Types I, II, and III Analyses in Linear Models

Scott K. Cooley

Department of Mathematical Sciences

Montana State University

December 11, 1997

A writing project submitted in partial fulfillment

of the requirements for the degree

Master of Science in Statistics

Approval

of writing project submitted by

Scott K. Cooley

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Dec 22 1997

Date

Robert J. Boik

Writing Project Director

Writing Project Outline

A. Introduction

1. Goals of Estimation & Hypothesis Testing
2. Estimability, Testability

B. Review of R-Notation

1. Interpretation
2. Formulas

C. Type I Analysis

1. Interpretation
2. Type I Sums of Squares
3. Type I Hypotheses
4. Type I Estimable Functions
5. Using SAS® to conduct a Type I Analysis
6. Role of Type I Analysis

D. Type II Analysis

1. Interpretation
2. Type II Sums of Squares

3. Type II Hypotheses

4. Type II Estimable Functions

5. Using SAS® to conduct a Type II Analysis

6. Role of Type II Analysis

E. Type III Analysis

1. Interpretation

2. Type III Sums of Squares

3. Type III Hypotheses

4. Type III Estimable Functions

5. Using SAS® to conduct a Type III Analysis

6. Role of Type III Analysis

F. Summary

G. References

H. Appendix

1. MATLAB Programs

2. General Form of F-statistic for Type I Analysis

Types I, II, and III Analyses in Linear Models

Introduction

The conventional linear model can be written as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad (1)$$

where $\mathbf{y}$ is the n-vector of responses (that is, $\mathbf{y}$: $n \times 1$ is the vector of responses), $\mathbf{X}$ is an $n \times p$ design matrix having rank $r \le p$, β is an unknown p-vector of regression coefficients, and ε is an n-vector of random errors. It is assumed in this paper that $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ so that $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, and that σ^2 is an unknown scalar. Statistical analyses of models having the form in (1) often have one of two primary goals (SAS/STAT, 1988). The first goal is estimating the elements of β (or linear combinations of the elements of β). The second goal is testing hypotheses about the elements of β (or linear combinations of the elements of β).

Denote the perpendicular projection operator onto the column space of $\mathbf{X}$ by ppo($\mathbf{X}$), or, alternatively, by $\mathbf{H}$. An explicit equation for $\mathbf{H}$ is

$$\mathbf{H} = \text{ppo}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}',$$

where $(\mathbf{X}'\mathbf{X})^{-}$ is any generalized inverse of $\mathbf{X}'\mathbf{X}$. If rank($\mathbf{X}$) = p, then $(\mathbf{X}'\mathbf{X})^{-}$ can be replaced by $(\mathbf{X}'\mathbf{X})^{-1}$. In some models it is convenient to partition $\mathbf{X}$ and β. For example, $\mathbf{X}$ and β might be conformably partitioned as

$$\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3) \text{ and } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix},$$

where $\mathbf{X}_i$ is $n \times p_i$ and β_i is $p_i \times 1$ for $i = 1,2,3$. We also require that $\sum_{i=1}^{3} p_i = p$. Each sub-matrix of $\mathbf{X}$ generates a projection operator:

$$\mathbf{H}_i = \text{ppo}(\mathbf{X}_i) = \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)\mathbf{X}_i'.$$

Let $\mathbf{L}\beta$ be a particular linear function of β of interest. If we wish to estimate $\mathbf{L}\beta$, we must first determine whether or not $\mathbf{L}\beta$ is estimable. By definition, $\mathbf{L}\beta$ is linearly estimable if and only if a linear combination of the $\mathbf{y}$'s exists which has an expected value of $\mathbf{L}\beta$, and this must hold for all β (SAS/STAT, 1988). That is, there must exist $\mathbf{A}$ and $\mathbf{K}$

such that $E(\mathbf{Ay+K}) = \mathbf{L}\beta$ for all β. But $E(\mathbf{Ay+K}) = \mathbf{A}E(\mathbf{y}) + E(\mathbf{K}) = \mathbf{AX}\beta + \mathbf{K}$, so $\beta = \mathbf{0}$, and $E(\mathbf{Ay+K}) = \mathbf{L}\beta = \mathbf{AX}\beta + \mathbf{K} = \mathbf{0}$ together imply that $\mathbf{K} = \mathbf{0}$. Thus, if $\mathbf{L}\beta$ is to be estimable, we must have $\mathbf{L}\beta = \mathbf{AX}\beta$ for some $\mathbf{A}$ and for all β. If we now let $\beta = \mathbf{e}_i$, the i^{th} column of $\mathbf{I}_p$ for $i = 1,2,\ldots, p$ we can conclude that $\mathbf{L}_i = (\mathbf{AX})_i$ and thus that $\mathbf{L} = \mathbf{AX}$ for some $\mathbf{A}$. This says that $\mathbf{L}\beta$ will be estimable if and only if a linear combination of the rows of $\mathbf{X}$ can be found which equals $\mathbf{L}$. That is, $\mathbf{L}'$ must be an element of the column space of $\mathbf{X}'$.

An understanding of estimablilty is important if the goal of the analysis is estimation, this is not surprising. However, it is also important that we understand estimability if the goal of the analysis is hypothesis testing because the definition of a testable hypothesis involves the concept of estimability. A testable hypothesis is a hypothesis which can be expressed in terms of estimable functions (Searle, 1971). So if we wish to test H_o: $\mathbf{L}\beta = \delta$, then $\mathbf{L}\beta$ must be estimable and $\delta \in R(\mathbf{L})$ must hold where $R(\mathbf{L})$ represents the column space of $\mathbf{L}$.

Review of R-Notation

For convenience of notation in the upcoming portions of this paper, it is worthwhile to briefly review some basics of R-Notation.

Some important sums of squares we need to review are SST, SSE, and SSR. The total sum of squares is $SST = \mathbf{y'y} = \sum_i y_i^2$. Notice that SST does not depend on the model we intend to use for the analysis. The sum of squares of the deviations (or sum of squared deviations) of the observed y's from their predicted values is $SSE = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y}-\mathbf{Hy})'(\mathbf{y}-\mathbf{Hy}) = \mathbf{y'(I-H)y} = \mathbf{y'y} - \mathbf{y'Hy}$ because $\hat{\mathbf{y}} = \hat{\mathbf{E}}(\mathbf{y}) = \mathbf{X}\hat{\beta} = \mathbf{X(X'X)^{-}X'y} = \mathbf{Hy}$ (Searle, 1982). Notice that because SSE involves $\mathbf{H}$, it reflects the choice of the model. Using $\mathbf{H}$ implies that we are using the entire parameter vector β, so we are saying that $E(\mathbf{y}) = \mathbf{X}\beta$. If we wanted to use only part of the parameter vector β, say β_1, that is, if we assume that $E(\mathbf{y}) = \mathbf{X}_1\beta_1$, then we would use $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1'$ rather than $\mathbf{H}$. So under the assumption that $E(\mathbf{y}) = \mathbf{X}_1\beta_1$, we would have $SSE_1 = \mathbf{y'(I-H}_1\mathbf{)y}$. Similar constructions of SSE's are possible using other parts of β along with their corresponding design matrices in place of β_1 and $\mathbf{X}_1$. In this way we see that SSE reflects the model we have selected for the analysis. SSE is also referred to as the residual sum of squares or sum of squares for error. The final sum of squares we need to mention is SSR, the sum of squares due to regression, the regression sum of squares, or the reduction in sum of squares (Searle, 1982). By definition, $SSR = SST - SSE = \mathbf{y'Hy}$, under the assumption that $E(\mathbf{y}) = \mathbf{X}\beta$. SSR represents the portion of SST attributable to having fit the model chosen for the analysis as indicated by the construction of SSE. So if we were working

under the assumption that $E(\mathbf{y}) = \mathbf{X}_1\beta_1$, we would have formed $SSE_1 = \mathbf{y}'(\mathbf{I}-\mathbf{H}_1)\mathbf{y}$ and hence we would have $SSR_1 = \mathbf{y}'\mathbf{H}_1\mathbf{y}$.

It is important to note that SST is a fixed quantity based only on the observed y_i values, not on the model. SSE is the sum of squared deviances $\sum_i (y_i - \hat{y}_i)^2$, thus it is a measure of the accuracy of the selected model. The smaller the value of SSE, the more accurate the model. So since SSR = SST – SSE, we see that larger values of SSR are better in the sense that they indicate a better fitting model, or in other words, larger values of SSR indicate that the selected model has accounted for a larger portion of the total sum of squares SST. In this way, we can compare various models based on our choice of the parameter vector we wish to use. We can try using the entire parameter vector β, or we can use only part of β such as β_1, β_2, β_3, $\beta_{12} = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$, $\beta_{13} = \begin{pmatrix}\beta_1\\ \beta_3\end{pmatrix}$, or $\beta_{23} = \begin{pmatrix}\beta_2\\ \beta_3\end{pmatrix}$. Note that if we wish to use $\beta_{12} = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$, we would need to use the corresponding design matrix $\mathbf{X}_{12} = (\mathbf{X}_1\ \mathbf{X}_2)$ and $\mathbf{H}_{12} = \mathbf{X}_{12}(\mathbf{X}_{12}'\mathbf{X}_{12})^{-}\mathbf{X}_{12}'$ when forming SSE_{12} and SSR_{12}.

With this review of SST, SSE, and SSR, we are ready to define the R-Notations R(·) and R(·|·)

R-Notation is intended to simplify the notation involved in representing the various sums of squares as we compare different models of interest. The R-Notation R(·) represents the reduction in sum of squares due to fitting a model having a specified parameter vector. So under $E(\mathbf{y}) = \mathbf{X}\beta$, $R(\beta) = SSR = \mathbf{y}'\mathbf{H}\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ as mentioned previously. Under $E(\mathbf{y}) = \mathbf{X}_1\beta_1$, $R(\beta_1) = \mathbf{y}'\mathbf{H}_1\mathbf{y}$ which we have denoted as SSR_1. Under $E(\mathbf{y}) = \mathbf{X}_{12}\beta_{12} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$, then $R(\beta_{12}) = \mathbf{y}'\mathbf{H}_{12}\mathbf{y}$ which we can denote as SSR_{12}, and so on. Note, $R(\beta_{12})$ is commonly denoted as $R(\beta_1, \beta_2)$.

The R-Notation R(·|·) is used to represent the reduction in sums of squares due to adding a specified term or terms to a model which currently contains certain other specified terms (Searle, 1982). For example, suppose we wish to compare the reduction in sums of squares due to fitting $E(\mathbf{y}) = \mathbf{X}_1\beta_1$ versus $E(\mathbf{y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 = \mathbf{X}_{12}\beta_{12}$. The difference between the reductions in sums of squares for these two models is $SSR_{12} - SSR_1 = (SST - SSE_{12}) - (SST - SSE_1) = SSE_1 - SSE_{12} = R(\beta_1, \beta_2) - R(\beta_1)$. It is this difference which we denote as $R(\beta_2|\beta_1)$. It represents the additional reduction in the sums of squares due to fitting $E(\mathbf{y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ above and beyond the reduction in sums of squares due to fitting $E(\mathbf{y}) = \mathbf{X}_1\beta_1$. Another way to interpret $R(\beta_2|\beta_1)$ is to say that it gives the amount by which the residual sums of squares can be reduced by including the term $\mathbf{X}_2\beta_2$ to a model which currently contains only $\mathbf{X}_1\beta_1$. The idea is that if this additional reduction in residual sums of squares is significant, then we would prefer the model which contains both terms $\mathbf{X}_1\beta_1$ and $\mathbf{X}_2\beta_2$. If including the term $\mathbf{X}_2\beta_2$ in the model does

not lead to a significant reduction in the residual sums of squares, then it is not deemed necessary to include the additional term $\mathbf{X}_2\beta_2$.

Similarly, we can check the effectiveness of adding the term $\mathbf{X}_1\beta_1$ to a model which currently contains only $\mathbf{X}_2\beta_2$. This would be accomplished by considering the difference $R(\beta_1 | \beta_2) = R(\beta_1, \beta_2) - R(\beta_2)$. This process applies in general to other models we may wish to consider, we need only be careful to properly account for the terms of interest.

Type I Analysis

The interpretation of the Type I Analysis is that it is a sequential analysis (Milliken and Johnson, 1992). For example, we could start with the model $y_{ijkl} = \mu + \varepsilon_{ijkl}$ where μ is the grand mean or intercept term. Then sequentially add terms to obtain the three-way model without interaction $y_{ijkl} = \mu + \alpha_i + \tau_j + \gamma_k + \varepsilon_{ijkl}$. The order in which the terms are added to the model is important in a Type I Analysis. So the Type I Analysis is model-order dependent (SAS/STAT, 1988), each effect is adjusted only for the preceding effects in the model. This can be seen when we consider the sums of squares involved in the Type I Analysis.

Suppose we had started with the model $y_{ijkl} = \mu + \varepsilon_{ijkl}$ and that the regression sum of squares was obtained for this model. Denote this regression sum of squares as SSR_1, or using R-Notation, we can denote this as $R(\mu)$. Now add Factor A to the model, represented by α_i, and obtain the regression sum of squares for the new model $y_{ijkl} = \mu + \alpha_i + \varepsilon_{ijkl}$. Denote this regression sum of squares as SSR_{12}, or using R-Notation, $R(\mu,\alpha)$. The difference $SSR_{12} - SSR_1$ is therefore $R(\alpha|\mu)$ and is also equal to $SSE_1 - SSE_{12}$. Thus, $R(\alpha|\mu)$ is the reduction in the residual sum of squares due to adding Factor A to the model which initially had only the intercept μ.

We could now add a second factor, Factor B, denoted by τ_j. So the model is now $y_{ijkl} = \mu + \alpha_i + \tau_j + \varepsilon_{ijkl}$. Again, we obtain the regression sum of squares for this model which we can denote as SSR_{123}. Note that $SSR_{123} = R(\mu,\alpha,\tau)$. The difference $SSR_{123} - SSR_{12}$ is $R(\tau|\mu,\alpha)$, which is also equal to $SSE_{12} - SSE_{123}$. Thus, $R(\tau|\mu,\alpha)$ is the reduction in residual sum of squares due to adding Factor B to a model which contained the intercept μ and Factor A.

