romano-British Pottery

                                The GLM Procedure
                         Multivariate Analysis of Variance

                              E = Error SSCP Matrix

              Al              Fe              Mg              Ca              Na

Al    48.288142857    7.0800714286    0.6080142857    0.1064714286    0.5889571429
Fe    7.0800714286    10.950845714    0.5270571429    -0.155194286    0.0667585714
Mg    0.6080142857    0.5270571429    15.429611429    0.4353771429    0.0276157143
Ca    0.1064714286    -0.155194286    0.4353771429    0.0514857143    0.0100785714
Na    0.5889571429    0.0667585714    0.0276157143    0.0100785714    0.1992928571


  Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 22             Al             Fe             Mg             Ca             Na

Al            1.000000       0.307889       0.022275       0.067526       0.189853
                             0.1529         0.9196         0.7595         0.3856

Fe            0.307889       1.000000       0.040547      -0.206685       0.045189
                0.1529                      0.8543         0.3440         0.8378

Mg            0.022275       0.040547       1.000000       0.488478       0.015748
                0.9196         0.8543                      0.0180         0.9431

Ca            0.067526      -0.206685       0.488478       1.000000       0.099497
                0.7595         0.3440         0.0180                      0.6515

Na            0.189853       0.045189       0.015748       0.099497       1.000000
                0.3856         0.8378         0.9431         0.6515
```

The PRINTH option produces the SSCP matrix for the hypotheses being tested (Site and the contrast); see Output 30.6.3. Since the Type III SS are the highest level SS produced by PROC GLM by default, and since the HTYPE= option is not specified, the SSCP matrix for Site gives the Type III $\mathbf{H}$ matrix. The diagonal elements of this matrix are the model sums of squares from the corresponding univariate analyses.

Four multivariate tests are computed, all based on the characteristic roots and vectors of $\mathbf{E}^{-1}\mathbf{H}$. These roots and vectors are displayed along with the tests. All four tests can be transformed to variates that have F distributions under the null hypothesis. Note that the four tests all give the same results for the contrast, since it has only one degree of freedom. In this case, the multivariate analysis matches the univariate results: there is an overall difference between the chemical composition of samples from different sites, and the samples from Llanederyn are different from the average of the other sites.

Output 30.6.4. Hypothesis SSCP Matrix and Multivariate Tests

```
                                    Romano-British Pottery

                                      The GLM Procedure
                              Multivariate Analysis of Variance

                              H = Type III SSCP Matrix for Site

                   Al                Fe                Mg                Ca                Na

 Al      175.61031868       -149.295533      -130.8097066      -5.889163736      -5.372264835
 Fe       -149.295533      134.22161582      117.74503516      4.8217865934      5.3259491209
 Mg      -130.8097066      117.74503516      103.35052703      4.2091613187      4.7105458242
 Ca      -5.889163736      4.8217865934      4.2091613187      0.2047027473       0.154782967
 Na      -5.372264835      5.3259491209      4.7105458242       0.154782967      0.2582456044


                  Characteristic Roots and Vectors of: E Inverse * H, where
                              H = Type III SSCP Matrix for Site
                                    E = Error SSCP Matrix

Characteristic           Characteristic Vector  V’EV=1
          Root  Percent            Al            Fe            Mg            Ca            Na

    34.1611140    96.39    0.09562211   -0.26330469   -0.05305978   -1.87982100   -0.47071123
     1.2500994     3.53    0.02651891   -0.01239715    0.17564390   -4.25929785    1.23727668
     0.0275396     0.08    0.09082220    0.13159869    0.03508901   -0.15701602   -1.39364544
     0.0000000     0.00    0.03673984   -0.15129712    0.20455529    0.54624873   -0.17402107
     0.0000000     0.00    0.06862324    0.03056912   -0.10662399    2.51151978    1.23668841


    MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Site Effect
                              H = Type III SSCP Matrix for Site
                                    E = Error SSCP Matrix

                                     S=3    M=0.5    N=8

       Statistic                        Value    F Value    Num DF    Den DF    Pr > F

       Wilks’ Lambda               0.01230091      13.09        15    50.091    <.0001
       Pillai’s Trace              1.55393619       4.30        15        60    <.0001
       Hotelling-Lawley Trace     35.43875302      40.59        15     29.13    <.0001
       Roy’s Greatest Root        34.16111399     136.64         5        20    <.0001

                 NOTE: F Statistic for Roy’s Greatest Root is an upper bound.
```

```
                                 Romano-British Pottery

                                   The GLM Procedure
                           Multivariate Analysis of Variance

                  H = Contrast SSCP Matrix for Llanederyn vs. the rest

                      Al                 Fe                 Mg                 Ca                 Na

 Al      58.583366402      -64.56230291      -57.57983466      -1.438395503      -3.698102513
 Fe      -64.56230291      71.151441323      63.456352116      1.5851961376      4.0755256878
 Mg      -57.57983466      63.456352116      56.593493386      1.4137558201      3.6347541005
 Ca      -1.438395503      1.5851961376      1.4137558201      0.0353168783      0.0907993915
 Na      -3.698102513      4.0755256878      3.6347541005      0.0907993915      0.2334444577


               Characteristic Roots and Vectors of: E Inverse * H, where
                  H = Contrast SSCP Matrix for Llanederyn vs. the rest
                                 E = Error SSCP Matrix

Characteristic           Characteristic Vector  V'EV=1
          Root  Percent            Al            Fe            Mg            Ca            Na

    16.1251646   100.00   -0.08883488    0.25458141    0.08723574    0.98158668    0.71925759
     0.0000000     0.00   -0.00503538    0.03825743   -0.17632854    5.16256699   -0.01022754
     0.0000000     0.00    0.00162771   -0.08885364   -0.01774069   -0.83096817    2.17644566
     0.0000000     0.00    0.04450136   -0.15722494    0.22156791    0.00000000    0.00000000
     0.0000000     0.00    0.11939206    0.10833549    0.00000000    0.00000000    0.00000000


             MANOVA Test Criteria and Exact F Statistics for the Hypothesis
                      of No Overall Llanederyn vs. the rest Effect
                  H = Contrast SSCP Matrix for Llanederyn vs. the rest
                                 E = Error SSCP Matrix

                                  S=1    M=1.5    N=8

       Statistic                        Value    F Value    Num DF    Den DF    Pr > F

       Wilks' Lambda               0.05839360      58.05         5        18    <.0001
       Pillai's Trace              0.94160640      58.05         5        18    <.0001
       Hotelling-Lawley Trace     16.12516462      58.05         5        18    <.0001
       Roy's Greatest Root        16.12516462      58.05         5        18    <.0001
```

Example 30.7. Repeated Measures Analysis of Variance

This example uses data from Cole and Grizzle (1966) to illustrate a commonly occurring repeated measures ANOVA design. Sixteen dogs are randomly assigned to four groups. (One animal is removed from the analysis due to a missing value for one dependent variable.) Dogs in each group receive either morphine or trimethaphan (variable Drug) and have either depleted or intact histamine levels (variable Depleted) before receiving the drugs. The dependent variable is the blood concentration of histamine at 0, 1, 3, and 5 minutes after injection of the drug. Logarithms are applied to these concentrations to minimize correlation between the mean and the variance of the data.

The following SAS statements perform both univariate and multivariate repeated measures analyses and produce Output 30.7.1 through Output 30.7.7:

