SAS/STAT® 14.3 User's Guide

The Four Types of Estimable Functions

This document is an individual chapter from *SAS/STAT® 14.3 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2017. *SAS/STAT® 14.3 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.3 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

September 2017

Chapter 15
The Four Types of Estimable Functions

Contents

Overview

Many regression and analysis of variance procedures in SAS/STAT label tests for various effects in the model as Type I, Type II, Type III, or Type IV. These four types of hypotheses might not always be sufficient for a statistician to perform all desired inferences, but they should suffice for the vast majority of analyses. This chapter explains the hypotheses involved in each of the four test types. For additional discussion, see Freund, Littell, and Spector (1991) or Milliken and Johnson (1984).

The primary context of the discussion is testing linear hypotheses in least squares regression and analysis of variance, such as with PROC GLM. In this context, tests correspond to hypotheses about linear functions of the true parameters and are evaluated using sums of squares of the estimated parameters. Thus, there will be frequent references to Type I, II, III, and IV (estimable) functions and corresponding Type I, II, III, and IV sums of squares, or simply SS.

Estimability

Given a response or dependent variable $\mathbf{Y}$, predictors or independent variables $\mathbf{X}$, and a linear expectation model $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ relating the two, a primary analytical goal is to estimate or test for the significance of certain linear combinations of the elements of $\boldsymbol{\beta}$. For least squares regression and analysis of variance, this is accomplished by computing linear combinations of the observed $\mathbf{Y}$s. An unbiased linear estimate of a specific linear function of the individual βs, say $\mathbf{L}\boldsymbol{\beta}$, is a linear combination of the $\mathbf{Y}$s that has an expected value of $\mathbf{L}\boldsymbol{\beta}$. Hence, the following definition:

> A linear combination of the parameters $\mathbf{L}\boldsymbol{\beta}$ is estimable if and only if a linear combination of the $\mathbf{Y}$s exists that has expected value $\mathbf{L}\boldsymbol{\beta}$.

Any linear combination of the $\mathbf{Y}$s, for instance $\mathbf{KY}$, will have expectation $\mathrm{E}[\mathbf{KY}] = \mathbf{KX}\boldsymbol{\beta}$. Thus, the expected value of any linear combination of the $\mathbf{Y}$s is equal to that same linear combination of the rows of $\mathbf{X}$ multiplied by $\boldsymbol{\beta}$. Therefore,

> $\mathbf{L}\boldsymbol{\beta}$ is estimable if and only if there is a linear combination of the rows of $\mathbf{X}$ that is equal to $\mathbf{L}$—that is, if and only if there is a $\mathbf{K}$ such that $\mathbf{L} = \mathbf{KX}$.

Thus, the rows of $\mathbf{X}$ form a generating set from which any estimable $\mathbf{L}$ can be constructed. Since the row space of $\mathbf{X}$ is the same as the row space of $\mathbf{X}'\mathbf{X}$, the rows of $\mathbf{X}'\mathbf{X}$ also form a generating set from which all estimable $\mathbf{L}$s can be constructed. Similarly, the rows of $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ also form a generating set for $\mathbf{L}$.

Therefore, if $\mathbf{L}$ can be written as a linear combination of the rows of $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$, or $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$, then $\mathbf{L}\boldsymbol{\beta}$ is estimable.

In the context of least squares regression and analysis of variance, an estimable linear function $\mathbf{L}\boldsymbol{\beta}$ can be estimated by $\mathbf{L}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$. From the general theory of linear models, the unbiased estimator $\mathbf{L}\widehat{\boldsymbol{\beta}}$ is, in fact, the *best* linear unbiased estimator of $\mathbf{L}\boldsymbol{\beta}$, in the sense of having minimum variance as well as maximum likelihood when the residuals are normal. To test the hypothesis that $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, compute the sum of squares

$$\mathrm{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = (\mathbf{L}\widehat{\boldsymbol{\beta}})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}\mathbf{L}\widehat{\boldsymbol{\beta}}$$

and form an F test with the appropriate error term. Note that in contexts more general than least squares regression (for example, generalized and/or mixed linear models), linear hypotheses are often tested by analogous sums of squares of the estimated linear parameters $(\mathbf{L}\widehat{\boldsymbol{\beta}})'(\mathrm{Var}[\mathbf{L}\widehat{\boldsymbol{\beta}}])^{-1}\mathbf{L}\widehat{\boldsymbol{\beta}}$.

General Form of an Estimable Function

This section demonstrates a shorthand technique for displaying the generating set for any estimable $\mathbf{L}$. Suppose

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ A_1 \\ A_2 \\ A_3 \end{bmatrix}$$

$\mathbf{X}$ is a generating set for $\mathbf{L}$, but so is the smaller set

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

$\mathbf{X}^*$ is formed from $\mathbf{X}$ by deleting duplicate rows.

