Chapter 8: The Multivariate General Linear Model

Requirements: Sections 3.4, 3.5 - 3.8, 4.3 Chapter 7

8.1 *Introduction*

The main difference between this chapter and the chapters on the General Linear Model; 5, 6 and 7; lies in the fact that here we are going to explicitly consider multiple dependent variables. Multiple dependent variables are to some extent discussed in Chapter 7 in the context of the analysis of variance. In that chapter, however, we made an assumption about the error distribution which allowed us to treat the problem as essentially univariate [see Equation (7.12)]. In this chapter, we will be dealing with multiple dependent variables in the most general way possible, namely the multivariate general linear model. Before we begin, it will be necessary to review some of the fundamentals of hypothesis testing, and then after, to introduce some mathematical details of use in this area.

8.2 *Testing Multiple Hypotheses*

In Chapter 6, we covered two different approaches to testing hypotheses about the coefficients of the linear model. In Equation (6.15) we had

$$\hat{t} = \frac{\mathbf{a}'\boldsymbol{\beta} - \mathrm{c}}{\sqrt{\mathrm{s}^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$$

that allows us to test one degree of freedom questions of the form $\mathbf{a}'\boldsymbol{\beta} = \mathrm{c}$, while in Equation (6.18) we have the test statistic

$$\hat{\mathrm{F}} = \frac{\mathrm{SS_H}/\mathrm{q}}{\mathrm{SS_{Error}}/\mathrm{n}-\mathrm{k}}$$

that allows us to test multiple degree of freedom questions $\mathbf{A}\boldsymbol{\beta} = \mathbf{C}$. In the former case we have n - k degrees of freedom, and in the latter, q and n - k degrees of freedom. In that chapter we made the implicit assumption that these tests had been planned *a priori*, and that they were relatively few in number. In the case of multiple dependent variables, this second assumption becomes far less tenable. We begin by discussing a way to test hypotheses even when there are a large number of them. We then discuss the case where this large number of hypotheses might even be post hoc.

8.3 *The Dunn-Bonferroni Correction*

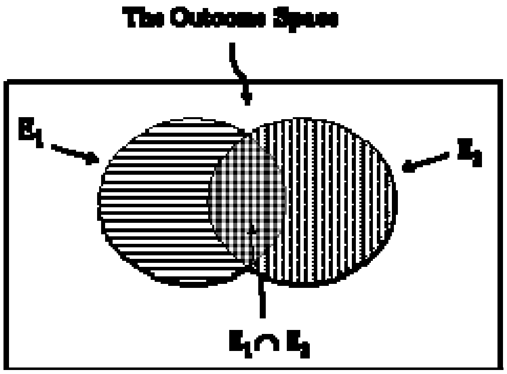
What can we do if we wish to test a large number of hypotheses, say, H_1, H_2, ···, H_r? For any particular hypothesis, we can limit the probability that we reject H_0 when it was indeed true of the population, that is we can limit

$$\Pr(\text{Type I Error on } H_i) = \alpha_i.$$

But what is the probability of at least one Type I error in a sequence of r hypotheses? To delve into this question it will be useful to utilize the notation of Set Theory, where $\cup$ symbolizes union and $\cap$ symbolizes intersection. The probability of at least one Type I error is

$$\alpha^* = \text{Pr(Type I error on } H_1 \cup \text{Type I Error on } H_2 \cup \cdots \cup \text{Type I Error on } H_r\text{).} \qquad (8.1)$$

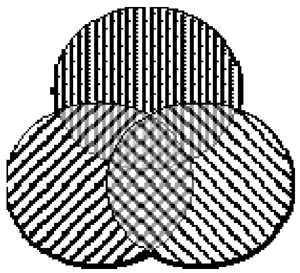
We can think of α^* as the overall α rate, the probability of at least on Type I Error. Define E_i as a Type I error event for H_i. From probability theory, with r = 2 hypotheses, lets say, the situation is illustrated below:



Two parts of the outcome space are shaded, the two parts that correspond to E_1 (a Type I Error on H_1) and E_2 (a similar result on H_2). There is some overlap, namely the part of the space comprising the intersection of E_1 and E_2. It can be shown that

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2).$$

Needless to say, one has to subtract out the $\Pr(E_1 \cap E_2)$ so that it is not counted twice when adding up $\Pr(E_1) + \Pr(E_2)$. For r = 3 hypotheses we have a diagram as below



and we can say

$$\Pr(E_1 \cup E_2 \cup E3) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3) -$$

$$\Pr(E_1 \cap E_2) - \Pr(E_1 \cap E_3) - \Pr(E_2 \cap E_3) + \Pr(E_1 \cap E_2 \cap E_3).$$

Here, we needed to subtract all of the two-way intersections but then we had to add back in the third way intersection which was subtracted once too often. In any case, it is clear that the simple sum of the probabilities, $\sum_i^r \Pr(E_i)$ is an upper bound on the probability of at least one Type I Error since we have not subtracted out any of the intersecting probabilities. We can then safely say that

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_r) \le \Pr(E_1) + \Pr(E_2) + \cdots + \Pr(E_r).$$

Now of course, $\Pr(E_i) = \alpha_i = \alpha$ for all i, so in that case we can state

$$\alpha^* = \Pr(E_1 \cup E_2 \cup \cdots \cup E_r) \le \sum_i^r \Pr(E_i) = \sum_i^r \alpha_i$$

in which case

$$\alpha^* \le r \cdot \alpha.$$

If we select α so that

$$\frac{\alpha^*}{r} \le \alpha$$

we set an upper limit on our overall α. For example, with r = 10 a priori hypotheses, if I want my overall Type I rate to be $\alpha^* = .05$, I would pick $\alpha = .05/10 = .005$ for each hypothesis.

This logic is of course flexible enough to be applicable to any sort of hypotheses whether they be about factor analysis loadings, differences between groups, or tests of betas. A problem with this approach becomes apparent when r gets big. It then becomes very conservative. At that point it is reasonable to use a different logic, a logic that is also applicable to post hoc hypotheses. We now turn to that.

