Notes on Analysis of Variance: Old School

John I. Marden

Copyright 2003
Chapter 1

Introduction to Linear Models

These notes are based on a course I taught using the text *Plane Answers to Complex Questions* by Ronald Christensen (Third edition, 2002, Springer-Verlag). Hence, everything throughout these pages implicitly uses that book as a reference. So keep a copy handy! But everything here is my own interpretation.

1.1 Dependent and explanatory variables

How is height related to weight? How are sex and age related to heart disease? What factors influence crime rate? Questions such as these have one dependent variable of interest, and one or more explanatory variables. The goal is to assess the relationship of the explanatory variables to the dependent variable. Examples:

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Explanatory Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>Height</td>
</tr>
<tr>
<td>Cholesterol level</td>
<td>Fat intake</td>
</tr>
<tr>
<td>Heart function</td>
<td>Age, sex</td>
</tr>
<tr>
<td>Crime rate</td>
<td>Population density, Average income, Educational level</td>
</tr>
<tr>
<td>Bacterial count</td>
<td>Drug</td>
</tr>
</tbody>
</table>

**Linear models** model the relationship by writing the mean of the dependent variable as a linear combination of the explanatory variables, or some representations of the explanatory variables. For example, a linear model relating cholesterol level to the percentage of fat in the diet would be

\[ cholesterol = \beta_0 + \beta_1(fat) + \text{residual}. \] (1.1)

The intercept \(\beta_0\) and slope \(\beta_1\) are parameters, usually unknown and to be estimated. One does not expect the cholesterol level to be an exact function of fat. Rather, there will be random variation: Two people with the same fat intake will likely have different cholesterol levels, just as two people of the same height will have different weights. The residual is the part of the dependent variable not explained by the linear function of the explanatory variables. As we go along, we will make other assumptions about the residuals, but the key
one at this point is that they have mean 0. That is, the dependent variable is on average equal to the linear function of the explanatory variables.

It is easy to think of more complicated models that are still linear, e.g.,

\[
\text{cholesterol} = \beta_0 + \beta_1(fat) + \gamma(\text{exercise}) + \text{residual}, \tag{1.2}
\]

or

\[
\text{cholesterol} = \beta_0 + \beta_1(fat) + \beta_3(fat)^2 + \text{residual}. \tag{1.3}
\]

"Wait!" you might say. That last equation is not linear, it is quadratic: the mean of cholesterol is a parabolic function of fat intake. Here is one of the strengths of linear models: The linearity is in the parameters, so that one or more representations of the explanatory variables can appear (e.g., here represented by \(fat\) and \(fat^2\)), as long as they are combined linearly. An example of a non-linear model:

\[
\text{cholesterol} = \beta_0 e^{\beta_1(fat)} + \text{residual}. \tag{1.4}
\]

This model is perfectly fine, just not a linear model.

A particular type of linear model, used when the explanatory variables are categorical, is the analysis of variance model, which is the main focus of this course. A categorical variable is one whose values are not-necessarily numerical. One study measured the bacterial count of leprosy patients, where each patient was given one of three treatment: Drug A, Drug D, or a placebo. The explanatory variable is the treatment, but it is not numerical. One way to represent treatment is with 0-1 variables, say \(a\), \(d\), and \(p\):

\[
a = \begin{cases} 
1 & \text{if treatment is Drug A} \\
0 & \text{if treatment is not Drug A}
\end{cases}, \quad d = \begin{cases} 
1 & \text{if treatment is Drug D} \\
0 & \text{if treatment is not Drug D}
\end{cases}, \tag{1.5}
\]

and

\[
p = \begin{cases} 
1 & \text{if treatment is the placebo} \\
0 & \text{if treatment is not the placebo}
\end{cases}. \tag{1.6}
\]

If someone received Drug A, that patient’s values for the representations would be \(a = 1, d = 0, p = 0\). Similarly, one receiving the placebo would have \(a = 0, d = 0, p = 1\). (Note that exactly one of \(a, d, p\) will be 1, the others 0.) One can go the other way: knowing \(a, d, p\), it is easy to figure out what the treatment is. In fact, you only need to know two of them, e.g., \(a\) and \(d\). A linear model constructed from these representations is then

\[
\text{bacterial count} = \beta_1 a + \beta_2 d + \beta_3 p + \text{residual}. \tag{1.7}
\]

If the residual has mean 0, then one can see that \(\beta_1\) is the mean of patients who receive Drug A, \(\beta_2\) is the mean of those who receive Drug D, and \(\beta_3\) is the mean for the control group.
1.2 Matrix notation

The values for the dependent variable will be denoted using $y$’s. The representations of the explanatory variables will usually be denoted using $x$’s, although other letters may show up. The $y$’s and $x$’s will have subscripts to indicate individual and variable. These subscripts may be single or may be multiple, whichever seems most useful at the time. The residuals use $e$’s, and the parameters are usually $\beta$’s, but can be any other Greek letter as well.

The next step is to write the model in a universal matrix notation. The dependent variable is an $n \times 1$ vector $\underline{y}$, where $n$ is the number of observations. The representations of the explanatory variables are in the $n \times p$ matrix $\underline{X}$, where the $j^{th}$ column of $\underline{X}$ contains the values for the $n$ observations on the $j^{th}$ representation. The $\beta$ is the $p \times 1$ vector of coefficients. Finally, $\underline{e}$ is the $n \times q$ vector of residuals. Putting these together, we have the model

$$\underline{y} = \underline{X}\underline{\beta} + \underline{e}. \tag{1.8}$$

(Generally, column vectors will be denoted by underlining, and matrices will be bold.) The following examples show how to set up this matrix notation.

**Simple linear regression.** Simple linear regression has one $x$, as in (1.1). If there are $n$ observations, then the linear model would be written

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n. \tag{1.9}$$

The $y$, $e$, and $\beta$ are easy to construct:

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \quad \text{and} \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}. \tag{1.10}$$

For $\underline{X}$, we need a vector for the $x_i$’s, but also a vector of 1’s, which are surreptitiously multiplying the $\beta_0$:

$$\underline{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}. \tag{1.11}$$

Check to see that putting (1.10) and (1.11) in (1.8) yields the model (1.9). Note that $p = 2$, that is, $\underline{X}$ has two columns even though there is only one $x$.

Another useful way to look at the model is to let $\underline{1}_n$ be the $n \times 1$ vector of 1’s, and $\underline{x}$ be the vector of the $x_i$’s, so that $\underline{X} = (\underline{1}_n, \underline{x})$, and

$$\underline{y} = \beta_0 \underline{1}_n + \beta_1 \underline{x} + \underline{e}. \tag{1.12}$$

(The text uses $J$ to denote $\underline{1}_n$, which is fine.)
Multiple linear regression. When there are more than one explanatory variables, as in (1.2), we need an extra subscript for $x$, so that $x_{i1}$ is the fat value and $x_{i2}$ is the exercise level for person $i$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad i = 1, \ldots, n.$$  

(1.13)

With $q$ variables, the model would be

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + e_i, \quad i = 1, \ldots, n.$$  

(1.14)

Notice that the quadratic model (1.3) is of this form with $x_{i1} = x_i$ and $x_{i2} = x_i^2$:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n.$$  

(1.15)

The general model (1.14) has the form (1.8) with a longer $\beta$ and wider $X$:

$$y = X\beta + \epsilon = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} + \epsilon.  

(1.16)

Here, $p = q + 1$, again there being an extra column in $X$ for the $1_n$ vector.

Analogous to (1.12), if we let $x_j$ be the vector of $x_{ij}$'s, which is the $(j+1)^{st}$ column of $X$, so that

$$X = (1_n, x_1, x_2, \ldots, x_q),$$  

(1.17)

we have that

$$y = \beta_0 1_n + \beta_1 x_1 + \cdots + \beta_q x_q + \epsilon.$$  

(1.18)

Analysis of variance. In analysis of variance, or ANOVA, explanatory variables are categorical. A one-way ANOVA has one categorical variable, as in the leprosy example (1.7). Suppose in that example, there are two observations for each treatment, so that $n = 6$. (The actual experiment had ten observations in each group.) The layout is

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug D</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{11}, y_{12}$</td>
<td>$y_{21}, y_{22}$</td>
<td>$y_{31}, y_{32}$</td>
</tr>
</tbody>
</table>

(1.19)

where now the dependent variable is denoted $y_{ij}$, where $i$ indicates the treatment, 1 = Drug A, 2 = Drug D, 3 = Control, and $j$ indicates the individual within the treatment. The linear model (1.7) translates to

$$y_{ij} = \beta_j + e_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2.$$  

(1.20)

To write the model in the matrix form (1.8), we first have to vectorize the $y_{ij}$'s (and $e_{ij}$'s), even though notationally they look like elements of a matrix. Any way you string them out is
fine, as long as you are consistent. We will do it systematically by grouping the observation by treatment, that is,

\[ y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \quad \text{and} \quad \varepsilon = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}. \]

Writing out the model in matrix form, we have

\[ \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \varepsilon. \]

Two-way ANOVA has two categorical explanatory variables. For example, the table below contains the leaf area/dry weight for some citrus trees, categorized by type of citrus fruit and amount of shade:

<table>
<thead>
<tr>
<th></th>
<th>Orange</th>
<th>Grapefruit</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>112</td>
<td>90</td>
<td>123</td>
</tr>
<tr>
<td>Half-shade</td>
<td>86</td>
<td>73</td>
<td>89</td>
</tr>
<tr>
<td>Shade</td>
<td>80</td>
<td>62</td>
<td>81</td>
</tr>
</tbody>
</table>

(From Table 11.2.1 in *Statistical Methods* by Snedecor and Cochran.) Each variable has 3 categories, which means there are 9 categories taking the two variables together. The dependent variable again has two subscripts, \( y_{ij} \), where now the \( i \) indicates the row variable (sun/shade) and the \( j \) represents the column variable (type of fruit). That is,

\[ \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \varepsilon. \]

One linear model for such data is the additive model, in which the mean for \( y_{ij} \) is the sum of an effect of the \( i \)\(^{th} \) row variable and an effect for the \( j \)\(^{th} \) column. That is, suppose \( \alpha_1, \alpha_2, \) and \( \alpha_3 \) are the effects attached to Sun, Half-shade, and Shade, respectively, and \( \beta_1, \beta_2, \) and \( \beta_3 \) are the effects attached to Orange, Grapefruit, and Mandarin, respectively. Then the additive model is

\[ y_{ij} = \alpha_i + \beta_j + e_{ij}. \]

The idea is that the two variables act separately. E.g., the effect of sun on \( y \) is the same for each fruit. The additive model places a restriction on the means of the cells, that is,

\[ \mu_{ij} = E(Y_{ij}) = \alpha_i + \beta_j. \]
For example, the following table of \( \mu_{ij} \)'s does follow an additive model:

\[
\begin{array}{ccc}
\text{Sun} & \text{Orange} & 110 \\
& \text{Grapefruit} & 90 \\
& \text{Mandarin} & 120 \\
\text{Half-shade} & 80 & 60 & 90 \\
Shade & 70 & 50 & 80
\end{array}
\]

Going from sun to half-shade subtracts 30, no matter which fruit; and going from half-shade to shade subtracts 10, again no matter which fruit. There are many sets of parameters that would fit those means. One is \( \alpha_1 = 0, \alpha_2 = -30, \alpha_3 = -40, \beta_1 = 110, \beta_2 = 90, \beta_3 = 120 \). An example of a non-additive model has means:

\[
\begin{array}{ccc}
\text{Sun} & \text{Orange} & 110 \\
& \text{Grapefruit} & 90 \\
& \text{Mandarin} & 120 \\
\text{Half-shade} & 85 & 60 & 85 \\
Shade & 60 & 50 & 80
\end{array}
\]

There are no sets of parameters that satisfy (1.26) for these \( \mu_{ij} \)'s. Note that going from sun to half-shade for Orange subtracts 25, while for Grapefruit it subtracts 30.

To write the additive model (1.25) in matrix form, we have to have 0-1 vectors for the rows, and 0-1 vectors for the columns. We will start with writing these vectors in table form. The row vectors:

\[
\begin{array}{c}
\mathbf{x}_1 : \\
1 1 1 \\
0 0 0 \\
0 0 0
\end{array}
\quad
\begin{array}{c}
\mathbf{x}_2 : \\
0 0 0 \\
1 1 1 \\
0 0 0
\end{array}
\quad
\begin{array}{c}
\mathbf{x}_3 : \\
0 0 0 \\
0 0 0 \\
1 1 1
\end{array}
\]

And the column vectors:

\[
\begin{array}{c}
\mathbf{x}_4 : \\
1 0 0 \\
1 0 0 \\
1 0 0
\end{array}
\quad
\begin{array}{c}
\mathbf{x}_5 : \\
0 1 0 \\
0 1 0 \\
0 1 0
\end{array}
\quad
\begin{array}{c}
\mathbf{x}_6 : \\
0 0 1 \\
0 0 1 \\
0 0 1
\end{array}
\]

As before, we write the \( y_{ij} \)'s, \( e_{ij} \)'s, and parameters in vectors:

\[
\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \\ e_{33} \end{pmatrix}, \quad \text{and} \quad \mathbf{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.
\]
For the $X$, we use the vectors in (1.29) and (1.30), making sure that the 0’s and 1’s in these vectors are correctly lined up with the observations in the $\mathbf{y}$ vector. That is, $X = (x_1, x_2, x_3, x_4, x_5, x_6)$, and the model is

$$y = \mathbf{X}\beta + \mathbf{e} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \mathbf{e}, \quad (1.32)$$

or

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta_1 x_4 + \beta_2 x_5 + \beta_3 x_6 + \mathbf{e} \quad (1.33)$$

If the additivity restriction (1.29), that $\mu_{ij} = \alpha_i + \beta_j$, is violated, then the model is said to have interaction. Specifically, the interaction term for each cell is defined by the difference

$$\gamma_{ij} = \mu_{ij} - \alpha_i - \beta_j, \quad (1.34)$$

and the model with interaction is

$$y_{ij} = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}. \quad (1.35)$$

To write out the model completely, we need to add the $\gamma_{ij}$’s to $\mathbf{y}$, and corresponding vectors to the $X$:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \mathbf{e}. \quad (1.36)$$
We will see later that we do not need that many vectors in \( X \), as there are many redundancies the way it is written now.

Analysis of covariance. It may be that the main interest is in comparing the means of groups as in analysis of variance, but there are other variables that may be effecting the \( y \).

For example, in the study comparing three drugs’ effectiveness in treating leprosy, there were bacterial measurements before and after treatment. The \( y_{ij} \)'s are the “after” measurement, and one would expect the “before” measurement, in addition to the drugs, to affect the after measurement. Letting \( z_{ij} \) represent the before measurements, the model modifies (1.20) to

\[
y_{ij} = \beta_j + \gamma z_{ij} + e_{ij},
\]  

or in matrix form, modifying (1.22),

\[
\begin{pmatrix}
y_{11} \\
y_{12} \\
y_{21} \\
y_{22} \\
y_{31} \\
y_{32}
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & z_{11} \\
1 & 0 & 0 & z_{12} \\
0 & 1 & 0 & z_{21} \\
0 & 1 & 0 & z_{22} \\
0 & 0 & 1 & z_{31} \\
0 & 0 & 1 & z_{32}
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\gamma
\end{pmatrix}
+ \mathcal{E}.
\]  

1.3 Vector spaces – Definition

The model (1.8), \( y = X\beta + e \) is very general. The \( X \) matrix can contain any numbers. The previous section gives some ideas of the scope of the model. In this section we look at the model slightly more abstractly. Letting \( \mu \) be the vector of means of the elements of \( y \), and \( X = (\begin{array}{c}
x_1 \\
x_2 \\
\vdots \\
x_p
\end{array}) \), the model states that

\[
\mu = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,
\]  

where the \( \beta_j \)'s can be any real numbers. Now \( \mu \) is a vector in \( \mathbb{R}^n \), and (1.39) shows that \( \mu \) is actually in a subset of \( \mathbb{R}^n \):

\[
\mu \in \mathcal{M} \equiv \{c_1 x_1 + c_2 x_2 + \cdots + c_p x_p \mid c_1 \in \mathbb{R}, \ldots, c_p \in \mathbb{R}\}.
\]  

Such a space of linear combinations of a set of vector is called a span.