Since all estimable **L**s must be linear functions of the rows of $\mathbf{X}^*$ for $\mathbf{L}\boldsymbol{\beta}$ to be estimable, an **L** for a single-degree-of-freedom estimate can be represented symbolically as

$$L1 \times (1\ 1\ 0\ 0) + L2 \times (1\ 0\ 1\ 0) + L3 \times (1\ 0\ 0\ 1)$$

or

$$\mathbf{L} = (L1 + L2 + L3,\ L1,\ L2,\ L3)$$

For this example, $\mathbf{L}\boldsymbol{\beta}$ is estimable if and only if the first element of **L** is equal to the sum of the other elements of **L** or if

$$\mathbf{L}\boldsymbol{\beta} = (L1 + L2 + L3) \times \mu + L1 \times A_1 + L2 \times A_2 + L3 \times A_3$$

is estimable for any values of *L1*, *L2*, and *L3*.

If other generating sets for **L** are represented symbolically, the symbolic notation looks different. However, the inherent nature of the rules is the same. For example, if row operations are performed on $\mathbf{X}^*$ to produce an identity matrix in the first 3×3 submatrix of the resulting matrix

$$\mathbf{X}^{**} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

then $\mathbf{X}^{**}$ is also a generating set for **L**. An estimable **L** generated from $\mathbf{X}^{**}$ can be represented symbolically as

$$\mathbf{L} = (L1,\ L2,\ L3,\ L1 - L2 - L3)$$

Note that, again, the first element of **L** is equal to the sum of the other elements.

With multiple generating sets available, the question arises as to which one is the best to represent **L** symbolically. Clearly, a generating set containing a minimum of rows (of full row rank) and a maximum of zero elements is desirable.

The generalized g_2-inverse $(\mathbf{X}'\mathbf{X})^-$ of $\mathbf{X}'\mathbf{X}$ computed by the modified sweep operation (Goodnight 1979) has the property that $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}$ usually contains numerous zeros. For this reason, in PROC GLM the nonzero rows of $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}$ are used to represent **L** symbolically.

If the generating set represented symbolically is of full row rank, the number of symbols $(L1, L2, \ldots)$ represents the maximum rank of any testable hypothesis (in other words, the maximum number of linearly independent rows for any **L** matrix that can be constructed). By letting each symbol in turn take on the value of 1 while the others are set to 0, the original generating set can be reconstructed.

Introduction to Reduction Notation

Reduction notation can be used to represent differences in sums of squares (SS) for two models. The notation $R(\mu, A, B, C)$ denotes the complete main-effects model for effects *A*, *B*, and *C*. The notation

$$R(A \mid \mu, B, C)$$

denotes the difference between the model SS for the complete main-effects model containing *A*, *B*, and *C* and the model SS for the reduced model containing only *B* and *C*.

In other words, this notation represents the differences in model SS produced by

```
proc glm;
   class a b c;
   model y = a b c;
run;
```

and

```
proc glm;
   class b c;
   model y = b c;
run;
```

As another example, consider a regression equation with four independent variables. The notation $R(\beta_3, \beta_4 \mid \beta_1, \beta_2)$ denotes the differences in model SS between

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

This is the difference in the model SS for the models produced, respectively, by

```
model y = x1 x2 x3 x4;
```

and

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

The following examples demonstrate the ability to manipulate the symbolic representation of a generating set. Note that any operations performed on the symbolic notation have corresponding row operations that are performed on the generating set itself.

Examples

A One-Way Classification Model

For the model

$$Y = \mu + A_i + \epsilon \qquad i = 1, 2, 3$$

the general form of estimable functions $\mathbf{L}\boldsymbol{\beta}$ is (from the previous example)

$$\mathbf{L}\boldsymbol{\beta} = L1 \times \mu + L2 \times A_1 + L3 \times A_2 + (L1 - L2 - L3) \times A_3$$

Thus,

$$\mathbf{L} = (L1, L2, L3, L1 - L2 - L3)$$

Tests involving only the parameters A_1, A_2, and A_3 must have an $\mathbf{L}$ of the form

$$\mathbf{L} = (0, L2, L3, -L2 - L3)$$

Since this $\mathbf{L}$ for the A parameters involves only two symbols, hypotheses with at most two degrees of freedom can be constructed. For example, letting $(L2, L3)$ be $(1, 0)$ and $(0, 1)$, respectively, yields

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

The preceding $\mathbf{L}$ can be used to test the hypothesis that $A_1 = A_2 = A_3$. For this example, any $\mathbf{L}$ with two linearly independent rows with column 1 equal to zero produces the same sum of squares. For example, a joint test for linear and quadratic effects of A

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

gives the same SS. In fact, for any $\mathbf{L}$ of full row rank and any nonsingular matrix $\mathbf{K}$ of conformable dimensions,

$$\text{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = 0) = \text{SS}(H_0\colon\ \mathbf{KL}\boldsymbol{\beta} = 0)$$

A Three-Factor Main-Effects Model

Consider a three-factor main-effects model involving the CLASS variables A, B, and C, as shown in Table 15.1.