8.4 *Union-Intersection Protection from Post Hoc Hypotheses*

This technique, also known as the Roy-Scheffé approach, is one that protects the marketing researcher from the worst data sniffing case possible, in other words, any post hoc hypothesis. As with the Dunn-Bonferroni test, it is applicable to any sort of hypothesis testing. And as with the Dunn-Bonferroni the overall probability of at least one Type I event is

$$\alpha^* = \Pr(\text{Type I error on } H_1 \cup \text{Type I Error on } H_2 \cup \cdots \cup \text{Type I Error on } H_r)$$

$$= \Pr(E_1 \cup E_2 \cup \cdots \cup E_r).$$

This probability is equivalent to 1 - Pr(No Type I Events). Define the complement of E_i as $\overline{E}_i$, a non-Type I event. We can then re-express the above equation, expressed as a union, as

$$\alpha^* = 1 - \Pr(\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_r). \tag{8.2}$$

which is instead expressed as an intersection. A commonality to all hypothesis testing situations is that $\overline{E}_i$ occurs when the calculated value of the test statistic, $\hat{\theta}_i$, exceeds a critical value, θ_i. Perhaps θ_i is a *t*, and F, or an eigenvector of $E^{-1}H$. In any of these cases,

$$\alpha^* = \Pr(\hat{\theta}_1 < \theta_0 \cap \hat{\theta}_2 < \theta_0 \cap \cdots \cap \hat{\theta}_r < \theta_0)$$
$$= 1 - \Pr(\hat{\theta}_{\max} < \theta_0) \tag{8.3}$$

where $\hat{\theta}_{\max}$ is the largest value of $\hat{\theta}$ that you could ever mine out of your data. Here is an example inspired from ANOVA. Suppose we wanted to test

$$H_0\text{: } \mathbf{c}'\boldsymbol{\mu} = 0$$

where $\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_k]'$ is the vector of population means from a one way univariate ANOVA, in other words the topic of Chapter 7 where the interest is on testing hypotheses about differences among the groups. Here we wish to be protected from

$$\hat{t}^2_{\max} = (\mathbf{c}'\bar{\mathbf{y}})^2 \Big/ \frac{s^2}{n}\mathbf{c}'\mathbf{c}.$$

Picking elements of the vector **c** so as to make this *t* as large as possible leads to the Scheffé (1959) post-hoc correction. More information on post hoc (and a priori) tests among means can be found in Keppel (1973).

8.5 *Details About the Trace Operator and It's Derivative*

The trace operator was introduced in Chapter 1. To briefly review, the trace of a square matrix, say **A**, is defined as Tr(**A**) = $\sum a_{ii}$, i.e. the sum of the diagonal elements. Some properties of Tr(·) follow. Assuming that **A** and **B** are square matrices we can say

Transpose $$\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}') \tag{8.4}$$

Additivity $$\text{Tr}(\mathbf{A}+\mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B}) \tag{8.5}$$

Then, for **A** m · n and **B** n · m we have

Commutative $$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}) \tag{8.6}$$

which further implies, for **C** m · m

Triple Product $$\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{CAB}) \tag{8.7}$$

In Chapter 3, we discuss the derivative of a scalar function of a vector, and a vector function of a vector. Here we want to look at the derivative of a scalar function of a matrix, that function being, of course, the trace of that matrix. To start off, note that by definition

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \frac{\partial f(\mathbf{X})}{\partial x_{12}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \frac{\partial f(\mathbf{X})}{\partial x_{21}} & \frac{\partial f(\mathbf{X})}{\partial x_{22}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \frac{\partial f(\mathbf{X})}{\partial x_{m2}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix}, \tag{8.8}$$

where f(**X**) is a scalar function of the matrix **X**. Now we can begin to talk about the Tr(·) function which is a scalar function of a square matrix. For **A** m · m we have

$$\frac{\partial \mathrm{Tr}(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{I} \tag{8.9}$$

For **A** m · n and **B** n · m we can say

$$\frac{\partial \mathrm{Tr}(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B}' \tag{8.10}$$

which also implies, from Equation (3.19)

$$\frac{\partial \mathrm{tr}(\mathbf{AB})}{\partial \mathbf{A}'} = \mathbf{B}\ . \tag{8.11}$$

Finally, assuming we have **A** m · m and **B** m · m,

$$\frac{\partial \mathrm{tr}(\mathbf{A}'\mathbf{BA})}{\partial \mathbf{A}} = (\mathbf{B} + \mathbf{B}')\mathbf{A}\ . \tag{8.12}$$

8.6 *The Kronecker Product*

We now review the definition of the *Kronecker product,* sometimes called the *Direct product*, with operator ⊗. By definition,

$$_{mp}\mathbf{C}_{nq} = {}_{m}\mathbf{A}_{n} \otimes {}_{p}\mathbf{B}_{q} = \{a_{ij}\mathbf{B}\}\ . \tag{8.13}$$

For example,

$$\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}\mathbf{B} \\ \hdashline a_{21}\mathbf{B} \end{bmatrix}.$$

Here are some properties of the Kronecker product. We can say that

Transpose $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'.$ (8.14)

For **A** m · n, **B** n · p and **C** p · q, it is the case that

Associative $\mathbf{AB} \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{BC}.$ (8.15)

For **A** and **B** m · n and **C** p · q,

Distributive $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}.$ (8.16)

For **A** m · n, **B** n · p and **C** q · r and **D** r · s,

$$(\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D}) = \mathbf{AB} \otimes \mathbf{CD}. \tag{8.17}$$

8.7 *The Vec Operator*

For a matrix **A**, lets say m by n, we define

$$\text{vec}(\mathbf{A}) = \text{vec}\begin{bmatrix} \mathbf{a}'_{1\cdot} \\ \mathbf{a}'_{2\cdot} \\ \cdots \\ \mathbf{a}'_{m\cdot} \end{bmatrix} = \begin{bmatrix} \mathbf{a}'_{1\cdot} & \mathbf{a}'_{2\cdot} & \cdots & \mathbf{a}'_{m\cdot} \end{bmatrix}. \tag{8.18}$$

While other definitions of Vec(·) are possible, this one, that does so one row at a time, will prove useful to us when we start to look at the multivariate GLM. In particular, the following theorem will be quite useful. For **A** m · n, **B** n · p and **C** p · q,

$$\text{Vec}(\mathbf{ABC}) = (\mathbf{A} \otimes \mathbf{C}')\ \text{Vec}(\mathbf{B}). \tag{8.19}$$

8.8 *Eigenstructure for Asymmetric Matrices*

Suppose we needed to maximize **x′Hx** subject to **x′Ex** = 1. Then

$$\text{f}(\mathbf{x}) = \mathbf{x}'\mathbf{H}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{E}\mathbf{x} - 1) \tag{8.20}$$

and to minimize we set

$$\frac{\partial \text{f}(\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{Hx} - 2\lambda\mathbf{Ex} = 0. \tag{8.21}$$

Rearranging a bit we have

$$(\mathbf{H} - \lambda\mathbf{E})\mathbf{x} = (\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{x} = 0. \tag{8.22}$$

You will note the eigenstructure discussed in Chapter 3 is a special case of the current discussion with **E** = **I**. In our case, as $\mathbf{E}^{-1}\mathbf{H}$ is asymmetric, the eigenvectors are not orthonormal [defined in Equation (3.33)]. Instead we have the relation

$$\mathbf{E}^{-1}\mathbf{H} = \mathbf{XLX}^{-1}. \tag{8.23}$$

For symmetric matrices we have had $\mathbf{X}^{-1} = \mathbf{X}'$, but not in this case.