**Definition 1** The span of the set of vectors \( \{x_1, \ldots, x_p\} \subset \mathbb{R}^n \) is

\[
\text{span}\{x_1, \ldots, x_p\} = \{c_1 x_1 + \cdots + c_p x_p \mid c_i \in \mathbb{R}, i = 1, \ldots, p\}.
\]  

Because the matrix notation (1.8) is heavily used in this course, we have notation for connecting the \( X \) to the \( \mathcal{M} \).
1.3. VECTOR SPACES – DEFINITION

Definition 2 For a \( n \times p \) matrix \( X \), \( C(X) \) denotes the column space of \( X \), which is the span of the columns of \( X \). That is, if \( X = (x_1, \ldots, x_p) \), then
\[
C(X) = \text{span}\{x_1, \ldots, x_p\}. \tag{1.42}
\]

Spans have nice properties. In fact, a span is a vector space. The formal definition of a vector space, at least for those that are subsets of \( \mathbb{R}^n \), follows. [Vector spaces are more general than those that are subsets of \( \mathbb{R}^n \), but since those are the only ones we need, we will stick with this restricted definition.]

Definition 3 A subset \( \mathcal{M} \subset \mathbb{R}^n \) is a vector space if
\[
\forall x, y \in \mathcal{M} \implies x + y \in \mathcal{M}, \quad \text{and} \quad c \in \mathbb{R}, x \in \mathcal{M} \implies cx \in \mathcal{M}. \tag{1.43, 1.44}
\]

Thus any linear combination of vectors in \( \mathcal{M} \) is also in \( \mathcal{M} \). Note that \( \mathbb{R}^n \) is itself a vector space, as is the set \( \{0_n\} \). [The \( n \times 1 \) vector of all 0’s is \( 0_n \).] Because \( c \) in (1.44) can be 0, any subspace must contain \( 0_n \). Any line through \( 0_n \), or plane through \( 0_n \), is a subspace. It is not hard to show that any span is a vector space. Take \( \mathcal{M} \) in (1.40). First, if \( x, y \in \mathcal{M} \), then there are \( a_i \)'s and \( b_i \)'s such that
\[
x = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p \quad \text{and} \quad y = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p, \tag{1.45}
\]
so that
\[
x + y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p, \quad \text{where} \quad c_i = a_i + b_i, \tag{1.46}
\]
hence \( x + y \in \mathcal{M} \). Second, for \( x \in \mathcal{M} \) as in (1.45) and real \( c \),
\[
 cx = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p, \quad \text{where} \quad c_i = ca_i, \tag{1.47}
\]
hence \( cx \in \mathcal{M} \).

Not only is any span a subspace, but any subspace is a span of some vectors. Thus a linear model (1.8) can equivalently be defined as one for which
\[
\mu \in \mathcal{M} \quad (\mu = E[Y]) \tag{1.48}
\]
for some vector space \( \mathcal{M} \).

Specifying a vector space through span is quite convenient, but not the only convenient way. Another is to give the form of elements directly. For example, the vector space of all vectors with equal elements can be given in the following two ways:
\[
\left\{ \begin{pmatrix} a \\ a \\ \vdots \\ a \end{pmatrix} \in \mathbb{R}^n \mid a \in \mathbb{R} \right\} = \text{span}\{1_0\}. \tag{1.49}
\]
When $n = 3$, the $x/y$ plane can be represented as
\[ \left\{ \begin{pmatrix} a \\ b \\ 0 \end{pmatrix} \mid a \in \mathbb{R}, b \in \mathbb{R} \right\} = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}. \tag{1.50} \]

A different plane is
\[ \left\{ \begin{pmatrix} a \\ a+b \\ b \end{pmatrix} \mid a \in \mathbb{R}, b \in \mathbb{R} \right\} = \text{span} \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}. \tag{1.51} \]

## 1.4 Linear independence and bases

Any subspace of $\mathbb{R}^n$ can be written as a span of at most $n$ vectors, although not in a unique way. For example,
\[
\text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\} = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\} = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\} = \text{span} \left\{ \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -7 \\ 0 \end{pmatrix}, \begin{pmatrix} 33 \\ 2 \\ 0 \end{pmatrix} \right\}. \tag{1.52} \]

Note that the space in (1.52) can be a span of two or three vectors, or a span of any number more than three as well. It cannot be written as a span of only one vector. The minimum number of vectors is called the rank of the space, which in this example is 2. Any set of two vectors which does span that space is called a basis. Notice that in the two sets of three vectors, there is a redundancy, that is, one of the vectors can be written as a linear combination of the other two: $(1,1,0)' = (1,0,0)' + (0,1,0)'$ and $(2,0,0)' = (4/(33 \ast 7))(0,-7,0)' + (2/33) \ast (33,2,0)'$. Such sets are called linearly dependent.

To formally define basis, we need to first define linear independence.

**Definition 4** The vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ in $\mathbb{R}^n$ are linear independent if
\[ a_1\mathbf{x}_1 + \cdots + a_p\mathbf{x}_p = \mathbf{0}, \quad \implies \quad a_1 = \cdots = a_p = 0. \tag{1.53} \]

Equivalently, the vectors are linearly independent if no one of them (as long as it is not $\mathbf{0}_n$) can be written as a linear combination of the others. That is, they are linear dependent if there is an $\mathbf{x}_i \neq \mathbf{0}_n$ and set of coefficients $a_i$ such that
\[ \mathbf{x}_i = a_1\mathbf{x}_1 + \cdots + a_{i-1}\mathbf{x}_{i-1} + a_{i+1}\mathbf{x}_{i+1} + \cdots + a_p\mathbf{x}_p. \tag{1.54} \]
They are not linearly dependent if and only if they are linearly independent. In (1.52), the sets with three vectors are linearly dependent, and those with two vectors are linearly independent. To see that latter fact for \(\{(1, 0, 0)', (1, 1, 0)\}'\), suppose that \(a_1(1, 0, 0)' + a_2(1, 1, 0)' = (0, 0, 0)'\). Then

\[
a_1 + a_2 = 0 \quad \text{and} \quad a_2 = 0 \implies a_1 = a_2 = 0,
\]

which verifies (1.53). If a set of vectors is linearly dependent, then one can remove one of the redundant vectors (1.54), and still have the same span. A basis is a set of vectors that has the same span but no dependencies.

**Definition 5** The set of vectors \(\{z_1, \ldots, z_d\}\) is a basis for the subspace \(\mathcal{M}\) if the vectors are linearly independent and \(\mathcal{M} = \text{span}\{z_1, \ldots, z_d\}\).

For estimating \(\beta\), the following lemma is useful.

**Lemma 1** If \(\{z_1, \ldots, z_p\}\) is a basis for \(\mathcal{M}\), then for \(\mathbf{z} \in \mathcal{M}\), there is a unique set of coefficients \(a_1, \ldots, a_p\) such that \(\mathbf{z} = a_1z_1 + \cdots + a_pz_p\).

Although a (nontrivial) subspace has many bases, each basis has the same number of elements, which is the rank.

**Definition 6** The rank of a subspace is the number of vectors in any of its bases.

A couple of useful facts about a vector space \(\mathcal{M}\) with rank \(d\) –

1. Any set of more than \(d\) vectors from \(\mathcal{M}\) is linearly dependent;

2. Any set of \(d\) linear independent vectors from \(\mathcal{M}\) forms a basis of \(\mathcal{M}\).

For example, consider the one-way ANOVA model in (1.22). The three vectors in \(\mathbf{X}\) are clearly linearly independent, hence the space \(C(\mathbf{X})\) has rank 3, and those vectors constitute a basis. On the other hand, the columns of \(\mathbf{X}\) in the two-way additive ANOVA model in (1.32) are not linearly independent: The first three add to \(\mathbf{0}_6\), as do the last three, hence

\[
\mathbf{z}_1 + \mathbf{z}_2 + \mathbf{z}_3 - \mathbf{z}_4 - \mathbf{z}_5 - \mathbf{z}_6 = \mathbf{0}_6.
\]

Removing any one of the vectors does leave a basis. The model (1.36) has many redundancies. For one thing, \(n = 9\), and there are 15 columns in \(\mathbf{X}\). One basis consists of the 9 interaction vectors (i.e., the last 9 vectors). Another consists of the columns of the following matrix,
obtained by dropping a judicious set of vectors from \( \mathbf{X} \):

\[
\begin{pmatrix}
\text{Rows} & \text{Columns} & \text{Interactions} \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

(1.57)

Not just any 9 vectors of \( \mathbf{X} \) will be a basis, though. For example, the first 9 are not linearly independent, as in (1.56).

### 1.5 Summary

This chapter introduced linear models, with some examples, and showed that any linear model can be expressed in a number of equivalent ways:

1. Each \( y_i \) is written as a linear combinations of \( x_i \)'s, plus residual (e.g., equation 1.13);

2. The vector \( \underline{y} \) is be written as a linear combination of \( \underline{x} \) vectors, plus vector of residuals (e.g., equation 1.18);

3. The vector \( \underline{y} \) is written as \( \mathbf{X}\underline{\beta} \), plus vector of residuals (as in equation 1.8);

4. The mean vector is restricted to a vector space, as in (1.48).

Each representation is useful in different situations, and it is important to be able to go from one to the others.

In the next chapter, we consider estimation of the mean \( E(\underline{Y}) \) and the parameters \( \underline{\beta} \). What can be estimated, and how, depends on the vectors in \( \mathbf{X} \), and whether they form a basis.
Chapter 2

Estimation and Projections

This chapter considers estimating $E[Y]$ and $\beta$. There are many estimation techniques, and which are best depends on the distribution assumptions made on the residuals. We start with minimal assumption on the residuals, and look at unbiased estimation using least squares.

The basic model is

$$y = X\beta + e,$$

where $E[e] = 0$.

(2.1)

The expected value of a vector is just the vector of expected values, so that

$$E[e] = E\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} E[e_1] \\ E[e_2] \\ \vdots \\ E[e_n] \end{pmatrix}.$$  

(2.2)

Then, as in (1.39), with $\mu \equiv E[Y] = X\beta + E[e]$, so that

$$\mu = X\beta.$$  

(2.3)

### 2.1 Projections and estimation of the mean

We first divorce ourselves from $\beta$, and just estimate $\mu \in \mathcal{M}$ (2.3) for a vector space $\mathcal{M}$. The idea in estimating $\mu$ is to pick an estimate $\hat{\mu} \in \mathcal{M}$ that is “close” to the observed $y$. The least squares principle is to find the estimate that has the smallest sum of squares from $y$. That is, $\hat{\mu}$ is the vector in $\mathcal{M}$ such that

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \leq \sum_{i=1}^{n} (y_i - a_i)^2 \text{ for any } a \in \mathcal{M}.$$  

(2.4)

The length of a vector $x \in \mathbb{R}^n$ is $\sqrt{\sum_{i=1}^{n} x_i^2}$, which is denoted by the norm $\|x\|$, so that

$$\|x\|^2 = \sum_{i=1}^{n} x_i^2 = x'x.$$  

(2.5)
Thus we can define the least squares estimate of \( \mu \in \mathcal{M} \) to be the \( \hat{\mu} \in \mathcal{M} \) such that

\[
\| y - \hat{\mu} \|^2 \leq \| y - a \|^2 \quad \text{for any } a \in \mathcal{M}.
\] (2.6)

Suppose \( n = 2 \), and the space \( \mathcal{M} = \text{span}\{ x \} \) for some nonzero vector \( x \). Then \( \mathcal{M} \) is a line through the origin. Take a point \( y \) somewhere in the space, and try to find the point on the line closest to \( y \). Draw a line segment from \( y \) to that point, and you will notice that that segment is perpendicular, or orthogonal, to the line \( \mathcal{M} \). This idea can be generalized to any vector space. First, we define orthogonality.

**Definition 7** For vectors and vector spaces in \( \mathbb{R}^n \):

1. Two vectors \( a \) and \( b \) are **orthogonal**, written \( a \perp b \), if \( a'b = 0 \).
2. A vector \( a \) is **orthogonal** to the vector space \( \mathcal{M} \), written \( a \perp \mathcal{M} \), if \( a \) is orthogonal to every \( x \in \mathcal{M} \).
3. Two vector spaces \( \mathcal{M} \) and \( \mathcal{N} \) are **orthogonal**, written \( \mathcal{M} \perp \mathcal{N} \), if \( a \in \mathcal{M}, b \in \mathcal{N} \implies a \perp b \).
4. The **orthogonal complement** of a vector space \( \mathcal{M} \), written \( \mathcal{M}^\perp \), is the set of all vectors \( a \in \mathbb{R}^n \) orthogonal to \( \mathcal{M} \), that is,

\[
\mathcal{M}^\perp = \{ a \in \mathbb{R}^n \mid a \perp \mathcal{M} \}.
\] (2.8)

Projecting a point on a vector space means dropping a perpendicular from the point to the space. The resulting vector in the space is the **projection**.

**Definition 8** For vector \( y \) and vector space \( \mathcal{M} \) in \( \mathbb{R}^n \), the **projection** of \( y \) onto \( \mathcal{M} \) is the point \( \hat{y} \in \mathcal{M} \) such that

\[
y - \hat{y} \perp \mathcal{M}.
\] (2.9)

It turns out that the projection solves the least squares problem.

**Proposition 1** For \( y \) and \( \mathcal{M} \) in \( \mathbb{R}^n \), the unique least squares estimate \( \hat{\mu} \) in (2.6) is \( \hat{y} \), the projection of \( y \) onto \( \mathcal{M} \).

**Proof.** Let \( a \) be any element of \( \mathcal{M} \). Then

\[
\| y - a \|^2 = \| (y - \hat{y}) + (\hat{y} - a) \|^2 = \| y - \hat{y} \|^2 + \| \hat{y} - a \|^2 + 2(y - \hat{y})'(\hat{y} - a). \] (2.10)

Now \( \hat{y} \in \mathcal{M} \) (because that’s where projections live) and \( a \in \mathcal{M} \) by assumption, so \( \hat{y} - a \in \mathcal{M} \) (Why?). But \( y - \hat{y} \perp \mathcal{M} \), because \( \hat{y} \) is the projection of \( y \), so in particular \( y - \hat{y} \perp a \), hence the last term in (2.10) is 0. Thus if \( a \neq \hat{y} \),

\[
\| y - a \|^2 > \| y - \hat{y} \|^2,
\] (2.11)

which means that \( \hat{\mu} = \hat{y} \), and it is unique. \(\square\)
2.1. PROJECTIONS AND ESTIMATION OF THE MEAN

2.1.1 Some simple examples

Suppose $M = \text{span}\{1_n\} = \{ (a, a, \ldots, a)' \mid a \in \mathbb{R} \}$. Then the projection of $y$ onto $M$ is a vector $(b, b, \ldots, b)' (\in M)$ such that $(y - (b, b, \ldots, b)') \perp M$, i.e.,

$$
\begin{pmatrix}
    y - \begin{pmatrix} b \\ b \\ \vdots \\ b \\ a \\ a \\ \vdots \\ a \\ a \\ \vdots \\ a \\ a \\ a \\ a \\
    \end{pmatrix}'
\end{pmatrix}
\begin{pmatrix}
    a \\
    a \\
    \vdots \\
    a \\
\end{pmatrix}
= 0 \text{ for all } a \in \mathbb{R}.
$$

(2.12)

Now (2.12) means that

$$
\sum_{i=1}^{n} (y_i - b)a = a(\sum_{i=1}^{n} y_i - nb) = 0 \text{ for all } a \in \mathbb{R}.
$$

(2.13)

The only way that last equality can hold for all $a$ is if

$$
\sum_{i=1}^{n} y_i - nb = 0,
$$

(2.14)

or

$$
b = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}.
$$

(2.15)

Thus the projection is

$$
\hat{y} = \begin{pmatrix}
    \bar{y} \\
    \bar{y} \\
    \vdots \\
    \bar{y}
\end{pmatrix}
$$

(2.16)

Extend this example to $M = \text{span}\{x\}$ for any fixed nonzero vector $x \in \mathbb{R}^n$. Because it is an element of $M$, $\hat{y} = cx$ for some $c$, and for $y - \hat{y}$ to be orthogonal to $M$, it must be orthogonal to $x$, that is,

$$
x'(y - cx) = 0.
$$

(2.17)

Solve for $c$:

$$
x'y = cx'x \implies c = \frac{x'y}{x'x},
$$

(2.18)

so that

$$
\hat{y} = \frac{x'y}{x'x} x.
$$

(2.19)

Next, consider the one-way ANOVA model (1.22),

$$
\begin{pmatrix}
    y_{11} \\
    y_{12} \\
    y_{21} \\
    y_{22} \\
    y_{31} \\
    y_{32}
\end{pmatrix} =
\begin{pmatrix}
    1 & 0 & 0 \\
    1 & 0 & 0 \\
    0 & 1 & 0 \\
    0 & 1 & 0 \\
    0 & 0 & 1 \\
    0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
    \beta_1 \\
    \beta_2 \\
    \beta_3
\end{pmatrix} + \varepsilon,
$$

(2.20)
so that
\[
M = \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \right\}. \tag{2.21}
\]

Now \( \hat{y} = (a, a, b, b, c, c)' \) for some \( a, b, c \). For \( y - \hat{y} \) to be orthogonal to \( M \), it is enough that it be orthogonal to the spanning vectors of \( M \).