Table 15.1 Three-Factor Main-Effects Model

Obs	*A*	*B*	*C*
1	1	2	1
2	1	1	2
3	2	1	3
4	2	2	2
5	2	2	2

The general form of an estimable function is shown in Table 15.2.

Table 15.2 General Form of an Estimable Function for Three-Factor Main-Effects Model

Parameter	Coefficient
μ (Intercept)	*L1*
A1	*L2*
A2	*L1 – L2*
B1	*L4*
B2	*L1 – L4*
C1	*L6*
C2	*L1 + L2 – L4 – 2 × L6*
C3	*–L2 + L4 + L6*

Since only four symbols (*L1*, *L2*, *L4*, and *L6*) are involved, any testable hypothesis will have at most four degrees of freedom. If you form an $\mathbf{L}$ matrix with four linearly independent rows according to the preceding rules, then testing $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ is equivalent to testing that $\text{E}[\mathbf{Y}]$ is uniformly 0. Symbolically,

$$\text{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = 0) = R(\mu, A, B, C)$$

In a main-effects model, the usual hypothesis of interest for a main effect is the equality of all the parameters. In this example, it is not possible to unambiguously test such a hypothesis because of confounding: any test for the equality of the parameters for any one of *A*, *B*, or *C* will necessarily involve the parameters for the other two effects. One way to proceed is to construct a maximum rank hypothesis (MRH) involving only the parameters of the main effect in question. This can be done using the general form of estimable functions. Note the following:

- To get an MRH involving only the parameters of *A*, the coefficients of $\mathbf{L}$ associated with μ, *B1*, *B2*, *C1*, *C2*, and *C3* must be equated to zero. Starting at the top of the general form, let *L1* = 0, then *L4* = 0, then *L6* = 0. If *C2* and *C3* are not to be involved, then *L2* must also be zero. Thus, *A1* – *A2* is not estimable; that is, the MRH involving only the *A* parameters has zero rank and $R(A \mid \mu, B, C) = 0$.
- To obtain the MRH involving only the *B* parameters, let *L1* = *L2* = *L6* = 0. But then to remove *C2* and *C3* from the comparison, *L4* must also be set to 0. Thus, *B1* – *B2* is not estimable and $R(B \mid \mu, A, C) = 0$.
- To obtain the MRH involving only the *C* parameters, let *L1* = *L2* = *L4* =0. Thus, the MRH involving only *C* parameters is

  $$C1 - 2 \times C2 + C3 = K \qquad \text{(for any } K\text{)}$$

  or any multiple of the left-hand side equal to *K*. Furthermore,

  $$\text{SS}(H_0\text{: } C1 - 2 \times C2 + C3 = 0) = R(C \mid \mu, A, B)$$

A Multiple Regression Model

Suppose

$$\text{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where the $\mathbf{X}'\mathbf{X}$ matrix has full rank. The general form of estimable functions is as shown in Table 15.3.

Table 15.3 General Form of Estimable Functions for a Multiple Regression Model When $\mathbf{X}'\mathbf{X}$ Matrix Is of Full Rank

Parameter	Coefficient
β_0	*L1*
β_1	*L2*
β_2	*L3*
β_3	*L4*

For example, to test the hypothesis that $\beta_2 = 0$, let *L1* = *L2* = *L4* = 0 and let *L3* = 1. Then $\text{SS}(\mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = R(\beta_2 \mid \beta_0, \beta_1, \beta_3)$. In this full-rank case, all parameters, as well as any linear combination of parameters, are estimable.

Suppose, however, that $X3 = 2x_1 + 3x_2$. The general form of estimable functions is shown in Table 15.4.

Table 15.4 General Form of Estimable Functions for a Multiple Regression Model When $\mathbf{X}'\mathbf{X}$ Matrix Is Not of Full Rank

Parameter	Coefficient
β_0	*L1*
β_1	*L2*
β_2	*L3*
β_3	$2 \times L2 + 3 \times L3$

For this example, it is possible to test H_0: $\beta_0 = 0$. However, β_1, β_2, and β_3 are not jointly estimable; that is,

$$\begin{aligned} R(\beta_1 \mid \beta_0, \beta_2, \beta_3) &= 0 \\ R(\beta_2 \mid \beta_0, \beta_1, \beta_3) &= 0 \\ R(\beta_3 \mid \beta_0, \beta_1, \beta_2) &= 0 \end{aligned}$$