8.9 *Eigenstructure for Rectangular Matrices*

For completeness, we note that any m · n matrix **A** or rank r can be decomposed into the triple product

$$\mathbf{A} = \mathbf{X}\mathbf{L}^{1/2}\mathbf{V}' \qquad (8.24)$$

where **X** is m · r, $\mathbf{L}^{1/2}$ is r · r and **V** is n · r. This is called *singular value decomposition*. The matrix **X** contains the *left eigenvectors* of **A** while **V** contains the *right eigenvectors* of **A**. Further, **V′V** = **I** and **X′X** = **I**. There are important relationships between the eigenvalues of a rectangular matrix and a cross product matrix. We have

$$\mathbf{A}'\mathbf{A} = (\mathbf{X}\mathbf{L}^{½}\mathbf{V}')\,(\mathbf{V}\mathbf{L}^{½}\mathbf{X}') \;=\; \mathbf{X}\mathbf{L}\mathbf{X}' \qquad (8.25)$$

and

$$\mathbf{A}\mathbf{A}' = (\mathbf{V}\mathbf{L}^{½}\mathbf{U}')\,(\mathbf{U}\mathbf{L}^{½}\mathbf{V}') \;=\; \mathbf{V}\mathbf{L}\mathbf{V}' \qquad (8.26)$$

If **A** is already symmetric then **A′A** = **AA′** so **X** = **V**.

8.10 *The Multivariate General Linear Model*

The multivariate general linear model is a straightforward generalization of the univariate case in Equation (5.3). Instead of having one dependent variable in one column of the vector **y**, we have a set of p dependent variables in the several columns of the matrix **Y**. The model is therefore

$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1p} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_{np} & \hat{y}_{np} & \cdots & \hat{y}_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k*} \\ 1 & x_{21} & \cdots & x_{2k*} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n2} & \cdots & x_{nk*} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{k*1} & \beta_{k*2} & \cdots & \beta_{k*p} \end{bmatrix} \qquad (8.27)$$

which, as you can see, implies that the number of columns of the **B** matrix match the number of columns of the **Y** matrix. Perhaps this concept is better represented using the dot subscript reduction operator (Section 1.1), which allows us to present the model as

$$[\hat{\mathbf{y}}_{\cdot 1} \quad \hat{\mathbf{y}}_{\cdot 2} \quad \cdots \quad \hat{\mathbf{y}}_{\cdot p}] = \mathbf{X}\,[\boldsymbol{\beta}_{\cdot 1} \quad \boldsymbol{\beta}_{\cdot 2} \quad \cdots \quad \boldsymbol{\beta}_{\cdot p}] \qquad (8.28)$$

with each column of **Y** entering into a regression equation with the corresponding column of **B** serving as the coefficient vector. We can express the model most succinctly by using

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}. \qquad (8.29)$$

Next we define the n · p error of prediction matrix as **ε**, i. e.

$$\boldsymbol{\varepsilon} = \hat{\mathbf{Y}} - \mathbf{Y}$$

so that

$$\mathbf{Y} = \mathbf{XB} + \varepsilon. \tag{8.30}$$

8.11 *A Least Squares Estimator for the MGLM*

How do we come up with estimators for the unknowns in the **B** matrix? When Y the error **e** was only a vector, as in Chapter 5, we could pick our objective function as **e′e**. The matrix **ε′ε** on the other hand, is not a scalar but a p · p sum of squares and cross products matrix. In this case what we do is to minimize the trace of ε′ε as we will now see. Our objective function is

$$f = \mathrm{Tr}[\varepsilon'\varepsilon] \tag{8.31}$$

which, according to Equation (8.30), can be expanded to

$$f = \mathrm{Tr}[(\mathbf{Y} - \mathbf{XB})' (\mathbf{Y} - \mathbf{XB})]. \tag{8.32}$$

Factoring the product leads to four components as below;

$$f = \mathrm{Tr}[\mathbf{Y'Y} - \mathbf{Y'XB} - \mathbf{B'X'Y} + \mathbf{B'X'XB}].$$

But since Equation (8.5) notes that the trace of a sum is equivalent to the sum of the traces, we can now say

$$f = \mathrm{Tr}(\mathbf{Y'Y}) - \mathrm{Tr}(\mathbf{Y'XB}) - \mathrm{Tr}(\mathbf{B'X'Y}) + \mathrm{Tr}(\mathbf{B'X'XB}).$$

More simplification is possible. From Equation (8.4) we note that Tr(**B′X′Y**) = Tr(**Y′XB**) and from Equation (8.7) we note that Tr(**Y′XB**) is equivalent to Tr(**BY′X**). We can now rewrite f as

$$f = \mathrm{Tr}(\mathbf{Y'Y}) - 2\mathrm{Tr}(\mathbf{BY'X}) + \mathrm{Tr}(\mathbf{B'X'XB}).$$

In order to make f as small as possible, it is necessary to find the ∂f/∂**B**. Using Equations (8.10) as well as (8.12), we have

$$\frac{\partial f}{\partial \mathbf{B}} = -2\mathbf{X'Y} + [\mathbf{X'X} + (\mathbf{X'X})']\mathbf{B}.$$

But since **X′X** is symmetric, we can simplify a bit more and have

$$\frac{\partial f}{\partial \mathbf{B}} = -2\mathbf{X'Y} + 2\mathbf{X'XB}. \tag{8.33}$$

After setting Equation (8.33) equal to zero, this now leads us to the multivariate analog of the normal equations [Equation (5.7)] as below:

$$\mathbf{X'XB} = \mathbf{X'Y} \tag{8.34}$$

so that

$$\hat{\mathbf{B}} = (\mathbf{X'X})\mathbf{X'Y} \tag{8.35}$$

Each column of $\hat{\mathbf{B}}$ has the same formula as the univariate model, i. e.