Finally, we could add the third factor to the model, Factor C, denoted by γ_k. This gives us the three-factor model $y_{ijkl} = \mu + \alpha_i + \tau_j + \gamma_k + \varepsilon_{ijkl}$ mentioned previously. We would again obtain the regression sum of squares for this model, denoted by SSR_{1234}.

Note that $SSR_{1234} = R(\mu,\alpha,\tau,\gamma)$. The difference $SSR_{1234} - SSR_{123}$ is $R(\gamma|\mu,\alpha,\tau)$, which is also equal to $SSE_{123} - SSE_{1234}$. Thus, $R(\gamma|\mu,\alpha,\tau)$ is the reduction in residual sum of squares due to adding Factor C to the modal which includes the intercept, Factor A, and Factor B.

The Type I Sum of Squares for the factors of this example are $R(\alpha|\mu)$, $R(\tau|\mu,\alpha)$, and $R(\gamma|\mu,\alpha,\tau)$. We emphasize that the sequential order in which the factors were added to the model is important. We added the factors in the order Factor A, Factor B, then Factor C. If we had added the factors in the order Factor C, Factor A, Factor B, then we would see different results. Under this new sequence, the Type I Sum of Squares would be $R(\gamma|\mu)$, $R(\alpha|\mu,\gamma)$, and $R(\tau|\mu,\gamma,\alpha)$. These sums of squares could be different than those obtained previously, they represent different models. We will denote the Type I Sum of Squares as Type I SS.

In the above discussion concerning Type I SS, it was mentioned that each Type I SS represents the reduction in residual sum of squares due to adding an additional term to the model. One may ask whether or not the reduction in residual sum of squares is significant. That is, we could ask whether or not it is important to add a specific term to a current, specified model. Formal hypotheses for this question are H_0: The additional term is not important to the model, versus H_a: The additional term is important to the model. By "important to the model", we mean that the additional term significantly reduces the residual sum of squares for the model thereby yielding a more accurate model in terms of prediction. The hypotheses described here are known as Type I Hypotheses or Type I Testable Hypotheses.

For an example of the formal symbolic presentation of the Type I Hypotheses, let us consider the two-factor model without interaction $y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$ and assume the data are balanced. We can represent this model as $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ as before. We would then have $\mathbf{X}$ partitioned as $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3)$ where $\mathbf{X}_1 = \mathbf{1}_n \otimes \mathbf{1}_a \otimes \mathbf{1}_b$, $\mathbf{X}_2 = \mathbf{1}_n \otimes \mathbf{I}_a \otimes \mathbf{1}_b$, and $\mathbf{X}_3 = \mathbf{1}_n \otimes \mathbf{1}_a \otimes \mathbf{I}_b$, assuming Factor A has a levels and Factor B has b levels and that the data are balanced. If we follow the convention that $\mathbf{X}$ is an $n \times p$ design matrix having rank r, then $\text{rank}(\mathbf{X}_1) = 1$, $\text{rank}(\mathbf{X}_2) = a$, $\text{rank}(\mathbf{X}_3) = b$, and $r = (a-1) + (b-1) + 1 = a+b-1$. Accordingly, β would be partitioned as

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \text{ where } \beta_1 = \mu,\ \beta_2 = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix}, \text{ and } \beta_3 = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_b \end{pmatrix}.$$

So then $E(\mathbf{y}) = \mathbf{X}\beta = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3$. Under this model, a test concerning the importance of Factor A over a model with only an intercept term is represented by H_o: $(\mathbf{H}_{12} - \mathbf{H}_1)E(\mathbf{y}) = \mathbf{0}$ versus H_a: $(\mathbf{H}_{12} - \mathbf{H}_1)E(\mathbf{y}) \neq \mathbf{0}$. Other equivalent forms for the null hypothesis include H_o: $\mathbf{X}_{2.1}'E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{X}_{2.1} = (\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2$, or H_o: $\mathbf{H}_{2.1}E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_{2.1} = \text{ppo}(\mathbf{X}_{2.1})$, or H_o: $\lambda = 0$ where $\lambda = \dfrac{E(\mathbf{y})'\mathbf{H}_{2.1}E(\mathbf{y})}{2\sigma^2}$. Another common representation for the null hypothesis which more closely follows the notation used with the idea of estimability is H_o: $\mathbf{L}\beta = \mathbf{0}$ where $\mathbf{L} = \mathbf{X}_{2.1}'\mathbf{X} = (\mathbf{X}_{2.1}'\mathbf{X}_1 \quad \mathbf{X}_{2.1}'\mathbf{X}_2 \quad \mathbf{X}_{2.1}'\mathbf{X}_3)$. Using this notation, we see that $\mathbf{L}\beta = \mathbf{0}$ simply becomes $\mathbf{X}_{2.1}'\mathbf{X}_1\beta_1 + \mathbf{X}_{2.1}'\mathbf{X}_2\beta_2 + \mathbf{X}_{2.1}'\mathbf{X}_3\beta_3 = \mathbf{X}_{2.1}'E(\mathbf{y}) = \mathbf{0}$ which was listed above as one of the equivalent forms of H_o. There are yet other equivalent forms of the null hypothesis, we mention the variety of equivalent forms because some are more convenient than others in certain circumstances. For example, the H_o: $\mathbf{L}\beta = \mathbf{0}$ form is convenient when working with SAS® output or when a practical interpretation of the test is desired. The H_o: $(\mathbf{H}_{12} - \mathbf{H}_1)E(\mathbf{y}) = \mathbf{0}$ form is convenient from a mathematical point of view, it is easily related to the test statistic as will be pointed out later. Finally, the H_o: $\mathbf{H}_{2.1}E(\mathbf{y}) = \mathbf{0}$ format is convenient when we wish to identify how each term is treated in the analysis. Although $\mathbf{H}_{12} - \mathbf{H}_1 = \mathbf{H}_{2.1}$, the subscript notation $\mathbf{H}_{2.1}$ is helpful because it classifies the terms contained in E(y) into three groups. The term listed before the "." is the term currently being tested. In this case, the second term is α_i which represents Factor A in the model. The term or terms listed after the "." are the terms currently included in the model. In this case the current model includes only the intercept term μ. The terms listed after the "." will be annihilated when multiplied by $(\mathbf{H}_@ - \mathbf{H}_\#)$ where $\mathbf{H}_@$ and $\mathbf{H}_\#$ are the perpendicular projection operators representing the alternative and null hypotheses respectively. In this example $(\mathbf{H}_@ - \mathbf{H}_\#)$ is $(\mathbf{H}_{12} - \mathbf{H}_1)$. Finally, any terms of E(y) which are not listed in the subscript may or may not affect the test. If the data are balanced, unlisted terms will be zeroed out by the orthogonality of the components of the design matrix **X**. If the data are not balanced but has no empty cells, unlisted terms will effect the test, they will be represented by non-zero coefficients in the matrix **L**. The matrix **L** will be discussed in more detail later.

The reference distribution for the test is an F distribution with noncentrality parameter λ. Under the null hypothesis, $\lambda = 0$, so that the F distribution would be a central F distribution. The form of the test statistic for this test is

$$F = \frac{\dfrac{R(\alpha|\mu)}{a-1}}{\dfrac{SSE_{123}}{n-r}}$$

where $R(\alpha|\mu) = SSR_{12} - SSR_1 = SSE_1 - SSE_{12}$. Recall that $SSE_1 = \mathbf{y}'(\mathbf{I}-\mathbf{H}_1)\mathbf{y}$, $SSE_{12} = \mathbf{y}'(\mathbf{I}-\mathbf{H}_{12})\mathbf{y}$, and $SSE_{123} = \mathbf{y}'(\mathbf{I}-\mathbf{H}_{123})\mathbf{y} = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}$ as described earlier. See appendix for a more general form of the test statistic.

Note that while we are not conducting a test concerning Factor B directly, for an unbalanced data set Factor B will effect the Type I Hypothesis matrix **L**. This occurs because for an unbalanced data set, $\mathbf{X}_3\beta_3$ is not annihilated by $(\mathbf{H}_{12} - \mathbf{H}_1)$. We will point this out again later when we review an example using SAS®.

Continuing with Type I Hypotheses, let us consider a test which examines the importance of Factor B in a model which currently contains an intercept term and Factor A. Our hypotheses are now H_o: The Factor B term is not important to the model currently containing the intercept and Factor A terms, versus Ha: The Factor B term is important to the model. Symbolically these hypotheses are H_o: $(\mathbf{H}_{123} - \mathbf{H}_{12})E(\mathbf{y}) = \mathbf{0}$ versus H_a: $(\mathbf{H}_{123} - \mathbf{H}_{12})E(\mathbf{y}) \neq \mathbf{0}$. Note that under the model $y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$, $\mathbf{H}_{123}$ is just **H**. Again, we can express the null hypothesis using various equivalent forms such as H_o: $\mathbf{X}_{3.12}'E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{X}_{3.12} = (\mathbf{I} - \mathbf{H}_{12})\mathbf{X}_3$, or H_o: $\mathbf{H}_{3.12}E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_{3.12} =$ ppo$(\mathbf{X}_{3.12})$, or H_o: $\lambda = 0$ where $\lambda = \dfrac{E(\mathbf{y})'\mathbf{H}_{3.12}E(\mathbf{y})}{2\sigma^2}$. Another common form is H_o: $E(\mathbf{y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ versus H_a: $E(\mathbf{y}) = \mathbf{X}\beta = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3$. As before, we can also express the null hypothesis as H_o: $\mathbf{L}\beta = \mathbf{0}$. For this test, $\mathbf{L} = \mathbf{X}_{3.12}'\mathbf{X} = (\mathbf{X}_{3.12}'\mathbf{X}_1 \quad \mathbf{X}_{3.12}'\mathbf{X}_2 \quad \mathbf{X}_{3.12}'\mathbf{X}_3)$. So once again, we see that $\mathbf{L}\beta = \mathbf{X}_{3.12}'\mathbf{X}_1\beta_1 + \mathbf{X}_{3.12}'\mathbf{X}_2\beta_2 + \mathbf{X}_{3.12}'\mathbf{X}_3\beta_3 = \mathbf{X}_{3.12}'E(\mathbf{y})$. As before, the reference distribution for the test is an F distribution with noncentrality parameter λ. Under the null hypothesis, $\lambda = 0$, so again the F distribution would be a central F distribution. The form of the test statistic for this test is now

$$F = \frac{\dfrac{R(\tau|\mu,\alpha)}{r-a}}{\dfrac{SSE_{123}}{n-r}}$$

where $R(\tau|\mu,\alpha) = SSR_{123} - SSR_{12} = SSE_{12} - SSE_{123}$, $SSE_{12} = \mathbf{y}'(\mathbf{I}-\mathbf{H}_{12})\mathbf{y}$ and $SSE_{123} = SSE = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}$ since $\mathbf{H}_{123} = \mathbf{H}$. Note that $r - a = b - 1$ which can be used in the expression for F.

We emphasize that with Type I Hypotheses, the order in which the terms are tested is key in the analysis. In the previous example, if we had tested the importance of Factor B over a model containing only an intercept term then proceeded to a test of the importance of Factor A in the presence of an intercept term and a Factor B term, the tests would be different than those described above.

The hypotheses concerning Factor B can be written as H_o: $(\mathbf{H}_{13} - \mathbf{H}_1)E(\mathbf{y}) = \mathbf{0}$ versus H_a: $(\mathbf{H}_{13} - \mathbf{H}_1)E(\mathbf{y}) \neq \mathbf{0}$. Some equivalent forms for the null hypothesis are H_o: $\mathbf{X}_{3.1}'E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{X}_{3.1} = (\mathbf{I} - \mathbf{H}_1)\mathbf{X}_3$, or H_o: $\mathbf{H}_{3.1}E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_{3.1} = \text{ppo}(\mathbf{X}_{3.1})$, or H_o: $\lambda = 0$ where

$\lambda = \frac{E(y)'\mathbf{H}_{3.1}E(y)}{2\sigma^2}$. Following the notation commonly used with the notion of estimability we could write the null hypothesis as H_o: $\mathbf{L}\beta = \mathbf{0}$ where $\mathbf{L} = \mathbf{X}_{3.1}'\mathbf{X} = (\mathbf{X}_{3.1}'\mathbf{X}_1 \quad \mathbf{X}_{3.1}'\mathbf{X}_2 \quad \mathbf{X}_{3.1}'\mathbf{X}_3)$ so that $\mathbf{L}\beta = \mathbf{X}_{3.1}'\mathbf{X}_1\beta_1 + \mathbf{X}_{3.1}'\mathbf{X}_2\beta_2 + \mathbf{X}_{3.1}'\mathbf{X}_3\beta_3 = \mathbf{X}_{3.1}'E(y)$. The form of the test statistic for this test is

$$F = \frac{\frac{R(\tau|\mu)}{b-1}}{\frac{SSE_{123}}{n-r}}$$

where $R(\tau|\mu) = SSR_{13} - SSR_1 = SSE_1 - SSE_{13}$, $SSE_1 = \mathbf{y}'(\mathbf{I}-\mathbf{H}_1)\mathbf{y}$ and $SSE_{13} = \mathbf{y}'(\mathbf{I}-\mathbf{H}_{13})\mathbf{y}$.