**Proposition 2** If \( M = \text{span}\{x_1, x_2, \ldots, x_p\} \), then \( a \perp M \) if and only if \( a \perp x_i \) for \( i = 1, \ldots, p \).

**Proof.** If \( a \perp M \), then \( a \perp x_i \) for each \( i \) because \( x \) is orthogonal to all vectors in \( M \). So suppose \( a \perp x_i \) for \( i = 1, \ldots, p \), and take any \( x \in M \). By definition of span, \( x = c_1 x_1 + \cdots + c_p x_p \), so that
\[
x^'a = (c_1 x_1 + \cdots + c_p x_p)^'a = c_1 x_1^'a + \cdots + c_p x_p^'a = 0, \tag{2.22}
\]
because each \( x^'a = 0 \).

Writing down the equations resulting from \( (y - (a, a, b, b, c, c)')'x \) for \( x \) being each of the spanning vectors in (2.21) yields
\[
y_{11} - a + y_{12} - a = 0 \\
y_{21} - b + y_{22} - b = 0 \\
y_{31} - c + y_{32} - c = 0. \tag{2.23}
\]

It is easy to solve for \( a, b, c \):
\[
a = \frac{y_{11} + y_{12}}{2} \equiv \overline{y}_1; \quad b = \frac{y_{21} + y_{22}}{2} \equiv \overline{y}_2; \quad c = \frac{y_{31} + y_{32}}{2} \equiv \overline{y}_3. \tag{2.24}
\]

These equations introduce the “dot” notation: When a variable has multiple subscripts, then replacing the subscript with a “\( \cdot \)”, and placing a bar over the variable, denotes the average of the variable over that subscript.

### 2.1.2 The projection matrix

Rather than figuring out the projection for every \( y \), one can find a matrix \( M \) that gives the projection.

**Definition 9** For vector space \( M \), the matrix \( M \) such that \( \hat{y} = My \) for any \( y \in \mathbb{R}^n \) is called the projection matrix.
2.1. PROJECTIONS AND ESTIMATION OF THE MEAN

The definition presumes that such an $M$ exists and is unique for any $M$, which is true. In Proposition 4, we will construct the matrix. We will first reprise the examples in Section 2.1.1, exhibiting the projection matrix.

For $M = \text{span}\{\mathbf{1}_n\}$, the projection (2.16) replaces the elements of $\bar{y}$ with their average, so the projection matrix must satisfy

$$M\bar{y} = \begin{pmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}. \quad (2.25)$$

The only matrix that will accomplish that feat is

$$M = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} = \frac{1}{n} \mathbf{1}_n\mathbf{1}_n'. \quad (2.26)$$

Notice that $\mathbf{1}_n\mathbf{1}_n'$ is the $n \times n$ matrix of all 1’s.

Next, consider $M = \text{span}\{\mathbf{x}\}$ ($\mathbf{x} \neq \mathbf{0}_n$). We have from (2.19), with a little rewriting,

$$\tilde{y} = \frac{x'x'}{x'x} \bar{y}. \quad (2.27)$$

There is the matrix, i.e.,

$$M = \frac{x'x'}{x'x}. \quad (2.28)$$

Note that if $\mathbf{x} = \mathbf{1}_n$, then (2.28) is the same as (2.26), because $\mathbf{1}_n\mathbf{1}_n' = n$.

For the $M$ in (2.21), we have that the projection (2.24) replaces each element of $\bar{y}$ with the average of its group. In this case, each group has just 2 elements, so that

$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} \frac{1}{2} \end{pmatrix}. \quad (2.29)$$

Next are some important properties of projection matrices. We need the $n \times n$ identity matrix, which is denoted $I_n$, and its columns $i_{nj}$:

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = (i_{n1}, i_{n2}, \ldots, i_{nn}). \quad (2.30)$$
Proposition 3 The projection matrix $M$ for vector space $\mathcal{M}$ has the following properties:

1. It exists;
2. It is unique;
3. It is symmetric, $M = M'$;
4. It is idempotent, $MM = M$;
5. For $x \in \mathcal{M}$, $Mx = x$;
6. For $x \perp \mathcal{M}$, $Mx = 0_n$;
7. $I_n - M$ is the projection matrix for $\mathcal{M}^\perp$.

Proof. Let $\hat{\lambda}_{nj}$ be the unique projection of $\hat{\lambda}_{nj}$ onto $\mathcal{M}$, $j = 1, \ldots, n$, and

$$\hat{I}_n = (\hat{\lambda}_{n1}, \hat{\lambda}_{n2}, \ldots, \hat{\lambda}_{nn}).$$  \hfill (2.31)

This $\hat{I}_n$ is a projection matrix: Take $y \in \mathbb{R}^n$ and let $\hat{y} = \hat{I}_ny$. By definition of projection (Def. 8), each $\hat{\lambda}_{nj} \in \mathcal{M}$, hence

$$\hat{y} = \hat{I}_ny = y_1\hat{\lambda}_{n1} + y_2\hat{\lambda}_{n2} + \cdots + y_n\hat{\lambda}_{nn} \in \mathcal{M}.$$  \hfill (2.32)

Also, each $\hat{\lambda}_{nj} - \hat{\lambda}_{nj}$ is orthogonal to $\mathcal{M}$, hence

$$y - \hat{y} = (I_n - \hat{I}_n)y = y_1(\hat{\lambda}_{n1} - \hat{\lambda}_{n1}) + y_2(\hat{\lambda}_{n2} - \hat{\lambda}_{n2}) + \cdots + y_n(\hat{\lambda}_{nn} - \hat{\lambda}_{nn}) \perp \mathcal{M}. $$  \hfill (2.33)

Thus $\hat{y}$ is the projection of $y$ onto $\mathcal{M}$, hence $\hat{I}_n$ is a projection matrix, proving item 1.

For item 2, suppose $M$ is a projection matrix. Then $M\hat{\lambda}_{nj} = \hat{\lambda}_{nj}$ for each $j$, that is, $MI_n = \hat{I}_n$, or $M = \hat{I}_n$. Thus the projection is unique.

For item 3, we know that the columns of $\hat{I}_n$ are orthogonal to the columns of $I_n - \hat{I}_n$, hence

$$(I_n - \hat{I}_n)\hat{I}_n = 0, \quad \text{which } \implies \hat{I}_n = \hat{I}_n\hat{I}_n.$$  \hfill (2.34)

The matrix $\hat{I}_n\hat{I}_n$ is symmetric, so $\hat{I}_n$, i.e., $M$ is symmetric.

Skipping to item 5, if $x \in \mathcal{M}$, then clearly the closest point to $x$ in $\mathcal{M}$ is $x$ itself, that is, $\hat{x} = x$, hence $Mx = x$. Then item 4 follows, because for any $x$, $Mx = \hat{x} \in \mathcal{M}$, hence $MMx = \hat{x}$, and the uniqueness of the projection matrix shows that $MM = M$.

Item 6 is easy: If $x \perp \mathcal{M}$, then $x - 0_n \perp \mathcal{M}$, and $0_n \in \mathcal{M}$, hence $\hat{x} = 0_n$. Item 7 follows because if $\hat{x}$ is the projection of $x$ onto $\mathcal{M}$, $x - \hat{x}$ is the projection onto $\mathcal{M}^\perp$ (since $x - \hat{x} \in \mathcal{M}^\perp$ and $x - (x - \hat{x}) \in (\mathcal{M}^\perp)^\perp = \mathcal{M}$).

Those are all useful properties, but we still would like to know how to construct $M$. The next proposition gives an linear equation to solve to obtain the projection matrix. It is a generalization of the procedure we used in these examples.
Proposition 4 Suppose $\mathcal{M} = C(X)$, where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ is a basis for $\mathcal{M}$, then

$$\mathbf{M} = X(X'X)^{-1}X'. \tag{2.35}$$

The proposition uses that $X'X$ is invertible if its columns are linearly independent. We will show that later. We do note that even if the columns are not linearly independent, $(X'X)^{-1}$ can be replaced by any generalized inverse, which we will mention later as well.

Proof of proposition. For any given $\mathbf{x}$, let $\hat{\mathbf{x}}$ be its projection onto $\mathcal{M}$, so that $\hat{\mathbf{x}} = X\mathbf{b}$ for some vector $\mathbf{b}$. Because $\mathbf{x} - \hat{\mathbf{x}} \perp \mathcal{M}$, $\mathbf{x} - \hat{\mathbf{x}} \perp \mathbf{x}_j$ for each $j$, so that $X'(\mathbf{x} - \hat{\mathbf{x}}) = 0_p$, hence

$$X'(\mathbf{x} - X\mathbf{b}) = 0_p, \quad \text{which} \quad \implies \quad X'\mathbf{x} = X'X\mathbf{b} \quad \implies \quad \mathbf{b} = (X'X)^{-1}X'\mathbf{x}. \tag{2.36}$$

and

$$\hat{\mathbf{x}} = X(X'X)^{-1}X'\mathbf{x}. \tag{2.37}$$

Thus (2.35) holds.

Compare (2.28) for $p = 1$ to (2.35). Also note that it is easy to see that this $\mathcal{M}$ is symmetric and idempotent. Even though the basis is not unique for a given vector space, the projection matrix is unique, hence any basis will yield the same $X(X'X)^{-1}X'$.

It is interesting that any symmetric idempotent matrix $\mathbf{M}$ is a projection matrix for some vector space, that vector space being

$$\mathcal{M} = \{\mathbf{M}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} \tag{2.38}$$

2.2 Estimating coefficients

2.2.1 Coefficients

Return to the linear model (2.1),

$$\underline{y} = X\underline{\beta} + \underline{e}, \quad \text{where} \quad E[\underline{e}] = \underline{0}_n. \tag{2.39}$$

In this section we consider estimating $\beta$, or linear functions of $\beta$. In some sense, this task is less fundamental than estimating $\mu = \underline{E}[\underline{Y}]$, since the meaning of any $\beta_j$ depends not only on its corresponding column in $\underline{X}$, but also what other columns happen to be in $X$. For example, consider these five equivalent models for $\underline{\mu}$ in the one-way ANOVA (1.22):

$$\underline{\mu} = X_1\underline{\beta}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}.$$
\[\begin{align*}
\mu &= X_2\beta_2 = \\
&= \begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\mu \\
\alpha_1 \\
\alpha_2 \\
\alpha_3
\end{pmatrix}
\end{align*}\]

\[\begin{align*}
\mu &= X_3\beta_3 = \\
&= \begin{pmatrix}
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 0 & 0 \\
1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\mu \\
\alpha_1 \\
\alpha_2
\end{pmatrix}
\end{align*}\]

\[\begin{align*}
\mu &= X_4\beta_4 = \\
&= \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & -1 \\
1 & 1 & -1 \\
1 & -2 & 0 \\
1 & -2 & 0
\end{pmatrix}
\begin{pmatrix}
\mu \\
\gamma_1 \\
\gamma_2
\end{pmatrix}
\end{align*}\]  \hspace{1cm} (2.40)

\[\begin{align*}
\mu &= X_5\beta_5 = \\
&= \mu_1 + \alpha_1 + \mu_2 + \alpha_2 + \mu_3 + \alpha_3
\end{align*}\]

The \(\mu_1, \mu_2, \mu_3\) in \(\beta_1\) are the means of the three groups, i.e., \(\mu_j = E[Y_{ij}]\). Comparing to \(\beta_2\), we see that \(\mu_j = \mu + \alpha_j\), but we still do not have a good interpretation for \(\mu\) and the \(\alpha\)'s. For example, if \(\mu = 0\), then \(\alpha_j = \mu_j\), the mean of the \(j^{th}\) group. But if \(\mu = \overline{\mu} \equiv (\mu_1 + \mu_2 + \mu_3)/3\), the overall average, then \(\alpha_j = \mu_j - \overline{\mu}\), the “effect” of group \(j\). Thus the presence of the \(1_6\) vector in \(X_2\) changes the meaning of the coefficients of the other vectors.

Now in \(\beta_4\), one may make the restriction that \(\alpha_3 = 0\). Then one has the third model, and \(\mu = \mu_3, \alpha_1 = \mu_1 - \mu_3, \alpha_2 = \mu_2 - \mu_3\). Alternatively, a common restriction is that \(\alpha_1 + \alpha_2 + \alpha_3 = 0\), so that the second formulation becomes the fourth:

\[\begin{align*}
\mu &= X_2\beta_2 = \\
&= \begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\mu \\
\alpha_1 \\
\alpha_2
\end{pmatrix}
\end{align*}\]

Now \(\mu = \overline{\mu}\) and \(\alpha_j = \mu_j - \mu\), the effect.
The final expression has
\[
\mu_1 = \mu + \gamma_1 + \gamma_2, \mu_2 = \mu + \gamma_1 - \gamma_2, \mu_3 = \mu - 2\gamma_1, \tag{2.42}
\]
from which can be derived
\[
\mu = \overline{\mu}, \gamma_1 = \frac{1}{3} \left( \frac{1}{2} (\mu_1 + \mu_2) - \mu_3 \right), \gamma_2 = \frac{1}{2} (\mu_1 - \mu_2). \tag{2.43}
\]
Then, e.g., for the leprosy example (1.19), \(\mu\) is the overall bacterial level, \(\gamma_1\) contrasts the average of the two drugs and the placebo, and \(\gamma_3\) contrasts the two drugs.

### 2.2.2 Least squares estimation of the coefficients

We know that the least squares estimate of the mean \(\mu = E[Y]\) is \(\hat{\mu} = \hat{\mu}\), the projection of \(\mu\) onto \(\mathcal{M} = C(X)\) in (2.39). It exists and is unique, because the projection is. A least squares estimate of \(\beta\) is one that yields the projection.

**Definition 10** In the model \(y = X\beta + e\), a least squares estimate \(\hat{\beta}\) of \(\beta\) is any vector for which
\[
\hat{\mu} = X\hat{\beta}, \tag{2.44}
\]
where \(\hat{\mu}\) is the projection of \(\mu\) into \(C(X)\).