$$\hat{\boldsymbol{\beta}}_{\cdot j} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{\cdot j}\,.$$

8.12 *Properties of the Error Matrix* $\boldsymbol{\varepsilon}$

In order to talk about the distribution of the error matrix ε, we will have to rearrange it somewhat using the Vec(·) function of Section 8.7. We will assume, in a multivariate analog to the Gauss Markov Assumption of Chapter 5, that the distribution of the n by p matrix ε is

$$\text{Vec}(\boldsymbol{\varepsilon}) \sim N({}_{np}\mathbf{0}_1,\ {}_n\mathbf{I}_n \otimes {}_p\boldsymbol{\Sigma}_p)\,. \tag{8.36}$$

The Vec operator has unpacked the ε matrix, one row at a time, in other words, one consumer's data at a time. Since there are n consumers with p measurements each, the mean vector of Vec(ε) is np by 1. The covariance matrix for Vec(ε), since the latter has np elements, must be np by np. This covariance matrix has a particular structure that logically, and visually, is reminiscent of the structure we assume in the univariate case presented in Equation (5.16), that of $\sigma^2\mathbf{I} = \mathbf{I} \cdot \sigma^2$. Here, instead we have the partitioned matrix

$$\mathbf{I} \otimes \boldsymbol{\Sigma} = \left[\begin{array}{c:c:c:c} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \hdashline \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \hdashline \cdots & \cdots & \cdots & \cdots \\ \hdashline \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{array}\right] \tag{8.37}$$

with each Σ and each null matrix **0** being p · p. The Σ in the ith diagonal partition represents the (homogeneous) variance matrix for observation i. The **0** in the i, jth position implies that rows i and j of ε, corresponding to subjects i and j, are independent.

8.13 *Properties of the* ***B*** *Matrix*

It is now timely to contemplate the expectation and variance of our estimator of Equation (8.35). Before proceeding, if you wish you can review some of the rules of expectations and variance presented in Section 4.1. The expectation will be straightforward, as

$$E(\hat{\mathbf{B}}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]$$

which for fixed **X** and Equation (4.5) leads to

$$E(\hat{\mathbf{B}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{B}.$$

In order to derive the $V(\hat{\mathbf{B}})$, we will need Theorem (4.9) as well as the more recent Theorem (8.19). OK, let us proceed by noting that

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{I}.$$

Now with $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ playing the role of "**A**", **Y** playing the role of "**B**", and the p by p identity matrix **I** playing the role of "**C**", we apply Theorem (8.19) to show that

$$\text{Vec}(\hat{\mathbf{B}}) = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \mathbf{I}]\text{Vec}(\mathbf{Y})$$

Now we just need to recall that Var[Vec(**Y**)] = $\mathbf{I} \otimes \mathbf{\Sigma}$ and to apply Theorem (4.9) and take it to the bank:

$$\text{Var}[\text{Vec}(\hat{\mathbf{B}})] = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \mathbf{I}](\mathbf{I} \otimes \mathbf{\Sigma})[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{I}]. \tag{8.38}$$

Note that in the above we have taken advantage of Equation (8.14) to express

$$[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \mathbf{I}]' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{I}.$$

Now applying Equation (8.17) two times to Equation (8.38) we can express it as

$$\text{Var}[\text{Vec}(\hat{\mathbf{B}})] = (\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{\Sigma}. \tag{8.39}$$

8.14 *The Multivariate General Linear Hypothesis*

In Chapter 6 we looked at q degree of freedom hypotheses of the form

$$H_0\text{: } \mathbf{A\beta} - \mathbf{c} = \mathbf{0},$$

where the matrix **A** had q rows and where **0** is a q by 1 column of zeroes. In this chapter, since the **B** matrix has multiple columns of possible interest, as compared to **β** which is a column vector, we allow ourselves the possibility to test linear hypotheses about these several columns of **B**. The general form of the hypothesis is then

$$H_0\text{: } \mathbf{ABM} - \mathbf{C} = \mathbf{0}. \tag{8.40}$$

The q rows of **A** test hypotheses concerning the k independent variables. **A** is therefore q · k with q ≤ k. The ℓ columns of **M** test hypotheses about the p dependent variables. **M** is necessarily p · ℓ with $\ell \leq$ p. Next, in Section 8.15 we will look at some examples of **A** and **M**.

8.15 *Some Examples of MGLM Hypotheses*

In our first example, we have k = 3 with $\mathbf{x}_{\cdot 0}$ being the usual column of 1's, $\mathbf{x}_{\cdot 1}$ being income, and then $\mathbf{x}_{\cdot 2}$ being education. On the dependent variable side, we have p = 2 with $\mathbf{y}_{\cdot 1}$ a measure of attitude towards a particular brand and $\mathbf{y}_{\cdot 2}$ being a likelihood of purchase measure. Imagine for a moment that we want to find out if education and income, taken jointly, impact the two dependent variables. Our hypothesis matrices would then take the form as shown below,

$$\mathbf{ABM} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In the second example, k = 1 and $\mathbf{x}_{\cdot 0}$ is the usual vector of n 1's. However, p = 4 where $\mathbf{y}_{\cdot 1}$ through $\mathbf{y}_{\cdot 4}$ are evaluations of four product concepts on a 10 point scale. In this second example, the question of interest is, "Do the product evaluations differ?" In this case, we will use the multivariate approach to repeated measures. The current approach is in contrast to the univariate approach covered in Section 7.7. Here we have

$$\mathrm{a}\mathbf{BM} = 1 \cdot [\beta_{01} \quad \beta_{02} \quad \beta_{03} \quad \beta_{04}] \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Since there are no real independent variables, the matrix **B** is actually a row vector with only the intercepts present. In an intercept only model (see Section 5.9), the β_0 values are simply the means of the dependent variables. The **M** hypothesis matrix transforms the four variable means into three mean-differences. Thus, the hypothesis is of three degress of freedom which test for equality among the levels of the four original dependent variables.