The hypotheses for testing the benefit of adding Factor A to a model which contains an intercept term and Factor B can be written as H_o: H_o: $(\mathbf{H}_{123} - \mathbf{H}_{13})E(\mathbf{y}) = \mathbf{0}$ versus H_a: $(\mathbf{H}_{123} - \mathbf{H}_{13})E(\mathbf{y}) \neq \mathbf{0}$. Again, under the model $y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$, $\mathbf{H}_{123}$ is simply $\mathbf{H}$. Equivalent forms of the null hypothesis include H_o: $\mathbf{X}_{2.13}'E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{X}_{2.13} = (\mathbf{I} - \mathbf{H}_{13})\mathbf{X}_2$, or H_o: $\mathbf{H}_{2.13}E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_{2.13} = ppo(\mathbf{X}_{2.13})$, or H_o: $\lambda = 0$ where $\lambda = \frac{E(y)\mathbf{H}_{2.13}E(y)}{2\sigma^2}$, or H_o: $\mathbf{L}\beta = \mathbf{0}$ where $\mathbf{L} = \mathbf{X}_{2.13}'\mathbf{X} = (\mathbf{X}_{2.13}'\mathbf{X}_1 \quad \mathbf{X}_{2.13}'\mathbf{X}_2 \quad \mathbf{X}_{2.13}'\mathbf{X}_3)$. So once again, we see that $\mathbf{L}\beta = \mathbf{X}_{2.13}'\mathbf{X}_1\beta_1 + \mathbf{X}_{2.13}'\mathbf{X}_2\beta_2 + \mathbf{X}_{2.13}'\mathbf{X}_3\beta_3 = \mathbf{X}_{2.13}'E(\mathbf{y})$. The form of the test statistic for this test is

$$F = \frac{\frac{R(\alpha|\mu,\tau)}{r-b}}{\frac{SSE_{123}}{n-r}}$$

where $R(\alpha|\mu,\tau) = SSR_{123} - SSR_{13} = SSE_{13} - SSE_{123}$, $SSE_{13} = \mathbf{y}'(\mathbf{I}-\mathbf{H}_{13})\mathbf{y}$ and $SSE_{123} = SSE = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}$ because $\mathbf{H}_{123} = \mathbf{H}$. It is common to use $a-1$ in place of $r-b$ in the above expression for F since $a-1 = r-b$.

Again, in a Type I Analysis the sequence in which the terms are considered and tested is of key importance. Also, the test statistic for each test involves the Type I SS associated with the term under consideration, see Table 1 below which refers to the three-factor example discussed earlier where the factors are added in the order Factor A, Factor B, and finally Factor C (the model did not include interactions, and we assume a balanced data set). In general, the larger the value of the Type I SS, the more important the proposed term is to the model (Milliken and Johnson, 1992). This is due to the relation between the Type I SS and the corresponding F-statistic, the larger the value of the Type I SS, the larger the value of F, and hence, the more likely we are to reject the null

hypothesis in favor of the alternative hypothesis which is that the additional term is important to the model.

Table 1

Source	df	SS	MS	F
Factor A	$a-1$	$R(\alpha\|\mu)$	$MSA=R(\alpha\|\mu)/(a-1)$	MSA/MSE_{1234}
Factor B	$b-1$	$R(\tau\|\mu,\alpha)$	$MSB=R(\tau\|\mu,\alpha)/(b-1)$	MSB/MSE_{1234}
Factor C	$c-1$	$R(\gamma\|\mu,\alpha,\tau)$	$MSC=R(\gamma\|\mu,\alpha,\tau)/(c-1)$	MSC/MSE_{1234}
Error	$n-a-b-c+2$	$\mathbf{y'y}-R(\mu,\alpha,\tau,\gamma)$	$MSE=SSE/(n-a-b-c+2)$	
Corrected Total	$n-1$	$\mathbf{y'y}-R(\mu)$		

Note that $SSE_{1234} = SSE = \mathbf{y'(I-H)y} = \mathbf{y'y} - R(\mu,\alpha,\tau,\gamma)$ since $\mathbf{H}_{1234} = \mathbf{H}$, so then $MSE_{1234} = MSE$. Also note that the Type I SS form a partition of the model sum of squares, where by "model sum of squares" we mean $SS_{Model} = SS_A + SS_B + SS_C$, and we see that $SS_{Corrected\ Total} = SS_{Model} + SS_{Error}$, where $SS_{Corrected\ Total} = \mathbf{y'y} - R(\mu)$. The quantity $R(\mu)$ is commonly called the correction for the mean (denoted CM) and is equal to $n\bar{y}^2$. To see this, recall that by construction, $R(\mu) = \mathbf{y'H_1y}$ where $\mathbf{H}_1 = ppo(\mathbf{X}_1)$. In this example, $\mathbf{X}_1 = \mathbf{1}_n$, an n-vector of ones. Therefore, $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1' = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-}\mathbf{1}_n'$. But $(\mathbf{1}_n'\mathbf{1}_n) = n$, so that $(\mathbf{1}_n'\mathbf{1}_n)^{-} = (\mathbf{1}_n'\mathbf{1}_n)^{-1} = 1/n$. Furthermore, $n\bar{y} = \mathbf{y'1}_n = \mathbf{1}_n\mathbf{y}'$. Finally then, $\mathbf{y'H_1y} = \mathbf{y'}\,\mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-}\mathbf{1}_n'\mathbf{y} = (n\bar{y}^2)/n = n\bar{y}^2$. To write these results using the sums of squares discussed earlier, SST, SSR, and SSE, we see that $SSE = SS_{Error}$, $SSR = SS_{Model} + CM$, $SST = \mathbf{y'y} = SS_{Corrected\ Total} + CM = SSR + SSE$ (Littell, Freund, and Spector, 1991 and Searle, 1982).

If a model contains higher degree polynomial terms or interaction terms, those terms are handled in the same way as other terms in the model. When dealing with interaction terms, it is of course necessary that the interaction term follow the main effect terms for factors involved in the interaction. For example, if we wish to test for interaction between Factor A and Factor B, the current model should at least contain Factor A and Factor B. For higher degree polynomial terms, it is conventional to include all lower degree terms in the model before considering the addition of a higher degree term. So for example, if we wanted to test a cubic term for a certain factor, the model should already contain the linear and quadratic terms for that factor.

When we consider the connection between the Type I SS and the Type I Hypotheses, it is easy to see why the names chosen for each are so similar. Another term often seen when conducting a Type I Analysis is Type I Estimable Functions. The term Type I Estimable Function combines the concept of a general estimable function $L\beta$ with the idea of a Type I Hypotheses. For example, suppose we return to a model which includes the intercept μ and Factor A, and that we are considering the addition of Factor B to this

model (Factor C is not involved in this example at all). We saw earlier that we could write the Type I Hypotheses in one of several equivalent forms. Another equivalent form that is often preferred from a practical point of view is H_o: $\tau_i = \tau_j$ for all $i \neq j$, versus H_a: $\tau_i \neq \tau_j$ for some $i \neq j$. For this format, Type I Estimable Functions must have the form $\mathbf{L}\beta$ where $\mathbf{L}$ is called the Type I Hypothesis matrix and β is the parameter vector $(\mu, \alpha_1, \alpha_2, ..., \alpha_a, \tau_1, \tau_2, ..., \tau_b)'$. For the convenience of an example, suppose that Factor A has 4 levels and that Factor B has 3 levels, and that we assume no AB interaction so that $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \tau_1, \tau_2, \tau_3)'$. In this case, $\mathbf{L}$ must be a matrix which is row equivalent to

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \qquad (2)$$

Thus we can rewrite our hypotheses as H_o: $\mathbf{L}\beta = \mathbf{0}$, versus H_a: $\mathbf{L}\beta \neq \mathbf{0}$ where $\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Note that the hypotheses H_o: $\mathbf{L}\beta = \mathbf{0}$ versus H_a: $\mathbf{L}\beta \neq \mathbf{0}$ is one of the equivalent forms of the Type I Hypotheses discussed earlier, and that H_o: $\mathbf{L}\beta = \mathbf{0}$ simply says that $\tau_i = \tau_j$ for all $i \neq j$, or equivalently that $\tau_i - \tau_j = 0$ for all $i \neq j$. The estimable functions for Factor B described by this example are $\tau_1 - \tau_3$, and $\tau_2 - \tau_3$. They are estimable because, for this example, $\mathbf{L} = \mathbf{X}_{3,12}'\mathbf{X}$, which says $\mathbf{L}'$ is an element of the column space of $\mathbf{X}'$. Furthermore, these estimable functions describe the Type I Hypotheses we wish to test concerning the addition of Factor B to a model which contains only an intercept term and a single factor, Factor A. The presence of interaction terms in the model will of course complicate the make-up of the Type I Estimable Functions. Examples involving interaction terms can be found in the literature (Milliken and Johnson, 1992).

A popular software package that can be used to conduct a Type I Analysis is SAS®. As part of PROC GLM, SAS® has the ability to calculate the Type I SS along with the F-statistic values associated with the Type I tests of hypothesis for each term in a specified model. The programs and printouts are very straight forward, but care must be taken when designating the model to insure that the desired sequence is followed. For example if we are working with a two-factor model without interaction, the command line

model y = A B;

in PROC GLM specifies that Factor A will be tested first against a model containing only an intercept term, then Factor B will be tested against a model containing the intercept term and a Factor A term. If the command line is changed to

model y = B A;,

then the terms are introduced in the opposite order. To have SAS® calculate and print the Type I SS along with their respective F-statistic, include the option SS1 after designating the model as follows:

model y = A B / SS1;.

The Type I SS will be listed immediately following the General Linear Models Procedure ANOVA table on the SAS® printout.

Another option of PROC GLM allows SAS® to represent **L**, the Type I Hypothesis matrix. The output does not present **L** in matrix form, but rather uses a format which allows the user to determine the Type I Estimable Functions (and thus the Type I Testable Hypotheses). To have SAS® represent the Type I Estimable Functions, type E1 after designating the model as follows:

model y = A B / E1;.

As an example of how to interpret the SAS® representation of the **L** matrix, let us return to the two-factor model without interaction and assume that Factor A has four levels and that Factor B has three levels. If the data are balanced and we plan to test Factor A first followed by Factor B, the printout will be as follows:

Type I Estimable Function for: A

Effect		Coefficients
INTERCEPT		0
A	1	L2
	2	L3
	3	L4
	4	–L2 – L3 – L4
B	1	0
	2	0
	3	0

Type I Estimable Function for: B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0
	3	0
	4	0

B	1	L6
	2	L7
	3	–L6 – L7.

To determine the Type I Estimable Functions matrix (in row-reduced form), substitute either ones or zeros for the L#'s. Note that by letting in turn L2 = 1, L3 = 1, L4 = 1, with all other coefficients set to zero in each case, SAS® is actually representing the Factor A Type I Estimable Functions matrix

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

which in turn represents the null hypothesis that $\alpha_i = \alpha_j$ for all $i \neq j$. Similar results hold for Factor B, the Factor B Type I Estimable Functions matrix is

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

Again, this is equivalent to the hypothesis that $\tau_i = \tau_j$ for all $i \neq j$ as mentioned earlier in (2).

If the data are not balanced but there are no empty cells, the output will differ depending on the cell counts. When the model contains no interaction terms, the main difference between the output for a balanced data set versus an unbalanced data set is that for the balanced data set we see that the only term with non-zero coefficients in the printout is the term currently being tested. This will not be the case for an unbalanced data set. For example, suppose that we continue with the two-factor design described above where Factor A has four levels and Factor B has three levels. Suppose further that each cell has two replicates with the exception of the first cell (level 1 of Factor a and level 1 of Factor B) which has only one replicate. We still plan to test Factor A first then Factor B. Then the SAS® printout will be as follows:

Type I Estimable Function for: A

Effect		Coefficients
INTERCEPT		0
A	1	L2
	2	L3
	3	L4
	4	–L2 – L3 – L4

B	1	–0.1333*L2
	2	0.0667*L2
	3	0.0667*L2

Type I Estimable Function for: B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0
	3	0
	4	0
B	1	L6
	2	L7
	3	–L6 – L7.

This printout represents the Factor A Type I Estimable Functions matrix

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & -0.1333 & 0.667 & 0.667 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix}.$$

Unlike the balanced data case, this is no longer equivalent to $\alpha_i = \alpha_j$ for all $i \neq j$. Instead we see that Factor B is involved in the test concerning Factor A. This is apparent because of the non-zero coefficients in the first row of the Factor B columns of the **L** matrix. For this example, the Type I Estimable Functions for Factor A are $\alpha_1 - \alpha_4 - 0.1333\tau_1 + 0.0667\tau_2 + 0.0667\tau_3$, $\alpha_2 - \alpha_4$, and $\alpha_3 - \alpha_4$. Considering the null hypothesis H_0: $\mathbf{L\beta} = \mathbf{0}$, the first row of $\mathbf{L\beta} = \mathbf{0}$ yields the equation $\alpha_1 - \alpha_4 - 0.1333\tau_1 + 0.0667\tau_2 + 0.0667\tau_3 = 0$ which represents the first of the Type I Estimable Functions for Factor A. This may not be a very meaningful hypothesis to the researcher, the terms involving the decimal factors of Factor B are most likely very difficult to interpret. This is typical of an unbalanced data set with no empty cells. Note that if empty cells do exist in the data set, a Type IV Analysis should be used. We will not discuss the Type IV Analysis in this paper, but discussions of Type IV Analysis can be found in several publications from the SAS® Institute as well as other sources in the literature. Staying with an unbalanced data set but changing the cell frequencies will lead to a different **L** matrix, one where Factor B still has some non-zero coefficients and thus is involved in the test of Factor A. For this reason, the Type I Analysis is often discarded when the data set is unbalanced. One last comment concerning this **L** matrix, if the underlying model as expressed by E(y) contained only the intercept term and the Factor A term, and Factor B was not included in the problem, then we would have the following **L** matrix:

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

In this case we would be testing whether or not $\alpha_i = \alpha_j$ for all $i \neq j$. This would be the case even if the data were unbalanced provided that Factor B was not involved in the problem.

Based on the SAS® printout, the Factor B Type I Estimable Functions matrix for the example where Factor A has four levels and Factor B has three levels, and where each cell has two replicates with the exception of the first cell which has only one replicate is

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

As with the balanced data, this is equivalent to testing whether or not $\tau_i = \tau_j$ for all $i \neq j$. So we again see the importance of the sequence in which the factors are to be added when conducting a Type I Analysis: In our example Factor A was listed first in the model statement of PROC GLM using SAS® and is therefore to be added to the model before Factor B. As a result of the unbalanced data set, the Type I Estimable Functions for Factor A were not very meaningful from a practical point of view, yet the Type I Estimable Functions for Factor B do turn out to be meaningful. These results would be reversed if we had listed Factor B first in the model statement of PROC GLM.

An important question to consider when discussing Type I Analysis is, "When is the Type I Analysis appropriate ?" The Type I Analysis is useful in a model building setting (Milliken and Johnson, 1992) where we eventually hope to predict the effects of various treatments. Appropriate model structures include (SAS/STAT, 1988) ANOVA models with balanced data, purely nested models, and polynomial regression models. As mentioned previously, terms must be added to the model in a specified sequence. This is true regardless of the model structure being used. The Type I Analysis does require that there are no empty cells in the data set (otherwise a Type IV Analysis may be appropriate).