```
data dogs;
   input Drug $12. Depleted $ Histamine0 Histamine1
         Histamine3 Histamine5;
   LogHistamine0=log(Histamine0);
   LogHistamine1=log(Histamine1);
   LogHistamine3=log(Histamine3);
   LogHistamine5=log(Histamine5);
   datalines;
Morphine      N  .04  .20  .10  .08
Morphine      N  .02  .06  .02  .02
Morphine      N  .07 1.40  .48  .24
Morphine      N  .17  .57  .35  .24
Morphine      Y  .10  .09  .13  .14
Morphine      Y  .12  .11  .10   .
Morphine      Y  .07  .07  .06  .07
Morphine      Y  .05  .07  .06  .07
Trimethaphan  N  .03  .62  .31  .22
Trimethaphan  N  .03 1.05  .73  .60
Trimethaphan  N  .07  .83 1.07  .80
Trimethaphan  N  .09 3.13 2.06 1.23
Trimethaphan  Y  .10  .09  .09  .08
Trimethaphan  Y  .08  .09  .09  .10
Trimethaphan  Y  .13  .10  .12  .12
Trimethaphan  Y  .06  .05  .05  .05
;
proc glm;
   class Drug Depleted;
   model LogHistamine0--LogHistamine5 =
         Drug Depleted Drug*Depleted / nouni;
   repeated Time 4 (0 1 3 5) polynomial / summary printe;
run;
```

The NOUNI option in the MODEL statement suppresses the individual ANOVA tables for the original dependent variables. These analyses are usually of no interest in a repeated measures analysis. The POLYNOMIAL option in the REPEATED statement indicates that the transformation used to implement the repeated measures analysis is an orthogonal polynomial transformation, and the SUMMARY option requests that the univariate analyses for the orthogonal polynomial contrast variables be displayed. The parenthetical numbers (0 1 3 5) determine the spacing of the orthogonal polynomials used in the analysis. The output is displayed in Output 30.7.1 through Output 30.7.7.

Output 30.7.1. Summary Information on Groups

```
                                       The GLM Procedure

                                    Class Level Information

                         Class         Levels    Values

                         Drug               2    Morphine Trimethaphan

                         Depleted           2    N Y


                                  Number of observations    16

NOTE: Observations with missing values will not be included in this analysis.  Thus, only 15
      observations can be used in this analysis.
```

The “Repeated Measures Level Information” table gives information on the repeated measures effect; it is displayed in Output 30.7.2. In this example, the within-subject (within-dog) effect is Time, which has the levels 0, 1, 3, and 5.

Output 30.7.2. Repeated Measures Levels

```
                        The GLM Procedure
              Repeated Measures Analysis of Variance

               Repeated Measures Level Information

                             Log        Log        Log        Log
Dependent Variable    Histamine0 Histamine1 Histamine3 Histamine5

     Level of Time             0          1          3          5
```

The multivariate analyses for within-subject effects and related interactions are displayed in Output 30.7.3. For the example, the first table displayed shows that the TIME effect is significant. In addition, the Time*Drug*Depleted interaction is significant, as shown in the fourth table. This means that the effect of Time on the blood concentration of histamine is different for the four Drug*Depleted combinations studied.

Output 30.7.3. Multivariate Tests of Within-Subject Effects

```
                                       The GLM Procedure
                               Repeated Measures Analysis of Variance

        Manova Test Criteria and Exact F Statistics for the Hypothesis of no Time Effect
                               H = Type III SSCP Matrix for Time
                                     E = Error SSCP Matrix

                                     S=1    M=0.5    N=3.5

        Statistic                        Value    F Value    Num DF    Den DF    Pr > F

        Wilks' Lambda               0.11097706      24.03         3         9    0.0001
        Pillai's Trace              0.88902294      24.03         3         9    0.0001
        Hotelling-Lawley Trace      8.01087137      24.03         3         9    0.0001
        Roy's Greatest Root         8.01087137      24.03         3         9    0.0001


     Manova Test Criteria and Exact F Statistics for the Hypothesis of no Time*Drug Effect
                            H = Type III SSCP Matrix for Time*Drug
                                     E = Error SSCP Matrix

                                     S=1    M=0.5    N=3.5

        Statistic                        Value    F Value    Num DF    Den DF    Pr > F

        Wilks' Lambda               0.34155984       5.78         3         9    0.0175
        Pillai's Trace              0.65844016       5.78         3         9    0.0175
        Hotelling-Lawley Trace      1.92774470       5.78         3         9    0.0175
        Roy's Greatest Root         1.92774470       5.78         3         9    0.0175


   Manova Test Criteria and Exact F Statistics for the Hypothesis of no Time*Depleted Effect
                          H = Type III SSCP Matrix for Time*Depleted
                                     E = Error SSCP Matrix

                                     S=1    M=0.5    N=3.5

        Statistic                        Value    F Value    Num DF    Den DF    Pr > F

        Wilks' Lambda               0.12339988      21.31         3         9    0.0002
        Pillai's Trace              0.87660012      21.31         3         9    0.0002
        Hotelling-Lawley Trace      7.10373567      21.31         3         9    0.0002
        Roy's Greatest Root         7.10373567      21.31         3         9    0.0002


 Manova Test Criteria and Exact F Statistics for the Hypothesis of no Time*Drug*Depleted Effect
                        H = Type III SSCP Matrix for Time*Drug*Depleted
                                     E = Error SSCP Matrix

                                     S=1    M=0.5    N=3.5

        Statistic                        Value    F Value    Num DF    Den DF    Pr > F

        Wilks' Lambda               0.19383010      12.48         3         9    0.0015
        Pillai's Trace              0.80616990      12.48         3         9    0.0015
        Hotelling-Lawley Trace      4.15915732      12.48         3         9    0.0015
        Roy's Greatest Root         4.15915732      12.48         3         9    0.0015
```

Output 30.7.4 displays tests of hypotheses for between-subject (between-dog) effects. This section tests the hypotheses that the different Drugs, Depleteds, and their interactions have no effects on the dependent variables, while ignoring the within-dog effects. From this analysis, there is a significant between-dog effect for Depleted (p-value=0.0229). The interaction and the main effect for Drug are not significant (p-values=0.1734 and 0.1281, respectively).

Output 30.7.4. Tests of Between-Subject Effects

```
                              The GLM Procedure
                    Repeated Measures Analysis of Variance
                Tests of Hypotheses for Between Subjects Effects

  Source                      DF     Type III SS     Mean Square    F Value    Pr > F

  Drug                         1      5.99336243      5.99336243       2.71    0.1281
  Depleted                     1     15.44840703     15.44840703       6.98    0.0229
  Drug*Depleted                1      4.69087508      4.69087508       2.12    0.1734
  Error                       11     24.34683348      2.21334850
```

Univariate analyses for within-subject (within-dog) effects and related interactions are displayed in Output 30.7.6. The results for this example are the same as for the multivariate analyses; this is not always the case. In addition, before the univariate analyses are used to make conclusions about the data, the result of the sphericity test (requested with the PRINTE option in the REPEATED statement and displayed in Output 30.7.5) should be examined. If the sphericity test is rejected, use the adjusted G-G or H-F probabilities. See the “Repeated Measures Analysis of Variance” section on page 1560 for more information.

Output 30.7.5. Sphericity Test

```
                              The GLM Procedure
                    Repeated Measures Analysis of Variance

                               Sphericity Tests

                                      Mauchly's
        Variables                DF    Criterion    Chi-Square    Pr > ChiSq

        Transformed Variates      5    0.1752641     16.930873        0.0046
        Orthogonal Components     5    0.1752641     16.930873        0.0046
```

Output 30.7.6. Univariate Tests of Within-Subject Effects

```
                              The GLM Procedure
                    Repeated Measures Analysis of Variance
              Univariate Tests of Hypotheses for Within Subject Effects

                                                                          Adj Pr > F
Source                     DF    Type III SS    Mean Square   F Value   Pr > F    G - G    H - F

Time                        3    12.05898677     4.01966226     53.44   <.0001   <.0001   <.0001
Time*Drug                   3     1.84429514     0.61476505      8.17   0.0003   0.0039   0.0008
Time*Depleted               3    12.08978557     4.02992852     53.57   <.0001   <.0001   <.0001
Time*Drug*Depleted          3     2.93077939     0.97692646     12.99   <.0001   0.0005   <.0001
Error(Time)                33     2.48238887     0.07522391