Our example number 3 includes k = 4 with an intercept term plus three attitude variables. For dependent variables, we have p = 3 behavioral measures. Our hypothesis will be an omnibus question designed to ask whether attitude influences behavior:

$$\mathbf{ABM} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \mathbf{I}.$$

Finally, in our fourth example, we have experimental data in which we had a 2 × 2 ANOVA with four groups of consumers. Half the groups saw a high price, and half a low price. Half the groups saw the presence of advertising with half seeing no advertising. There is also the potential interaction of these two factors. Two measures were $\mathbf{y}_{\cdot 1}$; an affective response and $\mathbf{y}_{\cdot 2}$; a cognitive response. The hypothesis concerns the one degree of freedom interaction between price and advertising. Does such an interaction occur for affect and cognition?

$$\mathbf{ABM} = [0 \quad 0 \quad 0 \quad 1] \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

8.16 *Hypothesis and Error Sums of Squares and Cross-Products*

In the univariate linear model, we calculate the hypothesis sum of squares, which is a scalar that corresponds to the single dependent variable. The following equation produces the sum of squares and cross products matrix for the hypothesis embodied in Equation (8.40). As such, it is the multivariate analog to the univariate version presented in Equation (6.17):

$$\mathbf{H} = (\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C})\,. \tag{8.41}$$

The result is ℓ by ℓ with ℓ being the number of columns of **M** and **C**, or in other words, the number of transformed dependent variables in the hypothesis in Equation (8.40). The error sums of squares and cross-products for the hypothesis, in contrast to the single sum of squares for the univariate version in Equation (5.22), is also an $\ell \cdot \ell$ matrix:

$$\mathbf{E} = \mathbf{M}'\,[\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{Y}]\,\mathbf{M}\,. \tag{8.42}$$

Again in the univariate case, in Equation (6.18) we formed an F-ratio using the sum of squares for the hypothesis, and the sum of squares for the error. Modifying the form of Equation (6.18) somewhat, we can express the calculated F as

$$\hat{F} = \frac{SS_h / q}{SS_{Error} / n - k} = \frac{h / q}{e / n - k} = e^{-1}h \cdot \frac{n - k}{q}.$$

In the multivariate case we will do something similar, but the degrees of freedom are absorbed into the multivariate tables. But more importantly, since $\mathbf{E}^{-1}\mathbf{H}$ is an $\ell \cdot \ell$ matrix, we must decide how to summarize all of those numbers in a way that allows us to make an all-or-nothing decision about the hypothesis in Equation (8.40).

Eigenstructure affords an optimal method for summarizing a matrix, and in Section 8.8 we studied the eigenstructure of asymmetric matrices like $\mathbf{E}^{-1}\mathbf{H}$. We are now ready to test our multivariate linear hypothesis.

8.17 *Statistics for Testing the Multivariate General Linear Hypothesis*

If we define s as the rank of $\mathbf{E}^{-1}\mathbf{H}$, we then have the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_s$ of the system

$$(\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{x} = 0. \tag{8.43}$$

In general, s = Min(q, ℓ), that is, whichever is smaller, the number of rows of **A** or the number of columns of **M**. The eigenstructure of $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ will be of interest also:

$$[\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1} - \theta\mathbf{I}]\mathbf{x} = 0 \tag{8.44}$$

with

$$\theta_i = \frac{\lambda_i}{1 + \lambda_i} \tag{8.45}$$

so that

$$\lambda_i = \frac{\theta_i}{1-\theta_i}. \tag{8.46}$$

In a logical sense, the λ_i are analogous to F ratios, being the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, while the θ_i are more analogous to squared multiple correlations, being the eigenvalues of $\mathbf{H}(\mathbf{H} + \mathbf{E}^{-1})$. Now there are four different ways to test the multivariate hypothesis, proposed by four different statisticians. In addition, there is an F approximation that is somewhat commonly used as well. The four are:

Hotelling-Lawley Trace

$$\mathrm{Tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_i^s \lambda_i \tag{8.47}$$

Roy's Largest Root

$$\theta_1 = \frac{\lambda_1}{1+\lambda_1} \tag{8.48}$$

Pillai's Trace

$$\mathrm{Tr}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_i^s \theta_i = \sum_i^s \frac{\lambda_i}{1+\lambda_i} \tag{8.49}$$

Wilk's Lambda

$$\Lambda = \frac{|\,\mathbf{H}\,|}{|\,\mathbf{H}+\mathbf{E}\,|} = \prod_i^s \frac{1}{1+\lambda_i} \tag{8.50}$$

An especially good set of tables for these statistics can be found in Timm (1975).

The F approximation is based on Wilk's determinantal criterion in Equation (8.50). That formula is

$$F' = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \cdot \frac{rt-2u}{\ell q} \tag{8.51}$$

where, as before, q is the number of rows or the rank of **A**, ℓ is the number of columns or the rank of **M**, but there are some other parameters. The values

$$u = \frac{\ell q - 2}{4},$$

$$r = n - k - \frac{\ell - q + 1}{2},$$

$$t = \begin{cases} \dfrac{\ell^2 q^2 - 4}{\ell^{2+} q^2 - 5} & \text{if } \ell^2 + q^2 - 5 > 0 \\[2ex] 1 & \text{if } \ell^2 + q^2 - 5 \le 0 \end{cases}$$

and n is the sample size while k is the number of columns of **X**. The degress of freedom for F′ are $\ell \cdot q$ in the numerator and rt - 2u in the denominator. The approximation is exact if s = Min(ℓ, q) ≤

2, which is to say that the rank of $\mathbf{E}^{-1}\mathbf{H}$ is 2 or less. You will note the eigenstructure discussed in Chapter 3 is a special case of the following discussion with $\mathbf{E} = \mathbf{I}$.

Earlier, in Section 8.4, we spoke of correcting a statistical test for having a large number of tests and also for post hoc data snooping. If we consider the hypothesis

$$H_0\text{: } \mathbf{a}'\mathbf{B}\mathbf{m} = 0$$

where we try to pick the elements in the vectors **a** and **m** to make the significance test as large as possible, then $\hat{\theta}_{\max}$, from Equation (8.3) is Roy's largest root. Unlike the Dunn-Bonferroni approach, the Union-Intersection approach controls for a high number of tests and also takes into account the correlations between the dependent variables. Another example would be where we try to maximize the correlation between a linear combination of x variables and a linear combination of the y variables. This is called canonical correlation.