There are also some drawbacks to the Type I Analysis. For one, as was pointed out in an earlier example, when the data are unbalanced, each equation in the hypotheses will involve parameters of the effect being tested as well as parameters from remaining terms which follow in the model statement (Goodnight, 1980). This complicates the interpretation of the hypotheses, possibly to the point that the interpretation is not of practical use. Another drawback which stems from unbalanced data sets is that the Type I Hypotheses will differ according to the cell counts. This implies that the Type I

Hypotheses are actually functions of the cell counts (Pendleton, Tress, and Bremer, 1986). This is rarely the intention of the researcher, quite often the cell counts are merely artifacts of the experimental or sampling process. The researcher normally has certain questions concerning the importance or effect of various factors or terms in the model, the cell counts are not of primary concern. In such circumstances the Type I Analysis would not be appropriate.

Type II Analysis

The Type II Sums of Squares are often called or interpreted as partial sums of squares (Littell, Freund, and Spector, 1991). As we shall see, the Type II SS for a main effect factor are equivalent to the Type I SS for that factor if it had been added last to the model. Recall that the Type I SS are sequential and order dependent. In a Type I Analysis, as a new term is added to the model, it is adjusted for the terms already contained in the model. This is not the case for Type II SS, they are not order dependent. As an example, suppose that we are working on a model which would potentially involve an intercept term and main effect terms for three factors (no interaction terms are being considered). Factor A can be represented by α_i, Factor B can be represented by τ_j, and Factor C can be represented by γ_k. If we denote the intercept term by μ, then the model could potentially be represented by $y_{ijkl} = \mu + \alpha_i + \tau_j + \gamma_k + \varepsilon_{ijkl}$. However, we may determine that some of these potential terms are not significantly important to the model and may be omitted. We can use R-Notation to develop the Type II SS for this example. The Type II SS for each factor are as follows:

Factor A	$R(\alpha\|\mu,\tau,\gamma)$
Factor B	$R(\tau\|\mu,\alpha,\gamma)$
Factor C	$R(\gamma\|\mu,\alpha,\tau)$.

So we see that each factor is considered in a model which already includes all other terms under consideration in the problem. That is, each potential term or effect is adjusted for all other effects. We also wish to emphasize that each of the Type II SS can be interpreted as the reduction in residual sum of squares due to adding the specified factor to a model containing all other potential factors. For example, suppose we obtain the regression sum of squares for the model $y_{ikl} = \mu + \alpha_i + \gamma_k + \varepsilon_{ikl}$ and denote this sum of squares as SSR_{124}. Next, obtain the regression sum of squares for the model $y_{ijkl} = \mu + \alpha_i + \tau_j + \gamma_k + \varepsilon_{ijkl}$ and denote this sum of squares as SSR_{1234}. Then $R(\tau|\mu,\alpha,\gamma) = SSR_{1234} - SSR_{124} = SSE_{124} - SSE_{1234}$. The reduction in residual sum of squares can then be tested for statistical significance.

Let us now consider the hypotheses behind the test concerning the statistical significance of a specified term in a Type II Analysis. As with the Type I Hypotheses,

we can write the Type II Hypotheses as H_o: The additional term is not important to the model, versus H_a: The additional term is important to the model. The difference between the two analyses is due to the difference in the assumed format of the model prior to the addition of the term under consideration. As was mentioned earlier, if the term under consideration is the last term in sequence to be added to the model, then the Type II SS are equivalent to the Type I SS, and thus the two analyses will be equivalent for that term.

. To see a more formal symbolic representation of the Type II Hypotheses, let us continue with the example involving an intercept term, main effect terms for three potential factors, but no interaction terms. We can write the model as $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. This requires us to partition $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3\ \mathbf{X}_4)$ where $\mathbf{X}_1 = \mathbf{1}_n \otimes \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c$, $\mathbf{X}_2 = \mathbf{1}_n \otimes \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c$, $\mathbf{X}_3 = \mathbf{1}_n \otimes \mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c$, and $\mathbf{X}_4 = \mathbf{1}_n \otimes \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{I}_c$ assuming Factor A has a levels and Factor B has b levels, Factor C has c levels, and that the data are balanced. If we again follow the convention that $\mathbf{X}$ is an $n \times p$ design matrix having rank r, then $\text{rank}(\mathbf{X}_1) = 1$, $\text{rank}(\mathbf{X}_2) = a$, $\text{rank}(\mathbf{X}_3) = b$, $\text{rank}(\mathbf{X}_3) = c$, and $r = (a-1) + (b-1) + (c-1) + 1 = a+b+c-2$. Accordingly, β would be partitioned as

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \text{ where } \beta_1 = \mu,\ \beta_2 = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix},\ \beta_3 = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_b \end{pmatrix},\ \text{and } \beta_4 = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_c \end{pmatrix}.$$

So then $E(\mathbf{y}) = \mathbf{X}\beta = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \mathbf{X}_4\beta_4$. Under this model, a test concerning the addition of Factor B to a model containing all other terms can be represented by the Type II Hypotheses H_o: $(\mathbf{H}_{1234} - \mathbf{H}_{124})E(\mathbf{y}) = \mathbf{0}$ versus H_a: $(\mathbf{H}_{1234} - \mathbf{H}_{124})E(\mathbf{y}) \neq \mathbf{0}$. As before, a variety of equivalent forms of these hypotheses exist. Equivalent forms of the null hypothesis include H_o: $\mathbf{X}'_{3.124}E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{X}_{3.124} = (\mathbf{I} - \mathbf{H}_{124})\mathbf{X}_3$, or H_o: $\mathbf{H}_{3.124}E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_{3.124} = \text{ppo}(\mathbf{X}_{3.124})$, or H_o: $\lambda = 0$ where $\lambda = \dfrac{E(\mathbf{y})'\mathbf{H}_{3.124}E(\mathbf{y})}{2\sigma^2}$. The equivalent form of the null hypothesis which emphasizes the idea of estimability is H_o: $\mathbf{L}\beta = \mathbf{0}$ where $\mathbf{L} = \mathbf{X}'_{3.124}\mathbf{X} = (\mathbf{X}'_{3.124}\mathbf{X}_1\ \ \mathbf{X}'_{3.124}\mathbf{X}_2\ \ \mathbf{X}'_{3.124}\mathbf{X}_3\ \mathbf{X}'_{3.124}\mathbf{X}_4)$. With this notation, we see that $\mathbf{L}\beta = \mathbf{0}$ simply says that $\mathbf{X}'_{3.124}\mathbf{X}_1\beta_1 + \mathbf{X}'_{3.124}\mathbf{X}_2\beta_2 + \mathbf{X}'_{3.124}\mathbf{X}_3\beta_3 + \mathbf{X}'_{3.124}\mathbf{X}_4\beta_4 = \mathbf{X}'_{3.124}E(\mathbf{y}) = \mathbf{0}$ which we already listed as an equivalent form of the null hypothesis.

The reference distribution for the above test is an F distribution with noncentrality parameter λ. Under the null hypothesis, $\lambda = 0$, so as with the Type I Analysis, the Type II Analysis assumes that the test statistic follows a central F distribution and has the general form

$$F = \frac{\frac{SSE_0 - SSE_a}{m-k}}{\frac{SSE_a}{n-m}}$$

where SSE_0 refers to the residual sum of squares for the model implied by the null hypothesis and SSE_a refers to the residual sum of squares for the model implied by the alternative hypothesis. We have already seen that $SSE_0 - SSE_a$ is equivalent to the $R(\cdot|\cdot)$ notation which is now the Type II SS for the term being tested. The quantity $m-k$ is obtained by considering the null hypothesis for the test. As was the case with the Type I Analysis, the null hypothesis of the Type II Analysis can be written as $(\mathbf{H}_@ - \mathbf{H}_\#)E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_@$ and $\mathbf{H}_\#$ are the perpendicular projection operators representing the alternative and null hypotheses respectively. So $m = \text{rank}(\mathbf{H}_@)$ and $k = \text{rank}(\mathbf{H}_\#)$, and n comes from the dimensions of $\mathbf{y}$ and $\mathbf{X}$, $\mathbf{y}: n \times 1$ and $\mathbf{X}: n \times p$. Note also that for a Type II Analysis, the SSE_a used in the calculation of the test statistic F is the residual sum of squares for a model containing all terms under consideration in the problem. So for a Type II Analysis, $m = \text{rank}(\mathbf{H}_@) = \text{rank}(\mathbf{X}) = r$. Thus, the test statistic could be written as

$$F = \frac{\frac{SSE_0 - SSE_a}{r-k}}{\frac{SSE_a}{n-r}}.$$

For our current three-factor example with an intercept term, $SSE_a = R(\mu,\alpha,\tau,\gamma)$. Also, $k = \text{rank}(\mathbf{H}_\#) = \text{rank}(\mathbf{X}_{124}) = 1 + (a-1) + (c-1) = a + c - 1$. Thus, for this current example where we are conducting a test concerning Factor B, the test statistic is

$$F = \frac{\frac{R(\tau|\mu,\alpha,\gamma)}{r-a-c+1}}{\frac{SSE_{1234}}{n-r}}.$$

Note that $r - a - c + 1 = b - 1$, and that SSE_{1234} can be written as simply SSE, so we could write the test statistic as

$$F = \frac{\frac{R(\tau|\mu,\alpha,\gamma)}{b-1}}{\frac{SSE}{n-r}}.$$

Recall also that $R(\tau|\mu,\alpha,\gamma) = SSR_{1234} - SSR_{124} = SSE_{124} - SSE_{1234}$, $SSE_{124} = \mathbf{y}'(\mathbf{I}-\mathbf{H}_{124})\mathbf{y}$, $\mathbf{H}_{124} = \text{ppo}(\mathbf{X}_{124})$, $\mathbf{X}_{124} = (\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_4)$, and $SSE_{1234} = SSE = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}$ where $\mathbf{H} = \text{ppo}(\mathbf{X})$.

In the preceding example, we concentrated on the Type II Hypotheses for Factor B. The Type II Hypotheses for the other factors are quite similar because, as stated earlier, these Type II Hypotheses are equivalent to Type I Hypotheses which add the specified factor as the last factor in sequence to the model.

Table 2 below shows the Type II SS for the three-factor example we have been discussing. Notice that the test statistic, which is the F-statistic, for each main effect test involves the Type II SS associated with the term being tested. This is similar to Table 1 which showed the Type I SS and their corresponding F-statistics.

Table 2

Source	df	SS	MS	F
Factor A	$a-1$	$R(\alpha\|\mu,\tau,\gamma)$	$MSA=R(\alpha\|\mu,\tau,\gamma)/(a-1)$	MSA/MSE_{1234}
Factor B	$b-1$	$R(\tau\|\mu,\alpha,\gamma)$	$MSB=R(\tau\|\mu,\alpha,\gamma)/(b-1)$	MSB/MSE_{1234}
Factor C	$c-1$	$R(\gamma\|\mu,\alpha,\tau)$	$MSC=R(\gamma\|\mu,\alpha,\tau)/(c-1)$	MSC/MSE_{1234}
Error	$n-a-b-c+2$	$\mathbf{y'y}-R(\mu,\alpha,\tau,\gamma)$	$MSE=SSE/(n-a-b-c+2)$	
Total	$n-1$			

Note that $SSE = \mathbf{y'y}-R(\mu,\alpha,\tau,\gamma) = \mathbf{y'(I-H)y}$, that since $r = (a-1)+(b-1)+(c-1)+1 = a+b+c-2$, then $n-r = n-a-b-c+2$. Thus, we could write $MSE_{1234} = MSE = SSE/(n-r)$. Also, note that unlike the Type I SS, the Type II SS do not form a partition of the model sum of squares, that is, $SS_{Model} \neq SS_A + SS_B + SS_C$. The Type II SS will not yield a partitioning of the model sum of squares unless the factors are mutually uncorrelated (Littell, Freund, and Spector, 1991).

As with the Type I analysis, we may wish to know the Type II Estimable Functions for a certain problem. As was the case with the Type I SS, the Type II SS combine the notions of general estimable functions and Type II Hypotheses. As before, we can write the Type II Estimable Functions as $\mathbf{L}\beta$ where $\mathbf{L}$ is called the Type II Hypothesis matrix and β is again the parameter vector for the problem. Determining the Type II Hypothesis matrix $\mathbf{L}$ is the key to determining the Type II Estimable Functions for a particular factor in a particular problem. We saw earlier that the Type I Hypothesis matrix $\mathbf{L}$ was a function, in part, of the cell frequencies. This can also be the case for the Type II Hypothesis matrix when the model used contains interaction terms.

Let us consider a specific example and use PROC GLM with SAS® to obtain the Type II Hypothesis matrix $\mathbf{L}$. We begin with a balanced data set in a problem involving two factors without interaction. The model for this example can be written as $y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$, hence the Type II SS for Factor A and Factor B are $R(\alpha|\mu,\tau)$ and $R(\tau|\mu,\alpha)$, respectively. We see that with the Type II SS, each factor is adjusted for the other. So unlike the Type I Hypothesis matrix, the Type II Hypothesis matrix will not be complicated by non-zero entries corresponding to factors of the model that are not

involved in the hypothesis being tested concerning the parameters of a particular factor. The importance of this is that the estimable functions for a Type II Analysis are easily interpreted, and this is true whether the data are balanced or not.

The model statement in PROC GLM is

model y = A B;.

Because the Type II Analysis is not order dependent like the Type I Analysis, we would get the same results for the Type II Analysis if we had used the model statement

model y = B A;.

As with the Type I Analysis, we can have SAS® produce the Type II SS and the Type II Estimable Functions by including the options SS2 and E2 in the model statement:

model y = A B / SS2 E2;.

For the purpose of this example, suppose that Factor A has 2 levels and Factor B has 3 levels. Then the Type II Estimable Functions are represented by the SAS® printout below.

Type II Estimable Function for: A

Effect		Coefficients
INTERCEPT		0
A	1	L2
	2	–L2
B	1	0
	2	0
	3	0

Type II Estimable Function for: B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0
B	1	L4
	2	L5
	3	–L4 – L5.

