Union-Intersection Multiple Comparison Tests

Jerry Brunner*

March 18, 2004

Abstract

The union-intersection principle yields a class of multiple comparison tests that include the classical Scheffé tests for analysis of variance, and are easily applied to any likelihood ratio test. Examples are given.

1 Introduction

This document describes a class of multiple comparison procedures that generalize some essential features of the classical Scheffé tests for analysis of variance (Scheffé 1953, 1959) so that they apply to an arbitrary hypothesis test. As in the Scheffé procedures, once a null hyothesis is rejected using an overall test, one can indulge in unlimited exploration within a family of component tests that are simultaneously protected against Type One error at joint signficance level α, the significance level of the overall test. In practice the method is sequential; the overall test is performed first, and if the null hypothesis is rejected, one explores with the component tests to see where the effect comes from. Consequently, in this document the overall test will be called the *initial test*, and the component tests will be called *follow-up tests*.

The main result is a method that can be used to construct a family of follow-up tests for any likelihood ratio test. The follow-ups are also likelihood ratio tests. If the null hypothesis of each follow-up test is implied by the null hypothesis of the initial test and if one uses the critical value of likelihood ratio statistic from the initial test, then all the follow-ups will be protected against Type One error at joint signficance level α.

This is not new at all. Roy (1953) is responsible for the union-intersection method of constructing overall tests from a family of component tests, and Roy and Bose (1953) note that the component tests are simultaneously protected at the significance level of the overall test . Gabriel (1969) is responsible for the result on likelihood ratio tests, and Hochberg and Tamhane (1987) provide an admirable account of multiple comparison methods in general, including the material presented here.

The present treatment may be accessible to a broader audience. The development is singlemindedly on *tests* rather than confidence regions, keeping in mind the image of a scientist who is only allowed to discuss findings that are statistically significant, and who is therefore much more interested in testing than in estimation. This focus makes it

*This work was supported by the Natural Sciences and Engineering Research Council of Canada, Grant OGPIN 014 .

possible to ignore a large amount of extraneous material that might be distracting to those who want to use the methods to explore a dataset using significance tests. In addition, details — trivial details, from the viewpoint of statistical research — are worked out for important tools such as multiple regression, multivariate analysis of variance (including repeated measures) and logistic regression. There is a collection of "recipies" that may be followed by scientists who use statistical methods as tools, but whose primary expertise is not in Statistics. Examples refer explicitly to the `SAS` statistical software package, and sample output is included.

One peculiarity of the examples is that they reflect a personal opinion. I believe that all significance tests based on a particular model should be simultaneously protected against Type One error. So, for example, in a three-factor analysis of variance, tests for the main effects, the two-way interactions and the three-way interaction should all be follow-ups to an initial test that indicated an overall difference among cell means. The examples include this sort of application as well as more traditional ones.

In spite of the applied intent of this document, the development is relatively complete, with most relevant definitions and proofs (there are not many) supplied in Section 2. Then go on to say what's in the other sections.

2 Union-Intersection Theory

We begin with the following general hypothesis-testing framework.

$$\begin{array}{l} Y \sim P_\theta,\ \theta \in \Theta, \\ H_0 : \theta \in \Theta_0 \text{ v.s. } H_A : \theta \in \Theta \cap \Theta_0^c, \\ \text{and critical region } C = \{y : H_0 \text{ is rejected when } Y \in C\} \\ \text{with } P_\theta\{Y \in C\} \leq \alpha \text{ for all } \theta \in \Theta_0 \end{array} \tag{1}$$

Of course the random quantity Y, the corresponding set of fixed values y and the parameter θ can be very large vectors. For example, in testing the parameters of a multivariate normal we might have $Y = \mathbf{Y}_1, \ldots, \mathbf{Y}_n$, $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and P_θ wuld posess a density $f(y;\theta)$, the product of n multivariate normal densities.

The null hypothesis Θ_0 and test C in (1) will be called the *initial null hypothesis* and the *initial test* respectively. They will be used to generate a family of *follow-up* null hypotheses and tests. All of these tests will be protected against Type One error at joint significance level α, and it will be impossible for any of them to be significant unless the initial test is significant.