Estimable Functions

Type I SS and Estimable Functions

In PROC GLM, the Type I SS and the associated hypotheses they test are byproducts of the modified sweep operator used to compute a generalized g_2-inverse of $\mathbf{X}'\mathbf{X}$ and a solution to the normal equations. For the model $\mathrm{E}[Y] = x_1\beta_1 + x_2\beta_2 + x_3\beta_3$, the Type I SS for each effect are as follows:

Effect	Type I SS
x_1	$R(\beta_1)$
x_2	$R(\beta_2 \mid \beta_1)$
x_3	$R(\beta_3 \mid \beta_1, \beta_2)$

Note that some other SAS/STAT procedures compute Type I hypotheses by sweeping $\mathbf{X}'\mathbf{X}$ (for example, PROC MIXED and PROC GLIMMIX), but their test statistics are not necessarily equivalent to the results of using those procedures to fit models that contain successively more effects.

The Type I SS are model-order dependent; each effect is adjusted only for the preceding effects in the model.

There are numerous ways to obtain a Type I hypothesis matrix $\mathbf{L}$ for each effect. One way is to form the $\mathbf{X}'\mathbf{X}$ matrix and then reduce $\mathbf{X}'\mathbf{X}$ to an upper triangular matrix by row operations, skipping over any rows with a zero diagonal. The nonzero rows of the resulting matrix associated with x_1 provide an $\mathbf{L}$ such that

$$\mathrm{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = R(\beta_1)$$

The nonzero rows of the resulting matrix associated with x_2 provide an $\mathbf{L}$ such that

$$\text{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = R(\beta_2 \mid \beta_1)$$

The last set of nonzero rows (associated with x_3) provide an $\mathbf{L}$ such that

$$\text{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = R(\beta_3 \mid \beta_1, \beta_2)$$

Another more formalized representation of Type I generating sets for x_1, x_2, and x_3, respectively, is

$$\begin{array}{lcccccc}
\mathbf{G}_1 & = & (& \mathbf{X}_1'\mathbf{X}_1 & \mid & \mathbf{X}_1'\mathbf{X}_2 & \mid & \mathbf{X}_1'\mathbf{X}_3 &) \\
\mathbf{G}_2 & = & (& 0 & \mid & \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2 & \mid & \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_3 &) \\
\mathbf{G}_3 & = & (& 0 & \mid & 0 & \mid & \mathbf{X}_3'\mathbf{M}_2\mathbf{X}_3 &)
\end{array}$$

where

$$\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1'$$

and

$$\mathbf{M}_2 = \mathbf{M}_1 - \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1$$

Using the Type I generating set $\mathbf{G}_2$ (for example), if an $\mathbf{L}$ is formed from linear combinations of the rows of $\mathbf{G}_2$ such that $\mathbf{L}$ is of full row rank and of the same row rank as $\mathbf{G}_2$, then $\text{SS}(H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = R(\beta_2 \mid \beta_1)$.

In the GLM procedure, the Type I estimable functions displayed symbolically when the E1 option is requested are

$$\begin{aligned}
\mathbf{G}_1^* &= (\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{G}_1 \\
\mathbf{G}_2^* &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{G}_2 \\
\mathbf{G}_3^* &= (\mathbf{X}_3'\mathbf{M}_2\mathbf{X}_3)^-\mathbf{G}_3
\end{aligned}$$

As can be seen from the nature of the generating sets $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$, only the Type I estimable functions for β_3 are guaranteed not to involve the β_1 and β_2 parameters. The Type I hypothesis for β_2 can (and often does) involve β_3 parameters, and likewise the Type I hypothesis for β_1 often involves β_2 and β_3 parameters.

There are, however, a number of models for which the Type I hypotheses are considered appropriate. These are as follows:

- balanced ANOVA models specified in proper sequence (that is, interactions do not precede main effects in the MODEL statement and so forth)
- purely nested models (specified in the proper sequence)
- polynomial regression models (in the proper sequence)

Type II SS and Estimable Functions

For main-effects models and regression models, the general form of estimable functions can be manipulated to provide tests of hypotheses involving only the parameters of the effect in question. The same result can also be obtained by entering each effect in turn as the last effect in the model and obtaining the Type I SS for that effect. These are the *Type II SS*. Using a modified reversible sweep operator, it is possible to obtain the Type II SS without actually refitting the model.

Thus, the **Type II SS correspond to the R notation in which each effect is adjusted for all other appropriate effects**. For a regression model such as

$$\mathrm{E}[Y] = x_1\beta_1 + x_2\beta_2 + x_3\beta_3$$

the Type II SS correspond to

Effect	Type II SS
x_1	$R(\beta_1 \mid \beta_2, \beta_3)$
x_2	$R(\beta_2 \mid \beta_1, \beta_3)$
x_3	$R(\beta_3 \mid \beta_1, \beta_2)$

For a main-effects model (A, B, and C as classification variables), the Type II SS correspond to

Effect	Type II SS
A	$R(A \mid B, C)$
B	$R(B \mid A, C)$
C	$R(C \mid A, B)$

As the discussion in the section "A Three-Factor Main-Effects Model" on page 261 indicates, for regression and main-effects models the Type II SS provide an MRH for each effect that does not involve the parameters of the other effects.