By letting in turn L2 = 1, L4 =1, L5 = 1, with all other coefficients set to zero in each case, SAS® is actually representing the Factor A and Factor B Type II Estimable Functions matrices

$$L = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix}, \text{ for Factor A, and}$$

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \text{ for Factor B.}$$

Thus, the Type II Estimable Function for Factor A is $\alpha_1 = \alpha_2$, and the Type II Estimable Functions for Factor B are $\tau_i = \tau_j$ for all $i \neq j$. These estimable functions have very intuitive and practical interpretations. Furthermore, these Type II Estimable Functions will be the same whether the data are balanced or unbalanced.

If we now change the model to include an A*B interaction term, the Type II SS and thus the Type II Estimable Functions will be affected. If we represent the interaction term as γ_{ij}, the model can now be written as $y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$. The Type II SS are now:

Factor A	$R(\alpha\|\mu,\tau)$
Factor B	$R(\tau\|\mu,\alpha)$
A*B Interaction	$R(\gamma\|\mu,\alpha,\tau)$.

We see that the Factor A and Factor B terms are not adjusted for the interaction term, therefore, the interaction term may cause the Type II Estimable Functions matrices for Factors A and B to have non-zero entries in the columns representing the interaction terms. This in turn will lead to estimable functions which involve interaction terms. When the data are unbalanced, the estimable functions can then be difficult to interpret. Staying with our example where Factor A has two levels and Factor B has three levels, the addition of the interaction term requires us to change the model statement in PROC GLM to

model y = A B A*B / SS2 E2; or to model y = A|B / SS2 E2;

(we have included the options SS2 and E2 in order to have SAS® list the Type II SS and the representation of the Type II Estimable Functions). If the data are balanced, the SAS® printout which represents the Type II Estimable Functions for this example is

Type II Estimable Function for: A

Effect	Coefficients
INTERCEPT	0

A	1	L2
	2	–L2
B	1	0
	2	0
	3	0
A*B	1 1	0.3333*L2
	1 2	0.3333*L2
	1 3	0.3333*L2
	2 1	–0.3333*L2
	2 2	–0.3333*L2
	2 3	–0.3333*L2

Type II Estimable Function for: B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0
B	1	L4
	2	L5
	3	–L4 – L5
A*B	1 1	0.5000*L4
	1 2	0.5000*L5
	1 3	–0.5000*L4 – 0.5000*L5
	2 1	0.5000*L4
	2 2	0.5000*L5
	2 3	–0.5000*L4 – 0.5000*L5

Type II Estimable Function for: A*B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0
B	1	0
	2	0
	3	0
A*B	1 1	L7
	1 2	L8
	1 3	–L7 – L8
	2 1	–L7
	2 2	–L8
	2 3	L7 + L8.

By setting in turn L2, L4, L5, L7, and L8 equal to one, with all other coefficients set to zero in each case, we can use this SAS® printout to find the Type II Estimable Functions matrices for Factor A, Factor B, and the A*B interaction term. They are:

$$L=\begin{pmatrix}0 & 1 & -1 & 0 & 0 & 0 & 0.3333 & 0.3333 & 0.3333 & -0.3333 & -0.3333 & -0.3333\end{pmatrix}$$

for Factor A,

$$L=\begin{pmatrix}0 & 0 & 0 & 1 & 0 & -1 & 0.500 & 0 & -0.500 & 0.500 & 0 & -0.500\\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0.500 & -0.500 & 0 & 0.500 & -0.500\end{pmatrix}$$

for Factor B, and

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

for the A*B interaction term.

The Type II Estimable Function for Factor A is $\alpha_1 - \alpha_2 + 0.3333\gamma_{11} + 0.3333\gamma_{12} + 0.3333\gamma_{13} - 0.3333\gamma_{21} - 0.3333\gamma_{22} - 0.3333\gamma_{23}$. So the null hypothesis that this estimable function equals zero does involve interaction terms. Another way to see how these hypotheses are obtained is displayed in Table 3 below.

Table 3

Expected Cell Means

		Factor B 1	2	3	Row Means
Factor A	1	$\mu+\alpha_1+\beta_1+\gamma_{11}$	$\mu+\alpha_1+\beta_2+\gamma_{12}$	$\mu+\alpha_1+\beta_3+\gamma_{13}$	$\mu_{1\cdot}=\mu+\alpha_1+\bar{\beta}+\bar{\gamma}_{1\cdot}$
	2	$\mu+\alpha_2+\beta_1+\gamma_{21}$	$\mu+\alpha_2+\beta_2+\gamma_{22}$	$\mu+\alpha_2+\beta_3+\gamma_{23}$	$\mu_{2\cdot}=\mu+\alpha_2+\bar{\beta}+\bar{\gamma}_{2\cdot}$
	Column Means	$\mu_{\cdot 1}=\mu+\bar{\alpha}+\beta_1+\bar{\gamma}_{\cdot 1}$	$\mu_2=\mu+\bar{\alpha}+\beta_2+\bar{\gamma}_{\cdot 2}$	$\mu_{\cdot 3}=\mu+\bar{\alpha}+\beta_3+\bar{\gamma}_{\cdot 3}$	

In Table 3, $\bar{\alpha}=\frac{1}{2}\sum_{i=1}^{2}\alpha_i$, $\bar{\beta}=\frac{1}{3}\sum_{j=1}^{3}\beta_j$, $\bar{\gamma}_{i\cdot}=\frac{1}{3}\sum_{j=1}^{3}\gamma_{ij}$, and $\bar{\gamma}_{\cdot j}=\frac{1}{2}\sum_{i=1}^{2}\gamma_{ij}$. So the Factor A testable hypothesis is $\mu_{1\cdot}=\mu_{2\cdot}$ or, equivalently, $\alpha_1+\bar{\gamma}_{1\cdot}=\alpha_2+\bar{\gamma}_{2\cdot}$. This testable hypothesis corresponds to the estimable function

$$\alpha_1+\frac{1}{3}(\gamma_{11}+\gamma_{12}+\gamma_{13})-\alpha_2-\frac{1}{3}(\gamma_{21}+\gamma_{22}+\gamma_{23}).$$

This agrees with the Type II Estimable Function for Factor A from the SAS® output. Similar results hold for the Type II Estimable Functions for Factor B. Note that although the interaction term does effect the tests for Factors A and B, it does so in a "balanced" way.

If we now change to an unbalanced data set where the cell counts are as follows,

		Factor B		
		Level 1	Level 2	Level 3
Factor A	Level 1	3	2	3
	Level 2	2	3	3

we will see that the cell counts also effect the results, the coefficients on the interaction terms will reflect the "imbalance" in the data set. With this unbalanced data set, the SAS® output changes to

Type II Estimable Function for: A

Effect		Coefficients
INTERCEPT		0
A	1	L2
	2	–L2
B	1	0
	2	0
	3	0
A*B	1 1	0.3077*L2
	1 2	0.3077*L2
	1 3	0.3846*L2
	2 1	–0.3077*L2
	2 2	–0.3077*L2
	2 3	–0.3846*L2

Type II Estimable Function for: B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0

B	1	L4
	2	L5
	3	–L4 – L5
A*B	1 1	0.5692*L4 + 0.0308*L5
	1 2	–0.0308*L4 + 0.4308*L5
	1 3	–0.5385*L4 – 0.4615*L5
	2 1	0.4308*L4 – 0.0308*L5
	2 2	0.0308*L4 + 0.5692*L5
	2 3	–0.4615*L4 – 0.5385*L5

Type II Estimable Function for: A*B

Effect		Coefficients
INTERCEPT		0
A	1	0
	2	0
B	1	0
	2	0
	3	0
A*B	1 1	L7
	1 2	L8
	1 3	–L7 – L8
	2 1	–L7
	2 2	–L8
	2 3	L7 + L8.

By setting in turn L2, L4, L5, L7, and L8 equal to one, with all other coefficients set to zero in each case, this SAS® printout is actually representing the Factor A, Factor B, and A*B interaction Type II Estimable Functions matrices

$$L=\begin{pmatrix}0 & 1 & -1 & 0 & 0 & 0 & 0.3077 & 0.3077 & 0.3846 & -0.3077 & -0.3077 & -0.3846\end{pmatrix}$$

for Factor A,

$$L=\begin{pmatrix}0 & 0 & 0 & 1 & 0 & -1 & 0.5692 & -0.0308 & -0.5385 & 0.4308 & 0.0308 & -0.4615\\ 0 & 0 & 0 & 0 & 1 & -1 & 0.0308 & 0.4308 & -0.4615 & -0.0308 & 0.5692 & -0.5385\end{pmatrix}$$

for Factor B, and

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

for the A*B interaction term.

Here there is potential for confusion when interpreting the estimable functions. For example, the Type II Estimable Functions for Factor B are now

$$\tau_1 - \tau_3 + 0.5692\gamma_{12} - 0.0308\gamma_{12} - 0.5385\gamma_{13} + 0.4308\gamma_{21} + 0.0308\gamma_{22} - 0.4615\gamma_{23} \text{ and}$$

$$\tau_2 - \tau_3 + 0.0308\gamma_{12} + 0.4308\gamma_{12} - 0.4615\gamma_{13} - 0.0308\gamma_{21} + 0.5692\gamma_{22} - 0.5385\gamma_{23}.$$

When the data are unbalanced, the addition of the interaction term clearly complicates the interpretation of the Type II Estimable Functions, possibly making the hypotheses meaningless to the researcher. Note that the Type II Estimable Functions for the interaction term are not effected by either Factor A or Factor B. This is because the Type II SS for the interaction is $R(\gamma|\mu,\alpha,\tau)$ which adjusts the interaction term γ for both Factor A and Factor B. Note also that Factor A does not effect the Factor B Type II Estimable Functions. Similarly, Factor B does not effect the Factor A Type II Estimable Functions. This again is due to the nature of the Type II SS for each factor. The Type II SS for Factors A and B are $R(\alpha|\mu,\tau)$ and $R(\tau|\mu,\alpha)$, respectively. We see that in both cases, each factor is adjusted for the other. Recall that this was not the case for Type I SS, they were sequential in nature.

Having discussed some of the properties of the Type II Analysis, let us discuss situations where the Type II Analysis would be appropriate, as well as some drawbacks to the Type II Analysis.

Appropriate uses of the Type II Analysis include model building situations where the goal is to eventually predict the effect of particular treatment combinations (Milliken and Johnson, 1992), purely nested ANOVA models, ANOVA models with balanced data sets, particularly those for main effect ANOVA tests in which no interaction terms are present, and full-rank regression settings (SAS/STAT, 1988, and Littell, Freund, and Spector, 1991). We saw earlier that the Type II Analysis can lead to difficulty when interpreting the estimable functions in an ANOVA model with interaction terms. This is not a problem in the full-rank regression setting because the crossproduct terms are viewed simply as additional independent variables without the usual concern for containment. So for a full-rank regression model which involves the independent variables x_1 and x_2 and also includes the crossproduct term x_1x_2, we could write the model as $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + \varepsilon$ or equivalently as $y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$. However, the Type II SS for the independent variables x_1 and x_2, are adjusted for the crossproduct term. So the Type II SS for x_1 and x_2 are $R(\alpha|\mu,\tau,\gamma)$ and $R(\tau|\mu,\alpha,\gamma)$, respectively. Thus the Type II Estimable Functions matrices for x_1 and x_2 will not be complicated by non-zero entries corresponding to the crossproduct term, thereby allowing the Type II Estimable Functions to be easily interpreted. It is also important to note that

the Type II F -tests are exactly equivalent to the parameter t -tests provided on regression printouts. In fact the Type II F statistic is equal to the square of the t statistic from the regression printout. That is to say, for most statistical software packages the t -tests for individual parameters in a regression model are equivalent to the Type II F -tests. For example, the t -tests obtained from PROC REG using SAS® are equivalent to the Type II F -tests obtained by using PROC GLM with SAS®. We emphasize that the Type II Analysis is not order dependent. If terms are to be added to the model in a particular sequence of interest, then the Type I Analysis should be used (assuming the other properties of the problem are consistent with the Type I Analysis).

Situations where the Type II Analysis may not be appropriate include main effect ANOVA tests for models which involve interaction terms when used with unbalanced data sets. We saw earlier that the interaction terms can complicate the interpretation of the estimable functions in a Type II Analysis. Also, as was the case with the Type I Analysis, the Type II SS are dependent on the cell frequencies. If we wish to use the imbalance in the data set to reflect similar proportions from the underlying population, then the Type II Analysis would be appropriate. However, with an unbalanced data set, the question being tested is "Is there a significant treatment effect when a particular set of weights is applied to these treatments ?" The weights applied to the treatments depend on the cell frequencies for the treatments (Pendleton, Von Tress, and Bremer, 1986). If we do not intend to use these weight to represent the proportions of the treatments in the underlying population, then the Type II Analysis for effects contained in other effects such as interaction terms are of questionable merit (SAS/STAT, 1988). Recall that the Type I Analysis had similar drawbacks when the data was unbalanced.

Type III Analysis

When working in a regression setting, the design matrix **X** is generally of full-rank. In such settings, the Type III SS can be represented using reduction notation as was discussed for the Type I and Type II SS. However, when working in an ANOVA setting, the design matrix **X** is typically of less than full-rank. This is the case with the overparameterized model given in (1) which we have been working with throughout this paper. In this case, we cannot use the usual reduction notation unless certain non-estimable constraints are imposed on the parameters in the vector β. For example, for the two factor model with interaction, we could impose the sum-to-zero-restriction, often referred to as the "usual assumptions"

$$\sum_i \alpha_i = \sum_j \tau_j = \sum_i (\alpha\tau)_i = \sum_j (\alpha\tau)_j = 0.$$

Under these restrictions, the resulting design matrix $\mathbf{X}^*$ has full column rank. The reparameterized model can be written as $\mathbf{y} = \mathbf{X}^*\beta^* + \varepsilon$, and the Type III SS can then be written as

Factor A	$R^*(\alpha\|\mu,\tau,\gamma)$
Factor B	$R^*(\tau\|\mu,\alpha,\gamma)$
A*B Interaction	$R^*(\gamma\|\mu,\alpha,\tau)$

where γ represents the interaction term and the notation R* refers to the reduction notation being applied to the reparameterized model. So for example, $R^*(\alpha|\mu,\tau,\gamma) = SSE_{reduced} - SSE_{full}$ where "reduced" refers to the reparameterized model which includes all terms except the one representing Factor A, and "full" refers to the reparameterized model which includes all terms under consideration. More details concerning the R-notation applied to the reparameterized model can be found in Speed, Hocking, and Hackney (1978) and other places in the literature. Under the reparameterized model, the Type III Sum of Squares can be viewed as a partial sum of squares because each main effect term is adjusted for all other terms in the model. So for a model involving two factors, A and B, with interaction term A*B included in the model, the Type III SS for Factors A and B would each be adjusted for the other as well as for the interaction term (Littell, Freund, and Spector, 1991). When the model does not contain interaction terms, the Type III SS equal the Type II SS. We will say more about this when we discuss the Type III Estimable Functions.