                              Greenhouse-Geisser Epsilon    0.5694
                              Huynh-Feldt Epsilon           0.8475
```

Output 30.7.7 is produced by the SUMMARY option in the REPEATED statement. If the POLYNOMIAL option is not used, a similar table is displayed using the default CONTRAST transformation. The linear, quadratic, and cubic trends for Time, labeled as ‘Time_1’, ‘Time_2’, and ‘Time_3’, are displayed, and in each case, the Source labeled ‘Mean’ gives a test for the respective trend.

Output 30.7.7. Tests of Between-Subject Effects for Transformed Variables

```
                                        The GLM Procedure
                               Repeated Measures Analysis of Variance
                             Analysis of Variance of Contrast Variables

Time_N represents the nth degree polynomial contrast for Time

Contrast Variable: Time_1


      Source                        DF     Type III SS     Mean Square    F Value    Pr > F

      Mean                           1      2.00963483      2.00963483      34.99    0.0001
      Drug                           1      1.18069076      1.18069076      20.56    0.0009
      Depleted                       1      1.36172504      1.36172504      23.71    0.0005
      Drug*Depleted                  1      2.04346848      2.04346848      35.58    <.0001
      Error                         11      0.63171161      0.05742833


Contrast Variable: Time_2


      Source                        DF     Type III SS     Mean Square    F Value    Pr > F

      Mean                           1      5.40988418      5.40988418      57.15    <.0001
      Drug                           1      0.59173192      0.59173192       6.25    0.0295
      Depleted                       1      5.94945506      5.94945506      62.86    <.0001
      Drug*Depleted                  1      0.67031587      0.67031587       7.08    0.0221
      Error                         11      1.04118707      0.09465337


Contrast Variable: Time_3


      Source                        DF     Type III SS     Mean Square    F Value    Pr > F

      Mean                           1      4.63946776      4.63946776      63.04    <.0001
      Drug                           1      0.07187246      0.07187246       0.98    0.3443
      Depleted                       1      4.77860547      4.77860547      64.94    <.0001
      Drug*Depleted                  1      0.21699504      0.21699504       2.95    0.1139
      Error                         11      0.80949018      0.07359002
```

Example 30.8. Mixed Model Analysis of Variance Using the RANDOM Statement

Milliken and Johnson (1984) present an example of an unbalanced mixed model. Three machines, which are considered as a fixed effect, and six employees, which are considered a random effect, are studied. Each employee operates each machine for either one, two, or three different times. The dependent variable is an overall rating, which takes into account the number and quality of components produced.

The following statements form the data set and perform a mixed model analysis of variance by requesting the TEST option in the RANDOM statement. Note that the machine*person interaction is declared as a random effect; in general, when an interaction involves a random effect, it too should be declared as random. The results of the analysis are shown in Output 30.8.1 through Output 30.8.4.

```
data machine;
   input machine person rating @@;
   datalines;
1 1 52.0  1 2 51.8  1 2 52.8  1 3 60.0  1 4 51.1  1 4 52.3
1 5 50.9  1 5 51.8  1 5 51.4  1 6 46.4  1 6 44.8  1 6 49.2
2 1 64.0  2 2 59.7  2 2 60.0  2 2 59.0  2 3 68.6  2 3 65.8
2 4 63.2  2 4 62.8  2 4 62.2  2 5 64.8  2 5 65.0  2 6 43.7
```

```
2 6 44.2  2 6 43.0  3 1 67.5  3 1 67.2  3 1 66.9  3 2 61.5
3 2 61.7  3 2 62.3  3 3 70.8  3 3 70.6  3 3 71.0  3 4 64.1
3 4 66.2  3 4 64.0  3 5 72.1  3 5 72.0  3 5 71.1  3 6 62.0
3 6 61.4  3 6 60.5
;

proc glm data=machine;
   class machine person;
   model rating=machine person machine*person;
   random person machine*person / test;
run;
```

The TEST option in the RANDOM statement requests that PROC GLM determine the appropriate F-tests based on person and machine*person being treated as random effects. As you can see in Output 30.8.4, this requires that a linear combination of mean squares be constructed to test both the machine and person hypotheses; thus, F-tests using Satterthwaite approximations are used.

Output 30.8.1. Summary Information on Groups

```
                          The GLM Procedure

                       Class Level Information

                 Class         Levels    Values

                 machine            3    1 2 3

                 person             6    1 2 3 4 5 6


                     Number of observations    44
```

Output 30.8.2. Fixed-Effect Model Analysis of Variance

```
                               The GLM Procedure

Dependent Variable: rating

                                        Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                       17    3061.743333     180.102549    206.41   <.0001

 Error                       26      22.686667       0.872564

 Corrected Total             43    3084.430000


              R-Square     Coeff Var      Root MSE    rating Mean

              0.992645      1.560754      0.934111       59.85000


 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 machine                      2    1648.664722     824.332361    944.72   <.0001
 person                       5    1008.763583     201.752717    231.22   <.0001
 machine*person              10     404.315028      40.431503     46.34   <.0001


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 machine                      2    1238.197626     619.098813    709.52   <.0001
 person                       5    1011.053834     202.210767    231.74   <.0001
 machine*person              10     404.315028      40.431503     46.34   <.0001
```

Output 30.8.3. Expected Values of Type III Mean Squares

```
                               The GLM Procedure

Source                 Type III Expected Mean Square

machine                Var(Error) + 2.137 Var(machine*person) + Q(machine)

person                 Var(Error) + 2.2408 Var(machine*person) + 6.7224
                       Var(person)

machine*person         Var(Error) + 2.3162 Var(machine*person)
```

Output 30.8.4. Mixed Model Analysis of Variance

```
                               The GLM Procedure
            Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: rating

 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 machine                      2    1238.197626     619.098813     16.57   0.0007

 Error                   10.036     375.057436      37.370384
 Error: 0.9226*MS(machine*person) + 0.0774*MS(Error)


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 person                       5    1011.053834     202.210767      5.17   0.0133

 Error                   10.015     392.005726      39.143708
 Error: 0.9674*MS(machine*person) + 0.0326*MS(Error)


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 machine*person              10     404.315028      40.431503     46.34   <.0001

 Error: MS(Error)            26      22.686667       0.872564
```

Note that you can also use the MIXED procedure to analyze mixed models. The following statements use PROC MIXED to reproduce the mixed model analysis of variance; the relevant part of the PROC MIXED results is shown in Output 30.8.5

```
proc mixed data=machine method=type3;
   class machine person;
   model rating = machine;
   random person machine*person;
run;
```

Output 30.8.5. PROC MIXED Mixed Model Analysis of Variance (Partial Output)

```
                              The Mixed Procedure

                          Type 3 Analysis of Variance

                                                  Sum of
             Source              DF          Squares     Mean Square

             machine              2      1238.197626      619.098813
             person               5      1011.053834      202.210767
             machine*person      10       404.315028       40.431503
             Residual            26        22.686667        0.872564

                          Type 3 Analysis of Variance

Source          Expected Mean Square

machine         Var(Residual) + 2.137 Var(machine*person) + Q(machine)
person          Var(Residual) + 2.2408 Var(machine*person) + 6.7224 Var(person)
machine*person  Var(Residual) + 2.3162 Var(machine*person)
Residual        Var(Residual)

                          Type 3 Analysis of Variance

                                                          Error
Source          Error Term                                   DF  F Value  Pr > F

machine         0.9226 MS(machine*person)                10.036    16.57  0.0007
                + 0.0774 MS(Residual)
person          0.9674 MS(machine*person)                10.015     5.17  0.0133
                + 0.0326 MS(Residual)
machine*person  MS(Residual)                                 26    46.34  <.0001
Residual        .                                             .        .      .
```

The advantage of PROC MIXED is that it offers more versatility for mixed models; the disadvantage is that it can be less computationally efficient for large data sets. See Chapter 41, “The MIXED Procedure,” for more details.

Example 30.9. Analyzing a Doubly-multivariate Repeated Measures Design

This example shows how to analyze a doubly-multivariate repeated measures design by using PROC GLM with an IDENTITY factor in the REPEATED statement. Note that this differs from previous releases of PROC GLM, in which you had to use a MANOVA statement to get a doubly repeated measures analysis.

Two responses, Y1 and Y2, are each measured three times for each subject (pretreatment, posttreatment, and in a later follow-up). Each subject receives one of three treatments; A, B, or the control. In PROC GLM, you use a REPEATED factor of type IDENTITY to identify the different responses and another repeated factor to identify the different measurement times. The repeated measures analysis includes multivariate tests for time and treatment main effects, as well as their interactions, across responses. The following statements produce Output 30.9.1 through Output 30.9.3.