A least squares estimate of a linear combination \(\lambda'\beta\), where \(\lambda \in \mathbb{R}^p\), is \(\lambda'\hat{\beta}\) for any least squares estimate \(\hat{\beta}\) of \(\beta\).

A least squares estimate of \(\beta\) always exists, but it may not be unique. The condition for uniqueness is direct.

**Proposition 5** The least squares estimate of \(\beta\) is unique if and only if the column of \(X\) are linearly independent.

The proposition follows from Lemma 1, because if the columns of \(X\) are linearly independent, they form a basis for \(C(X)\), hence there is a unique set of \(\hat{\beta}_j\)'s that will solve
\[
\hat{\mu} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p. \tag{2.45}
\]
And if the columns are not linearly independent, there are many sets of coefficients that will yield the \(\hat{\mu}\).

If columns of \(X\) are linearly independent, then \(X'X\) is invertible. In that case, as in the proof of Proposition 4, equation (2.36), we have that
\[
\hat{y} = X\hat{\beta} \quad \text{for} \quad \hat{\beta} = (X'X)^{-1}X'y, \tag{2.46}
\]
which means that the unique least squares estimate of \(\beta\) is
\[
\hat{\beta} = (X'X)^{-1}X'y. \tag{2.47}
\]
For an example, go back to the one-way ANOVA model (2.41). We know from (2.24) that
\[ \hat{\mu} = \frac{1}{3} (\bar{y}_1 + \bar{y}_2 - \bar{y}_3), \quad \hat{\gamma}_2 = \frac{1}{2} (\bar{y}_1 - \bar{y}_2). \] (2.49)

Consider the second expression,
\[ \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix} = X_2 \hat{\beta}_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{pmatrix}. \] (2.50)

The columns of \( X_2 \) are not linearly independent, so there are many possible estimates for the parameters, e.g.,
\[ \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix}, \begin{pmatrix} \bar{y}_1 - \bar{y} \\ \bar{y}_1 - \bar{y}_3 \\ \bar{y}_2 - \bar{y}_3 \\ 0 \end{pmatrix}, \begin{pmatrix} 83.72 \\ \bar{y}_1 - 83.72 \\ \bar{y}_2 - 83.72 \\ \bar{y}_3 - 83.72 \end{pmatrix}. \] (2.51)

It is interesting that even if \( \beta \) does not have a unique estimate, some linear combinations of it do. In the example above, (2.51), \( \mu \) and \( \alpha_1 \) do not have unique estimates, but \( \mu + \alpha_1 = (1, 1, 0, 0) \) does, i.e., \( \bar{y}_1 \). That is, all least squares estimates of \( \beta \) have the same \( (1, 1, 0, 0) \).

Any contrast among the \( \alpha_i \)'s also has a unique estimate, e.g., \( \alpha_1 - \alpha_2 \) has unique estimate \( \bar{y}_1 - \bar{y}_2 \), and \( \alpha_1 + \alpha_2 - 2\alpha_3 \) has unique estimate \( \bar{y}_1 + \bar{y}_2 - 2\bar{y}_3 \).

This discussion leads to the notion of \textbf{estimability}, which basically means that there is a unique least squares estimate. Here is the formal definition.

\textbf{Definition 11} A linear combination \( X' \beta \) \textbf{is estimable} if it has an unbiased linear estimator, that is, there exists an \( n \times 1 \) vector \( a \) such that
\[ E[a'Y] = X' \beta \quad \text{for all } \beta \in \mathbb{R}^p. \] (2.52)
2.2. ESTIMATING COEFFICIENTS

With the above example, \( \mu + \alpha_1 \) has \( \underline{\lambda} = (1, 1, 0, 0)' \), and taking \( \underline{a} = (1/2, 1/2, 0, 0, 0, 0)' \), so that \( \underline{a}'\underline{y} = \underline{y}_1 \), we have that

\[
E[\underline{a}'\underline{Y}] = \frac{1}{2} (E[\underline{Y}_{11}] + E[\underline{Y}_{12}]) = \frac{1}{2} (\mu + \alpha_1 + \mu + \alpha_1) = \mu + \alpha_1
\]  

(2.53)

no matter what \( \mu \) and the \( \alpha_i \)'s are. That is not the unique estimator. Note that \( Y_{11} \) alone works, also.

On the other hand, consider \( \mu = \underline{\lambda}'\underline{\beta} \) with \( \underline{\lambda} = (1, 0, 0, 0)' \). Can we find \( \underline{a} \) so that

\[
E[\underline{a}'\underline{Y}] = E[(a_{11}, a_{12}, a_{21}, a_{22}, a_{31}, a_{32})\underline{Y}]
\]

\[
= (a_{11} + a_{12}) (\mu + \alpha_1) + (a_{21} + a_{22}) (\mu + \alpha_2) + (a_{31} + a_{32}) (\mu + \alpha_3)
\]

\[
\equiv \mu?
\]  

(2.54)

For that to occur, we need

\[
a_{11} + a_{12} + a_{21} + a_{22} + a_{31} + a_{32} = 1, \quad a_{11} + a_{12} = 0, \quad a_{21} + a_{22} = 0, \quad a_{31} + a_{32} = 0.
\]  

(2.55)

Those equations are impossible to solve, since the last three imply that the sum of all \( a_{ij} \)'s is 0, not 1. Thus \( \mu \) is not estimable.

The next proposition systematizes how to check for estimability.

**Proposition 6** In the model \( \underline{y} = \underline{X}\underline{\beta} + \underline{e} \), with \( E[\underline{e}] = \underline{0}_n \), \( \underline{\lambda}'\underline{\beta} \) is estimable if and only if there exists an \( n \times 1 \) vector \( \underline{a} \) such that

\[
\underline{a}'\underline{X} = \underline{\lambda}'.
\]  

(2.56)

**Proof.** By Definition 11, since \( E[\underline{Y}] = \underline{X}\underline{\beta} \) in (2.52), \( \underline{\lambda}'\underline{\beta} \) is estimable if and only if there exists \( \underline{a} \) such that \( \underline{a}'\underline{X}\underline{\beta} = \underline{\lambda}'\underline{\beta} \) for all \( \underline{\beta} \in \mathbb{R}^p \). But that equality is equivalent to \( \underline{a}'\underline{X} = \underline{\lambda}' \). \( \square \)

Note that the condition (2.56) means that \( \underline{\lambda} \) is a linear combination of the rows of \( \underline{X} \), i.e., \( \underline{\lambda} \in C(\underline{X}') \). (Or we could introduce the notation \( R(\underline{X}) \) to denote the span of the rows.)

To see how that works in the example (2.50), look at

\[
C(\underline{X}') = \operatorname{span} \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}.
\]  

(2.57)

(The rows of \( \underline{X} \) have some duplicate vectors.) Then \( \lambda + \alpha_1 \) has \( \underline{\lambda} = (1, 1, 0, 0)' \), which is clearly in \( C(\underline{X}') \), since it is one of the basis vectors. (Similarly for \( \mu + \alpha_2 \) and \( \mu + \alpha_3 \).) The contrast \( \alpha_1 - 2\alpha_2 + \alpha_3 \) has \( \underline{\lambda} = (0, 1, -2, 1) \). That vector is also in \( C(\underline{X}') \), since \( \underline{\lambda}' = (1, 1, 0, 0)' - 2(1, 0, 1, 0)' + (1, 0, 0, 1)' \). On the other hand, consider estimating \( \mu \), which
has $\lambda = (1, 0, 0, 0)'$. To see that that vector is not in $C(X')$, we have to show that there are no $a, b, c$ so that

$$
\begin{pmatrix}
1 \\
0 \\
0 \\
0
\end{pmatrix} = a \begin{pmatrix}
1 \\
1 \\
0 \\
0
\end{pmatrix} + b \begin{pmatrix}
1 \\
0 \\
1 \\
0
\end{pmatrix} + c \begin{pmatrix}
0 \\
0 \\
0 \\
1
\end{pmatrix} = \begin{pmatrix}
a + b + c \\
a \\
b \\
c
\end{pmatrix}.
$$

(2.58)

Clearly equation (2.58) requires that $a = b = c = 0$, so $a + b + c$ cannot equal 1. Thus $\mu$ is not estimable. Also, the individual $\alpha_i$’s are not estimable.

If $\lambda'\beta$ is estimable, then there is a unique least squares estimate of it.

**Proposition 7** If $\lambda'\beta$ is estimable, the unique least squares estimate is $\lambda'\hat{\beta}$ for any least squares estimate $\hat{\beta}$ of $\beta$.

Of course, if there is a unique least squares estimate $\hat{\beta}$ of $\beta$, then the unique least square estimate of any $\lambda'\beta$ is $\lambda'\hat{\beta}$.

**Proof of proposition.** If $\lambda'\beta$ is estimable, then by Proposition 6 there exists an $a$ such that $a'X = \lambda'$. If $\hat{\beta}_1$ and $\hat{\beta}_2$ are least squares estimates of $\beta$, then by Definition 10, $\hat{\mu} = X\hat{\beta}_1 = X\hat{\beta}_2$, hence $a'X\hat{\beta}_1 = a'X\hat{\beta}_2$, which implies that $\lambda'\hat{\beta}_1 = \lambda'\hat{\beta}_2$. Thus the least squares estimate of $\lambda'\beta$ is unique.

Later, we show that the least squares estimators are optimal under certain conditions.

If one can find an unbiased estimate of $\lambda'\beta$, then the least squares estimate uses the same linear combination, but of the $\hat{y}$.

**Proposition 8** If $a'y$ is an unbiased estimator of $\lambda'\beta$, then $a'y$ is the least squares estimate of $\lambda'\beta$.

**Proof.** As in (2.56), $a'X = \lambda'$. By (2.44), any least squares estimate $\hat{\beta}$ of $\beta$ satisfies $\hat{y} = X\hat{\beta}$. Thus $a'\hat{y} = a'X\hat{\beta} = \lambda'\hat{\beta}$, which is the least squares estimate of $\lambda'\beta$.

For example, consider the one-way ANOVA model (2.41). The $\mu + \alpha_1$ is estimable, and, e.g., $y_{i1}$ is an unbiased estimate, which has $a = (1, 0, 0, 0, 0)'$. So by the proposition, $a'y$ is the least squares estimate. From (2.48), we know that $\hat{y}_{ij} = \overline{y}_i$, hence $a'y = \overline{y}_1$. is the least squares estimate of $\mu + \alpha_1$. Note that there are other $a$’s, e.g., $y_{i2}$ is also an unbiased estimate, and has $a* = (0, 1, 0, 0, 0)'$. Although this $a*$ is different from $a$, $a*y = \overline{y}_1$, too. That has to be true, since by Proposition 7 the least squares estimate is unique, but we also see that starting with any unbiased $a'y$, we can find the least squares estimate by replacing $y$ with $\hat{y}$.
2.2. ESTIMATING COEFFICIENTS

2.2.3 Example: Leprosy

Below are data on leprosy patients (from Snedecor and Cochran, *Statistical Methods*). There were 30 patients, randomly allocated to three groups of 10. The first group received drug A, the second drug D, and the third group received a placebo. Each person had the bacterial count taken before and after receiving the treatment.

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug D</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>15</td>
</tr>
</tbody>
</table>

First, consider the one-way ANOVA, with the “after” measurements as the $y$’s, ignoring the “before” measurements. The model is

\[
\begin{pmatrix}
  y_{11} \\
  y_{12} \\
  \vdots \\
  y_{1,10} \\
  y_{21} \\
  y_{22} \\
  \vdots \\
  y_{2,10} \\
  y_{31} \\
  y_{32} \\
  \vdots \\
  y_{3,10}
\end{pmatrix}
= 
\begin{pmatrix}
  1 & 1 & 0 & 0 \\
  1 & 1 & 0 & 0 \\
  \vdots & \vdots & \vdots & \vdots \\
  1 & 1 & 0 & 0 \\
  1 & 0 & 1 & 0 \\
  1 & 0 & 1 & 0 \\
  \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 1 & 0 \\
  1 & 0 & 0 & 1 \\
  1 & 0 & 0 & 1 \\
  \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
  \mu \\
  \alpha_1 \\
  \alpha_2 \\
  \alpha_3
\end{pmatrix}
+ \varepsilon. \tag{2.59}
\]

The sample means are $\overline{y}_1 = 5.3, \overline{y}_2 = 6.1, \overline{y}_3 = 12.3$. Suppose we are interested in the two contrasts $\alpha_1 - \alpha_2$, comparing the two drugs, and $(\alpha_1 + \alpha_2) - \alpha_3$, comparing the placebo to the average of the two drugs. The least squares estimates are found by taking the same contrasts of the sample means:

\[
\hat{\alpha}_1 - \hat{\alpha}_2 = 5.3 - 6.1 = -0.8, \quad (\alpha_1 + \hat{\alpha}_2) - \alpha_3 = (5.3 + 6.1)/2 - 12.3 = -6.6. \tag{2.60}
\]
There doesn’t appear to be much difference between the two drugs, but their average seems better than the placebo. Is it significantly different? That question will be addressed in the next chapter. We need standard errors for the estimates.

What about the “before” measurements? The sample means of the before measurements are 9.3, 10, and 12.9, respectively. Thus, by chance, the placebo group happened to get people who were slightly worse off already, so it might be important to make some adjustments. A simple one would be to take the \( y \)'s as after before, which is very reasonable in this case. Instead, we will look at analysis of covariance model (1.38), which adds in the before measurements (the covariates) as \( z_{ij} \)'s:

\[
\begin{pmatrix}
  y_{11} \\
y_{12} \\
\vdots \\
y_{1,10} \\
y_{21} \\
y_{22} \\
\vdots \\
y_{2,10} \\
y_{31} \\
y_{32} \\
\vdots \\
y_{3,10}
\end{pmatrix}
= 
\begin{pmatrix}
  1 & 1 & 0 & 0 & z_{11} \\
  1 & 1 & 0 & 0 & z_{12} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 1 & 0 & 0 & z_{1,10} \\
  1 & 0 & 1 & 0 & z_{21} \\
  1 & 0 & 1 & 0 & z_{22} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 1 & 0 & z_{2,10} \\
  1 & 0 & 0 & 1 & z_{31} \\
  1 & 0 & 0 & 1 & z_{32} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 0 & 1 & z_{3,10}
\end{pmatrix}
\begin{pmatrix}
  \mu \\
  \alpha_1 \\
  \alpha_2 \\
  \alpha_3 \\
  \gamma
\end{pmatrix}
+ \begin{pmatrix}
  1 & 1 & 0 & 0 & 11 \\
  1 & 1 & 0 & 0 & 8 \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 1 & 0 & 0 & 3 \\
  1 & 0 & 1 & 0 & 6 \\
  1 & 0 & 1 & 0 & 6 \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 1 & 0 & 15 \\
  1 & 0 & 0 & 1 & 16 \\
  1 & 0 & 0 & 1 & 13 \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 0 & 1 & 12
\end{pmatrix}
\]  

(2.61)

We want to estimate the same contrasts. Start with \( \alpha_1 - \alpha_2 \). Is it estimable? We need to show that there is a \( \beta \) such that \( \beta'X = (0, 1, -1, 0, 0) \). We can do it with just three rows of \( X \), the first two and the eleventh. That is, we want to find \( a, b, c \) so that

\[ a(1, 1, 0, 0, 11) + b(1, 1, 0, 0, 8) + c(1, 0, 1, 0, 6) = (0, 1, -1, 0, 0), \]

(2.62)

or

\[
\begin{align*}
  a + b + c &= 0 \\
a + b &= 1 \\
c &= -1 \\
11a + 8b + 6c &= 0
\end{align*}
\]

(2.63)

The second and third equations imply the first, and the third is that \( c = -1 \), hence using \( b = 1 - a \), the fourth equation yields \( a = -2/3 \), hence \( b = 5/3 \). Thus

\[ -\frac{2}{3} y_{11} + \frac{5}{3} y_{12} - y_{21} \] is an unbiased estimate of \( \alpha_1 - \alpha_2 \).