Definition 1 *Assume a null hypothesis $H_0 : \theta \in \Theta_0$ and critical region C. A family of null hypotheses $\{H_\ell : \theta \in \Theta_\ell\}$ and corresponding family of tests with critical regions $\{C_\ell\}$, where ℓ belongs to some index set $\mathcal{L}$, is said to obey the* union-intersection principle *provided*

$$\Theta_0 = \bigcap_{\ell \in \mathcal{L}} \Theta_\ell \text{ and } C = \bigcup_{\ell \in \mathcal{L}} C_\ell$$

for each $\ell \in \mathcal{L}$.

It may be helpful to restate this in words.

Definition 2 *Given an initial null hypothesis and test, a family of tests is said to obey the* union-intersection principle *provided*

- *The null hypothesis of the initial test is true if and only if the null hypotheses of all the follow-up tests are true.*
- *The null hypothesis of the initial test is rejected if and only if the null hypothesis of at least one follow-up test is rejected.*

In 1 below, we see that if an initial test and family of follow-ups obey the union-intersection principle, then all the follow-ups are simultaneously protected against Type One Error at the significance level of the initial test. Note that in the proof, it is not necessary for the critical region of the initial test to exactly *equal* the union of the critical region of the follow-up tests. Containment is enough; that is, all we need is that rejection of the initial null hypothesis is implied by rejection of any follow-up null hypothesis.

Result 1 *Let the family $\mathcal{L}$ of null hypotheses and tests obey the union-intersection principle. Then $P_\theta\{Y \in \bigcup_{\ell\in\mathcal{L}} C_\ell\} \leq \alpha$ for all $\theta \in \bigcap_{\ell\in\mathcal{L}} \Theta_\ell$.*

Proof. Let $\theta \in \bigcap_{\ell\in\mathcal{L}} \Theta_\ell = \Theta_0$. Then $\bigcup_{\ell\in\mathcal{L}} C_\ell \subseteq C \Rightarrow P_\theta\{Y \in \bigcup_{\ell\in\mathcal{L}} C_\ell\} \leq P_\theta\{Y \in C\} \leq \alpha$.

The intersection part of the union-intersection principle says that if the null hypothesis of the initial test is true, then the null hypotheses of all the follow-up tests are true. Therefore, if even one follow-up null hypothesis is false, then the initial null hypothesis is false also. Therefore, union-intersection follow-up tests are very well suited for exploring *ways* in which an initial null hypothesis might be false.

Most of the time, the class of follow-ups to an initial test is very large, typically with infinitely many members. But here is an interesting exception.

Example: two-sided tests Consider a t-test, say for a regression coefficient β_k. Suppose $|t|$ exceeds some critical value $t_{1-\alpha/2}$, so that $H_0 : \beta_k = 0$ is rejected. This conclusion actually does not tell us whether $\beta_k > 0$ or $\beta_k < 0$. It is tempting to just look at the sign of the computed t-statistic and decide that way — but is such a practice justified? As evidence that this is a serious matter, Lehmann (1986, p.152) raises the question without providing a direct answer, instead referring the reader to literature on three-decision problems.

Union-intersection follow-up tests provide reassurance with very little effort. The follow-up family has two members, with null hypotheses $H_1 : \beta_k \leq 0$ and $H_2 : \beta_k \geq 0$. Immediately we have $\Theta_0 = \Theta_1 \bigcap \Theta_2$. Obtaining $C = C_1 \bigcup C_2$ is also quite easy; just rejecting H_1 if $t > t_{1-\alpha/2}$ and rejecting H_2 if $t < -t_{1-\alpha/2}$. Thus, the union-intersection follow-ups to a two-sided test are the corresponding one-sided tests, but they use the critical values of the two-sided test.

Recipie 1 *Suppose a two-sided t-test leads to the conclusion that a regression coefficient or a difference between means is different from zero. Look at the sign of the t statistic to decide whether the difference is positive or negative.*

This is what people do anyway, implicitly doing a very quick follow-up to a significant two-sided t-test. The union-intersection argument establishes that this does not inflate Type One error. It also holds so-called "Type Three Error" to level α.