In order to see what effects are appropriate to adjust for in computing Type II estimable functions, note that for models involving interactions and nested effects, in the absence of a priori parametric restrictions, it is not possible to obtain a test of a hypothesis for a main effect free of parameters of higher-level interactions effects with which the main effect is involved. It is reasonable to assume, then, that any test of a hypothesis concerning an effect should involve the parameters of that effect and only those other parameters with which that effect is involved. The concept of effect containment helps to define this involvement.

Contained Effect

Given two effects *F1* and *F2*, *F1* is said to be *contained in F2* provided that the following two conditions are met:

- Both effects involve the same continuous variables (if any).
- *F2* has more CLASS variables than *F1* does, and if *F1* has CLASS variables, they all appear in *F2*.

Note that the intercept effect μ is contained in all pure CLASS effects, but it is not contained in any effect involving a continuous variable. No effect is contained by μ.

Type II, Type III, and Type IV estimable functions rely on this definition, and they all have one thing in common: the estimable functions involving an effect *F1* also involve the parameters of all effects that contain *F1*, and they do not involve the parameters of effects that do not contain *F1* (other than *F1*).

Hypothesis Matrix for Type II Estimable Functions

The Type II estimable functions for an effect *F1* have an $\mathbf{L}$ (before reduction to full row rank) of the following form:

- All columns of $\mathbf{L}$ associated with effects not containing *F1* (except *F1*) are zero.
- The submatrix of $\mathbf{L}$ associated with effect *F1* is $(\mathbf{X}_1'\mathbf{M}\mathbf{X}_1)^-(\mathbf{X}_1'\mathbf{M}\mathbf{X}_1)$.
- Each of the remaining submatrices of $\mathbf{L}$ associated with an effect *F2* that contains *F1* is $(\mathbf{X}_1'\mathbf{M}\mathbf{X}_1)^-(\mathbf{X}_1'\mathbf{M}\mathbf{X}_2)$.

In these submatrices,

$$\begin{aligned}
\mathbf{X}_0 &= \text{the columns of } \mathbf{X} \text{ whose associated effects do not contain } F1 \\
\mathbf{X}_1 &= \text{the columns of } \mathbf{X} \text{ associated with } F1 \\
\mathbf{X}_2 &= \text{the columns of } \mathbf{X} \text{ associated with an } F2 \text{ effect that contains } F1 \\
\mathbf{M} &= \mathbf{I} - \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^-\mathbf{X}_0'
\end{aligned}$$

For the model

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

the Type II SS correspond to

$$R(A \mid \mu, B), \quad R(B \mid \mu, A), \quad R(A * B \mid \mu, A, B)$$

for effects A, B, and A * B, respectively. For the model

```
class A B C;
model Y =  A B(A) C(A B);
```

the Type II SS correspond to

$$R(A \mid \mu), \quad R(B(A) \mid \mu, A), \quad R(C(AB) \mid \mu, A, B(A))$$

for effects A, $B(A)$ and $C(A\ B)$, respectively. For the model

```
model Y = x x*x;
```

the Type II SS correspond to

$$R(X \mid \mu, X * X) \quad \text{and} \quad R(X * X \mid \mu, X)$$

for x and $x * x$, respectively.

Note that, as in the situation for Type I tests, PROC MIXED and PROC GLIMMIX compute Type I hypotheses by sweeping $\mathbf{X}'\mathbf{X}$, but their test statistics are not necessarily equivalent to the results of sequentially fitting with those procedures models that contain successively more effects; while PROC TRANSREG computes tests labeled as being Type II by leaving out each effect in turn, but the specific linear hypotheses associated with these tests might not be precisely the same as the ones derived from successively sweeping $\mathbf{X}'\mathbf{X}$.

Example of Type II Estimable Functions

For a 2×2 factorial with w observations per cell, the general form of estimable functions is shown in Table 15.5. Any nonzero values for *L2*, *L4*, and *L6* can be used to construct $\mathbf{L}$ vectors for computing the Type II SS for *A*, *B*, and *A* * *B*, respectively.

Table 15.5 General Form of Estimable Functions for 2×2 Factorial

Effect	Coefficient
μ	*L1*
A1	*L2*
A2	*L1 – L2*
B1	*L4*
B2	*L1 – L4*
AB11	*L6*
AB12	*L2 – L6*
AB21	*L4 – L6*
AB22	*L1 – L2 – L4 + L6*

For a balanced 2×2 factorial with the same number of observations in every cell, the Type II estimable functions are shown in Table 15.6.