(2.64)
The least squares estimate replaces the \( y_{ij} \)'s with their hats, the \( \hat{y}_{ij} \)'s. We know in principle the projection \( \hat{y} \) from Question 4 of HW #2 (where now we have 10 instead of 2 observations in each group). That is, the projection vector has elements of the form
\[
\hat{y}_{1j} = a + d z_{1j}, \quad \hat{y}_{2j} = b + d z_{2j}, \quad \hat{y}_{3j} = c + d z_{3j}.
\]
(2.65)

The constants are
\[
a = \overline{y}_1 - d \overline{z}_1, \quad b = \overline{y}_2 - d \overline{z}_2, \quad c = \overline{y}_3 - d \overline{z}_3,
\]
(2.66)
hence
\[
\hat{y}_{ij} = \bar{y}_i + d (z_{ij} - \bar{z}_i).
\]
(2.67)

The \( d \) is
\[
d = \frac{\sum_{i=1}^{3} \sum_{j=1}^{10} (y_{ij} - \bar{y}_i) z_{ij}}{\sum_{i=1}^{3} \sum_{j=1}^{10} (z_{ij} - \bar{z}_i) z_{ij}}.
\]
(2.68)

Plugging in the data, we obtain that \( d = \frac{585.4}{593} = 0.987 \).

Back to estimating \( \alpha_1 - \alpha_2 \), substitute the \( \hat{y}_{ij} \)'s of (2.67) for the \( y_{ij} \)'s in (2.64) to get the least squares estimate
\[
-\frac{2}{3} \hat{y}_{11} + \frac{5}{3} \hat{y}_{12} - \hat{y}_{21} = -\frac{2}{3} (\bar{y}_1 + d (z_{11} - \bar{z}_1)) + \frac{5}{3} (\bar{y}_1 + d (z_{12} - \bar{z}_1))
\]
\[
- (\bar{y}_2 + d (z_{21} - \bar{z}_2))
\]
\[
= (\bar{y}_1 - d \bar{z}_1) - (\bar{y}_2 - d \bar{z}_2) + d(\frac{2}{3} z_{11} + \frac{5}{3} z_{12} - z_{21})
\]
\[
= (\bar{y}_1 - d \bar{z}_1) - (\bar{y}_2 - d \bar{z}_2)
\]
\[
= 5.3 - 0.987(9.3) - (6.1 - 0.987(10))
\]
\[
= -0.109.
\]
(2.69)

The third line comes from the second line since \(-\frac{2}{3} z_{11} + \frac{5}{3} z_{12} - z_{21} = -\frac{2}{3} z_{11} + \frac{5}{3} z_{12} - 6 = 0 \).

Notice that the least squares estimate is the same as that without covariates, but using adjusted (\( \bar{y}_i - d \bar{z}_i \))'s instead of plain \( \bar{y}_i \)'s.

The unadjusted estimate (not using the covariate) was \(-0.8\), so the adjusted estimate is even smaller.

A similar procedure will show that the least squares estimate of \((\alpha_1 + \alpha_2)/2 - \alpha_3\) is
\[
\frac{(\bar{y}_1 - d \bar{z}_1) + (\bar{y}_2 - d \bar{z}_2)}{2} - (\bar{y}_3 - d \bar{z}_3) = -3.392.
\]
(2.70)

This value is somewhat less (in absolute value) than the unadjusted estimate \(-4.60\).
CHAPTER 2. ESTIMATION AND PROJECTIONS
Chapter 3

Variance and Covariances

The previous chapter focused on finding unbiased estimators for $\mu$, $\beta$, and $\lambda'\beta$. In this section, we tackle standard errors and variances, and in particular look for estimators with small variance. The variance of a random variable $Z$ is $\text{Var}(Z) = E[(Z - \mu_Z)^2]$, where $\mu_Z = E[Z]$. [Note: These are true if the expectations exist.] With two variables, $Y_1$ and $Y_2$, there is the covariance:

$$
\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)], \quad \text{where} \quad \mu_1 = E(Y_1) \text{ and } \mu_2 = E(Y_2). 
$$

The covariance of a variable with itself is the variance.

### 3.1 Covariance matrices for affine transformations

Defining the mean of a vector or of a matrix is straightforward: it is just the vector or matrix of means. That is, as in (2.2), for vector $Y = (Y_1, \ldots, Y_n)'$,

$$
E(Y) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix}, 
$$

and for an $n \times p$ matrix $W$,

$$
E[W] = E\left[ \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1p} \\ W_{21} & W_{22} & \cdots & W_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{np} \end{pmatrix} \right] = \begin{pmatrix} E(W_{11}) & E(W_{12}) & \cdots & E(W_{1p}) \\ E(W_{21}) & E(W_{22}) & \cdots & E(W_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(W_{n1}) & E(W_{n2}) & \cdots & E(W_{np}) \end{pmatrix}. 
$$

Turning to variances, a $n \times 1$ vector $Y = (Y_1, \ldots, Y_n)'$ has $n$ variances (the $\text{Var}(Y_i)$’s), and several covariances $\text{Cov}(Y_i, Y_j)$. These are all conveniently arranged in the covariance matrix:
matrix, defined for vector $Y$ to be the $n \times n$ matrix $Cov(Y)$ whose $ij$\textsuperscript{th} element is $Cov(Y_i, Y_j)$. It is often denoted $\Sigma$:

$$
Cov(Y) = \Sigma = \begin{pmatrix}
\text{Var}(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_n) \\
Cov(Y_2, Y_1) & \text{Var}(Y_2) & \cdots & Cov(Y_2, Y_n) \\
\vdots & \vdots & \ddots & \vdots \\
Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & \cdots & \text{Var}(Y_n)
\end{pmatrix}.
$$

(3.4)

The diagonals are the variances, and the matrix is symmetric because $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$. Analogous to the definition of variance, an equivalent definition of covariance is

$$
Cov(Y) = E[(Y - \mu)(Y - \mu)^\prime], \quad \text{where} \quad \mu = E(Y).
$$

(3.5)

The variances and covariances of linear combinations are often needed in this course, e.g., the least squares estimates are linear combinations of the $y_i$'s. Fortunately, the means and (co)variances of linear combinations are easy to obtain from those of the originals. With just one variable $Z$, we know that for any $a$ and $b$,

$$
E[a + b Z] = a + b E[Z] \quad \text{and} \quad Var[a + b Z] = b^2 Var[Z].
$$

(3.6)

Note that the constant $a$ does not effect the variation.

Turn to an $n \times 1$ vector, $Z$, and consider the affine transformation

$$
\mathbf{W} = \mathbf{a} + \mathbf{BZ}
$$

(3.7)

for some $m \times 1$ vector $\mathbf{a}$ and $m \times n$ matrix $\mathbf{B}$, so that $\mathbf{W}$ is $m \times 1$. (A linear transformation would be $\mathbf{BZ}$. The word “affine” pops up because of the additional constant vector $\mathbf{a}$.) Expected values are linear, so

$$
E[\mathbf{W}] = E[\mathbf{a} + \mathbf{BZ}] = \mathbf{a} + \mathbf{B} E[Z].
$$

(3.8)

For the covariance, start by noting that

$$
\mathbf{W} - E[\mathbf{W}] = (\mathbf{a} + \mathbf{BZ}) - (\mathbf{a} + \mathbf{B} E[Z]) = \mathbf{B}(Z - E[Z]),
$$

(3.9)

so that

$$
Cov[\mathbf{W}] = Cov[\mathbf{a} + \mathbf{BZ}] = E[(\mathbf{W} - E[\mathbf{W}])(\mathbf{W} - E[\mathbf{W}])^\prime] = E[\mathbf{B}(Z - E[Z])(Z - E[Z])^\prime \mathbf{B}^\prime] = \mathbf{B} E[(Z - E[Z])(Z - E[Z])^\prime] \mathbf{B}^\prime = \mathbf{B} Cov[Z] \mathbf{B}^\prime.
$$

(3.10)

Compare these formulas to the univariate ones, (3.6).
3.2 Covariances of estimates

Whether estimating \( \mu, \beta \), or \( \lambda' \beta \), it is important to also estimate the corresponding standard errors. (The standard error of an estimator \( \hat{\theta} \) is its standard deviation, \( se(\hat{\theta}) = \sqrt{Var(\hat{\theta})} \).) So far we have assumed that \( E[\epsilon] = \mathbf{0}_n \), but nothing about the variances or covariances of the residuals. For most (if not all) of the course we will assume that the residuals are independent and have equal variances. Later we will also add the assumption of normality. As with all assumptions, one should perform diagnostic checks.

If two variables are independent, then their covariance is 0. Thus under the assumption that \( e_1, \ldots, e_n \) are independent, and have the same variance \( \sigma^2_e = Var(e_i) \), the linear model becomes

\[
\mathbf{y} = \mathbf{X} \beta + \epsilon, \quad E[\epsilon] = \mathbf{0}_n, \quad Cov[\epsilon] = \sigma^2_e \mathbf{I}_n. \tag{3.11}
\]

With this assumption, it is easy to write down the desired covariances. First, for the least squares estimate \( \hat{\mu} \) of \( \mu \), which is the projection \( \hat{\mu} = \mathbf{M} \mathbf{y} \), where \( \mathbf{M} \) is the projection matrix for \( C(\mathbf{x}) \),

\[
Cov(\hat{\mu}) = \text{Cov}[\mathbf{M} \mathbf{y}] = \text{Cov}[\mathbf{M} \epsilon] = \mathbf{M} \text{Cov}[\epsilon] \mathbf{M}' \quad \text{because } \mathbf{X} \beta \text{ is a constant}
= \mathbf{M} \sigma^2_e \mathbf{I}_n \mathbf{M}' \quad \text{by (3.10)}
= \sigma^2_e \mathbf{M} \quad \text{because } \mathbf{M} \mathbf{M}' = \mathbf{M} \tag{3.12}
\]

Now if \( \lambda' \beta \) is estimable, then, as in (2.52), there exists an \( \mathbf{a} \) such that \( E[\mathbf{a}' \mathbf{Y}] = \lambda' \beta \). The variance of this estimator is then

\[
Var(\mathbf{a}' \mathbf{Y}) = \mathbf{a}' \text{Cov}(\mathbf{Y}) \mathbf{a} = \mathbf{a}' \sigma^2_e \mathbf{I}_n \mathbf{a}' = \sigma^2_e \| \mathbf{a} \|^2. \tag{3.13}
\]

This \( \mathbf{a} \) need not be unique, but there is only one \( \mathbf{a} \) that yields the least squares estimate. Below we find the variance of that estimator in general, but first do it in the case that \( \mathbf{X}' \mathbf{X} \) is invertible. In that case, we know that from (2.47), the least squares estimate of \( \beta \) is

\[
\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}, \tag{3.14}
\]

hence

\[
Cov(\hat{\lambda}' \beta) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' [\sigma^2_e \mathbf{I}_n] \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} = \sigma^2_e (\mathbf{X}' \mathbf{X})^{-1}. \tag{3.15}
\]

Then

\[
Var(\lambda' \hat{\beta}) = \sigma^2_e \lambda' (\mathbf{X}' \mathbf{X})^{-1} \lambda. \tag{3.16}
\]

Consider the one-way ANOVA model again, with three groups and 2 observations per group: \( y_{ij} = \mu + \alpha_i + e_{ij}, \ i = 1, 2, 3, \ j = 1, 2 \). Let \( \lambda' \beta = (\alpha_1 + \alpha_2)/2 - \alpha_3 \), so that \( \lambda' = (0, 1/2, 1/2, -1) \). This linear combination is estimable, e.g., \( (y_{11} + y_{21})/2 - y_{31} \) is an unbiased estimate, which has \( \mathbf{a}' = (1/2, 0, 1/2, 0, -1, 0) \). Thus

\[
Var[(Y_{11} + Y_{21})/2 - Y_{31}] = \sigma^2_e \| \mathbf{a} \|^2 = \sigma^2_e ((1/2)^2 + 0^2 + (1/2)^2 + 0^2 + (-1)^2 + 0^2) = \frac{3}{2} \sigma^2_e. \tag{3.17}
\]
This estimate is not the least squares estimate, but we know from Proposition 8 that if \(a'y\) is unbiased, then \(\hat{a}'\hat{y}\) is the least squares estimate. The projection is \(\hat{y}_{ij} = \overline{y}_i\), so the least squares estimate is \((\overline{y}_1 + \overline{y}_2)/2 - \overline{y}_3\), and

\[
Var[(\overline{y}_1 + \overline{y}_2)/2 - \overline{y}_3] = \frac{1}{4} Var[\overline{y}_1] + \frac{1}{4} Var[\overline{y}_2] + Var[\overline{y}_3] = \frac{1}{4} + \frac{1}{4} + \frac{1}{2}\sigma_e^2 = \frac{3}{4} \sigma_e^2, \tag{3.18}
\]

because \(Var[\overline{y}_i] = \sigma_e^2/2\). Note that this variance is half that of the other estimator (3.17).

The vector for the least squares estimate can be seen to be \(a^* = (1/4, 1/4, 1/4, 1/4, -1/2, -1/2)'\), which indeed has \(\|a^*\|^2 = 4(1/4)^2 + 2(1/2)^2 = 3/4\). Now compare the two \(a\)'s:

\[
a = \begin{pmatrix}
\frac{1}{2} & 0 \\
\frac{1}{2} & 0 \\
-1 & 0 
\end{pmatrix}, \quad a^* = \begin{pmatrix}
\frac{1}{4} & -\frac{1}{4} \\
-\frac{1}{4} & -\frac{1}{2}
\end{pmatrix}. \tag{3.19}
\]

What relationship is there? Notice that \(a^*\) is in \(C(X)\), but \(a\) is not. In fact, \(a^*\) is the projection of \(a\) onto \(C(X)\). That is not just a coincidence, as shown in the next result.

**Proposition 9** If \(X\beta\) is estimable, and \(a'y\) is an unbiased estimate, then the least squares estimate is \(\hat{a}'\hat{y}\), where \(\hat{a}\) is the projection of \(a\) onto \(C(X)\).

**Proof.** By Proposition 8, \(a'y\) is the least squares estimate, but \(\hat{y} = My\), where \(M\) is the projection matrix for \(C(X)\), hence

\[
a'y = a'My = (Ma)'
\]

Because for an estimable function, the least squares estimate is unique, \(\hat{a}\) must be unique, that is, any \(a\) that yields an unbiased estimate has the same projection. Also, we noted that the least squares estimate had a lower variance than the alternative in the above example. This feature is also general.

**Definition 12** A linear estimator of a parameter \(\theta\) is the best linear unbiased estimator, or BLUE, if it unbiased, and any other unbiased linear estimator has a variance at least as large.

**Theorem 1** Gauss-Markov. In the model (3.11) with \(\sigma_e^2 > 0\), if \(X\beta\) is estimable, then least squares estimate is the unique BLUE.

**Proof.** For any \(a \in \mathbb{R}^n\),

\[
\|a\|^2 = \|\hat{a} + (a - \hat{a})\|^2 = \|\hat{a}\|^2 + \|a - \hat{a}\|^2, \tag{3.21}
\]
3.3. ESTIMATING THE VARIANCE

because \( \hat{a} \perp (a - \hat{a}) \), hence
\[
\|a\|^2 > \|\hat{a}\|^2 \quad \text{unless} \quad a = \hat{a}.
\] (3.22)

Thus if \( a'y \) is an unbiased estimator of \( A'\beta \) that is not the least squares estimate, then \( a \neq \hat{a} \).
\( \hat{a}'y \) is the least squares estimate by Proposition 9, and
\[
\text{Var}(a'y) = \sigma^2_e \|a\|^2 > \sigma^2_e \|\hat{a}\|^2 = \text{Var}(\hat{a}'y).
\] (3.23)

That is, any unbiased linear estimate has larger variance than the least squares estimate. □

Thus we have established that the least squares estimate is best in terms of variance.