We will now see that any likelihood ratio test has a natural set of union-intersection follow-ups, and in practice it is easy to construct them.

Result 2 *Let the initial null hypothesis and test be as in (1), with the critical region C based on a likelihood ratio test. Define the family of follow-up null hypotheses to be all the statements about θ that are implied by the null hypothesis of the initial test, and let the corresponding tests be likelihood ratio tests. If all the follow-up tests use the same critical value for the likelihood ratio that the initial test uses, the family will have joint significance level α.*

Proof Let $\{H_\ell : \theta \in \Theta_\ell, \ell \in \mathcal{L}\}$ be the set of null hypotheses implied by H_0. That is, $\Theta_0 \subseteq \Theta_\ell$ for each $\ell \in \mathcal{L}$, and we have $\Theta_0 \subseteq \bigcap_{\ell\in\mathcal{L}} \Theta_\ell$. Since the initial null hypothesis implies itself (the initial test is a member of the family), we also have $\theta \in \bigcap_{\ell\in\mathcal{L}} \Theta_\ell \Rightarrow \theta \in \Theta_0$, so that $\bigcap_{\ell\in\mathcal{L}} \Theta_\ell \subseteq \Theta_0$, yielding $\Theta_0 = \bigcap_{\ell\in\mathcal{L}} \Theta_\ell$. Thus the intersection part of Definition 1 is satisfied. It remains to establish the union part. Denoting by $f(\mathbf{y};\theta)$ the (Radon-Nikodym) derivative of P_θ, the critical region of the initial test has the form

$$C = \left\{ \mathbf{y} : \frac{\sup_{\theta\in\Theta_0} f(\mathbf{y};\theta)}{\sup_{\theta\in\Theta} f(\mathbf{y};\theta)} \leq k \right\} = \{\mathbf{y} : \lambda(\mathbf{y}) \leq k\}.$$

Now let

$$C_\ell = \left\{ \mathbf{y} : \frac{\sup_{\theta\in\Theta_\ell} f(\mathbf{y};\theta)}{\sup_{\theta\in\Theta} f(\mathbf{y};\theta)} \leq k \right\} = \{\mathbf{y} : \lambda_\ell(\mathbf{y}) \leq k\}.$$

Since $\Theta_0 \subseteq \Theta_\ell$, $\sup_{\theta\in\Theta_0} f(\mathbf{y};\theta) \leq \sup_{\theta\in\Theta_\ell} f(\mathbf{y};\theta)$. Therefore,

$$\frac{\sup_{\theta\in\Theta_0} f(\mathbf{y};\theta)}{\sup_{\theta\in\Theta} f(\mathbf{y};\theta)} \leq \frac{\sup_{\theta\in\Theta_\ell} f(\mathbf{y};\theta)}{\sup_{\theta\in\Theta} f(\mathbf{y};\theta)};$$

that is, $\lambda(\mathbf{y}) \leq \lambda_\ell(\mathbf{y})$. Now let $\mathbf{y} \in C_\ell \Rightarrow \lambda_\ell(\mathbf{y}) \leq k \Rightarrow \lambda(\mathbf{y}) \leq k \Rightarrow \mathbf{y} \in C$, so that $C_\ell \subseteq C$. Because the initial test is in the family, we have equality (though we don't need it). The conclusion now follows by Result 1. ■

Recipie 2 *Suppose one is using large-sample likelihood ratio tests, in which -2 times the natural log of the likelihood ratio has a chi-square distribution — say with k degrees of freedom for the initial test. Then one would just use the critical value of a chi-square with k degrees of freedom for all the follow-up tests. This covers generalized linear models (including logistic regression), log-linear models for categorical data and structural equation models of the LISREL variety, among other methods.*

3 Univariate Linear Models: The Scheffé Tests

3.1 Details

This section is concerned with the usual univariate linear model with independent normal errors. That is, let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$

is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, with $\sigma^2 > 0$ an unknown constant. It will be assumed that the rank of $\mathbf{X}$ is p, so the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$, and the MLE of σ^2 is $\hat{\sigma}^2 = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n}$.