The Type III sums of squares given above (using the reparameterized model) are the same as those produced by the PROC GLM model statement

model y = A B A*B; or model y = A|B;.

They are also equal to the sums of squares developed and used in the Yates' weighted squares-of-means method.

Other non-estimable constraints can be used to reparameterize the less than full-rank model. Other commonly used constraints include the set-to-zero restrictions

$$\alpha_i = \beta_i = \gamma_{ij} = \gamma_{ji} = 0 \text{ for all } i \text{ and } j, \text{ or}$$

$$\alpha_a = \beta_b = \gamma_{aj} = \gamma_{ib} = 0 \text{ for all } i \text{ and } j.$$

It is important to note that imposing different restrictions will result in different Type III SS. Thus caution must be taken when trying to represent the Type III SS using reduction notation when the design matrix is less than full-rank (Littell, Freund, and Spector, 1991).

We mentioned earlier that the Type I and Type II Analyses were often considered inappropriate for unbalanced data because the Types I and II Hypotheses are functions of

the cell frequencies. The Types I and II Hypotheses are therefore considered weighted hypotheses, the weights being determined by the cell frequencies. We saw earlier that these weighted hypotheses were often difficult to interpret and therefore had little meaning for the researcher. This is not the case for the Type III Hypotheses. The Type III Hypotheses are unweighted hypotheses regardless of whether the data are balanced or not. Furthermore, the Type III Hypotheses correspond to those hypotheses which the researcher is most often interested in testing. For a two-factor model without interaction, the researcher is usually interested in testing hypotheses such as

H_o: all Factor A effects are equal, and
H_o: all Factor B effects are equal.

In a two-factor model with interaction, the researcher is usually interested in testing hypotheses such as

H_o: there is no interaction effect,
H_o: all row means are equal, and
H_o: all column means are equal .

The Type III Analysis will test these hypotheses in a way which is not weighted according to the cell frequencies.

Perhaps the easiest way to precisely determine the Type III Hypotheses is to use the cell means model rather than the overparameterized model we have considered thus far in this paper. The cell means model for the two-factor example is

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \text{ for } i = 1,2,...,a \text{ , } j = 1,2,...,b, \text{ and } k = 1,2,...,n_{ij} > 0. \quad (3)$$

The parameter μ_{ij} is called the population cell mean for the cell involving the i^{th} level of Factor A and the j^{th} level of Factor B. The cell means are related to the parameters used in the effects or overparameterized model by

$$\mu_{ij} = \mu + \alpha_i + \tau_j + \gamma_{ij} \quad (4)$$

where μ is the grand mean and γ represents the interaction term. Before we discuss the nature of the Type III Hypotheses based on the cell means model, it may be helpful to briefly review the structure of the data set as explained under the model in (3). If we assume that Factor A has three levels and Factor B has four levels, then the design can be displayed using Table 4 below.

Table 4

		Factor B				
		1	2	3	4	
Factor A	1	μ_{11}	μ_{12}	μ_{13}	μ_{14}	$\mu_{1\cdot}$
	2	μ_{21}	μ_{22}	μ_{23}	μ_{24}	$\mu_{2\cdot}$
	3	μ_{31}	μ_{32}	μ_{33}	μ_{34}	$\mu_{3\cdot}$
		$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot 3}$	$\mu_{\cdot 4}$	$\mu_{\cdot\cdot}$

Note that $\mu_{i\cdot} = \sum_{j=1}^{b} \mu_{ij}$, $\mu_{\cdot j} = \sum_{i=1}^{a} \mu_{ij}$, and $\mu_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} \mu_{ij}$.

As mentioned earlier, the parameters μ_{ij} are called the population cell means. We can also form the parameters

$$\overline{\mu}_{i\cdot} = \frac{\sum_{j=1}^{b} \mu_{ij}}{b} = \frac{\mu_{i\cdot}}{b} \text{ and } \overline{\mu}_{\cdot j} = \frac{\sum_{i=1}^{a} \mu_{ij}}{a} = \frac{\mu_{\cdot j}}{a}.$$

We will refer to the parameters $\overline{\mu}_{i\cdot}$ and $\overline{\mu}_{\cdot j}$ as the population marginal means (for, in general, $i = 1,2,\ldots,a$, and $j = 1,2,\ldots,b$). It is also important to recall that we are assuming that there are no empty cells in the data set, that is, $n_{ij} > 0$ for all i and j.

General formulas for the marginal means can also be written based on the overparameterized model as

$$\overline{\mu}_{i\cdot} = \mu + \alpha_i + \frac{\sum_{j=1}^{b} \tau_j}{b} + \frac{\sum_{j=1}^{b} \gamma_{ij}}{b} \text{ and } \overline{\mu}_{\cdot j} = \mu + \frac{\sum_{i=1}^{a} \alpha_i}{a} + \tau_j + \frac{\sum_{i=1}^{a} \gamma_{ij}}{a}.$$

Note that these formulas are simply the result of averaging the cell means as explained by (4) over the various rows and columns of the design when the design is displayed as in Table 4. Regardless of which version of the formula we use for the population marginal means, we see that they are not functions of the cell frequencies. For our example where Factor A has three levels and Factor B has four levels, just use 3 in place of a and 4 in place of b in the above formulas. We stress that all parameters displayed in Table 4 are population parameters, and thus so are the population marginal means. Most often, inferences must be made based on only a sample, not on the complete population. Had the complete population been available, exact parameter values could be determined and we would not need to make inferences. So in practice, the population marginal means will not be known and must therefore be estimated from the sample data. When no

empty cells are present, the population marginal means are estimable, and the best estimates for the population marginal means are what we will refer to as the least squares means. SAS® calls the least squares means the LSMEANS, they are obtained using the command LSMEANS with PROC GLM. It should be noted that the LSMEANS are not the same as the means obtained using the command MEANS in PROC GLM. Note also that the names "population marginal means" and "least squares means" are not used universally throughout the literature which is the cause of some confusion. General formulas for the least squares means are

$$\hat{\bar{\mu}}_{i\cdot} = \hat{\mu} + \hat{\alpha}_i + \frac{\sum_{j=1}^{b}\hat{\tau}_j}{b} + \frac{\sum_{j=1}^{b}\hat{\gamma}_{ij}}{b} \quad \text{and} \quad \hat{\bar{\mu}}_{\cdot j} = \hat{\mu} + \frac{\sum_{i=1}^{a}\hat{\alpha}_i}{a} + \hat{\tau}_j + \frac{\sum_{i=1}^{a}\hat{\gamma}_{ij}}{a}$$

where $\hat{\mu}, \hat{\alpha}_i, \hat{\tau}_j$, and $\hat{\gamma}_{ij}$ are estimates found by solving the normal equations $\mathbf{X'X}\beta = \mathbf{X'y}$ for β. The solution is denoted as $\hat{\beta}$, and is given by

$$\hat{\beta} = \left(\hat{\mu} \quad \hat{\alpha}_1 \quad \hat{\alpha}_2 \quad \cdots \quad \hat{\alpha}_a \quad \hat{\tau}_1 \quad \hat{\tau}_2 \quad \cdots \quad \hat{\tau}_b \quad \hat{\gamma}_{11} \quad \hat{\gamma}_{12} \quad \cdots \quad \hat{\gamma}_{ab}\right)'.$$

Although this solution is not unique, the population marginal means are estimable, so the least squares means are unique. We can also express the formulas for the least squares means using the cell means model. For Factor A the formula is

$$\hat{\bar{\mu}}_{i\cdot} = \frac{\sum_{j=1}^{b}\frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}}}{b} = \frac{\hat{\mu}_{i\cdot}}{b} \text{ where } \hat{\mu}_{i\cdot} = \sum_{j=1}^{b}\hat{\mu}_{ij} \text{ and } \hat{\mu}_{ij} = \frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}}.$$

Note that $\hat{\mu}_{ij}$ is just the estimated cell mean for cell *ij*. Similarly, for Factor B the formula is

$$\hat{\bar{\mu}}_{\cdot j} = \frac{\sum_{i=1}^{a}\frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}}}{a} = \frac{\hat{\mu}_{\cdot j}}{a} \text{ where } \hat{\mu}_{\cdot j} = \sum_{i=1}^{a}\hat{\mu}_{ij} \text{ and } \hat{\mu}_{ij} = \frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}} \text{ as before.}$$

With this review of the cell means model, we can restate the hypotheses listed earlier as those most often of interest to the researcher. In terms of the cell means model, these hypotheses are

$$H_o: \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot}, \text{ and}$$
$$H_o: \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \cdots = \bar{\mu}_{\cdot b}$$

for the two-factor model without interaction, and

$$H_o: \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot},$$
$$H_o: \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \cdots = \bar{\mu}_{\cdot b}, \text{ and}$$
$$H_o: \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0 \text{ for all } i \neq i' \text{ and } j \neq j'$$

for the two-factor model with interaction (Milliken and Johnson, 1992).

In the two-factor model without interaction, these hypotheses can be stated in terms of the effects model as

$$H_o: \alpha_1 = \alpha_2 = \cdots = \alpha_a, \text{ and}$$
$$H_o: \tau_1 = \tau_2 = \cdots = \tau_b,$$

and in the two-factor model with interaction as

$$H_o: \alpha_1 + \frac{\sum_{j=1}^{b}\gamma_{1j}}{b} = \alpha_2 + \frac{\sum_{j=1}^{b}\gamma_{2j}}{b} = \cdots = \alpha_a + \frac{\sum_{j=1}^{b}\gamma_{aj}}{b},$$

$$H_o: \tau_1 + \frac{\sum_{i=1}^{a}\gamma_{i1}}{a} = \tau_2 + \frac{\sum_{i=1}^{a}\gamma_{i2}}{a} = \cdots = \tau_b + \frac{\sum_{i=1}^{a}\gamma_{ib}}{a}, \text{ and}$$

$$H_o: \gamma_{ij} - \frac{\sum_{j=1}^{b}\gamma_{ij}}{b} - \frac{\sum_{i=1}^{a}\gamma_{ij}}{a} + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}}{ab} = 0 \text{ for all } i \text{ and } j$$

(Milliken and Johnson, 1992). If we consider the population marginal means as representing the levels of main effects, then one advantage of using the cell means model over the effects model becomes apparent: Under the cell means model, the Type III Hypotheses for main effects are the same whether interaction is present or not.

Again, these are the hypotheses tested by the Type III Analysis, and they are tested in an unweighted manner in that the cell frequencies do not effect the hypotheses being tested. When empty cells are present in the design, the Type III Hypotheses will depend on which cells are empty and which are not empty, but they still won't depend on the cell frequencies. As mentioned previously, when empty cells are present, the Type IV

Analysis should probably be used. However, the Type IV Analysis is not a subject of this paper. To see a summary of the weighted nature of the Type I and Type II Analyses as opposed to the unweighted nature of the Type III Analysis, refer to page 160 of Littell, Freund, and Spector (1991) or to page 106 of Speed, Hocking, and Hackney (1978).

As with the Type I and Type II Analyses, the Type III Hypotheses are closely related to the Type III Estimable Functions. If we can determine the Type III Estimable Functions, then we can determine the hypotheses being tested, and visa versa. So when we consider the Type III Hypotheses given above, it is easy to see that a basis set of Type III Estimable functions under the cell means model is

$$\bar{\mu}_{i\cdot} - \bar{\mu}_{i'\cdot} \text{ for Factor A where } i \neq i', \text{ and}$$
$$\bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot j'} \text{ for Factor B where } j \neq j'$$

if we assume the two-factor model without interaction. If we assume the two-factor model with interaction, then a basis set of Type III Estimable Functions under the cell means model is

$$\bar{\mu}_{i\cdot} - \bar{\mu}_{i'\cdot} \text{ for Factor A where } i \neq i',$$
$$\bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot j'} \text{ for Factor B where } j \neq j', \text{ and}$$
$$\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \text{ for AB interaction where } i \neq i' \text{ and } j \neq j'.$$

In terms of the effects model, for the two-factor model without interaction a basis set of Type III Estimable Functions can be written as

$$\alpha_i - \alpha_{i'} \text{ for Factor A where } i \neq i', \text{ and}$$
$$\tau_j - \tau_{j'} \text{ for Factor B where } j \neq j'.$$

If we assume the two-factor model with interaction, then a basis set of Type III Estimable Functions under the effects model is

$$\alpha_i - \alpha_{i'} + \frac{\sum_{j=1}^{b} \gamma_{ij} - \sum_{j=1}^{b} \gamma_{i'j}}{b} \text{ for Factor A where } i \neq i',$$
$$\tau_j - \tau_{j'} + \frac{\sum_{i=1}^{a} \gamma_{ij} - \sum_{i=1}^{a} \gamma_{ij'}}{a} \text{ for Factor B where } j \neq j', \text{ and} \qquad (5)$$
$$\gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'} \text{ for AB interaction where } i \neq i' \text{ and } j \neq j'.$$

Note that these functions can be written as

$(\alpha_i + \bar{\gamma}_{i\cdot}) - (\alpha_{i'} + \bar{\gamma}_{i'\cdot})$ for Factor A where $i \neq i'$,

$(\tau_j + \bar{\gamma}_{\cdot j}) - (\tau_{j'} + \bar{\gamma}_{\cdot j'})$ for Factor B where $j \neq j'$, and

$\gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'}$ for AB interaction where $i \neq i'$ and $j \neq j'$.