The next section develops estimates of the variance.

3.3 Estimating the variance

In the model (3.11), \( \sigma^2_e \) is typically a parameter to be estimated. Because the residuals have mean zero, \( E(e_i) = 0 \), we have that \( \sigma^2_e = \text{Var}(e_i) = E(e_i^2) \), so that
\[
E\left[ \frac{\sum_{i=1}^n e_i^2}{n} \right] = \sigma^2_e.
\] (3.24)

Unfortunately, we do not observe the actual \( e_i \)'s, because \( e = y - X\beta \), and \( \beta \) is not observed. Thus we have to estimate \( e \), which we can do by plugging in the (or a) least squares estimate of \( \beta \):
\[
\hat{e} = y - X\hat{\beta} = y - \hat{y} = (I_n - M)y,
\] (3.25)

where \( M \) is the projection matrix for \( C(X) \). Note that \( \hat{e} \) is the projection of \( y \) onto \( C(X)^\perp \).
See Proposition 3. Because \( E[y] = E[\hat{y}] = X\beta \), \( E[\hat{e}] = 0_n \), hence as for the \( e_i \)'s, \( \text{Var}[\hat{e}_i] = E[\hat{e}_i^2] \), but unlike the \( e_i \)'s, the \( \hat{e}_i \)'s do not typically have variance \( \sigma^2_e \). Rather, the variance of \( \hat{e}_i \) is \( \sigma^2_e \) times the \( i^{th} \) diagonal of \( (I_n - M) \). Then
\[
E[\sum_{i=1}^n \hat{e}_i^2] = \sum_{i=1}^n \text{Var}[\hat{e}_i] = \sigma^2_e \times \text{[sum of diagonals of} (I_n - M)] = \sigma^2_e \text{trace}(I_n - M).
\] (3.26)

Thus an unbiased estimator of \( \sigma^2_e \) is
\[
\hat{\sigma}_e^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{\text{trace}(I_n - M)} = \frac{\|\hat{e}\|^2}{\text{trace}(I_n - M)}.
\] (3.27)

It is easy enough to calculate the trace of a matrix, but the trace of a projection matrix is actually the rank of the corresponding vector space.

**Proposition 10** The rank of a vector space \( \mathcal{M} \) is \( \text{trace}(\mathcal{M}) \), where \( \mathcal{M} \) is the projection matrix for \( \mathcal{M} \).
**Proof.** Suppose \( \text{rank}(\mathcal{M}) = p \), and let \( \{ \mathbf{z}_1, \ldots, \mathbf{z}_p \} \) be a basis for \( \mathcal{M} \). Then with \( \mathbf{X} = (\mathbf{z}_1, \ldots, \mathbf{z}_p) \) (so that \( \mathcal{M} = C(\mathbf{X}) \)), the projection matrix is \( \mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \). The \( \mathbf{X}'\mathbf{X} \) is invertible because the columns of \( \mathbf{X} \) are linearly independent. But then using the fact that \( \text{trace}(AB) = \text{trace}(BA) \),

\[
\text{trace}(\mathbf{M}) = \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \text{trace}((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})) = \text{trace}(I_p) = p,
\]

(3.28)
since \( \mathbf{X}'\mathbf{X} \) is \( p \times p \).

Thus we have from (3.27) that

\[
\hat{\sigma}^2_e = \frac{||\hat{\mathbf{e}}||^2}{n - p}, \quad p = \text{rank}(C(\mathbf{X})).
\]

(3.29)

Now everything is set to estimate the standard error of estimates.

### 3.4 Example: Leprosy

#### 3.4.1 Without covariate

Continue the example from Section 2.2.3. Start with the model without the covariate, (2.59), and consider the first contrast \( \alpha_1 - \alpha_2 \), which compares the two drugs. From (2.60), the least squares estimate is

\[
\hat{\alpha}_1 - \hat{\alpha}_2 = \overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2 = -0.8.
\]

(3.30)

This estimate is \( \hat{\mathbf{a}}'\mathbf{y} \) with \( \mathbf{a} = (1/10, \ldots, 1/10, -1/10, \ldots, -1/10, 0, \ldots, 0)' \), where each value is repeated ten times. Then

\[
\text{Var}[\hat{\alpha}_1 - \hat{\alpha}_2] = \sigma^2_e ||\hat{\mathbf{a}}||^2 = \sigma^2_e (10(\frac{1}{10})^2 + 10(\frac{1}{10})^2) = \frac{1}{5} \sigma^2_e.
\]

(3.31)

To estimate \( \sigma^2_e \), we need \( ||\hat{\mathbf{e}}||^2 \) and \( p \), the rank of \( C(\mathbf{X}) \). The columns of \( \mathbf{X} \) in (2.59) are not linearly independent, because the first, \( \mathbf{1}_n \), is the sum of the other three. Thus we can eliminate the first, and the remaining are linearly independent, so \( p = 3 \). The \( \hat{\mathbf{y}} = (\overline{\mathbf{y}}_1, \ldots, \overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2, \ldots, \overline{\mathbf{y}}_3, \ldots, \overline{\mathbf{y}}_3)' \), so that

\[
\hat{\sigma}^2_e = \frac{||\hat{\mathbf{e}}||^2}{n - p} = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n - p} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{10}(y_{ij} - \overline{y}_i)^2}{n - p} = \frac{995.1}{30 - 3} = 36.856.
\]

(3.32)

Now the estimate of the standard error of the estimate \( \hat{\alpha}_1 - \hat{\alpha}_2 \) is \( \sqrt{(1/5)\hat{\sigma}^2_e} = \sqrt{36.856/5} = 2.715 \). Thus, with the estimate given in (3.30), we have an approximate 95% confidence interval for \( \alpha_1 - \alpha_2 \) being

\[
(\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2 \text{ se}) = (-0.8 \pm 2(2.715)) = (-6.23, 4.63).
\]

(3.33)
This interval is fairly wide, containing 0, which suggests that there is no evidence of a difference between the two drugs. Equivalently, we could look at the approximate z-statistic,

\[
\frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\text{se}} = \frac{-0.8}{2.715} = -0.295,
\]

and note that it is well less than 2 in absolute value. (Later, we will refine these inferences by replacing the “2” with a t-value, at least when assuming normality of the residuals.)

For the contrast comparing the average of the two drugs to the control, \((\alpha_1 + \alpha_2)/2 - \alpha_3\), we again have from (2.60) that

\[
(\alpha_1 + \alpha_2) - \alpha_3 = (\bar{y}_1 + \bar{y}_2)/2 - \bar{y}_3 = (5.3 + 6.1)/2 - 12.3 = -6.6.
\]

The \(a\) for this estimate is \(a = (1/20, \ldots , 1/20, -1/10, \ldots , -1/10)'\), where there are 20 1/20’s and 10 1/10’s. Thus \(\|a\|^2 = 20/20^2 + 10/10^2 = 0.15\), and \(se = \sqrt{36.856 \times 0.15} = 2.35\), and the approximate confidence interval is

\[
((\hat{\alpha}_1 + \hat{\alpha}_2) - \alpha_3 \pm 2 \text{se}) = (-6.6 \pm 2(2.35)) = (-11.3, -1.9).
\]

This interval is entirely below 0, which suggests that the drugs are effective relative to the placebo. (Or look at \(z = -6.6/2.35 = 2.81\).)

### 3.4.2 With covariate

The hope is that by using the before measurements, the parameters can be estimated more accurately. In this section we use model (2.61), \(y_{ij} = \mu + \alpha_i + \gamma z_{ij} + e_{ij}\). Because the patients were randomly allocated to the three groups, and the \(z_{ij}\)’s were measured before treatment, the \(\mu\) and \(\alpha_i\)’s have the same interpretation in both models (with and without covariates). However, the \(\sigma^2_e\) is not the same.

The projections now are, from (2.67), \(\hat{y}_{ij} = \bar{y}_i + d(z_{ij} - \bar{z}_i)\), where \(d = \hat{\gamma} = 0.987\). The dimension of \(X\) is now \(p = 4\), because, after removing the \(1_n\) vector, the remaining four are linearly independent. (They would not be linearly independent if the \(z_{ij}\)’s were the same within each group, but clearly they are not.) Then

\[
\hat{\sigma}^2_e = \frac{\sum_{i=1}^3 \sum_{j=1}^{10} (y_{ij} - \bar{y}_i - 0.987 (z_{ij} - \bar{z}_i))^2}{30 - 26} = 16.046.
\]

Note how much smaller this estimate is than the 36.856 in (3.32).

If we were to follow the procedure in the previous section, we would need to find the \(a\)’s for the estimates, then their lengths, in order to find the standard errors. Instead, we will go the other route, using (3.16) to find the variances: \(\hat{\sigma}^2_e \Lambda' (X'X)^{-1} \Lambda\). In order to proceed, we need \(X'X\) to be invertible, which at present is not true. We need to place a restriction on the parameters so that the matrix is invertible, but also in such a way that the meaning
of the contrasts is the same. One way is to simply set \( \mu = 0 \), so that

\[
y = X\beta + \varepsilon = \begin{pmatrix} 1 & 0 & 0 & 11 \\ 1 & 0 & 0 & 8 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 3 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 6 \\ 0 & 1 & 0 & 6 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 15 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 16 \\ 0 & 0 & 1 & 13 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 12 \\ \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \gamma \end{pmatrix} + \varepsilon. \tag{3.38}
\]

Now

\[
X'X = \begin{pmatrix} 10 & 0 & 0 & 93 \\ 0 & 10 & 0 & 100 \\ 0 & 0 & 10 & 129 \\ 93 & 100 & 129 & 4122 \end{pmatrix} \quad \text{and} \quad (X'X)^{-1} = \begin{pmatrix} 0.2459 & 0.1568 & 0.2023 & -0.0157 \\ 0.1568 & 0.2686 & 0.2175 & -0.0169 \\ 0.2023 & 0.2175 & 0.3806 & -0.0218 \\ -0.0157 & -0.0169 & -0.0218 & 0.0017 \end{pmatrix}. \tag{3.39}
\]

For \( \alpha_1 - \alpha_2 \), we have from (2.69) the estimate \(-0.109\). For this \( \bar{\beta}, \bar{\lambda} = (1, -1, 0, 0)' \), hence

\[
Var(\alpha_1 - \alpha_2) = \sigma_e^2 \Delta'(X'X)^{-1} \Delta = \sigma_e^2 (0.2459 - 2 \times 0.1568 + 0.2686) = \sigma_e^2 (0.201). \tag{3.40}
\]

Using the estimate in (3.37), we have that

\[
se(\alpha_1 - \alpha_2) = \sqrt{16.046 \times 0.201} = 1.796, \tag{3.41}
\]

hence

\[
z = \frac{-1.09}{1.796} = -0.61, \tag{3.42}
\]

which is again quite small, showing no evidence of a difference between drugs.

For the drug versus control contrast, from (2.70), we have the estimate \((\alpha_1 + \alpha_2)/2 - \alpha_3 = -3.392\). Now \( \bar{\lambda} = (1/2, 1/2, -1, 0)' \), hence

\[
se = \sqrt{16.046 \times (1/2, 1/2, -1, 0)(X'X)^{-1}(1/2, 1/2, -1, 0)'} = \sqrt{16.046 \times 0.1678} = 1.641. \tag{3.43}
\]

Compare this \( se \) to that without the covariate, 2.35. It is substantially smaller, showing that the covariate does help improve accuracy in this example.
3.4. EXAMPLE: LEPROSY

Now

\[ z = \frac{-3.392}{1.641} = -2.07. \]  \hspace{1cm} (3.44)

This is marginally significant, suggesting there may be a drug effect. However, it is a bit smaller (in absolute value) than the \( z = -2.81 \) calculated without covariates. Thus the covariate is also important in adjusting for the fact that the controls had somewhat less healthy patients initially.
Chapter 4

Distributions: Normal, $\chi^2$, and $t$

The previous chapter presented the basic estimates of parameters in the linear models. In this chapter we add the assumption of normality to the residuals, which allows us to provide more formal confidence intervals and hypothesis tests. The central distribution is the multivariate normal, from which the $\chi^2$, $t$, and $F$ are derived.

4.1 Multivariate Normal

The standard normal distribution for random variable $Z$ is the familiar bell-shaped curve. The density is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2}. \quad (4.1)$$

The mean is 0 and variance is 1. The more general normal distribution, with arbitrary mean $\mu$ and variance $\sigma^2$ is written $X \sim N(\mu, \sigma^2)$, and has density

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4.2)$$

when $\sigma^2 > 0$. If $\sigma^2 = 0$, then $X$ equals $\mu$ with probability 1, that is, $X$ is essentially a constant.

The normal works nicely with affine transformations. That is,

$$X \sim N(\mu, \sigma^2) \implies a + bX \sim N(a + b\mu, b^2\sigma^2). \quad (4.3)$$

We already know that $E[a + bX] = a + b\mu$ and $Var[a + bX] = b^2\sigma^2$, but the added property in (4.3) is that if $X$ is normal, so is $a + bX$. It is not hard to show using the change-of-variable formula.

We will assume that the residuals $e_i$ are independent $N(0, \sigma^2_e)$ random variables, which will imply that the $y_i$’s are independent normals as well. We also need the distributions of the vectors such as $\hat{y}$, $\hat{e}$ and $\hat{\beta}$. It will turn out that under the assumptions, the individual components of these vectors are normal, but they are typically not independent (because
CHAPTER 4. DISTRIBUTIONS: NORMAL, \( \chi^2 \), AND T

their covariance matrices are not diagonal). Thus we need a distribution for the entire vector. This distribution is the multivariate normal, which we now define as an affine transformation of independent standard normals.

**Definition 13** An \( n \times 1 \) vector \( \mathbf{W} \) has a multivariate normal distribution if for some \( n \times 1 \) vector \( \mathbf{a} \) and \( n \times q \) matrix \( \mathbf{B} \),

\[
\mathbf{W} = \mathbf{a} + \mathbf{BZ},
\]

where \( \mathbf{Z} = (Z_1, \ldots, Z_q)' \), the \( Z_i \)'s being independent standard normal random variables.

If (4.4) holds, then \( \mathbf{W} \) is said to be multivariate normal with mean \( \mu \) and covariance \( \Sigma \), where \( \mu = \mathbf{a} \) and \( \Sigma = \mathbf{BB}' \), written

\[
\mathbf{W} \sim N_n(\mu, \Sigma). \tag{4.5}
\]

The elements of the vector \( \mathbf{Z} \) in the definition all have \( E[Z_i] = 0 \) and \( \text{Var}[Z_i] = 1 \), and they are independent, hence \( E[\mathbf{Z}] = \mathbf{0}_q \) and \( \text{Cov}[\mathbf{Z}] = \mathbf{I}_q \). Thus, as in (3.8), that \( \mu = E[\mathbf{W}] = E[\mathbf{a} + \mathbf{BZ}] = \mathbf{a} \) and \( \Sigma = \text{Cov}[\mathbf{W}] = \text{Cov}[\mathbf{a} + \mathbf{BZ}] = \mathbf{BB}' \) follows from (3.8) and (3.10). The added fillip is the multivariate normality. Note that by taking \( \mathbf{a} = \mathbf{0}_q \) and \( \mathbf{B} = \mathbf{I}_q \), we have that

\[
\mathbf{Z} \sim N_q(\mathbf{0}_q, \mathbf{I}_q). \tag{4.6}
\]

The definition presumes that the distribution is well-defined. That is, two different \( \mathbf{B} \)'s could yield the same \( \Sigma \), so how can one be sure the distributions are the same? For example, suppose \( n = 2 \), and consider the two matrices

\[
\mathbf{B}_1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{B}_2 = \begin{pmatrix} \sqrt{\frac{3}{2}} & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{2} \end{pmatrix}. \tag{4.7}
\]

Certainly

\[
\mathbf{B}_1 \mathbf{B}'_1 = \mathbf{B}_2 \mathbf{B}'_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \mathbf{\Sigma}, \tag{4.8}
\]

but is it clear that

\[
\mathbf{a} + \mathbf{B}_1 \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} \quad \text{and} \quad \mathbf{a} + \mathbf{B}_2 \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \tag{4.9}
\]

have the same distribution? Not only are they different linear combinations, but they are linear combinations of different numbers of standard normals. So it certainly is not obvious, but they do have the same distribution. This result depends on the normality of the \( Z_i \)'s. It can be proved using moment generating functions. Similar results do not hold for other \( Z_i \)'s, e.g., Cauchy or exponential.