```
data Trial;
   input Treatment $ Repetition PreY1 PostY1 FollowY1
                                PreY2 PostY2 FollowY2;
   datalines;
A        1  3  13  9  0  0  9
A        2  0  14 10  6  6  3
A        3  4   6 17  8  2  6
A        4  7   7 13  7  6  4
A        5  3  12 11  6 12  6
A        6 10  14  8 13  3  8
B        1  9  11 17  8 11 27
B        2  4  16 13  9  3 26
B        3  8  10  9 12  0 18
B        4  5   9 13  3  0 14
B        5  0  15 11  3  0 25
B        6  4  11 14  4  2  9
Control  1 10  12 15  4  3  7
Control  2  2   8 12  8  7 20
Control  3  4   9 10  2  0 10
Control  4 10   8  8  5  8 14
Control  5 11  11 11  1  0 11
Control  6  1  5  15  8  9 10
;

proc glm data=Trial;
   class Treatment;
   model PreY1 PostY1 FollowY1
         PreY2 PostY2 FollowY2 = Treatment / nouni;
   repeated Response 2 identity, Time 3;
run;
```

Output 30.9.1. A Doubly-multivariate Repeated Measures Design

```
                              The GLM Procedure

                           Class Level Information

                      Class          Levels    Values

                      Treatment           3    A B Control


                          Number of observations    18
```

The levels of the repeated factors are displayed in Output 30.9.2. Note that RESPONSE is 1 for all the Y1 measurements and 2 for all the Y2 measurements, while the three levels of Time identify the pretreatment, posttreatment, and follow-up measurements within each response. The multivariate tests for within-subject effects are displayed in Output 30.9.3.

Output 30.9.2. Repeated Factor Levels

```
                                 The GLM Procedure
                      Repeated Measures Analysis of Variance

                       Repeated Measures Level Information

  Dependent Variable       PreY1   PostY1 FollowY1    PreY2   PostY2 FollowY2

   Level of Response           1        1        1        2        2        2
       Level of Time           1        2        3        1        2        3
```

Output 30.9.3. Within-subject Tests

```
                                 The GLM Procedure
                      Repeated Measures Analysis of Variance

    Manova Test Criteria and Exact F Statistics for the Hypothesis of no Response Effect
                       H = Type III SSCP Matrix for Response
                               E = Error SSCP Matrix

                                 S=1    M=0    N=6

      Statistic                         Value    F Value    Num DF    Den DF    Pr > F

      Wilks' Lambda                0.02165587     316.24         2        14    <.0001
      Pillai's Trace               0.97834413     316.24         2        14    <.0001
      Hotelling-Lawley Trace      45.17686368     316.24         2        14    <.0001
      Roy's Greatest Root         45.17686368     316.24         2        14    <.0001


Manova Test Criteria and F Approximations for the Hypothesis of no Response*Treatment Effect
                     H = Type III SSCP Matrix for Response*Treatment
                               E = Error SSCP Matrix

                                S=2    M=-0.5    N=6

      Statistic                         Value    F Value    Num DF    Den DF    Pr > F

      Wilks' Lambda                0.72215797       1.24         4        28    0.3178
      Pillai's Trace               0.27937444       1.22         4        30    0.3240
      Hotelling-Lawley Trace       0.38261660       1.31         4    15.818    0.3074
      Roy's Greatest Root          0.37698780       2.83         2        15    0.0908

                NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                       NOTE: F Statistic for Wilks' Lambda is exact.


 Manova Test Criteria and Exact F Statistics for the Hypothesis of no Response*Time Effect
                       H = Type III SSCP Matrix for Response*Time
                               E = Error SSCP Matrix

                                 S=1    M=1    N=5

      Statistic                         Value    F Value    Num DF    Den DF    Pr > F

      Wilks' Lambda                0.14071380      18.32         4        12    <.0001
      Pillai's Trace               0.85928620      18.32         4        12    <.0001
      Hotelling-Lawley Trace       6.10662362      18.32         4        12    <.0001
      Roy's Greatest Root          6.10662362      18.32         4        12    <.0001


                    Manova Test Criteria and F Approximations for the
                     Hypothesis of no Response*Time*Treatment Effect
                   H = Type III SSCP Matrix for Response*Time*Treatment
                               E = Error SSCP Matrix

                                S=2    M=0.5    N=5

      Statistic                         Value    F Value    Num DF    Den DF    Pr > F

      Wilks' Lambda                0.22861451       3.27         8        24    0.0115
      Pillai's Trace               0.96538785       3.03         8        26    0.0151
      Hotelling-Lawley Trace       2.52557514       3.64         8        15    0.0149
      Roy's Greatest Root          2.12651905       6.91         4        13    0.0033

                NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                       NOTE: F Statistic for Wilks' Lambda is exact.
```

The table for Response*Treatment tests for an overall treatment effect across the two responses; likewise, the tables for Response*Time and Response*Treatment*Time test for time and the treatment-by-time interaction, respectively. In this case, there is a strong main effect for time and possibly for the interaction, but not for treatment.

In previous releases (before the IDENTITY transformation was introduced), in order to perform a doubly repeated measures analysis, you had to use a MANOVA statement with a customized transformation matrix M. You might still want to use this approach to see details of the analysis, such as the univariate ANOVA for each transformed variate. The following statements demonstrate this approach by using the MANOVA statement to test for the overall main effect of time and specifying the SUMMARY option.

```
proc glm data=Trial;
   class Treatment;
   model PreY1 PostY1 FollowY1
         PreY2 PostY2 FollowY2 = Treatment / nouni;
   manova  h=intercept  m=prey1 - posty1,
                          prey1 - followy1,
                          prey2 - posty2,
                          prey2 - followy2 / summary;
run;
```

The M matrix used to perform the test for time effects is displayed in Output 30.9.4, while the results of the multivariate test are given in Output 30.9.5. Note that the test results are the same as for the Response*Time effect in Output 30.9.3.

Output 30.9.4. M Matrix to Test for Time Effect (Repeated Measure)

```
                                         The GLM Procedure
                                 Multivariate Analysis of Variance

                               M Matrix Describing Transformed Variables

                PreY1           PostY1        FollowY1           PreY2          PostY2        FollowY2

MVAR1               1               -1               0               0               0               0
MVAR2               1                0              -1               0               0               0
MVAR3               0                0               0               1              -1               0
MVAR4               0                0               0               1               0              -1
```

Output 30.9.5. Tests for Time Effect (Repeated Measure)

```
                                        The GLM Procedure
                                  Multivariate Analysis of Variance

                      Characteristic Roots and Vectors of: E Inverse * H, where
                                H = Type III SSCP Matrix for Intercept
                                        E = Error SSCP Matrix

                           Variables have been transformed by the M Matrix

  Characteristic               Characteristic Vector  V'EV=1
            Root    Percent          MVAR1           MVAR2           MVAR3           MVAR4