For a particular design, the Type III Estimable Functions can also be obtained directly from the general form of estimable functions. We begin by describing a method for obtaining the general form of estimable functions. If **X** is the design matrix, then **C'**β is estimable if and only if **C** $\in$ $\mathcal{R}$(**X'**), the column space of **X'**. Various spanning sets of estimable functions can be formed, one of which is **X'X**β. Note **X'X**β that is a spanning set of estimable functions of β because **X'X** is a spanning set for $\mathcal{R}$(**X'**). Other spanning sets for $\mathcal{R}$(**X'**) include **TX'X** for any nonsingular matrix **T** (**T**: $p \times p$). However, if we are working under the assumption that **X** is an $n \times p$ design matrix having rank $r < p$, then **X'X** and **TX'X** are $p \times p$ with rank r, and are therefore not basis sets for $\mathcal{R}$(**X'**) (recall that a basis set is a minimal spanning set). We can form a basis set for $\mathcal{R}$(**X'**) by forming a matrix consisting of the non-zero rows of the reduced row echelon form of **X'X** or **TX'X**. Call this matrix **K'**, then **K'** is $r \times p$ with rank r, and **K'** is a basis set for $\mathcal{R}$(**X'**). Pre-multiplying **K'** by any nonsingular matrix **L** (**L**: $r \times r$) will produce another basis set **LK'** of $\mathcal{R}$(**X'**). Thus, **LK'**β will be a basis set of estimable functions of β. The matrix **LK'** is referred to as the general form of estimable functions (Boik, 1996). As with the Type I and Type II Analyses, we can use SAS® to determine appropriate **LK'** matrices for the Type III Estimable Functions. This will be discussed momentarily, but we first show how the Type III Estimable Functions may be obtained from the general form of estimable functions.

To obtain the Type III Estimable Functions, we first require the matrix **LK'** mentioned in the preceding paragraph to be a diagonal matrix. We can label the diagonal entries in **L** as l_i for $i = 1,2,\ldots,r$. We then choose the diagonal elements of **L** so that the coefficients of estimable functions for each lower order effect are orthogonal to the coefficients of estimable functions for higher order effects which contain the effect in question (Boik, 1996, and Goodnight, 1980). So Type III Estimable Functions involve only parameters of the effect in question and parameters of effects which contain the effect in question, and the coefficients of lower order effects are orthogonal to the coefficients of higher order effects which contain the lower order effects. These requirements are more easily explained using a specific example and using SAS® to help us determine the general form of estimable functions. Before continuing to an example involving the use of SAS® to obtain the Type III Estimable Functions, note that if the model contains no higher order effects such as interaction terms, then the Type III Estimable Functions coincide with the Type II Estimable Functions. This is true whether the data are balanced or unbalanced, recall that when the model contains only main effects (no interaction is present in the model), the Type II SS are the same whether the data are balanced or not.

Let us consider the two-factor model with interaction. Suppose that Factor A has two levels and Factor B has three levels. The model statement

model y = A B A*B / SS3 E E3;

in PROC GLM of SAS® will conduct an ANOVA test using the Type III SS and will also print out a representation of the Type III Estimable Functions as well as the general form of estimable functions. For this example, the representation of estimable functions is given below.

General Form of Estimable Functions

Effect	Coefficients	
INTERCEPT		L1
A	1	L2
	2	L1 – L2
B	1	L4
	2	L5
	3	L1 – L4 – L5
A*B	1 1	L7
	1 2	L8
	1 3	L2 – L7 – L8
	2 1	L4 – L7
	2 2	L5 – L8
	2 3	L1 – L2 – L4 – L5 + L7 + L8

Type III Estimable Functions for: A

Effect	Coefficients	
INTERCEPT		0
A	1	L2
	2	– L2
B	1	0
	2	0
	3	0
A*B	1 1	0.3333*L2
	1 2	0.3333*L2
	1 3	0.3333*L2
	2 1	–0.3333*L2
	2 2	–0.3333*L2
	2 3	–0.3333*L2

Type III Estimable Functions for: B

Effect	Coefficients	
INTERCEPT		0
A	1	0
	2	0
B	1	L4
	2	L5
	3	–L4 – L5
A*B	1 1	0.5*L4
	1 2	0.5*L5
	1 3	–0.5*L4 – 0.5*L5
	2 1	0.5*L4
	2 2	0.5*L5
	2 3	–0.5*L4 – 0.5*L5

Type III Estimable Functions for: A*B

Effect	Coefficients	
INTERCEPT		0
A	1	0
	2	0
B	1	0
	2	0
	3	0
A*B	1 1	L7
	1 2	L8
	1 3	–L7 – L8
	2 1	–L7
	2 2	–L8
	2 3	L7 + L8

By letting in turn each L# (for # = 1, 2, ..., 8) equal one while setting all other L#'s equal to zero, we can obtain the general form of estimable functions as well as the Type III Estimable Functions. We can also see that the Type III Estimable Functions we obtain agree with those presented in (5).

To see a full explanation of how to obtain the Type III Estimable Functions from the general form of estimable functions refer to pages 100 and 101 of SAS/STAT, 1988 where a three-step process is presented. Following the three-step process to obtain the Type III Estimable Functions for Factor A from the general form of estimable functions for this example, we

1. set L1 = L4 = L5 = 0,

2. not necessary for this example,

3. set L7 = $k_{7,2}$ L2 and L8 = $k_{8,2}$L2, then solve for $k_{7,2}$ and $k_{8,2}$ so that the Type III coefficients for Factor A are orthogonal to the Type III coefficients for A*B.

For now, we can write the Type III coefficients for Factor A as:

Effect	Coefficients	
INTERCEPT		0
A	1	L2
	2	– L2
B	1	0
	2	0
	3	0
A*B	1 1	$k_{7,2}$ L2
	1 2	$k_{8,2}$L2
	1 3	$(1 - k_{7,2} - k_{8,2})$L2
	2 1	$-k_{7,2}$ L2
	2 2	$-k_{8,2}$L2
	2 3	$(-1 + k_{7,2} + k_{8,2})$L2

Solutions for $k_{7,2}$ and $k_{8,2}$ are developed below. Note that the Type III coefficients for A*B are obtained by

1. setting L1 = L2 = L4 = L5 = 0,

2. not necessary for this example,

3. no action needed since A*B is not contained in any other effect.

Thus the Type III coefficients for A*B are:

Effect	Coefficients	
INTERCEPT		0
A	1	0
	2	0
B	1	0
	2	0
	3	0

A*B	1 1	L7
	1 2	L8
	1 3	–L7 – L8
	2 1	–L7
	2 2	–L8
	2 3	L7 + L8.

Let $\mathbf{C}_A$ be the column matrix containing the coefficients for Factor A, and let $\mathbf{C}_{A*B}$ be the matrix of coefficients for A*B. Then the orthogonality requirement can be expressed as $\mathbf{C}_A'\mathbf{C}_{A*B} = \mathbf{0}$. For our example, this says that

$$k_{7,2}\,L2{\cdot}L7 + k_{8,2}L2{\cdot}L8 - L2{\cdot}L7 - L2{\cdot}L8 + k_{7,2}\,L2{\cdot}L7 + k_{8,2}L2{\cdot}L8 + k_{7,2}L2{\cdot}L8 + k_{8,2}L2{\cdot}L7 + k_{7,2}L2{\cdot}L7 + k_{8,2}L2{\cdot}L8 - L2{\cdot}L7 - L2{\cdot}L8 + k_{7,2}\,L2{\cdot}L7 + k_{7,2}L2{\cdot}L8 + k_{8,2}L2{\cdot}L7 + k_{8,2}L2{\cdot}L8 = 0 \text{ for all choices of L2, L7, and L8.} \quad (6)$$

So, if we set L2 = L7 = 1 and L8 = 0, then (6) says that $2k_{7,2} + k_{8,2} = 1$. If we then set L2 = L8 = 1 and L7 = 0, we get the equation $k_{7,2} + 2k_{8,2} = 1$. Solving these equations for $k_{7,2}$ and $k_{8,2}$, we find $k_{7,2} = k_{8,2} = 1/3$. This result is consistent with the SAS® printout for the Factor A Type III coefficients. This process can also be used to find the Type III coefficients for Factor B.

Having obtained the representation for Type III Estimable Functions using SAS®, we note that the Type III coefficients for Factors A and B do follow the requirement of orthogonality, inner products of Factor A coefficients with A*B coefficients equal zero as do inner products of Factor B coefficients with A*B coefficients. For example, set L2 of the Factor A coefficients to one, then $\mathbf{C}_A$ is

$$\mathbf{C}_A = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{-1}{3} & \frac{-1}{3} & \frac{-1}{3} \end{pmatrix}'.$$

Next, set L7 to one and L8 to zero for the A*B coefficients. This produces one column of the matrix $\mathbf{C}_{A*B}$. Then set L7 to zero and L8 to one, this produces the second column of $\mathbf{C}_{A*B}$. So that $\mathbf{C}_{A*B}$ is

$$\mathbf{C}_{A*B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}'.$$

Finally then, $\mathbf{C}_A'\mathbf{C}_{A*B} = \mathbf{0}$. Where $\mathbf{0} = (0\ 0)$. Similar results hold for Factor B, if we set L4 of the Factor B coefficients to 1 and L5 to zero, we get a column of coefficients for the matrix $\mathbf{C}_B$. We then set L4 to zero and L5 to one and get the second column of coefficients for the matrix $\mathbf{C}_B$. The result is

$$\mathbf{C}_B = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & -1 & \frac{1}{2} & 0 & \frac{-1}{2} & \frac{1}{2} & 0 & \frac{-1}{2} \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & \frac{1}{2} & \frac{-1}{2} & 0 & \frac{1}{2} & \frac{-1}{2} \end{pmatrix}'.$$

Using the same matrix $\mathbf{C}_{A*B}$ as constructed above, we get

$$\mathbf{C}_B'\mathbf{C}_{A*B} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \mathbf{0}.$$

Having discussed the nature and properties of the Type III Analysis, let us mention situations where the Type III Analysis would be appropriate, as well as some situations where the Type III Analysis would not be appropriate.

We saw earlier that the Type I and Type II Analyses were appropriate for model building situations. In the model building setting, different effects (main effects and higher order effects such as interaction effects) are considered and tested in order to establish a "best" model, probably to be used for making predictions. The exact nature of the hypotheses being tested is usually not of primary interest. Recall for example that the Type I and Type II Hypotheses for unbalanced data and a design which includes interaction terms can be very difficult to interpret from a practical point of view. Yet the Type I and Type II Analyses can be appropriate in such settings because the Type I and Type II Analyses are more concerned with testing the importance of the various terms in the model than with the exact hypothesis being tested. This is generally not the case for the Type III Analysis. For the Type III Analysis, the "primary interest is in testing interpretable functions of the parameters in a given model. That is, the model is known; there is no model building. Instead, one concentrates on testing functions of the known model" (Boik, 1996). Put another way, the primary objective of the Type III Analysis is comparing the effects of different treatments. The Type III Analysis is especially versatile in that the interpretations of the tests are the same whether the data are balanced or not, and, under the cell means model, the main effects tests are the same whether interaction terms are present or not. For this reason, the Type III Analysis is generally considered the appropriate analysis to use when the goal of the analysis is to compare main effects in a model which contains interaction.

In general we can say that the Type III Analysis is the appropriate analysis to use for ANOVA designs when the data are unbalanced (assuming no empty cells are present). As was mentioned at the time we presented the Type III Hypotheses and the Type III Estimable Functions, the Type III Analysis tests an unweighted hypothesis concerning various effects of interest. These Type III Hypotheses did not depend on the cell frequencies. This can be considered appropriate if the cell frequencies are not intended to represent corresponding population proportions. For example, it may be that the data set was intended to be balanced but some data was lost or otherwise had to be discarded form

the data set. In such circumstances, it seems reasonable to say that the resulting imbalance in the data set should not change the hypotheses the researcher originally intended to test. However, there may be instances where just the opposite is true. We may at times want to conduct a weighted analysis to preserve, as part of the analysis, the population proportions of the different treatments as contained in the underlying population. In such cases the Type II Analysis should be used, or the Type I Analysis if the sequence of the terms is also important to the analysis. The cell frequencies will then represent the treatment proportions of the underlying population.

Summary

By way of summary, the following guidelines are suggested when conducting analyses where the data set contains no empty cells:

Type I Analysis — To be used in model building situations where effects are to be considered sequentially. Unbalanced data will complicate the hypotheses being tested possibly resulting in hypotheses which do not correspond to those of interest to the researcher.

Type II Analysis — To be used in model building situations where the effects are to be tested in the presence of all other effects under consideration. Unbalanced data will complicate the hypotheses being tested if the model contains higher order terms such as interactions. The resulting hypotheses may not be of interest to the researcher. Therefore, the Type II Analysis is generally not considered appropriate for unbalanced data sets when interaction is included in the model. The Type II Hypotheses will be the same as the Type III Hypotheses when the model contains only main effects.

Type III Analysis — Generally used when the researcher desires to compare effects in a specified model, not in a model building situation. The Type III Analysis is an unweighted analysis and is therefore the method most often recommended for unbalanced data sets. The Type III Hypotheses are the same whether the data are balanced or not.

References

Bolgiano, D. C. (1981). Making Sense of Types I, II, III, and IV Sums of Squares in SAS GLM. *Proceedings of the SAS Users Group International Annual Conference.* 6: 85–88.

Boik, R. J. (1996). Course Notes for STAT 506, Linear Models, Montana State University. 14–21.

Freund, R. J. & Littell, R. C. (1991). *SAS® System for Regression* (Second Edition). Cary, NC: SAS Institute Inc.

Goodnight, J. H. (1980). Tests of Hypotheses in Fixed Effects Linear Models. *Communications in Statistics*. A9(2): 167–180.

Goodnight, J. H. & Harvey, W. R. (1978). Least Squares Means in the Fixed Effects General Model. *SAS® Technical Report R–103*. Cary, NC: SAS Institute Inc.

Graybill, F. A. (1983). *Matrices with Applications in Statistics* (Second Edition). Belmont, CA: Wadsworth.

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

Milliken G. A. & Johnson, D. E. (1992). *Analysis of Messy Data, Volume I: Designed Experiments*. New York: Chapman & Hall.

Miller, R. L. (1981). The Cell Means Model as an Analytical Tool for Evaluating SAS GLM Type III and IV Sums of Squares for Linear Models with Missing Cells. *Proceedings of the SAS Users Group International Annual Conference*. 6: 547–554.

Myers, R. H. & Milton, J. S. (1991). *A First Course in the Theory of Linear Statistical Models*. Boston: PWS-Kent Publishing Co.