Table 15.6 Type II Estimable Functions for Balanced 2×2 Factorial

	Coefficients for Effect		
Effect	***A***	***B***	***A * B***
μ	0	0	0
A1	*L2*	0	0
A2	*–L2*	0	0
B1	0	*L4*	0
B2	0	*–L4*	0
AB11	$0.5 \times L2$	$0.5 \times L4$	*L6*
AB12	$0.5 \times L2$	$-0.5 \times L4$	*–L6*
AB21	$-0.5 \times L2$	$0.5 \times L4$	*–L6*
AB22	$-0.5 \times L2$	$-0.5 \times L4$	*L6*

Now consider an unbalanced 2×2 factorial with two observations in every cell except the *AB22* cell, which contains only one observation. The general form of estimable functions is the same as if it were balanced, since the same effects are still estimable. However, the Type II estimable functions for *A* and *B* are not the same as they were for the balanced design. The Type II estimable functions for this unbalanced 2×2 factorial are shown in Table 15.7.

Table 15.7 Type II Estimable Functions for Unbalanced 2×2 Factorial

	Coefficients for Effect		
Effect	***A***	***B***	***A * B***
μ	0	0	0
A1	*L2*	0	0
A2	*–L2*	0	0
B1	0	*L4*	0
B2	0	*–L4*	0
AB11	$0.6 \times L2$	$0.6 \times L4$	*L6*
AB12	$0.4 \times L2$	$-0.6 \times L4$	*–L6*
AB21	$-0.6 \times L2$	$0.4 \times L4$	*–L6*
AB22	$-0.4 \times L2$	$-0.4 \times L4$	*L6*

By comparing the hypothesis being tested in the balanced case to the hypothesis being tested in the unbalanced case for effects *A* and *B*, you can note that the Type II hypotheses for *A* and *B* are dependent on the cell frequencies in the design. For unbalanced designs in which the cell frequencies are not proportional to the background population, the Type II hypotheses for effects that are contained in other effects are of questionable value.

However, if an effect is not contained in any other effect, the Type II hypothesis for that effect is an MRH that does not involve any parameters except those associated with the effect in question.

Thus, Type II SS are appropriate for the following models:

- any balanced model
- any main-effects model
- any pure regression model
- an effect not contained in any other effect (regardless of the model)

In addition to the preceding models, Type II SS are generally accepted by most statisticians for purely nested models.

Type III and IV SS and Estimable Functions

When an effect is contained in another effect, the Type II hypotheses for that effect are dependent on the cell frequencies. The philosophy behind both the Type III and Type IV hypotheses is that the hypotheses tested for any given effect should be the same for all designs with the same general form of estimable functions.

To demonstrate this concept, recall the hypotheses being tested by the Type II SS in the balanced 2×2 factorial shown in Table 15.6. Those hypotheses are precisely the ones that the Type III and Type IV hypotheses employ for all 2×2 factorials that have at least one observation per cell. The Type III and Type IV hypotheses for a design without missing cells usually differ from the hypothesis employed for the same design with missing cells since the general form of estimable functions usually differs.

Many SAS/STAT procedures can perform tests of Type III hypotheses, but only PROC GLM offers Type IV tests as well.

Type III Estimable Functions

Type III hypotheses are constructed by working directly with the general form of estimable functions. The following steps are used to construct a hypothesis for an effect *F1*:

1. For every effect in the model except *F1* and those effects that contain *F1*, equate the coefficients in the general form of estimable functions to zero.

   If *F1* is not contained in any other effect, this step defines the Type III hypothesis (as well as the Type II and Type IV hypotheses). If *F1* is contained in other effects, go on to step 2. (See the section "Type II SS and Estimable Functions" on page 265 for a definition of when effect *F1* is contained in another effect.)

2. If necessary, equate new symbols to compound expressions in the *F1* block in order to obtain the simplest form for the *F1* coefficients.

3. Equate all symbolic coefficients outside the *F1* block to a linear function of the symbols in the *F1* block in order to make the *F1* hypothesis orthogonal to hypotheses associated with effects that contain *F1*.

By once again observing the Type II hypotheses being tested in the balanced 2×2 factorial, it is possible to verify that the *A* and *A* * *B* hypotheses are orthogonal and also that the *B* and *A* * *B* hypotheses are orthogonal. This principle of orthogonality between an effect and any effect that contains it holds for all balanced designs. Thus, construction of Type III hypotheses for any design is a logical extension of a process that is used for balanced designs.

The Type III hypotheses are precisely the hypotheses being tested by programs that reparameterize using the usual assumptions (for example, constraining all parameters for an effect to sum to zero). When no missing cells exist in a factorial model, Type III SS coincide with Yates' weighted squares-of-means technique. When cells are missing in factorial models, the Type III SS coincide with those discussed in Harvey (1960) and Henderson (1953).