      6.10662362     100.00    -0.00157729      0.04081620     -0.04210209      0.03519437
      0.00000000       0.00     0.00796367      0.00493217      0.05185236      0.00377940
      0.00000000       0.00    -0.03534089     -0.01502146     -0.00283074      0.04259372
      0.00000000       0.00    -0.05672137      0.04500208      0.00000000      0.00000000


MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Intercept Effect
                   on the Variables Defined by the M Matrix Transformation
                           H = Type III SSCP Matrix for Intercept
                                    E = Error SSCP Matrix

                                      S=1    M=1    N=5

       Statistic                         Value    F Value    Num DF    Den DF    Pr > F

       Wilks' Lambda                0.14071380      18.32         4        12    <.0001
       Pillai's Trace               0.85928620      18.32         4        12    <.0001
       Hotelling-Lawley Trace       6.10662362      18.32         4        12    <.0001
       Roy's Greatest Root          6.10662362      18.32         4        12    <.0001
```

The SUMMARY option in the MANOVA statement creates an ANOVA table for each transformed variable as defined by the M matrix. MVAR1 and MVAR2 contrast the pretreatment measurement for Y1 with the posttreatment and follow-up measurements for Y1, respectively; MVAR3 and MVAR4 are the same contrasts for Y2. Output 30.9.6 displays these univariate ANOVA tables and shows that the contrasts are all strongly significant except for the pre-versus-post difference for Y2.

Output 30.9.6. Summary Output for the Test for Time Effect

```
                                The GLM Procedure
                         Multivariate Analysis of Variance

Dependent Variable: MVAR1

      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      Intercept                    1     512.0000000     512.0000000      22.65    0.0003
      Error                       15     339.0000000      22.6000000



                                The GLM Procedure
                         Multivariate Analysis of Variance

Dependent Variable: MVAR2

      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      Intercept                    1     813.3888889     813.3888889      32.87    <.0001
      Error                       15     371.1666667      24.7444444



                                The GLM Procedure
                         Multivariate Analysis of Variance

Dependent Variable: MVAR3

      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      Intercept                    1      68.0555556      68.0555556       3.49    0.0814
      Error                       15     292.5000000      19.5000000



                                The GLM Procedure
                         Multivariate Analysis of Variance

Dependent Variable: MVAR4

      Source                      DF     Type III SS     Mean Square    F Value    Pr > F

      Intercept                    1     800.0000000     800.0000000      26.43    0.0001
      Error                       15     454.0000000      30.2666667
```

Example 30.10. Testing for Equal Group Variances

This example demonstrates how you can test for equal group variances in a one-way design. The data come from the University of Pennsylvania Smell Identification Test (UPSIT), reported in O'Brien and Heft (1995). The study is undertaken to explore how age and gender are related to sense of smell. A total of 180 subjects 20 to 89 years old are exposed to 40 different odors: for each odor, subjects are asked to choose which of four words best describes the odor. The Freeman-Tukey modified arcsine transformation (Bishop, Feinberg, and Holland 1975) is applied to the proportion of correctly identified odors to arrive at an olfactory index. For the following analysis, subjects are divided into five age groups:

$$\textsf{agegroup} = \begin{cases} 1 & \text{if} & & & \text{age} & \leq & 25 \\ 2 & \text{if} & 25 & < & \text{age} & \leq & 40 \\ 3 & \text{if} & 40 & < & \text{age} & \leq & 55 \\ 4 & \text{if} & 55 & < & \text{age} & \leq & 70 \\ 5 & \text{if} & 70 & < & \text{age} & & \end{cases}$$

The following statements create a data set named upsit, containing the age group and olfactory index for each subject.

```
data upsit;
   input agegroup smell @@;
   datalines;
1 1.381  1 1.322  1 1.162  1 1.275  1 1.381  1 1.275  1 1.322
1 1.492  1 1.322  1 1.381  1 1.162  1 1.013  1 1.322  1 1.322
1 1.275  1 1.492  1 1.322  1 1.322  1 1.492  1 1.322  1 1.381
1 1.234  1 1.162  1 1.381  1 1.381  1 1.381  1 1.322  1 1.381
1 1.322  1 1.381  1 1.275  1 1.492  1 1.275  1 1.322  1 1.275
1 1.381  1 1.234  1 1.105
2 1.234  2 1.234  2 1.381  2 1.322  2 1.492  2 1.234  2 1.381
2 1.381  2 1.492  2 1.492  2 1.275  2 1.492  2 1.381  2 1.492
2 1.322  2 1.275  2 1.275  2 1.275  2 1.322  2 1.492  2 1.381
2 1.322  2 1.492  2 1.196  2 1.322  2 1.275  2 1.234  2 1.322
2 1.098  2 1.322  2 1.381  2 1.275  2 1.492  2 1.492  2 1.381
2 1.196
3 1.381  3 1.381  3 1.492  3 1.492  3 1.492  3 1.098  3 1.492
3 1.381  3 1.234  3 1.234  3 1.129  3 1.069  3 1.234  3 1.322
3 1.275  3 1.230  3 1.234  3 1.234  3 1.322  3 1.322  3 1.381
4 1.322  4 1.381  4 1.381  4 1.322  4 1.234  4 1.234  4 1.234
4 1.381  4 1.322  4 1.275  4 1.275  4 1.492  4 1.234  4 1.098
4 1.322  4 1.129  4 0.687  4 1.322  4 1.322  4 1.234  4 1.129
4 1.492  4 0.810  4 1.234  4 1.381  4 1.040  4 1.381  4 1.381
4 1.129  4 1.492  4 1.129  4 1.098  4 1.275  4 1.322  4 1.234
4 1.196  4 1.234  4 0.585  4 0.785  4 1.275  4 1.322  4 0.712
4 0.810
5 1.322  5 1.234  5 1.381  5 1.275  5 1.275  5 1.322  5 1.162
5 0.909  5 0.502  5 1.234  5 1.322  5 1.196  5 0.859  5 1.196
5 1.381  5 1.322  5 1.234  5 1.275  5 1.162  5 1.162  5 0.585
5 1.013  5 0.960  5 0.662  5 1.129  5 0.531  5 1.162  5 0.737
5 1.098  5 1.162  5 1.040  5 0.558  5 0.960  5 1.098  5 0.884
5 1.162  5 1.098  5 0.859  5 1.275  5 1.162  5 0.785  5 0.859
;
```

Older people are more at risk for problems with their sense of smell, and this should be reflected in significant differences in the mean of the olfactory index across the different age groups. However, many older people also have an excellent sense of smell, which implies that the older age groups should have greater variability. In order to test this hypothesis and to compute a one-way ANOVA for the olfactory index that is robust to the possibility of unequal group variances, you can use the HOVTEST and WELCH options in the MEANS statement for the GLM procedure, as shown in the following code.

```
proc glm data=upsit;
   class agegroup;
   model smell = agegroup;
   means agegroup / hovtest welch;
run;
```

Output 30.10.1, Output 30.10.2, and Output 30.10.3 display the usual ANOVA test for equal age group means, Levene's test for equal age group variances, and Welch's test for equal age group means, respectively. The hypotheses of age effects for mean and variance of the olfactory index are both confirmed.

Output 30.10.1. Usual ANOVA Test for Age Group Differences in Mean Olfactory Index

```
                               The GLM Procedure

Dependent Variable: smell

 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 agegroup                     4     2.13878141     0.53469535     16.65   <.0001
```

Output 30.10.2. Levene's Test for Age Group Differences in Olfactory Variability

```
                               The GLM Procedure

                Levene's Test for Homogeneity of smell Variance
                 ANOVA of Squared Deviations from Group Means

                                 Sum of        Mean
        Source           DF     Squares      Square    F Value    Pr > F

        agegroup          4      0.0799      0.0200       6.35    <.0001
        Error           175      0.5503     0.00314
```

Output 30.10.3. Welch's Test for Age Group Differences in Mean Olfactory Index

```
                               The GLM Procedure

                            Welch's ANOVA for smell

                   Source            DF    F Value    Pr > F

                   agegroup      4.0000      13.72    <.0001
                   Error        78.7489
```

Example 30.11. Analysis of a Screening Design

Yin and Jillie (1987) describe an experiment on a nitride etch process for a single wafer plasma etcher. The experiment is run using four factors: cathode power (power), gas flow (flow), reactor chamber pressure (pressure), and electrode gap (gap). Of interest are the main effects and interaction effects of the factors on the nitride etch rate (rate). The following statements create a SAS data set named HalfFraction, containing the factor settings and the observed etch rate for each of eight experimental runs.