The null hypothesis of the initial test will be $H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\gamma}$, where $\mathbf{L}$ is a $d \times p$ matrix of row rank $d \leq p$. Denoting by $\hat{\hat{\boldsymbol{\beta}}}$ the MLE of $\boldsymbol{\beta}$ subject to the constraint of H_0, we have $\hat{\hat{\sigma}}^2 = \frac{(\mathbf{Y}-\mathbf{X}\hat{\hat{\boldsymbol{\beta}}})'(\mathbf{Y}-\mathbf{X}\hat{\hat{\boldsymbol{\beta}}})}{n}$. It is convenient to call the unconstrained model the "full model," and the model constrained by H_0 the "reduced model." Thus we may write $\hat{\sigma}^2 = \frac{SSE_F}{n}$ and $\hat{\hat{\sigma}}^2 = \frac{SSE_R}{n}$. Here "SSE" stands for Sum of Squares Error.

The critical region of the likelihood ratio test is

$$\begin{aligned}
C &= \left\{ \mathbf{y} : \frac{\prod_{i=1}^{n} \frac{1}{\hat{\hat{\sigma}}\sqrt{2\pi}} \exp -\frac{1}{2\hat{\hat{\sigma}}^2}(y_i - \mathbf{x}_i'\hat{\hat{\boldsymbol{\beta}}})^2}{\prod_{i=1}^{n} \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp -\frac{1}{2\hat{\sigma}^2}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2} \leq k_1 \right\} \\
&= \left\{ \mathbf{y} : \frac{(\frac{1}{\hat{\hat{\sigma}}^2 2\pi})^{\frac{n}{2}} \exp -\frac{1}{2\hat{\hat{\sigma}}^2}(\mathbf{y} - \mathbf{X}\hat{\hat{\boldsymbol{\beta}}})'(\mathbf{y} - \mathbf{X}\hat{\hat{\boldsymbol{\beta}}})}{(\frac{1}{\hat{\sigma}^2 2\pi})^{\frac{n}{2}} \exp -\frac{1}{2\hat{\sigma}^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})} \leq k_1 \right\} \\
&= \left\{ \mathbf{y} : \frac{(\hat{\sigma}^2)^{\frac{n}{2}} e^{-\frac{n}{2}}}{(\hat{\hat{\sigma}}^2)^{\frac{n}{2}} e^{-\frac{n}{2}}} \leq k_1 \right\} = \left\{ \mathbf{y} : (\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2})^{\frac{n}{2}} \leq k_1 \right\} \\
&= \left\{ \mathbf{y} : \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} \leq k_2 \right\} = \left\{ \mathbf{y} : \Lambda(\mathbf{y}) \leq k_2 \right\},
\end{aligned}$$

where $\Lambda = \Lambda(\mathbf{y}) = \frac{SSE_F}{SSE_R}$ is the multivariate test statistic Wilks' Lambda for this univariate case. Defining $g(x)$ by $g(x) = \frac{1-x}{x}$, we continue to write the critical region as follows:

$$\begin{aligned}
C &= \{ \mathbf{y} : g(\Lambda(\mathbf{y})) \geq g(k_2) = k_3 \} \\
&= \left\{ \mathbf{y} : \frac{1 - \Lambda}{\Lambda} \geq k_3 \right\} = \left\{ \mathbf{y} : \frac{\frac{SSE_R}{SSE_R} - \frac{SSE_F}{SSE_R}}{\frac{SSE_F}{SSE_R}} \geq k_3 \right\} \\
&= \left\{ \mathbf{y} : \frac{SSE_R - SSE_F}{SSE_F} \geq k_3 \right\} = \left\{ \mathbf{y} : \frac{(SSE_R - SSE_F)/d}{SSE_F/(n-p)} \geq \frac{n-p}{d} k_3 = k_4 \right\} \\
&= \left\{ \mathbf{y} : \frac{SSE_R - SSE_F}{d\, MSE_F} \geq k_4 \right\}.
\end{aligned}$$