The next question is, “What \( \mu \)'s and \( \Sigma \)'s are valid?” Any \( \mu \in \mathbb{R}^n \) is possible, since \( \mathbf{a} \) is arbitrary. But the possible matrices \( \Sigma \) are restricted. For one, \( \mathbf{BB}' \) is symmetric for any \( \mathbf{B} \), so \( \Sigma \) must be symmetric, but we already knew that because all covariance matrices are symmetric. We also need \( \Sigma \) to be nonnegative definite, which we deal with next.
4.1. MULTIVARIATE NORMAL

Definition 14 A symmetric \( n \times n \) matrix \( A \) is nonnegative definite if
\[
x'Ax \geq 0 \quad \text{for all } x \in \mathbb{R}^n. \tag{4.10}
\]
The matrix is positive definite if
\[
x'Ax > 0 \quad \text{for all } x \in \mathbb{R}^n, x \neq 0. \tag{4.11}
\]

All covariance matrices are nonnegative definite. To see this fact, suppose \( \text{Cov}(Y) = \Sigma \). Then for any vector \( x \) (of the right dimension),
\[
x'\Sigma x = \text{Var}(x'Y) \geq 0, \tag{4.12}
\]
because variances are always nonnegative. Not all covariances are positive definite, though. For example, we know that for our models, \( \text{Cov}(\hat{y}) = \sigma^2 M \), where \( M \) is the projection matrix onto \( M \). Now for any \( x \), because \( M \) is symmetric and idempotent,
\[
x'Mx = x'M'Mx = \hat{x}'\hat{x} = \|\hat{x}\|^2. \tag{4.13}
\]
Certainly \( \|\hat{x}\|^2 \geq 0 \), but is it always strictly positive? No. If \( x \perp M \), then \( \hat{x} = 0_n \). Thus if there are any vectors besides \( 0_n \) that are orthogonal to \( M \), then \( M \) is not positive definite. There always are such vectors, unless \( M = \mathbb{R}^n \), in which case \( M = I_n \).

If \( \text{Cov}(Y) = \Sigma \) is not positive definite, then there is a linear combination of the \( Y_i \)’s, \( x'Y \), that has variance 0. That is, \( x'Y \) is essentially a constant.

The nonnegative definiteness of covariance matrices implies the covariance inequality that follows, which in turn implies that the correlation between any two random variables is in the range \([-1, 1]\).

Lemma 2 Cauchy-Schwarz Inequality. For any two random variables \( Y_1 \) and \( Y_2 \) with finite variances,
\[
\text{Cov}(Y_1, Y_2) \leq \sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}. \tag{4.14}
\]
Thus, if the variances are positive,
\[
-1 \leq \text{Corr}(Y_1, Y_2) \leq 1, \quad \text{where } \text{Corr}(Y_1, Y_2) = \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}}. \tag{4.15}
\]
is the correlation between \( Y_1 \) and \( Y_2 \).

Proof. Let \( \Sigma = \text{Cov}((Y_1, Y_2)') \). Because \( \Sigma \) is nonnegative definite, \( x'\Sigma x \geq 0 \) for any \( x \). Two such \( x \)’s are \( (\sigma_{22}, -\sigma_{12})' \) and \( (-\sigma_{12}, \sigma_{11})' \), which yield the inequalities
\[
\sigma_{22}(\sigma_{11}\sigma_{22} - \sigma_{12}^2) \geq 0 \quad \text{and} \quad \sigma_{11}(\sigma_{11}\sigma_{22} - \sigma_{12}^2) \geq 0, \tag{4.16}
\]
respectively. If either $\sigma_{11}$ or $\sigma_{22}$ is positive, then at least one of the equations shows that $\sigma_{12}^2 \leq \sigma_{11}\sigma_{22}$, which implies (4.14). If $\sigma_{11} = \sigma_{22} = 0$, then it is easy to see that $\sigma_{12} = 0$, e.g., by looking at $(1,1)\Sigma(1,1)' = 2\sigma_{12} \geq 0$ and $(1,-1)\Sigma(1,-1)' = -2\sigma_{12} \geq 0$, which imply that $\sigma_{12} = 0$.

Back to matrices $BB'$. All such matrices are nonnegative definite, because $\underline{x}'(BB')\underline{x} = (B'\underline{x})'B'\underline{x} = \|B'\underline{x}\|^2 \geq 0$. Thus the $\Sigma$ in Definition 13 must be nonnegative definite. But again, all covariance matrices are nonnegative definite. Thus the question is, “Are all nonnegative definite symmetric matrices equal to $BB'$ for some $B$?” The answer is, “Yes.” There are many possibilities, but the next subsection shows that there always exists a lower-triangular matrix $L$ with $\Sigma = LL'$. Note that if $L$ works, so does $L\Gamma$ for any $n \times n$ orthogonal matrix $\Gamma$.

### 4.1.1 Cholesky decomposition

There are thousands of matrix decompositions. One that exhibits a $B$ as above is the Cholesky decomposition, for which the matrix $B$ is lower triangular.

**Definition 15** An $n \times n$ matrix $L$ is **lower triangular** if $l_{ij} = 0$ for $i < j$:

$$
L = \begin{pmatrix}
  l_{11} & 0 & 0 & \cdots & 0 \\
  l_{21} & l_{22} & 0 & \cdots & 0 \\
  l_{31} & l_{32} & l_{33} & \cdots & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn}
\end{pmatrix}.
$$

(4.17)

Some properties:

1. The product of two lower triangular matrices is also lower triangular.

2. The lower triangular matrix $L$ is invertible if and only if the diagonals are nonzero, $l_{ii} \neq 0$. If it exists, the inverse is also lower triangular, and the diagonals are $1/l_{ii}$.

The main property we need is the following.

**Proposition 11** If $\Sigma$ is symmetric and nonnegative definite, then there exists a lower triangular matrix with diagonals $l_{ii} \geq 0$ such that

$$
\Sigma = LL'.
$$

(4.18)

The $L$ is unique if $\Sigma$ is positive definite. In addition, $\Sigma$ is positive definite if and only if the $L$ in (4.18) has all diagonals $l_{ii} > 0$. 

4.1. MULTIVARIATE NORMAL

Proof. We will use induction on \( n \). The first step is to prove it works for \( n = 1 \). In that case \( \Sigma = \sigma^2 \) and \( L = l \), so the equation (4.18) is \( \sigma^2 = l^2 \), which is solved by taking \( l = +\sqrt{\sigma^2} \). This \( l \) is nonnegative, and positive if and only if \( \sigma^2 > 0 \).

Now assume the decomposition works for any \((n - 1) \times (n - 1)\) symmetric nonnegative definite matrix, and write the \( n \times n \) matrix \( \Sigma \) as

\[
\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \Sigma_{22} \end{pmatrix},
\]

(4.19)

where \( \Sigma_{22} \) is \((n - 1) \times (n - 1)\), and \( \sigma_{12} \) is \((n - 1) \times 1\). Partion the lower-triangular matrix \( L \) similarly, that is,

\[
L = \begin{pmatrix} l_{11} & 0_{n-1} \\ l_{12} & L_{22} \end{pmatrix},
\]

(4.20)

where \( L_{22} \) is an \((n - 1) \times (n - 1)\) lower-triangular matrix, and \( l_{12} \) is \((n - 1) \times 1\). We want to find such an \( L \) that satisfies (4.18), which translates to the equations

\[
\begin{align*}
\sigma_{11} &= l_{11}^2 \\
\sigma_{12} &= l_{11} l_{12} \\
\Sigma_{22} &= L_{22} L_{22}' + l_{12} l_{12}'.
\end{align*}
\]

(4.21)

It is easy to see that \( l_{11} = +\sqrt{\sigma_{11}} \). To solve for \( l_{12} \), we have to know whether \( \sigma_{11} \), hence \( l_{11} \), is positive. So there are two cases.

- \( \sigma_{11} > 0 \): Then \( l_{11} > 0 \), and using the second equation in (4.21), the unique solution is \( l_{12} = (1/l_{11}) \sigma_{12} = (1/\sqrt{\sigma_{11}}) \sigma_{12} \).

- \( \sigma_{11} = 0 \): By the covariance inequality in Lemma 2, the \( \sigma_{11} = 0 \) implies that the covariances between the first variable and the others are all 0, that is, \( \sigma_{12} = 0_{n-1} \). Thus in this case we can take \( l_{11} = 0 \) and \( l_{12} = 0_{n-1} \) in the second equation of (4.21), although any \( l_{12} \) will work.

Now to solve for \( L_{22} \). If \( \sigma_{11} = 0 \), then as above, \( \sigma_{12} = 0_{n-1} \), so that by the induction hypothesis we have that there does exist a lower-triangular \( L_{22} \) with \( \Sigma_{22} = L_{22} L_{22}' \). If \( \sigma_{11} > 0 \), then the third line in (4.21) becomes

\[
\Sigma_{22} - \frac{1}{\sigma_{11}} \sigma_{12} \sigma_{12}' = L_{22} L_{22}'.
\]

(4.22)

By the induction hypothesis, that equation can be solved if the left-hand side is nonnegative definite. For any \( m \times n \) matrix \( B \), \( B \Sigma B' \) is also symmetric and nonnegative definite. (Why?) Consider the \((n - 1) \times n\) matrix

\[
B = (-\frac{1}{\sigma_{11}} \sigma_{12}, I_{n-1}).
\]

(4.23)
Multiplying out shows that
\[ B \Sigma B' = \Sigma_{22} - \frac{1}{\sigma_{11}} \Sigma_{12} \Sigma_{12}'. \]  
(4.24)

Therefore (4.22) can be solved with a lower-triangular \( L_{22} \), which by induction proves that any nonnegative definite matrix \( \Sigma \) can be written as in (4.18) with a lower-triangular \( L \).

There are a couple of other parts to this proposition. We won’t give all the details, but suppose \( \Sigma \) is positive definite. Then \( \sigma_{11} > 0 \), and the above proof shows that \( l_{11} \) is positive and \( l_{12} \) is uniquely determined. Also, \( \Sigma_{22} \) will be positive definite, so that by induction the diagonals \( l_{22}, \ldots, l_{nn} \) will also be positive, and the off-diagonals of \( L_{22} \) will be unique.

\[ \square \]

### 4.2 Some properties of the multivariate normal

The multivariate normal has many properties useful for analyzing linear models. Three of them follow.

**Proposition 12**

1. **Affine transformations.** If \( \underline{W} \sim N_n(\underline{\mu}, \Sigma) \), \( \underline{\xi} \) is \( m \times 1 \) and \( \underline{D} \) is \( m \times n \), then
\[ \underline{\xi} + \underline{D} \underline{W} \sim N_m(\underline{\xi} + \underline{D} \underline{\mu}, \underline{D} \Sigma \underline{D}'. \]  
(4.25)

2. **Marginals.** Suppose \( \underline{W} \sim N_n(\underline{\mu}, \Sigma) \) is partitioned as
\[ \underline{W} = \begin{pmatrix} \underline{W}_1 \\ \underline{W}_2 \end{pmatrix}, \]  
(4.26)

where \( \underline{W}_1 \) is \( n_1 \times 1 \) and \( \underline{W}_2 \) is \( n_2 \times 1 \), and the parameters are similarly partitioned:
\[ \underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \]  
(4.27)

where \( \underline{\mu}_i \) is \( n_i \times 1 \) and \( \Sigma_{ij} \) is \( n_i \times n_j \). Then
\[ \underline{W}_1 \sim N_{n_1}(\underline{\mu}_1, \Sigma_{11}), \]  
(4.28)

and similarly for \( \underline{W}_2 \). In particular, the individual components have \( W_i \sim N(\mu_i, \sigma_{ii}) \), where \( \sigma_{ii} \) is the \( i \)th diagonal of \( \Sigma \).

3. **Independence.** Partitioning \( \underline{W} \) as in part 2, \( \underline{W}_1 \) and \( \underline{W}_2 \) are independent if and only if \( \Sigma_{12} = 0 \). In particular, \( W_i \) and \( W_j \) are independent if and only if \( \sigma_{ij} = 0 \).

Part 1 follows from the Definition 13 of multivariate normality. That is, if \( \underline{W} \sim N_n(\underline{\mu}, \Sigma) \), then \( \underline{W} = \underline{\mu} + \underline{B} \underline{Z} \) where \( \underline{B} \underline{B}' = \Sigma \) and \( \underline{Z} \) is a vector of independent standard normals. Thus
\[ \underline{\xi} + \underline{D} \underline{W} = (\underline{\xi} + \underline{D} \underline{\mu}) + (\underline{D} \underline{B}) \underline{Z}, \]  
(4.29)
which by definition is \( N_m(\xi + D\mu, (DB)(DB)' ) \), and \( (DB)(DB)' = D\Sigma D' = D\Sigma' \). Part 2 follow from part 1 by taking \( \xi = 0_n \) and \( D = (I_n, 0) \), where the 0 is \( n_2 \times n \).

To see part 3, let \( B_1 \) and \( B_2 \) be matrices so that \( B_1B_1' = \Sigma_{11} \) and \( B_2B_2' = \Sigma_{22} \). Consider

\[
W^* = \begin{pmatrix} W_1^* \\ W_2^* \end{pmatrix} = \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},
\]

where the \( Z = (Z_1, Z_2)' \) is a vector of independent standard normals. Now \( W_1^* = \mu_1 + B_1 Z_1 \) and \( W_2^* = \mu_2 + B_2 Z_2 \) are independent because \( Z_1 \) and \( Z_2 \) are. Also, \( W^* \sim N_n(\mu, \Sigma) \), where

\[
\Sigma = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}' = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.
\]

That is, \( W^* \) has the same distribution as \( W \) when \( \Sigma_{12} = 0 \), hence \( W_1 \) and \( W_2 \) (as \( W_1^* \) and \( W_2^* \)) are independent.

### 4.3 Distribution of linear estimates

We now add the normality assumption to the residuals, which allows development of more distributional results. As for all assumptions, in practice these are wrong, hence one should check to see if they are at least reasonable. The assumption is that the \( e_i \)'s are independent \( N(0, \sigma^2_e) \), which means that

\[
y = X\beta + \xi, \quad \xi \sim N_n(0_n, \sigma^2_e I_n),
\]

or, equivalently,

\[
y \sim N_n(X\beta, \sigma^2_e I_n).
\]

The distribution of linear estimates then follows easily from previous work. Thus if \( X' \beta \) is estimable, and \( a'y \) is an unbiased estimate, then because \( a'y \) is an affine transformation of a multivariate normal, part 1 of Proposition 12 shows that

\[
a'y \sim N(a'\beta, \sigma^2_e \|a\|^2).
\]

More generally, if \( XX \) is invertible, then \( \beta = (XX)^{-1}X'y \), hence

\[
\beta \sim N_p(\beta, \sigma^2_e (XX)^{-1}).
\]

Projections are linear transformations as well, \( \tilde{y} = My \) and \( \tilde{\xi} = (I_n - M)y \), hence multivariate normal. An important result is that two projections on orthogonal spaces, such as \( \tilde{y} \) and \( \tilde{\xi} \), are independent. To show this result, consider the \( 2n \times 1 \) vector that strings out the two projections, \( (\tilde{y}', \tilde{\xi}')' \), which is a linear transformation of \( y \):

\[
\begin{pmatrix} \tilde{y} \\ \tilde{\xi} \end{pmatrix} = \begin{pmatrix} M \\ (I_n - M) \end{pmatrix} y.
\]
Then
\[
\text{Cov} \left( \left( \begin{array}{c} \hat{y} \\ \tilde{e} \end{array} \right) \right) = \left( \begin{array}{c} M \\ (I_n - M) \end{array} \right) \sigma_e^2 \left( \begin{array}{c} M \\ (I_n - M) \end{array} \right)^t
\]
\[
= \sigma_e^2 \left( \begin{array}{cc} MM' & M(I_n - M)' \\ (I_n - M)M' & (I_n - M)(I_n - M)' \end{array} \right)
\]
\[
= \sigma_e^2 \left( \begin{array}{cc} M & 0 \\ 0 & I_n - M \end{array} \right),
\]
(4.37)
because \( M \) is idempotent and symmetric (so, e.g., \( M(I_n - M)' = M - MM = M - M = 0 \)).
Thus the covariance between \( \hat{y} \) and \( \tilde{e} \) is 0, which means that \( \hat{y} \) and \( \tilde{e} \) are independent by part 3 of Proposition 12. (If the residuals are not normal, then these projections will not be independent in general, but just uncorrelated.)