The following discussion illustrates the construction of Type III estimable functions for a 2×2 factorial with no missing cells.

To obtain the *A* * *B* interaction hypothesis, start with the general form and equate the coefficients for effects μ, *A*, and *B* to zero, as shown in Table 15.8.

Table 15.8 Type III Hypothesis for $A * B$ Interaction

Effect	General Form	$L1 = L2 = L4 = 0$
μ	$L1$	0
$A1$	$L2$	0
$A2$	$L1 - L2$	0
$B1$	$L4$	0
$B2$	$L1 - L4$	0
$AB11$	$L6$	$L6$
$AB12$	$L2 - L6$	$-L6$
$AB21$	$L4 - L6$	$-L6$
$AB22$	$L1 - L2 - L4 + L6$	$L6$

The last column in Table 15.8 represents the form of the MRH for $A * B$.

To obtain the Type III hypothesis for A, first start with the general form and equate the coefficients for effects μ and B to zero (let $L1 = L4 = 0$). Next let $L6 = K \times L2$, and find the value of K that makes the A hypothesis orthogonal to the $A * B$ hypothesis. In this case, $K = 0.5$. Each of these steps is shown in Table 15.9.

In Table 15.9, the fourth column (under $L6 = K \times L2$) represents the form of all estimable functions not involving μ, $B1$, or $B2$. The prime difference between the Type II and Type III hypotheses for A is the way K is determined. Type II chooses K as a function of the cell frequencies, whereas Type III chooses K such that the estimable functions for A are orthogonal to the estimable functions for $A * B$.

Table 15.9 Type III Hypothesis for A

Effect	General Form	$L1 = L4 = 0$	$L6 = K \times L2$	$K = 0.5$
μ	$L1$	0	0	0
$A1$	$L2$	$L2$	$L2$	$L2$
$A2$	$L1 - L2$	$-L2$	$-L2$	$-L2$
$B1$	$L4$	0	0	0
$B2$	$L1 - L4$	0	0	0
$AB11$	$L6$	$L6$	$K \times L2$	$0.5 \times L2$
$AB12$	$L2 - L6$	$L2 - L6$	$(1 - K) \times L2$	$0.5 \times L2$
$AB21$	$L4 - L6$	$-L6$	$-K \times L2$	$-0.5 \times L2$
$AB22$	$L1 - L2 - L4 + L6$	$-L2 + L6$	$-(1 - K) \times L2$	$-0.5 \times L2$

An example of Type III estimable functions in a 3×3 factorial with unequal cell frequencies and missing diagonals is given in Table 15.10 (N_1 through N_6 represent the nonzero cell frequencies).

Table 15.10 3×3 Factorial Design with Unequal Cell Frequencies and Missing Diagonals

		B		
		1	2	3
A	1		N_1	N_2
	2	N_3		N_4
	3	N_5	N_6	

For any nonzero values of N_1 through N_6, the Type III estimable functions for each effect are shown in Table 15.11.

Table 15.11 Type III Estimable Functions for 3×3 Factorial Design with Unequal Cell Frequencies and Missing Diagonals

Effect	*A*	*B*	*A* * *B*
μ	0	0	0
A1	*L2*	0	0
A2	*L3*	0	0
A3	$-L2 - L3$	0	0
B1	0	*L5*	0
B2	0	*L6*	0
B3	0	$-L5 - L6$	0
AB12	$0.667 \times L2 + 0.333 \times L3$	$0.333 \times L5 + 0.667 \times L6$	*L8*
AB13	$0.333 \times L2 - 0.333 \times L3$	$-0.333 \times L5 - 0.667 \times L6$	$-L8$
AB21	$0.333 \times L2 + 0.667 \times L3$	$0.667 \times L5 + 0.333 \times L6$	$-L8$
AB23	$-0.333 \times L2 + 0.333 \times L3$	$-0.667 \times L5 - 0.333 \times L6$	*L8*
AB31	$-0.333 \times L2 - 0.667 \times L3$	$0.333 \times L5 - 0.333 \times L6$	*L8*
AB32	$-0.667 \times L2 - 0.333 \times L3$	$-0.333 \times L5 + 0.333 \times L6$	$-L8$

Type IV Estimable Functions

By once again looking at the Type II hypotheses being tested in the balanced 2×2 factorial (see Table 15.6), you can see another characteristic of the hypotheses employed for balanced designs: the coefficients of lower-order effects are averaged across each higher-level effect involving the same subscripts. For example, in the *A* hypothesis, the coefficients of *AB11* and *AB12* are equal to one-half the coefficient of *A1*, and the coefficients of *AB21* and *AB22* are equal to one-half the coefficient of *A2*. With this in mind, the basic concept used to construct Type IV hypotheses is that the coefficients of any effect, say *F1*, are distributed equitably across higher-level effects that contain *F1*. When missing cells occur, this same general philosophy is adhered to, but care must be taken in the way the distributive concept is applied.