8.18 *Canonical Correlation*

In the multivariate general linear model, since there are p elements to the **y** vector and the k variables in the **x** vector, we face an embarrassment of riches in trying to summarize the relationship between the two sets of variables. Shown below, we see the partitioned matrix of all the variables, partitioned into y and x sets:

$$R = \left[\begin{array}{c|c} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \hline \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{array}\right]$$

The p · k matrix $\mathbf{R}_{yx}$ certainly has information in it about the relationship between the two sets of variables, containing as it does, the correlations between the sets. But in order to summarize the relationship between the two sets, we want a scalar. One obvious approach is to create new two new scores, one from the x set and one from the y set such that the correlation between the two scores is as high as possible. In essence, the problem is to pick the p elements of $\mathbf{c}'$ in

$$u = \mathbf{c}'\mathbf{z}_y \tag{8.52}$$

and the k elements of $\mathbf{d}'$ in

$$v = \mathbf{d}'\mathbf{z}_x \tag{8.53}$$

such that

$$\rho^2 = \frac{(\mathbf{c}\mathbf{R}_{yx}\mathbf{d})^2}{\mathbf{c}'\mathbf{R}_{yy}\mathbf{c} \cdot \mathbf{d}'\mathbf{R}_{xx}\mathbf{d}} \tag{8.54}$$

is maximized. This leads to two different eigenvector problems,

$$[\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} - \rho^2 I]\mathbf{c} = 0 \tag{8.55}$$

and

$$[\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} - \rho^2 \mathrm{I}]\mathbf{d} = 0. \tag{8.56}$$

We can pick the smaller problem to solve and then deduce the other eigenvector using either

$$\mathbf{d} = \frac{1}{\rho^2}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{c} \tag{8.57}$$

or

$$\mathbf{c} = \frac{1}{\rho^2}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{d}. \tag{8.58}$$

The canonical correlation can be thought of as a linear hypothesis of the form of Equation (8.40) with

$${}_{k}\mathbf{A}_{k*} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = [\,{}_{k}\mathbf{0}_{1} \mid {}_{k*}\mathbf{I}_{k*}\,]$$

and $\mathbf{M} = {}_{p}\mathbf{I}_{p}$. The number of canonical correlations and eigenvector combinations depends on s, which in this case is simply whichever is smaller, k or p. The first canonical correlation squared corresponds to Roy's Largest Root in Equation (8.48), which can be used to test the hypothesis that the canonical correlation is zero. One can also use Pillai's Trace [Equation (8.49)] to test whether all of the canonical correlations are zero, i. e.

$$H_0\colon \rho_1^2 = \rho_2^2 = \cdots = \rho_s^2 = 0.$$

Placing each of the eigenvectors, $\mathbf{a}_{\cdot 1}$, $\mathbf{a}_{\cdot 2}$, $\cdots$, $\mathbf{a}_{\cdot s}$ into columns of the matrix **A** (not the hypothesis matrix), we have rows of **A** that correspond to y variables and columns of **A** that correspond to different canonical variables from Equation (8.52). We can standardize the elements of **A** using

$$\mathbf{C}_s = \mathbf{C}(\mathbf{C}'\mathbf{R}_{yy}\mathbf{C})^{-1/2}$$

and for the x set we have

$$\mathbf{D}_s = \mathbf{D}(\mathbf{D}'\mathbf{R}_{xx}\mathbf{D})^{-1/2}.$$

It is also instructive to look at the correlations between each of the canonical variables in Equation (8.53) and the variables of the x set, and the canonical variables in Equation (8.52) and the variables of the y set. We have for each combination

$$\mathrm{Cov}(\mathbf{u}, \mathbf{z}_y) = \mathbf{C}_s'\mathbf{R}_{yy},$$

$$\mathrm{Cov}(\mathbf{v}, \mathbf{z}_y) = \mathbf{D}_s'\mathbf{R}_{xx},$$

$$\mathrm{Cov}(\mathbf{u}, \mathbf{z}_x) = \mathbf{C}_s'\mathbf{R}_{yx},$$

$$\text{Cov}(\mathbf{v}, \mathbf{z}_y) = \mathbf{D}_s'\mathbf{R}_{xy.}$$

8.19 *MANOVA*

We will begin with an example with a purely between subjects design, and two different dependent variables. Imagine that we have four groups of subjects, each group having seen a different advertisement. Thus, k = 4 with $\mathbf{x}_{\cdot 0}$ being the usual vector of constants and $\mathbf{x}_{\cdot 1}$, $\mathbf{x}_{\cdot 2}$ and $\mathbf{x}_{\cdot 3}$ coding for group membership. To keep things simple, lets say that $\mathbf{y}_{\cdot 1}$ contains the respondent's answer to the question, "How much do you like the product?" while $\mathbf{y}_{\cdot 2}$ has data on "Intention to buy." In summary, **Y** is n · 2, **X** is n · 4 and **B** is 4 · 2 with

$$\hat{\mathbf{Y}} = \mathbf{XB}.$$

It would be natural to test the hypothesis of no group differences for the two dependent variables. This hypothesis is much the same as canonical correlation, its just that the emphasis is slightly different. We calculate the hypothesis sum of squares and cross product matrix

$$\mathbf{H} = (\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C}),$$

with

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

and $\mathbf{M} = {}_2\mathbf{I}_2$ and the error sum of squares and cross products matrix,

$$\mathbf{E} = \mathbf{M}' \, [\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \, \mathbf{Y}] \, \mathbf{M},$$

invert this latter matrix in order to find the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, calculate the four criteria and the F approximation, and see to the fate of H_0. In addition, the eigenvectors for the y set,

$$\mathbf{v} = \mathbf{d}'\mathbf{y},$$

can tell us the optimal combination of y's for detecting group differences. Similarly, the eigenvectors for the x set reveal the best possible contrast among the group means.

8.20 *MANOVA and Repeated Measures*

To start off this section, we will pick an example with no grouping variables, just one group of consumers who rate a product using the same scale under p = 3 different scenarios. The multivariate model is then

$$\hat{\mathbf{Y}} = \mathbf{XB}$$

$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1p} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_{n1} & \hat{y}_{n2} & \cdots & \hat{y}_{np} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} [\beta_{01} \quad \beta_{02} \quad \beta_{03}].$$

To test the hypothesis that all scenarios lead to equal ratings, we use

$$H_0\text{: } \mathbf{ABM} = \mathbf{C}$$

$$H_0\text{: } 1 \cdot [\beta_{01} \quad \beta_{02} \quad \beta_{03}] \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = [0 \quad 0].$$

We can conceptualize the process here a little bit differently. For each subject, you could transform the scores prior to the analysis by applying the **M** hypothesis matrix directly to the **Y** matrix. In that case, you could simply test whether the β_0 values of the transformed measures were zero. So if we define

$$\tilde{\mathbf{Y}} = \mathbf{XB}$$

where **M** is exactly as before, and now we test to see if

$$H_0\text{: } 1 \cdot [\tilde{\beta}_{01} \quad \tilde{\beta}_{02}] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = [0 \quad 0]$$

where the parameters $\tilde{\beta}_{01}$ and $\tilde{\beta}_{02}$ would be estimated from $\tilde{\mathbf{Y}}$ instead of **Y**. Both approaches are equivalent because the hypotheses

$$H_0\text{: } \mu_{\tilde{y}_1} = \mu_{\tilde{y}_2} = 0 \tag{8.59}$$

and

$$H_0\text{: } \mu_{y_1} = \mu_{y_2} = \mu_{y_3} \tag{8.60}$$

are equivalent. Using the transformed dependent variable matrix $\tilde{\mathbf{Y}}$ and testing the Hypothesis of Equation (8.59) is an example of Hotelling's T^2 (pronounced Tao Squared), which is the multivariate analog of the household variety *t*-statistic. The T^2 is used to test hypotheses of the form

$$H_0\text{: } \boldsymbol{\mu}_y = \mathbf{c}$$

with μ_y being the vector of population means for the dependent variables. Hotelling's T^2can also be used to test multivariable mean differences across two groups, just as the *t* does where there is but one dependent variable.