Choosing k_4 to be the $1 - \alpha$ quantile of the F distribution with d and $n - p$ degrees of freedom, we have the usual test based on

$$F = \frac{SSE_R - SSE_F}{d\, MSE_F}. \tag{2}$$

One benefit of going through all these details is that we have the relationship between the F statistic and Wilks' lambda, which will come in handy later when we want to perform univariate tests as union-intersection followups to multivariate tests. It is recorded here for future reference.

$$\Lambda = \frac{n-p}{n-p+dF} \Leftrightarrow F = \left(\frac{1-\Lambda}{\Lambda} \right) \left(\frac{n-p}{d} \right) \tag{3}$$

If the initial null hypothesis $H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\gamma}$ is rejected, we will follow up with tests whose null hypotheses are all implied by the initial null hypothesis. Here, attention will be confined to null hypotheses of the form $H_\ell : \mathbf{C}_\ell\boldsymbol{\beta} = \mathbf{h}_\ell$, where $\mathbf{C}_\ell = \mathbf{A}_\ell\mathbf{L}$ and $\mathbf{h}_\ell = \mathbf{A}_\ell\boldsymbol{\gamma}$. The matrix $\mathbf{A}_\ell$ is a $q \times d$ matrix of row rank $q = 1, \ldots, d$, and $\ell \in \mathcal{L}$, an index set corresponding to the set of all such matrices. Thus, H_0 implies H_ℓ in a particularly simple way: $\mathbf{L}\boldsymbol{\beta} = \boldsymbol{\gamma} \Rightarrow \mathbf{A}_\ell\mathbf{L}\boldsymbol{\beta} = \mathbf{A}_\ell\boldsymbol{\gamma}$. Note that each row of $\mathbf{C}_\ell$ is a linear combination of the rows of $\mathbf{L}$, making it easy to formally verify that a follow-up null hypothesis is implied by the initial null hypothesis — if formal verifiation is necessary.

The class of linear follow-ups just described corresponds exactly to the classical Scheffé tests for multiple regression. This class is enormously rich, and sufficient for almost any imaginable application. For example, the typical textbook account of Scheffé procedures is confined to confidence intervals and tests for single linear combinations of treatment means in a one-way design, leaving the practitioner in the dark about what to do if there are covariates. Here, however, we can easily perform tests on *collections* of linear combinations, and do so in the presence of covariates.

Incidentally, there are uncountably many *nonlinear* follow-ups that are also automatically protected at the same joint significance level along with the Scheffé tests, but these appear to be mostly curiosities. For example, $\beta_2 = \beta_3 = \beta_4 = 0$ implies $\log(1 + \beta_2) = (14 - \beta_3)^{\beta_4} - 1$, and we could obtain maximum likelihood estimates of the parameters subject to this odd constraint, but who cares?

Derivation of the likelihood ratio test of H_ℓ proceeds exactly as in the case of the initial test. Again we arrive at test statistic (2), with the understanding that the quantity SSE_R now refers to sum of squares error from a reduced model in which $\boldsymbol{\beta}$ is constrained by H_ℓ rather than H_0. The quantity MSE_F is, of course, the same, since *follow-ups always assume the same model as the initial test*. Following Result 2, the critical value of the follow-up test will also be the same as that of the primary test. To clarify, the union-intersection follow-up test has the form

$$F_s = \frac{SSE_R - SSE_F}{d\,MSE_F} \tag{4}$$

(the s in F_s stands for Scheffé).

Using statistical software such as SAS, it is routine to produce a test of H_ℓ (in some contexts it might help to think of it as a "planned comparison"). But what the software will produce is not (4), but the test statistic $F_\ell = \frac{SSE_R - SSE_F}{q\,MSE_F}$, and a p-value that is appropriate for an initial test. However, it is easy to compute

$$F_s = \frac{q}{d} F_\ell \tag{5}$$

and compare the result to the $1-\alpha$ quantile of an F distribution with d and $n-p$ degrees of freedom.

This illustrates the main advantage of having gone through the derivation of the likelihood ratio test. A careless application of Result 2 might have suggested comparing F_ℓ to the $F(d, n-p)$ critical value, but the correct Scheffé test statistic is $F_s = \frac{q}{d}F_\ell$. It is encapsulated in the following recipie.