```
data HalfFraction;
   input power flow pressure gap rate;
   datalines;
0.8   4.5 125 275     550
0.8   4.5 200 325     650
0.8 550.0 125 325     642
0.8 550.0 200 275     601
1.2   4.5 125 325     749
1.2   4.5 200 275    1052
1.2 550.0 125 275    1075
1.2 550.0 200 325     729
;
```

Notice that each of the factors has just two values. This is a common experimental design when the intent is to screen from the many factors that *might* affect the response the few that actually *do*. Since there are $2^4 = 16$ different possible settings of four two-level factors, this design with only eight runs is called a “half fraction.” The eight runs are chosen specifically to provide unambiguous information on main effects at the cost of confounding interaction effects with each other.

One way to analyze this data is simply to use PROC GLM to compute an analysis of variance, including both main effects and interactions in the model. The following statements demonstrate this approach.

```
proc glm data=HalfFraction;
   class power flow pressure gap;
   model rate=power|flow|pressure|gap@2;
run;
```

The ‘@2’ notation on the model statement includes all main effects and two-factor interactions between the factors. The output is shown in Output 30.11.1.

Output 30.11.1. Analysis of Variance for Nitride Etch Process Half Fraction

```
                               The GLM Procedure

                            Class Level Information

                        Class         Levels    Values

                        power              2    0.8 1.2

                        flow               2    4.5 550

                        pressure           2    125 200

                        gap                2    275 325


                          Number of observations    8


                               The GLM Procedure

Dependent Variable: rate

                                       Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                        7    280848.0000     40121.1429         .      .

 Error                        0         0.0000          .

 Corrected Total              7    280848.0000


               R-Square     Coeff Var      Root MSE     rate Mean

               1.000000            .               .     756.0000


 Source                      DF      Type I SS    Mean Square   F Value   Pr > F

 power                        1    168780.5000    168780.5000         .      .
 flow                         1       264.5000       264.5000         .      .
 power*flow                   1       200.0000       200.0000         .      .
 pressure                     1        32.0000        32.0000         .      .
 power*pressure               1      1300.5000      1300.5000         .      .
 flow*pressure                1     78012.5000     78012.5000         .      .
 gap                          1     32258.0000     32258.0000         .      .
 power*gap                    0         0.0000          .             .      .
 flow*gap                     0         0.0000          .             .      .
 pressure*gap                 0         0.0000          .             .      .


 Source                      DF    Type III SS    Mean Square   F Value   Pr > F

 power                        1    168780.5000    168780.5000         .      .
 flow                         1       264.5000       264.5000         .      .
 power*flow                   0         0.0000          .             .      .
 pressure                     1        32.0000        32.0000         .      .
 power*pressure               0         0.0000          .             .      .
 flow*pressure                0         0.0000          .             .      .
 gap                          1     32258.0000     32258.0000         .      .
 power*gap                    0         0.0000          .             .      .
 flow*gap                     0         0.0000          .             .      .
 pressure*gap                 0         0.0000          .             .      .
```

Notice that there are no error degrees of freedom. This is because there are 10 effects in the model (4 main effects plus 6 interactions) but only 8 observations in the data set. This is another cost of using a fractional design: not only is it impossible to estimate all the main effects and interactions, but there is also no information left to estimate the underlying error rate in order to measure the significance of the effects that are estimable.

Another thing to notice in Output 30.11.1 is the difference between the Type I and Type III ANOVA tables. The rows corresponding to main effects in each are the same, but no Type III interaction tests are estimable, while some Type I interaction tests are estimable. This indicates that there is *aliasing* in the design: some interactions are completely confounded with each other.

In order to analyze this confounding, you should examine the aliasing structure of the design using the ALIASING option in the MODEL statement. Before doing so, however, it is advisable to *code* the design, replacing low and high levels of each factor with the values -1 and +1, respectively. This puts each factor on an equal footing in the model and makes the aliasing structure much more interpretable. The following statements code the data, creating a new data set named Coded.

```
   data Coded; set HalfFraction;
      power    = -1*(power   =0.80) + 1*(power   =1.20);
      flow     = -1*(flow    =4.50) + 1*(flow    =550 );
      pressure = -1*(pressure=125 ) + 1*(pressure=200 );
      gap      = -1*(gap     =275 ) + 1*(gap     =325 );
   run;
```

The following statements use the GLM procedure to reanalyze the coded design, displaying the parameter estimates as well as the functions of the parameters that they each estimate.

```
   proc glm data=Coded;
      model rate=power|flow|pressure|gap@2 / solution aliasing;
   run;
```

The parameter estimates table is shown in Output 30.11.2.

Output 30.11.2. Parameter Estimates and Aliases for Nitride Etch Process Half Fraction

```
                                       The GLM Procedure

Dependent Variable: rate

                                       Standard
Parameter             Estimate            Error   t Value   Pr > |t|   Expected Value

Intercept          756.0000000                .         .          .      Intercept
power              145.2500000                .         .          .      power
flow                 5.7500000                .         .          .      flow
power*flow          -5.0000000 B              .         .          .      power*flow + pressure*gap
pressure             2.0000000                .         .          .      pressure
power*pressure     -12.7500000 B              .         .          .      power*pressure + flow*gap
flow*pressure      -98.7500000 B              .         .          .      flow*pressure + power*gap
gap                -63.5000000                .         .          .      gap
power*gap            0.0000000 B              .         .          .
flow*gap             0.0000000 B              .         .          .
pressure*gap         0.0000000 B              .         .          .

NOTE: The X’X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter ’B’ are not
      uniquely estimable.
```

Looking at the “Expected Value” column, notice that, while each of the main effects is unambiguously estimated by its associated term in the model, the expected values of the interaction estimates are more complicated. For example, the relatively large effect (-98.75) corresponding to flow*pressure actually estimates the combined effect of flow*pressure and power*gap. Without further information, it is impossible to disentangle these aliased interactions; however, since the main effects of both power and gap are large and those for flow and pressure are small, it is reasonable to suspect that power*gap is the more “active” of the two interactions.

Fortunately, eight more runs are available for this experiment (the other half fraction.) The following statements create a data set containing these extra runs and add it to the previous eight, resulting in a full $2^4 = 16$ run replicate. Then PROC GLM displays the analysis of variance again.