Now we put together an example where there are different groups of subjects as well as repeated measurements. As before, we assume that all subjects rate a product under p = 3 different scenarios. But now there are actually four different treatment groups, each group having seen a different advertisement for the product. In that case, k = 4 so that the **B** matrix is 4 by 3. Each column of **B** corresponds to one of the three rating scenarios. The first row of **B** contains the intercept terms, while the next three rows pertain to group differences.

Is there an impact of advertisement? In the univariate approach, we add up the three measures to create for each subject i, $\tilde{y} = y_1 + y_2 + y_3$. We test the hypothesis using

$$H_0\text{: } \mathbf{A}\tilde{\boldsymbol{\beta}} = \mathbf{c}$$

$$H_0\text{: } \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

which is covered in Chapter 7. In the multivariate approach covered in this chapter, we do not transform the dependent variables, we leave them as they are. We have

$$H_0\text{: } \mathbf{ABM} = \mathbf{C}$$

$$H_0\text{: } \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \mathbf{I} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This approach confounds the main effect of group with the simple main effect of advertisement on $\mathbf{y}_{\cdot 1}$, on $\mathbf{y}_{\cdot 2}$ and on $\mathbf{y}_{\cdot 3}$. In other words, from column 1 of **Y** we look to see what effect there is of group membership, we do the same thing with columns 2 and 3. But this claims some of the variance that would ordinarily be considered part of the advertisement × scenario interaction. The main effect of advertisement would generally look only at a summary of the group differences holding the scenario constant.

Is there an effect of scenario? Here we start with the univariate approach. If we define

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

and assume that

$$\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}$$

as we did in Equation (7.12), we can utilize the univariate approach to repeated measures and use the F-test discussed in Section 7.7 with an error term of subjects × scenario interaction. In the univariate approach all scores are placed in a single column vector. In contrast, in the multivariate case each scenario constitutes a different column of **Y** and we test

$$H_0\text{:}\ \ \mathbf{ABM} = \mathbf{C}$$

$$H_0\text{:}\ \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\beta_{01} & \beta_{02} & \beta_{03}\\ \beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & \beta_{23}\\ \beta_{31} & \beta_{32} & \beta_{33}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 0\\ 0 & -1\end{bmatrix} = \begin{bmatrix}0 & 0\end{bmatrix}.$$

There also exists an approach in between the univariate and multivariate methods. One could Test

$$H_0\text{:}\ \ \mathbf{M'\Sigma M} = \sigma^2\mathbf{I}$$

and pick the univariate approach if you fail to reject and the multivariate approach if you reject. Another approach was proposed by Greenhouse and Geisser (1959) who suggested that we could correct the univariate F to the degree that

$$\mathbf{M'SM} \neq \hat{\sigma}^2\mathbf{I}. \qquad (8.61)$$

Here in Equation (8.61) we have replaced **Σ** with it's estimator, **S**.

If we wish to test the advertisement × scenario interaction according to the univariate approach, we would need to assume that $\mathbf{M'\Sigma M} = \sigma^2\mathbf{I}$, place all scores in the vector **y**, and use the interaction of subjects × scenario as the error term.

In order to test the advertisement × scenario interaction according to the multivariate model, we can combine the **A** matrix from the advertisement main effect and the **M** matrix from the scenario main effect. In that case we have

$$H_0\text{:}\ \ \mathbf{ABM} = \mathbf{C}$$

$$H_0\text{:}\ \begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\beta_{01} & \beta_{02} & \beta_{03}\\ \beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & \beta_{23}\\ \beta_{31} & \beta_{32} & \beta_{33}\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 0\\ 0 & -1\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & 0\\ 0 & 0\end{bmatrix}.$$

8.21 *Classification*

To motivate this section, which will discuss the technique known as the discriminant function, we begin the discussion with a little two group example. Imagine we are trying to decide who to include in direct mailing. Our goal is to classify our customers into two groups based on whether they will, or will not, respond to the mailout. From a sample of our customer base, we have collected some data which we will get to in just a minute. For now, we note that the cost, or disutility, of misclassifying someone in group i, mistakenly placing them in group j is c_{ij}. Given our two groups, we might then tabulate the cost matrix as

		Classification Decision	
		Group 1	Group 2
Reality	Group 1	0	c_{12}
	Group 2	c_{21}	0

For each individual we have a p element row vector from the matrix $\mathbf{Y}$, $\mathbf{y}'_{i\cdot}$, containing numeric variables. The probability density for the individuals in group j is $f_j(\mathbf{y}_{i\cdot})$, while π_j is the relative size of group j, also called the *prior probability*. The *conditional probability* an individual with vector $\mathbf{y}_{i\cdot}$ comes from group j is

$$\Pr(j \mid \mathbf{y}_{i\cdot}) = \frac{\pi_j f_j(\mathbf{y}_{i\cdot})}{\sum_m \pi_m f_m(\mathbf{y}_{i\cdot})}$$

We want to minimize our expected cost which in the two group case is given by

$$\Pr(1 \mid \mathbf{y}_{i\cdot}) c_{12} + \Pr(2 \mid \mathbf{y}_{i\cdot}) c_{21}$$

and we can decide that individual is in group 1 if

$$f_1(\mathbf{y}_{i\cdot}) \cdot \pi_1 \cdot c_{12} > f_2(\mathbf{y}_{i\cdot}) \cdot \pi_2 \cdot c_{21}$$

or rearranging we can say that we should decide that the individual is in group 1 if

$$\frac{f_1(\mathbf{y}_{i\cdot})}{f_2(\mathbf{y}_{i\cdot})} > \frac{\pi_2}{\pi_1} \cdot \frac{c_{21}}{c_{12}}. \qquad (8.62)$$

If the π_j are unknown or assumed to be equal, and $c_{21} = c_{12}$, then it is only the right hand side of the above Equation (8.62) and what matters is the relative height of the two densities. The crossover point of Equation (8.62) would be the place where the densities themselves cross over.