Construction of Type IV hypotheses begins as does the construction of the Type III hypotheses. That is, for an effect *F1*, equate to zero all coefficients in the general form that do not belong to *F1* or to any other effect containing *F1*. If *F1* is not contained in any other effect, then the Type IV hypothesis (and Type II and III) has been found. If *F1* is contained in other effects, then simplify, if necessary, the coefficients associated with *F1* so that they are all free coefficients or functions of other free coefficients in the *F1* block.

To illustrate the method of resolving the free coefficients outside the *F1* block, suppose that you are interested in the estimable functions for an effect *A* and that *A* is contained in *AB*, *AC*, and *ABC*. (In other words, the main effects in the model are *A*, *B*, and *C*.)

With missing cells, the coefficients of intermediate effects (here they are *AB* and *AC*) do not always have an equal distribution of the lower-order coefficients, so the coefficients of the highest-order effects are determined first (here it is *ABC*). Once the highest-order coefficients are determined, the coefficients of intermediate effects are automatically determined.

The following process is performed for each free coefficient of *A* in turn. The resulting symbolic vectors are then added together to give the Type IV estimable functions for *A*.

1. Select a free coefficient of *A*, and set all other free coefficients of *A* to zero.
2. If any of the levels of *A* have zero as a coefficient, equate all of the coefficients of higher-level effects involving that level of *A* to zero. This step alone usually resolves most of the free coefficients remaining.
3. Check to see if any higher-level coefficients are now zero when the coefficient of the associated level of *A* is not zero. If this situation occurs, the Type IV estimable functions for *A* are not unique.
4. For each level of *A* in turn, if the *A* coefficient for that level is nonzero, count the number of times that level occurs in the higher-level effect. Then equate each of the higher-level coefficients to the coefficient of that level of *A* divided by the count.

An example of a 3 × 3 factorial with four missing cells (N_1 through N_5 represent positive cell frequencies) is shown in Table 15.12.

Table 15.12 3 × 3 Factorial Design with Four Missing Cells

		B		
		1	2	3
A	1	N_1	N_2	
	2	N_3	N_4	
	3			N_5

The Type IV estimable functions are shown in Table 15.13.

Table 15.13 Type IV Estimable Functions for 3 × 3 Factorial Design with Four Missing Cells

Effect	*A*	*B*	*A* * *B*
μ	0	0	0
A1	–*L3*	0	0
A2	*L3*	0	0
A3	0	0	0
B1	0	*L5*	0
B2	0	–*L5*	0
B3	0	0	0
AB11	–0.5 × *L3*	0.5 × *L5*	*L8*
AB12	–0.5 × *L3*	–0.5 × *L5*	–*L8*
AB21	0.5 × *L3*	0.5 × *L5*	–*L8*
AB22	0.5 × *L3*	–0.5 × *L5*	*L8*
AB33	0	0	0

A Comparison of Type III and Type IV Hypotheses

For the vast majority of designs, Type III and Type IV hypotheses for a given effect are the same. Specifically, they are the same for any effect *F1* that is not contained in other effects for any design (with or without missing cells). For factorial designs with no missing cells, the Type III and Type IV hypotheses coincide for all effects. When there are missing cells, the hypotheses can differ. By using the GLM procedure, you

can study the differences in the hypotheses and then decide on the appropriateness of the hypotheses for a particular model.

The Type III hypotheses for three-factor and higher completely nested designs with unequal *N*s in the lowest level differ from the Type II hypotheses; however, the Type IV hypotheses do correspond to the Type II hypotheses in this case.

When missing cells occur in a design, the Type IV hypotheses might not be unique. If this occurs in PROC GLM, you are notified, and you might need to consider defining your own specific comparisons.

References

Freund, R. J., Littell, R. C., and Spector, P. C. (1991). *SAS System for Linear Models*. Cary, NC: SAS Institute Inc.

Goodnight, J. H. (1978). *Tests of Hypotheses in Fixed-Effects Linear Models*. Technical Report R-101, SAS Institute Inc., Cary, NC.

Goodnight, J. H. (1979). "A Tutorial on the Sweep Operator." *American Statistician* 33:149–158.

Harvey, W. R. (1960). *Least-Squares Analysis of Data with Unequal Subclass Frequencies*. Technical Report ARS 20-8, Agriculture Research Service, U.S. Department of Agriculture.

Henderson, C. R. (1953). "Estimation of Variance and Covariance Components." *Biometrics* 9:226–252.

Milliken, G. A., and Johnson, D. E. (1984). *Designed Experiments.* Vol. 1 of Analysis of Messy Data. Belmont, CA: Lifetime Learning Publications.