```
data OtherHalf;
   input power flow pressure gap rate;
   datalines;
0.8   4.5 125 325     669
0.8   4.5 200 275     604
0.8 550.0 125 275     633
0.8 550.0 200 325     635
1.2   4.5 125 275    1037
1.2   4.5 200 325     868
1.2 550.0 125 325     860
1.2 550.0 200 275    1063
;
data FullRep;
   set HalfFraction OtherHalf;
run;

proc glm data=FullRep;
   class power flow pressure gap;
   model rate=power|flow|pressure|gap@2;
run;
```

The results are displayed in Output 30.11.3.

Output 30.11.3. Analysis of Variance for Nitride Etch Process Full Replicate

```
                               The GLM Procedure

                            Class Level Information

                        Class         Levels    Values

                        power              2    0.8 1.2

                        flow               2    4.5 550

                        pressure           2    125 200

                        gap                2    275 325

                          Number of observations    16


                               The GLM Procedure

Dependent Variable: rate

                                       Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F

 Model                      10    521234.1250     52123.4125     25.58   0.0011

 Error                       5     10186.8125      2037.3625

 Corrected Total            15    531420.9375


               R-Square     Coeff Var      Root MSE     rate Mean

               0.980831      5.816175      45.13715      776.0625


 Source                     DF      Type I SS    Mean Square   F Value   Pr > F

 power                       1    374850.0625    374850.0625    183.99   <.0001
 flow                        1       217.5625       217.5625      0.11   0.7571
 power*flow                  1        18.0625        18.0625      0.01   0.9286
 pressure                    1        10.5625        10.5625      0.01   0.9454
 power*pressure              1         1.5625         1.5625      0.00   0.9790
 flow*pressure               1      7700.0625      7700.0625      3.78   0.1095
 gap                         1     41310.5625     41310.5625     20.28   0.0064
 power*gap                   1     94402.5625     94402.5625     46.34   0.0010
 flow*gap                    1      2475.0625      2475.0625      1.21   0.3206
 pressure*gap                1       248.0625       248.0625      0.12   0.7414