Pendleton, O. J., Von Tress, M. & Bremer R. (1986). Interpretation of the Four Types of Analysis of Variance Tables in SAS. *Communications in Statistics*. 15(9): 2785–2808.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley & Sons, Inc.

Searle, S. R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.

Speed, F. M., Hocking, R. R. & Hackney, O. P. (1978). Methods of Analysis of Linear Models with Unbalanced Data. *Journal of the American Statistical Association*. 73: 105–112.

SAS/STAT® User's Guide, Release 6.03 Edition (1988). Cary NC: SAS Institute, Inc.

Appendix

MATLAB programs

The following MATLAB programs will calculate the L matrices needed to obtain the various estimable functions. Each program is based on a two-factor design. The presence or absence of interaction will be indicated.

The following program determines the Type I Estimable Functions for balanced data where the design has no interaction term. In this case we are assuming Factor A has four levels and Factor B has three levels, and that there is only one replicate per cell. The value n mentioned in the program may be changed to allow for other cell frequencies provided that the data remain balanced. Note that Factor A is added to the model first, then Factor B.

```
a=4;
b=3;
n=1;
x1=kron(kron(ones(n),ones(b,1)),ones(a,1))
x2=kron(kron(ones(n),ones(b,1)),eye(a))
x3=kron(kron(ones(n),eye(b)),ones(a,1))
x=[x1 x2 x3]

h1=x1*pinv(x1'*x1)*x1'
x2_1p=x2'*(eye(a*b)-h1)
lstar_a=[x2_1p*x1 x2_1p*x2 x2_1p*x3]
eps=1000*eps;
l_a=rref(lstar_a)

x12=[x1 x2]
h12=x12*pinv(x12'*x12)*x12'
x3_12p=x3'*(eye(a*b)-h12)
lstar_b=[x3_12p*x1 x3_12p*x2 x3_12p*x3]
l_b=rref(lstar_b)
```

The following program determines the Type I Estimable Functions for unbalanced data where the design has no interaction term. In this case we are assuming Factor A has three levels and Factor B has four levels, and that there are two replicates in each cell except the cell where both factors are at level one. In that cell there is only one observation. The matrix N mentioned in the program can be changed to allow other numbers of levels for the factors as well as other cell frequencies. Note that Factor A is added to the model first, then Factor B.

```
N=[1 2 2 2;2 2 2 2;2 2 2 2];
[a,b]=size(N);
x=[];
s=[];
ea=eye(a);
eb=eye(b);

for j=1:b
for i=1:a
for k=1:N(i,j)
s(k,:)=[1 ea(i,:) eb(j,:)];
end;
x=[x' s']';
s=[];
end;
end;

x

x1=x(:,1);
x2=x(:,2:2+(a-1));
x3=x(:,a+2:a+b+1);

h1=x1*pinv(x1'*x1)*x1';
[r,t]=size(h1);
x2_1p=x2'*(eye(r)-h1);
lstar_a=[x2_1p*x1 x2_1p*x2 x2_1p*x3];
eps=1000*eps;
l_a=rref(lstar_a)

x12=[x1 x2];
h12=x12*pinv(x12'*x12)*x12';
[r,t]=size(h12);
x3_12p=x3'*(eye(r)-h12);
lstar_b=[x3_12p*x1 x3_12p*x2 x3_12p*x3];
l_b=rref(lstar_b)
```

The following program determines the Type I Estimable Functions for unbalanced data where the design does have an interaction term. In this case we are assuming Factor A has three levels and Factor B has four levels, and that there are two replicates in each cell except the cell where both factors are at level one. In that cell there is only one observation. The matrix N mentioned in the program can be changed to allow other numbers of levels for the factors as well as other cell frequencies. Note that Factor A is added to the model first, then Factor B, then finally the interaction term.

```
N=[1 2 2 2;2 2 2 2;2 2 2 2];
[a,b]=size(N);
x=[];
s=[];
ea=eye(a);
eb=eye(b);

for j=1:b
for i=1:a
for k=1:N(i,j)
s(k,:)=[1 ea(i,:) eb(j,:)];
end;
x=[x' s']';
s=[];
end;
end;

x

x1=x(:,1);
x2=x(:,2:2+(a-1));
x3=x(:,a+2:a+b+1);

x123=x;
x4=[];
k=1;
for i=1:a;
for j=1:b;
x4(:,k)=x2(:,i).*x3(:,j);
k=k+1;
end;
end;
x=[x x4];

h1=x1*pinv(x1'*x1)*x1';
[r,t]=size(h1);
x2_1p=x2'*(eye(r)-h1);
lstar_a=[x2_1p*x1 x2_1p*x2 x2_1p*x3 x2_1p*x4];
eps=1000*eps;
l_a=rref(lstar_a)
```

```
x12=[x1 x2];
h12=x12*pinv(x12'*x12)*x12';
[r,t]=size(h12);
x3_12p=x3'*(eye(r)-h12);
lstar_b=[x3_12p*x1 x3_12p*x2 x3_12p*x3 x3_12p*x4];
l_b=rref(lstar_b)

h123=x123*pinv(x123'*x123)*x123';
[r,t]=size(h123);
x4_123p=x4'*(eye(r)-h123);
lstar_ab=[x4_123p*x1 x4_123p*x2 x4_123p*x3 x4_123p*x4];
l_ab=rref(lstar_ab)
```

The following program determines the Type II Estimable Functions for balanced data where the design has no interaction term. In this case we are assuming Factor A has four levels and Factor B has three levels, and that there is only one replicate per cell. The value n mentioned in the program may be changed to allow for other cell frequencies provided that the data remain balanced.

```
a=4;
b=3;
n=1;
x1=kron(kron(ones(n),ones(b,1)),ones(a,1))
x2=kron(kron(ones(n),ones(b,1)),eye(a))
x3=kron(kron(ones(n),eye(b)),ones(a,1))
x=[x1 x2 x3]

x13=[x1 x3]
h13=x13*pinv(x13'*x13)*x13'
x2_13p=x2'*(eye(a*b)-h13)
lstar_a=[x2_13p*x1 x2_13p*x2 x2_13p*x3]
l_a=rref(lstar_a)

x12=[x1 x2]
h12=x12*pinv(x12'*x12)*x12'
x3_12p=x3'*(eye(a*b)-h12)
lstar_b=[x3_12p*x1 x3_12p*x2 x3_12p*x3]
l_b=rref(lstar_b)
```

The following program determines the Type II Estimable Functions for unbalanced data where the design has no interaction term. In this case we are assuming Factor A has three levels and Factor B has four levels, and that there are two replicates in each cell except the cell where both factors are at level one. In that cell there is only one observation. The matrix N mentioned in the program can be changed to allow other numbers of levels for the factors as well as other cell frequencies.

```
N=[1 2 2 2;2 2 2 2;2 2 2 2];
[a,b]=size(N);
x=[];
s=[];
ea=eye(a);
eb=eye(b);

for j=1:b
for i=1:a
for k=1:N(i,j)
s(k,:)=[1 ea(i,:) eb(j,:)];
end;
x=[x' s']';
s=[];
end;
end;

x

x1=x(:,1);
x2=x(:,2:2+(a-1));
x3=x(:,a+2:a+b+1);

x13=[x1 x3];
h13=x13*pinv(x13'*x13)*x13';
[r,t]=size(h13);
x2_13p=x2'*(eye(r)-h13);
lstar_a=[x2_13p*x1 x2_13p*x2 x2_13p*x3];
eps=1000*eps;
l_a=rref(lstar_a)

x12=[x1 x2];
h12=x12*pinv(x12'*x12)*x12';
[r,t]=size(h12);
x3_12p=x3'*(eye(r)-h12);
lstar_b=[x3_12p*x1 x3_12p*x2 x3_12p*x3];
l_b=rref(lstar_b)
```

The following program determines the Type II Estimable Functions for unbalanced data where the design does have an interaction term. In this case we are assuming Factor A has three levels and Factor B has four levels, and that there are two replicates in each cell except the cell where both factors are at level one. In that cell there is only one observation. The matrix N mentioned in the program can be changed to allow other numbers of levels for the factors as well as other cell frequencies.

```
N=[1 2 2 2;2 2 2 2;2 2 2 2];
[a,b]=size(N);
x=[];
s=[];
ea=eye(a);
eb=eye(b);

for j=1:b
for i=1:a
for k=1:N(i,j)
s(k,:)=[1 ea(i,:) eb(j,:)];
end;
x=[x' s']';
s=[];
end;
end;

x

x1=x(:,1);
x2=x(:,2:2+(a-1));
x3=x(:,a+2:a+b+1);

x123=x;
x4=[];
k=1;
for i=1:a;
for j=1:b;
x4(:,k)=x2(:,i).*x3(:,j);
k=k+1;
end;
end;
x=[x x4];

x13=[x1 x3];
h13=x13*pinv(x13'*x13)*x13';
[r,t]=size(h13);
x2_13p=x2'*(eye(r)-h13);
lstar_a=[x2_13p*x1 x2_13p*x2 x2_13p*x3 x2_13p*x4];
eps=1000*eps;
l_a=rref(lstar_a)
```

```
x12=[x1 x2];
h12=x12*pinv(x12'*x12)*x12';
[r,t]=size(h12);
x3_12p=x3'*(eye(r)-h12);
lstar_b=[x3_12p*x1 x3_12p*x2 x3_12p*x3 x3_12p*x4];
l_b=rref(lstar_b)

h123=x123*pinv(x123'*x123)*x123';
[r,t]=size(h123);
x4_123p=x4'*(eye(r)-h123);
lstar_ab=[x4_123p*x1 x4_123p*x2 x4_123p*x3 x4_123p*x4];
l_ab=rref(lstar_ab)
```

Designs which do not contain interaction will lead to Type III Estimable Functions which coincide with the Type II Estimable Functions whether the data are balanced or not, so we do not give separate programs for Type III Estimable Functions under such conditions.

The following program determines the Type III Estimable Functions for unbalanced data where the design does have an interaction term. In this case we are assuming Factor A has three levels and Factor B has four levels, and that there are two replicates in each cell except the cell where both factors are at level one. In that cell there is only one observation. The matrix N mentioned in the program can be changed to allow other numbers of levels for the factors as well as other cell frequencies.

```
N=[3 2 3;2 3 3];
ets=ets*0.00001;
[a,b]=size(N);
x=[];
s=[];
ea=eye(a);
eb=eye(b);

for j=1:b
for i=1:a
for k=1:N(i,j)
s(k,:)=[1 ea(i,:) eb(j,:)];
end;
x=[x' s']';
s=[];
end;
end;
```

```
x1=x(:,1);
x2=x(:,2:2+(a-1));
x3=x(:,a+2:a+b+1);

x123=x;
x4=[];
k=1;
for i=1:a;
for j=1:b;
x4(:,k)=x2(:,i).*x3(:,j);
k=k+1;
end;
end;
x=[x x4]

L=rref(x'*x)';
L=L(:,1:rank(L))

[n,p1]=size([x1 x2]);
[n,p2]=size(x3);
[n,p3]=size(x4);
r1=rank([x1 x2]);
r2=rank([x1 x2 x3])-r1;
r3=rank(x)-r1-r2;
G=rref(x'*x);
G11=G(2:r1,2:p1);
G22=G(r1+1:r1+r2,p1+1:p1+p2);
G13=G(2:r1,p1+p2+1:p1+p2+p3);
G23=G(r1+1:r1+r2,p1+p2+1:p1+p2+p3);
G33=G(r1+r2+1:r1+r2+r3,p1+p2+1:p1+p2+p3);
HG33p=G33'*pinv(G33*G33')*G33;
N=[G11 G13-G13*HG33p];
M=[G22 G23-G23*HG33p];
t3a=rref(N)';
t3b=rref(M)';
l_a=t3a(:,1:rank(t3a));
l_b=t3b(:,1:rank(t3b));
[r,t]=size(l_a);
za=zeros(b,t);
l_a=[zeros(t,1) l_a(1:a,:)' za' l_a(a+1:r,:)']'
[r,t]=size(l_b);
zb=zeros(a,t);
l_b=[zeros(t,1) zb' l_b']'

h=x*pinv(x'*x)*x';
h123=x123*pinv(x123'*x123)*x123';
T=(h-h123)*x;
K3ab=rref(T'*T)';
l_ab=K3ab(:,1:rank(K3ab))
```

The General form of the F-statistic for a Type I Analysis

While it is not the intention of this paper to develop the distribution of the test statistic, it should be beneficial to at least mention a point or two concerning the construction of the F-statistic as it relates to the test of hypothesis being conducted. When the F-statistic is viewed in the context of a test of hypothesis which tests a null hypothesis of the form

$$H_o: E(\mathbf{y})=\mathbf{X}^*\beta^*$$

versus an alternative hypothesis of the form

$$H_a: E(\mathbf{y})=\mathbf{X}\beta$$

where $\mathbf{X}^*\beta^*$ is a nested reduction of $\mathbf{X}\beta$, and $\mathbf{X}\beta$ represents the complete model which involves all terms under consideration in the problem, then the F-statistic takes on the following format (Boik, 1996):

$$F = \frac{\dfrac{SSE_o - SSE_a}{m-k}}{\dfrac{SSE_a}{n-m}}$$

where SSE_o refers to the residual sum of squares for the model implied by the null hypothesis and SSE_a refers to the residual sum of squares for the model implied by the alternative hypothesis. We have already seen that $SSE_o - SSE_a$ is equivalent to the R(|) notation which is the Type I SS for the term being tested. An easy way to think of the quantity $m-k$ is to look at the null hypothesis for the test. As presented in the preceding example, the null hypothesis can be written as $(\mathbf{H}_@ - \mathbf{H}_\#)E(\mathbf{y}) = \mathbf{0}$ where $\mathbf{H}_@$ and $\mathbf{H}_\#$ are the perpendicular projection operators representing the alternative and null hypotheses respectively. With this notation, $m = \text{rank}(\mathbf{H}_@)$ and $k = \text{rank}(\mathbf{H}_\#)$. Finally, n comes from the dimensions of $\mathbf{y}$ and $\mathbf{X}$, $\mathbf{y}: n\times 1$ and $\mathbf{X}: n\times p$. These results parallel those described for the F-statistic of a Type II Analysis as presented on page 17.