The usual assumption is that an observation vector from group j

$$\mathbf{y}_{i\cdot} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

which implies from Equation (4.17) that

$$f_i(\mathbf{y}_{i\cdot}) = \frac{1}{|\boldsymbol{\Sigma}_j|^{1/2} (2\pi)^{p/2}} \exp\left[-(\mathbf{y}_{i\cdot} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_{i\cdot} - \boldsymbol{\mu}_j)/2\right].$$

Taking Equation (8.62) and taking logs to both sides, we would then place a case in group 1 if

$$\ln \frac{f_1(\mathbf{y}_{i\cdot})}{f_2(\mathbf{y}_{i\cdot})} > \ln \frac{\pi_2 c_{21}}{\pi_1 c_{12}}$$

$$\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}-\frac{1}{2}\left[(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_1)-(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_2)\right]>\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}.$$

If we assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ the above expression simplifies to

$$\left[y_{i\cdot}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]-\frac{1}{2}\left[(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mu_1-\boldsymbol{\mu}_2)\right]>\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}. \tag{8.63}$$

To get to this point it helps to realize that ($\mathbf{a}$ - $\mathbf{b}$)′$\mathbf{C}$($\mathbf{a}$ - $\mathbf{b}$) = $\mathbf{a}'\mathbf{Ca}$ - 2$\mathbf{a}'\mathbf{Cb}$ + $\mathbf{b}'\mathbf{Cb}$ and that ($\mathbf{a}$ + $\mathbf{b}$)′$\mathbf{C}$($\mathbf{a}$ - $\mathbf{b}$) = $\mathbf{a}'\mathbf{Ca}$ - $\mathbf{b}'$C$\mathbf{b}$. Noting also that $\ln\frac{a}{b}=-\ln\frac{b}{a}$, if we substract $\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}$ from both sides of the above equation we get

$$\left[y_{i\cdot}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]-\frac{1}{2}\left[(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]-\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}>0. \tag{8.64}$$

Define the left hand side of this last equation as ν_{12}. Our decision to place a case in group 1 is made if

$$\nu_{12} > 0.$$

For population 1 we have

$$\nu_{12} \sim N\left(\ln\frac{\pi_1 c_{12}}{\pi_2 c_{21}}+\Delta_{12}^2,\ \Delta_{12}^2\right)$$

where

$$\Delta_{12}^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

which is known as the Mahalanobis distance between the mean vectors of the two populations. Knowing the distribution of ν_{12} allows us to estimate the probability and the total costs of misclassification. We also define the raw discriminant function as

$$\nu = \mathbf{d}'\mathbf{y}_{i\cdot}$$

where

$$\mathbf{d} = \Sigma^{-1}(\mu_1 - \mu_2).$$

We can also standardize the function using

$$\nu_s=\Delta_{12}^{-1}\mathbf{d}'\mathbf{y}_{i\cdot}=\mathbf{d}_s'\mathbf{y}_{i\cdot}$$

Back to the decision,

$$\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}} + \left[y'_{i\cdot}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] - \frac{1}{2}\left[(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] > 0$$

$$\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}} + \mathbf{y}'_{i\cdot}\mathbf{d} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{d} > 0.$$

Rearranging, our decision "1" is taken if

$$\nu = \mathbf{y}'_{i\cdot}\mathbf{d} > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{d} - \ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}$$

or in the standardized version

$$\nu_s = \mathbf{y}'_{i\cdot}\mathbf{d}_s > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{d}_s - \Delta_{12}^{-1}\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}.$$

The discriminant function maximizes the separation between the values $\bar{\nu}_1$ and $\bar{\nu}_2$, the means for the two groups on the discriminant scores. When we don't know the μ_j or Σ, we split our samples into validation and holdout samples.

8.22 *Multiple Group Discriminant Function*

The problem can be approached as a special case of MANOVA. For example, assuming that we have k = 4 groups with p discriminating dependent variables, and the general linear hypothesis

$$H_0\text{: } \mathbf{ABM} = 0,$$

we would use the hypothesis matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with **M** = **I**. Just as we did before in Equations (8.41) and (8.42), we would calculate the hypothesis and error sum of squares matrices **H** and **E**. In order to find a score, $\nu = \mathbf{d}'\mathbf{y}_{i\cdot}$, with ν have as large a between groups sum of squares as possible, we will utilize the eigenstructure of $\mathbf{E}^{-1}\mathbf{H}$ as before. We pick values in the vector **d** such that our F test for group differences on ν is as large as possible. In other words, we maximize the between groups sum of squares for ν divided by it's within groups sum of squares, that is to say $\frac{\mathbf{d}'\mathbf{Hd}}{\mathbf{d}'\mathbf{Ed}}$, over all possible values of **a**. It is customary to scale **a** such that the within-group variance (mean square) is

$$\frac{\mathbf{d}'\mathbf{Ed}}{n-k} = \mathbf{d}'\mathbf{Sd} = 1.$$

References

Greenhouse, Samuel W. and S. Geisser (1959) On Methods in the Analysis of Profile Data. *Psychometrika*, 24, 95-112.

Hair, Joseph F., Rolph E. Anderson, Ronald L. Tatham and William C. Black (1995) *Multivariate Data Analysis. Fourth Edition*.. Englewood Cliffs, NJ: Prentice-Hall.

Keppel, Geoffrey (1973) *Design and Analysis: A Researcher's Handbook.* Englewood Cliffs, New Jersey: Prentice-Hall.

Lattin, James, J. Douglas Carroll and Paul E. Green (2003) *Analyzing Multivariate Data.* Pacific Grove, CA: Brooks/Cole.

Marascuilo, Leonard A. and Joel R. Levin (1983) *Multivariate Statistics in the Social Sciences*. Monterey, CA: Brooks/Cole.

Scheffé, Henri (1959) *The Analysis of Variance*. New York: Wiley.

Sharma, Subhash (1996) *Applied Multivariate Techniques*. New York: Wiley

Tatsuoka, Maurice M. (1971) *Multivariate Analysis*. New York: Wiley.

Timm, N. H. (1975) *Multivariate Analysis with Applications in Education and Psychology*. Monterey, CA: Brooks/Coles.