 Source                     DF    Type III SS    Mean Square   F Value   Pr > F

 power                       1    374850.0625    374850.0625    183.99   <.0001
 flow                        1       217.5625       217.5625      0.11   0.7571
 power*flow                  1        18.0625        18.0625      0.01   0.9286
 pressure                    1        10.5625        10.5625      0.01   0.9454
 power*pressure              1         1.5625         1.5625      0.00   0.9790
 flow*pressure               1      7700.0625      7700.0625      3.78   0.1095
 gap                         1     41310.5625     41310.5625     20.28   0.0064
 power*gap                   1     94402.5625     94402.5625     46.34   0.0010
 flow*gap                    1      2475.0625      2475.0625      1.21   0.3206
 pressure*gap                1       248.0625       248.0625      0.12   0.7414
```

With sixteen runs, the analysis of variance tells the whole story: all effects are estimable and there are five degrees of freedom left over to estimate the underlying error. The main effects of power and gap and their interaction are all significant, and no other effects are. Notice that the Type I and Type III ANOVA tables are the same; this is because the design is orthogonal and all effects are estimable.

This example illustrates the use of the GLM procedure for the model analysis of a screening experiment. Typically, there is much more involved in performing an experiment of this type, from selecting the design points to be studied to graphically assessing significant effects, optimizing the final model, and performing subsequent experimentation. Specialized tools for this are available in SAS/QC software, in particular the ADX Interface and the FACTEX and OPTEX procedures. Refer to *SAS/QC User's Guide* for more information.

References

Afifi, A.A. and Azen, S.P. (1972), *Statistical Analysis: A Computer-Oriented Approach*, New York: Academic Press, Inc.

Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons, Inc.

Bartlett, M.S. (1937), "Properties of Sufficiency and Statistical Tests," *Proceedings of the Royal Society of London, Series A*, 160, 268–282.

Begun, J.M. and Gabriel, K.R. (1981), "Closure of the Newman-Keuls Multiple Comparisons Procedure," *Journal of the American Statistical Association*, 76, 374.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley & Sons, Inc.

Bishop, Y., Feinberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Box, G.E.P. (1953), "Non-normality and Tests on Variance," *Biometrika*, 40, 318–335.

Box, G.E.P. (1954), "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification," *Annals of Mathematical Statistics*, 25, 484–498.

Brown, M.B. and Forsythe, A.B. (1974), "Robust Tests for Equality of Variances," *Journal of the American Statistical Association*, 69, 364–367.

Carmer, S.G. and Swanson, M.R. (1973), "Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte-Carlo Methods," *Journal of the American Statistical Association*, 68, 66–74.

Cochran, W.G. and Cox, G.M. (1957), *Experimental Designs*, Second edition. New York: John Wiley & Sons, Inc.

Cole, J.W.L. and Grizzle, J.E. (1966), "Applications of Multivariate Analysis of Variance to Repeated Measures Experiments," *Biometrics*, 22, 810–828.

Conover, W.J., Johnson, M.E., and Johnson, M.M. (1981), "A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data," *Technometrics*, 23, 351–361.

Cornfield, J. and Tukey, J.W. (1956), "Average Values of Mean Squares in Factorials," *Annals of Mathematical Statistics*, 27, 907–949.

Draper, N.R. and Smith, H. (1966), *Applied Regression Analysis*, New York: John Wiley & Sons, Inc.

Duncan, D.B. (1975), "t Tests and Intervals for Comparisons Suggested by the Data," *Biometrics*, 31, 339–359.

Dunnett, C.W. (1955), "A Multiple Comparisons Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, 50, 1096–1121.

Dunnett, C.W. (1980), "Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case," *Journal of the American Statistical Association*, 75, 789–795.

Edwards, D. and Berry, J.J. (1987), "The Efficiency of Simulation-Based Multiple Comparisons," *Biometrics*, 43, 913–928.

Einot, I. and Gabriel, K.R. (1975), "A Study of the Powers of Several Methods of Multiple Comparisons," *Journal of the American Statistical Association*, 70, 351.

Freund, R.J., Littell, R.C., and Spector, P.C. (1986), *SAS System for Linear Models, 1986 Edition*, Cary, NC: SAS Institute Inc.

Gabriel, K.R. (1978), "A Simple Method of Multiple Comparisons of Means," *Journal of the American Statistical Association*, 73, 364.

Games, P.A. (1977), "An Improved t Table for Simultaneous Control on g Contrasts," *Journal of the American Statistical Association*, 72, 531–534.

Goodnight, J.H. (1976), "The New General Linear Models Procedure," *Proceedings of the First International SAS Users' Conference*, Cary, NC: SAS Institute Inc.

Goodnight, J.H. (1978), *Tests of the Hypotheses in Fixed-Effects Linear Models*, SAS Technical Report R-101, Cary, NC: SAS Institute Inc.

Goodnight, J.H. (1979), "A Tutorial on the Sweep Operator," *American Statistician*, 33, 149–158. (Also available as *The Sweep Operator: Its Importance in Statistical Computing*, SAS Technical Report R-106.)

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

Goodnight, J.H. and Speed, F.M. (1978), *Computing Expected Mean Squares*, SAS Technical Report R-102, Cary, NC: SAS Institute Inc.

Graybill, F.A. (1961), *An Introduction to Linear Statistical Models, Volume I*, New York: McGraw-Hill Book Co.

Greenhouse, S.W. and Geisser, S. (1959), “On Methods in the Analysis of Profile Data,” *Psychometrika*, 32, 95–112.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994), *A Handbook of Small Data Sets*, London: Chapman & Hall.

Hartley, H.O. and Searle, S.R. (1969), “On Interaction Variance Components in Mixed Models,” *Biometrics*, 25, 573–576

Harvey, W.R. (1975), *Least-squares Analysis of Data with Unequal Subclass Numbers*, USDA Report ARS H-4.

Hayter, A.J. (1984), “A Proof of the Conjecture that the Tukey-Kramer Method is Conservative,” *The Annals of Statistics*, 12, 61–75.

Hayter, A.J. (1989), “Pairwise Comparisons of Generally Correlated Means,” *Journal of the American Statistical Association*, 84, 208–213.

Heck, D.L. (1960), “Charts of Some Upper Percentage Points of the Distribution of the Largest Characteristic Root,” *Annals of Mathematical Statistics*, 31, 625–642.

Hochberg, Y. (1974), “Some Conservative Generalizations of the T-Method in Simultaneous Inference,” *Journal of Multivariate Analysis*, 4, 224–234.

Hocking, R.R. (1976), “The Analysis and Selection of Variables in a Linear Regression,” *Biometrics*, 32, 1–50.

Hocking, R.R. (1985), *The Analysis of Linear Models*, Belmont, CA: Brooks/Cole Publishing Co.

Hsu, J.C. (1992), “The Factor Analytic Approach to Simultaneous Confidence Interval for Multiple Comparisons with the Best,” *Journal of Computational Statistics and Graphics*, 1, 151–168.

Hsu, J.C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.

Hsu, J.C. and Nelson, B. (1998), “Multiple Comparisons in the General Linear Model,” *Journal of Computational and Graphical Statistics*, 7(1), 23–41.

Huynh, H. and Feldt, L. S. (1970), “Conditions under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions,” *Journal of the American Statistical Association*, 65, 1582–1589.

Huynh, H. and Feldt, L.S. (1976), “Estimation of the Box Correction for Degrees of Freedom from Sample Data in the Randomized Block and Split Plot Designs,” *Journal of Educational Statistics*, 1, 69–82.

Kennedy, W.J., Jr. and Gentle, J.E. (1980), *Statistical Computing*, New York: Marcel Dekker, Inc.

Kramer, C.Y. (1956), “Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications,” *Biometrics*, 12, 307–310.

Krishnaiah, P.R. and Armitage, J.V. (1966), “Tables for Multivariate t Distribution,” *Sankhya, Series B*, 31–56.

Kutner, M.H. (1974), “Hypothesis Testing in Linear Models (Eisenhart Model),” *American Statistician*, 28, 98–100.

LaTour, S.A. and Miniard, P.W. (1983), “The Misuse of Repeated Measures Analysis in Marketing Research,” *Journal of Marketing Research*, XX, 45–57.

Levene, H. (1960), “Robust Tests for the Equality of Variance,” in *Contributions to Probability and Statistics*, ed. I. Olkin, Palo Alto, CA: Stanford University Press, 278–292.

Marcus, R., Peritz, E., and Gabriel, K.R. (1976), “On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance,” *Biometrika*, 63, 655–660.

McLean, R.A., Sanders, W.L., and Stroup, W.W. (1991), “A Unified Approach to Mixed Linear Models,” *The American Statistician*, 45, 54–64.

Miller, R.G., Jr. (1981), *Simultaneous Statistical Inference*, New York: Springer-Verlag.

Milliken, G.A. and Johnson, D.E. (1984), *Analysis of Messy Data, Volume I: Designed Experiments*, Belmont, CA: Lifetime Learning Publications.

Morrison, D.F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill Book Co.

Nelder, J.A. (1994), “The Statistics of Linear Models: Back to Basics,” *Statistics and Computing*, 4.

O’Brien, R.G. (1979), “A General ANOVA Method for Robust Tests of Additive Models for Variances,” *Journal of the American Statistical Association*, 74, 877–880.

O’Brien, R.G. (1981), “A Simple Test for Variance Effects in Experimental Designs,” *Psychological Bulletin*, 89(3), 570–574.

O’Brien, R.G. and Heft, M.W. (1995), “New Discrimination Indexes and Models for Studying Sensory Functioning in Aging,” *Journal of Applied Statistics*, 22, 9–27.

Olejnik, S.F. and Algina, J. (1987), “Type I Error Rates and Power Estimates of Selected Parametric and Non-parametric Tests of Scale,” *Journal of Educational Statistics*, 12, 45–61.

Petrinovich, L.F. and Hardyck, C.D. (1969), “Error Rates for Multiple Comparison Methods: Some Evidence Concerning the Frequency of Erroneous Conclusions,” *Psychological Bulletin*, 71, 43–54.

Pillai, K.C.S. (1960), *Statistical Tables for Tests of Multivariate Hypotheses*, Manila: The Statistical Center, University of the Philippines.

Pringle, R.M. and Raynor, A.A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing Co.

Ramsey, P.H. (1978), “Power Differences Between Pairwise Multiple Comparisons,” *Journal of the American Statistical Association*, 73, 363.

Rao, C.R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons, Inc.

Rodriguez, R., Tobias, R., and Wolfinger, R. (1995), "Comments on J.A. Nelder 'The Statistics of Linear Models: Back to Basics,'" *Statistics and Computing*, 5, 97–101.

Ryan, T.A. (1959), "Multiple Comparisons in Psychological Research," *Psychological Bulletin*, 56, 26–47.

Ryan, T.A. (1960), "Significance Tests for Multiple Comparison of Proportions, Variances, and Other Statistics," *Psychological Bulletin*, 57, 318–328.

SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1996), *SAS/STAT Software: Changes and Enhancements through Release 6.12*, Cary, NC: SAS Institute Inc.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114.

Schatzoff, M. (1966), "Exact Distributions of Wilks' Likelihood Ratio Criterion," *Biometrika*, 53, 347–358.

Scheffé, H. (1953), "A Method for Judging All Contrasts in the Analysis of Variance," *Biometrika*, 40, 87–104.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley & Sons, Inc.

Searle, S.R. (1971), *Linear Models*, New York: John Wiley & Sons, Inc.

Searle, S.R. (1987), *Linear Models for Unbalanced Data*, New York: John Wiley & Sons, Inc.

Searle, S.R. (1995), "Comments on J.A. Nelder 'The Statistics of Linear Models: Back to Basics,'" *Statistics and Computing*, 5, 103–107.

Searle, S. R., Casella, G., and McCulloch, C.E. (1992), *Variance Components*, New York: John Wiley & Sons, Inc.

Searle, S.R., Speed, F.M., and Milliken, G.A. (1980), "Populations Marginal Means in the Linear Model: An Alternative to Least Squares Means," *The American Statistician*, 34, 216–221.

Sidak, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *Journal of the American Statistical Association*, 62, 626–633.

Snedecor, G.W. and Cochran, W.G. (1967), *Statistical Methods*, Ames, IA: Iowa State University Press.

Steel, R.G.D. and Torrie, J.H. (1960), *Principles and Procedures of Statistics*, New York: McGraw-Hill Book Co.

Tubb, A., Parker, A.J., and Nickless, G. (1980), "The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry," *Archaeometry*, 22, 153–171.

Tukey, J.W. (1952), "Allowances for Various Types of Error Rates," unpublished IMS address, Chicago, IL.

Tukey, J.W. (1953), "The Problem of Multiple Comparisons," unpublished manuscript.

Waller, R.A. and Duncan, D.B. (1969), "A Bayes Rule for the Symmetric Multiple Comparison Problem," *Journal of the American Statistical Association*, 64, 1484–1499, and (1972) "Corrigenda," 67, 253–255.

Waller, R.A. and Duncan, D.B. (1972), "Corrigenda," *Journal of the American Statistical Association*, 67, 253–255.

Waller, R.A. and Kemp, K.E. (1976), "Computations of Bayesian t-Values for Multiple Comparisons," *Journal of Statistical Computation and Simulation*, 75, 169–172.

Welch, B.L. (1951), "On the Comparison of Several Mean Values: An Alternative Approach," *Biometrika*, 38, 330–336.

Welsch, R.E. (1977), "Stepwise Multiple Comparison Procedures," *Journal of the American Statistical Association*, 72, 359.

Westfall, P.J. and Young, S.S. (1993), *Resampling-based Multiple Testing*, New York: John Wiley & Sons, Inc.

Winer, B. J. (1971), *Statistical Principles in Experimental Design*, Second Edition, New York: McGraw-Hill Book Co.

Wolfinger, R.D. and Chang, M. (1995), "Comparing the SAS GLM and MIXED Procedures for Repeated Measures," *Proceedings of the Twentieth Annual SAS Users Group Conference*, SAS Institute Inc., Cary, NC.

Yin, G.Z. and Jillie, D.W. (1987), "Orthogonal Design for Process Optimization and its Application in Plasma Etching," *Solid State Technology*, May, 127–132.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.