Recipie 3 *Suppose the initial test is an F-test from multiple regression, based on d and $n-p$ degrees of freedom. All the follow-up tests will employ the same critical value as the* *initial test. For any potential follow-up test, first verify that its null hypothesis is implied by the null hypothesis of the initial test. Then use statistical software to calculate the F statistic for the followup test in the usual manner; call this quantity F_ℓ. It will appear to be based on q and $n-p$ degrees of freedom, where $q < d$. Compute $F_s = \frac{q}{d}F_\ell$ with a calculator, and declare the Scheffé test significant if F_s is greater than the critical value of F with d and $n-p$ degrees of freedom.*

Most textbook treatments of multiple comparisons, including classics like Miller (1981) and Hochberg and Tamhane (1987) follow Scheffé's original argument, and present confidence intervals for single linear combinations of regression parameters. But by going back to first principles, we have been able to obtain Scheffé tests for *collections* of linear combinations. These tests were part of the same family as the tests for single linear combinations, and are simultaneously protected at the same significance level at no extra expense.Therefore, we are able, for example, to easily treat the test of a main effect in a multi-factor ANOVA as a union-intersection (Scheffé) followup to an initial test for equality of all the treatment means.

Usually, it is obvious when the initial null hypothesis implies a given follow-up null hypothesis. Most of the time, the initial null hypothesis is that some collection of regression coefficients all equal zero, and the follow-up null hypothesis is that members of some subset of the collection are zero, or that some subset of the collection of regression coefficients are all equal to each other. In such cases, no formal demonstration is necessary. In rare cases it may not be so clear, and a formal proof may be desirable. Alternatively, software may be used to check that each row of $\mathbf{C}_\ell$ is linearly dependent upon the rows of $\mathbf{L}$. An example using SAS will be given later.

Penalty $\frac{q}{d}$, projection

3.2 A Textbook One-Factor Example

3.3 A Regression with Quantitative Measured Variables

3.4 An Unbalanced Two-factor Example

3.5 Power and Sample Size

The relation $F_s = \frac{q}{d}F_\ell$ (Equation 5) means that the Scheffé procedure exacts a severe penalty for small q (like the very common $q = 1$), especially if d is large. What this means is that if a null hypothesis H_ℓ is false, a Scheffé test will be less likely to reject it than an ordinary one-at-a-time F-test. That is, the *power* of Scheffé tests is lower.

4 Multivariate Linear Models

4.1 Details

The multivariate linear model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times m$ vector of unknown constants, and the rows of the $n \times m$ random matrix $\boldsymbol{\epsilon}$ are independent multivatiate normals with mean zero and unknown variance-covariance matrix $\boldsymbol{\Sigma}$. The rank of $\mathbf{X}$ is p, the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and the MLE of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta})}{n}$. SAS `proc reg` allows testing null hupotheses of the form $H_0 : \mathbf{L}\boldsymbol{\beta}\mathbf{M} = \mathbf{0}$, so that will be our initial null hypothesis. The matrix of regression coefficients estimated under the constraints of H_0 will be denoted by $\hat{\hat{\boldsymbol{\beta}}}$, and the restricted MLE of $\boldsymbol{\Sigma}$ is $\hat{\hat{\Sigma}} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\hat{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\hat{\beta}})}{n}$.

Tests will be based on Wilks' lambda:

$$\Lambda = \frac{|\hat{\hat{\boldsymbol{\Sigma}}}|}{|\hat{\boldsymbol{\Sigma}}|} = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}, \tag{6}$$

where

$$\mathbf{H} = (\mathbf{L}\widehat{\boldsymbol{\beta}}\mathbf{M})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}\mathbf{M}) \text{ and } \mathbf{E} = \mathbf{M}'(\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}})\mathbf{M}$$

This is a very direct likelihood ratio statistic; in fact it is just the $\frac{n}{2}$ root of the likelihood ratio. This means every union-intersection follow-up test will reject its the null hypothesis if Wilks' lambda is *less* than the critical value for the initial test (yes, we are using the left tail of the distribution).

So we need to determine the critical value of Wilks' lambda for the initial test. Tables are available in some older books, but it would be necessary to interpolate in the tables; it is better to use software. The p-values that SAS gives for multivariate tests are based on F approximations – approximations that are very good, and quite often exact in the case of Wilks' lambda.

Now I will transcribe the formulas from the SAS manual (need the reference). This is complex enough so that I will just use their notation (that of Rao 1973, I believe), employing script-style (calligraphic) letters to minimize incompatibilites and overlaps with the symbols in the rest of this document. Let

- $\mathcal{P}$ be the rank of $(\mathbf{E}+\mathbf{H})$ ($\leq$ number columns of $\mathbf{M}$)
- $\mathcal{Q}$ be the rank of $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$ (the number of rows in $\mathbf{L}$, if $\mathbf{L}$ is of full row rank and $\mathbf{X}$ is of full column rank)
- $\mathcal{V}$ be the degrees of freedom for error
- $\mathcal{R} = \mathcal{V} - \frac{\mathcal{P}-\mathcal{Q}+1}{2}$
- $\mathcal{U} = \frac{\mathcal{P}\mathcal{Q}-2}{4}$
- $\mathcal{T} = \sqrt{\frac{\mathcal{P}^2\mathcal{Q}^2-4}{\mathcal{P}^2+\mathcal{Q}^2-5}}$ if $\mathcal{P}^2 + \mathcal{Q}^2 > 5$, or one otherwise.

Then under the null hypothesis,

$$F = \frac{1-\Lambda^{1/\mathcal{T}}}{\Lambda^{1/\mathcal{T}}}\left(\frac{\mathcal{R}\mathcal{T}-2\mathcal{U}}{\mathcal{P}\mathcal{Q}}\right) \tag{7}$$

has an approximate F distribution with degrees of freedom $\mathcal{P}\mathcal{Q}$ and $\mathcal{R}\mathcal{T}-2\mathcal{U}$. If $\min(\mathcal{P},\mathcal{Q}) \leq 2$ it's exact (Rao 1973).

For the case of a single dependent variable, the matrix $\mathbf{M}$ has just one column, containing a one corresponding to the dependent variable of interest, and all the rest zeros. Thus, $\mathcal{P}$=1, $\mathcal{Q} = d$, $\mathcal{V} = n-p$, $\mathcal{R} = n-p-1+\frac{d}{2}$, $\mathcal{T}$=1, and (7) agrees with (3).

To get an (approximate) critical value for Λ, denote by $F_{1-\alpha}(\mathcal{PQ}, \mathcal{RT} - 2\mathcal{U})$ the $1 - \alpha$ quantile of the F distribution with $\mathcal{PQ}$ and $\mathcal{RT} - 2\mathcal{U}$ degrees of freedom. Setting this equal to F in (7) and solving for Λ, we arrive at a critical value for Wilks' Lambda equal to

$$\Lambda_{crit} = \frac{\mathcal{RT} - 2\mathcal{U}}{\mathcal{RT} - 2\mathcal{U} + \mathcal{PQ}F_{1-\alpha}(\mathcal{PQ}, \mathcal{RT} - 2\mathcal{U})}\,. \tag{8}$$

References

[Gabriel 1969] Gabriel, K. R. (1969). "Simultaneous test procedures — some theory of multiple comparisons." *Ann. Math. Statist.*, 40, 224–250.

[Hochberg and Tamhane 1987] Hochberg, Y., and Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

[Miller 1981] Miller, R. G. Jr. (1981). *Simultaneous statistical inference.* 2nd ed., New York: Springer-Verlag.

[Rao 1973] Rao, C. R. (1973). *Linear statistical inference and its applications*, Second Edition, New York: Wiley.

[Roy 1953] Roy, S. N. (1953). "On a heuristic method of test construction and its use in multivariate analysis." *Ann. Math. Statist.*, 24, 220-238.

[Roy and Bose 1953] Roy, S. N. and Bose, R. C. (1953). "Simultaneous confidence interval estimation." *Ann. Math. Statist.*, 24, 513-536.

[Scheffé 1953] Scheffé, H. (1953). "A method for judging all contrasts in the analysis of variance." *Biometrika*, 40, 87–104.

[Scheffé 1959] Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.