We will need the next fact for confidence intervals and confidence regions.

**Proposition 13** Under model (4.32), if \( X'X \) is invertible, \( \hat{\beta} \) and \( \hat{\varepsilon} \) are independent.

The proof is easy once you realize that \( \hat{\beta} \) is a function of \( \hat{y} \), which follows either by recalling that \( \hat{\beta} \) is found by satisfying \( \hat{y} = X\hat{\beta} \), or by using the formula \( \hat{\beta} = (X'X)^{-1}X'y \), and noting that \( X'M = X' \), hence \( (X'X)^{-1}X'y = (X'X)^{-1}X'My = (X'X)^{-1}X'M\hat{y} \), or just writing it out:
\[
(X'X)^{-1}X'y = (X'X)^{-1}X'(X(X'X)^{-1}X'y) = (X'X)^{-1}X'y = \hat{\beta}.
\]
(4.38)

### 4.4 Chi-squares

Under the normality assumption, if \( a'y \) is an unbiased estimate of \( \lambda'\hat{\beta} \),
\[
\frac{a'y - \lambda'\hat{\beta}}{\sigma_{\varepsilon}\|a\|} \sim N(0, 1).
\]
(4.39)

To derive an exact confidence interval for \( \lambda'\hat{\beta} \), start with
\[
P \left[ -z_{\alpha/2} < \frac{a'y - \lambda'\hat{\beta}}{\sigma_{\varepsilon}\|a\|} < z_{\alpha/2} \right] = 1 - \alpha,
\]
(4.40)
where \( z_{\alpha/2} \) is the upper \( (\alpha/2)^{th} \) cutoff point for the \( N(0, 1) \), i.e., \( P[-z_{\alpha/2} < N(0, 1) < z_{\alpha/2}] = 1 - \alpha \). Then rewriting the inequalities in (4.40) so that \( \lambda'\hat{\beta} \) is in the center shows that an exact 100 \( \times (1 - \alpha)\)% confidence interval for \( \lambda'\hat{\beta} \) is
\[
a'y \pm z_{\alpha/2}\sigma_{\varepsilon}\|a\|.
\]
(4.41)
Unfortunately, the \( \sigma_{\varepsilon} \) is still unknown, so we must estimate it, which then destroys the exact normality in (4.39). It turns out that Student’s \( t \) is the correct way to adjust for this estimation, but first we need to obtain the distribution of \( \sigma_{\varepsilon}^2 \). Which brings us to the \( \chi^2 \) (chi-squared) distribution.
Definition 16 If $Z \sim N_\nu(0, I_\nu)$, then $\|Z\|^2$ is distributed as a chi-squared random variable with $\nu$ degrees of freedom, written

$$\|Z\|^2 \sim \chi^2_\nu. \quad (4.42)$$

A more familiar but equivalent way to write the definition is $Z_1^2 + \cdots + Z_\nu^2 \sim \chi^2_\nu$, where the $Z_i$’s are independent $N(0, 1)$’s. Because $E[Z_i] = 0$, $E[Z_i^2] = Var(Z_i) = 1$, so that $E[\chi^2_\nu] = \nu$.

The chi-squared distribution is commonly used for the distributions of quadratic forms of normals, where a quadratic form is $(y - c)'D^{-1}(y - c)$ for some vector $c$ and symmetric matrix $D$. E.g., the $\|Z\|^2$ in (4.42) is a quadratic form with $c = 0$ and $D = I_\nu$. Not all quadratic forms are chi-squared, by any means, but two popular ones are given next. The first is useful for simultaneous confidence intervals on $\lambda'\beta$’s, and the second for the squared norms of projections, such as $\|\hat{e}\|^2$.

Proposition 14 1. If $Y \sim N_n(\mu, \Sigma)$, and $\Sigma$ is invertible, then

$$(Y - \mu)'\Sigma^{-1}(Y - \mu) \sim \chi^2_n. \quad (4.43)$$

2. If $W \sim N_n(0_n, M)$, where $M$ is (symmetric and) idempotent, then

$$\|W\|^2 \sim \chi^2_{\text{trace}(M)}. \quad (4.44)$$

Proof. 1. Use the Cholesky decomposition, Proposition 11, to find lower triangular $L$ such that $\Sigma = LL'$. Because $\Sigma$ is positive definite, the diagonals of $L$ are positive, hence $L^{-1}$ exists. By part 1 of Proposition 12,

$$W = L^{-1}Y - L^{-1}\mu \sim N_n(L^{-1}\mu - L^{-1}\mu, L^{-1}LL'(L^{-1})') = N_n(0_n, I_n), \quad (4.45)$$

so that

$$\|W\|^2 = W'W = (Y - \mu)'(L^{-1})'(Y - \mu) = (Y - \mu)'\Sigma^{-1}(Y - \mu), \quad (4.46)$$

because $(L^{-1})'L^{-1} = (LL')^{-1} = \Sigma^{-1}$. But by Definition 16 of the chi-square, $\|W\|^2 \sim \chi^2_n$, hence (4.43) holds.

2. If $M$ is symmetric and idempotent, then it is a projection matrix for some vector space $M (= C(M))$. Suppose the rank of $M$ is $p$, and let $x_1, \ldots, x_p$ be an orthonormal basis, which is an orthogonal basis where each vector has length 1, $\|x_i\| = 1$. (There always is one: take any basis, and use Gram-Schmidt to obtain an orthogonal basis. Then divide each vector by its length.) Then

$$X \equiv (x_1, \ldots, x_p) \text{ is } n \times p, \text{ and } X'X = I_p, \quad (4.47)$$

because $x_i'x_j = 0$ if $i \neq j$, $x_i'x_i = 1$, hence

$$M = X(X'X)^{-1}X' = XX'. \quad (4.48)$$
Now
\[ Z = X'W \sim N_p(0, X'MX) = N_p(0, X'XX'X) = N_p(0, I_p) \implies \|Z\|^2 \sim \chi^2_p. \quad (4.49) \]
Equation (4.48), and the fact that \( M \) is idempotent, shows that
\[ \|Z\|^2 = \|X'W\|^2 = W'XX'W = W'MW = (MW)'(MW). \quad (4.50) \]
Finally, \( MW \sim N_n(0, M) \), so \( MW \) and \( W \) have the same distribution, and (4.44) holds, because \( p = \text{trace}(M) \). \( \square \)

4.4.1 Distribution of \( \|\hat{e}\|^2 \)

In the model (4.32), we have that \( \hat{e} \sim N_n(0, \sigma^2_e(I_n - M)) \), hence \( (1/\sigma_e)\hat{e} \sim N_n(0, I_n - M) \). Thus by (4.44), since \( \text{trace}(I_n - M) = n - p \),
\[ \frac{1}{\sigma^2_e} \|\hat{e}\|^2 \sim \chi^2_{n-p}, \quad (4.51) \]
hence
\[ \|\hat{e}\|^2 \sim \sigma^2_e \chi^2_{n-p}, \quad (4.52) \]
and
\[ \hat{\sigma}_e^2 = \frac{1}{n-p} \|\hat{e}\|^2 \sim \frac{\sigma^2_e}{n-p} \chi^2_{n-p}. \quad (4.53) \]

4.5 Exact confidence intervals: Student’s \( t \)

We now obtain the distribution of (4.39) with \( \sigma^2_e \) replaced by its estimate. First, we need to define Student’s \( t \).

**Definition 17** If \( Z \sim N(0, 1) \) and \( U \sim \chi^2_v \), and \( Z \) and \( U \) are independent, then
\[ T = \frac{Z}{\sqrt{U}} \quad (4.54) \]
has a Student’s \( t \) distribution on \( v \) degrees of freedom, written
\[ T \sim t_v. \quad (4.55) \]

**Proposition 15** Under the model (4.39) with \( \sigma^2_e > 0 \), if \( \lambda'\beta \) is estimable and \( \hat{a}'\hat{y} \) is the least squares estimate, then
\[ \frac{\hat{a}'\hat{y} - \lambda'\beta}{\hat{\sigma}_e \|\hat{a}\|} \sim t_{n-p}. \quad (4.56) \]
4.5. EXACT CONFIDENCE INTERVALS: STUDENT’S T

Now is a good time to remind you that if $XX'$ is invertible, then $a'y = X\hat{\beta}$ and $\|\hat{a}\|^2 = \lambda'(XX')^{-1}\lambda$.

**Proof.** We know that $a'y = \hat{a}'y \sim N(\lambda'\beta, \sigma^2_e\|\hat{a}\|^2)$, hence as in (4.39),

$$Z = \frac{a'y - \lambda'\beta}{\sigma_e\|\hat{a}\|} \sim N(0, 1). \tag{4.57}$$

From (4.51),

$$U = \frac{1}{\sigma^2_e} \|\hat{e}\|^2 \sim \chi^2_{n-p}. \tag{4.58}$$

Furthermore, $\hat{y}$ and $\hat{e}$ are independent by (4.37), hence the $Z$ in (4.57) and $U$ in (4.58) are independent. Plugging them into the formula for $T$ in (4.54) yields

$$T = \frac{a'y - \lambda'\beta}{\sqrt{\frac{1}{\sigma^2_e} \|\hat{e}\|^2}} = \frac{a'y - \lambda'\beta}{\sigma_e\|\hat{a}\|}. \tag{4.59}$$

This statistic is that in (4.56), hence by definition (4.55), it is $t_\nu$ with $\nu = n - p$. \hfill $\square$

To obtain a confidence interval for $\lambda'\beta$, proceed as in (4.40) and (4.40), but use the Student’s $t$ instead of the Normal, that is, an exact $100 \times (1 - \alpha)$% confidence interval is

$$a'y \pm t_{\nu, \alpha/2} \sigma_e\|\hat{a}\|, \tag{4.60}$$

where the $t_{\nu, \alpha/2}$ is found in a $t$ table so that

$$P[-t_{\nu, \alpha/2} < t_\nu < t_{\nu, \alpha/2}] = 1 - \alpha. \tag{4.61}$$

**Example.** Consider the contrast $(\alpha_1 + \alpha_2)/2 - \alpha_3$ from the Leprosy example, using the covariate, in Section 3.4.2. The least squares estimate is $-3.392$, and $se = \sqrt{\sigma^2_e \Delta'(XX)^{-1}\Delta} = 1.641$. With the covariate, $p = 4$, and $n = 30$, hence $\nu = n - p = 26$. Finding a $t$-table, $t_{26, 0.025} = 2.056$, hence the 95% confidence interval is

$$(-3.392 \pm 2.056 \times 1.641) = (-6.77, -0.02). \tag{4.62}$$

This interval just barely misses 0, so the effectiveness of the drugs is marginally significant.

**Note.** This confidence interval is exact if the assumptions are exact. Because we do not really believe that $\epsilon$ or $y$ are multivariate normal, in reality even the $t$ interval is approximate. But in general, if the data are not too skewed or have large outliers, the approximation is fairly good.
Chapter 5

Nested Models

5.1 Introduction

The previous chapters introduced inference on single parameters or linear combinations $X'\beta$. Analysis of variance is often concerned with combined effects, e.g., the treatment effect in the leprosy example (1.26), or the sun/shade effect, fruit effect, or interaction effect in the two-way ANOVA example (1.23). Such effects are often not representable using one parameter, but rather by several parameters, or more generally, by nested vector spaces.

For example, consider the one-way ANOVA model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, 2, 3, j = 1, 2$:

$$
\begin{pmatrix}
    y_{11} \\
    y_{12} \\
    y_{21} \\
    y_{22} \\
    y_{31} \\
    y_{32}
\end{pmatrix}
= X\beta + \epsilon
= \begin{pmatrix}
    1 & 1 & 0 & 0 \\
    1 & 1 & 0 & 0 \\
    1 & 0 & 1 & 0 \\
    1 & 0 & 1 & 0 \\
    1 & 0 & 0 & 1 \\
    1 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
    \mu \\
    \alpha_1 \\
    \alpha_2 \\
    \alpha_3
\end{pmatrix}
+ \epsilon
\tag{5.1}
$$

We now how to assess single contrasts, e.g., $\alpha_1 - \alpha_2$ or $(\alpha_1 + \alpha_2)/2 - \alpha_3$, but one may wish to determine whether there is any difference among the groups at all. If there is no difference, then the six observations are as from one large group, in which case the model would be

$$
y = \mathbf{1}_6\mu + \epsilon
\tag{5.2}
$$

Letting $M_A = C(X)$ be the vector space for model (5.1) and $M_0 = \text{span}\{\mathbf{1}_6\}$ be that for model (5.2), we have that $M_0 \subset M_A$. Such spaces are said to be nested. Note that model (5.2) can be obtained from model (5.1) by setting some parameters to zero, $\alpha_1 = \alpha_2 = \alpha_3 = 0$. It is not necessary that that be the case, e.g., we could have represented (5.1) without the
vector,

\[
\begin{pmatrix}
y_{11} \\
y_{12} \\
y_{21} \\
y_{22} \\
y_{31} \\
y_{32}
\end{pmatrix}
= X^* \beta^* + \varepsilon =
\begin{pmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3
\end{pmatrix}
+ \varepsilon .
\]

(5.3)

in which case we still have \( M_0 \subset M_A = C(X^*) \), but setting any of the \( \beta_i \)'s would not yield \( M_0 \). We could do it by setting \( \beta_1 = \beta_2 = \beta_3 \), though.

Using the hypothesis testing formulation, we are interested in testing the smaller model as the null hypothesis, and the larger model as the alternative. Thus with \( \mu = E[y] \), we are testing

\[
H_0 : \mu \in M_0 \text{ versus } H_A : \mu \in M_A .
\]

(5.4)

(Formally, we should not let the two hypotheses overlap, so that \( H_A \) should be \( \mu \in M_A - M_0 \).)

The ANOVA approach to comparing two nested models is to consider the squared lengths of the projections onto \( M_0 \) and \( M_A \), the idea being that the length of a projection represents the variation in the data \( y \) that is “explained” by the vector space. The basic decomposition of the squared length of \( y \) based on vector space \( M_0 \) is

\[
\| y \|^2 = \| \hat{y}_0 \|^2 + \| y - \hat{y}_0 \|^2 ,
\]

(5.5)

where \( \hat{y}_0 \) is the projection of \( y \) onto \( M_0 \). (Recall that this decomposition is due to \( \hat{y}_0 \) and \( y - \hat{y}_0 \) being orthogonal. It is the Pythagorean Theorem.) The equation in (5.5) is expressed as

Total variation = Variation due to \( M_0 \) + Variation unexplained by \( M_0 \).

(5.6)

A similar decomposition using the projection onto \( M_A \), \( \hat{y}_A \), is

\[
\| y \|^2 = \| \hat{y}_A \|^2 + \| y - \hat{y}_A \|^2 ,
\]

Total variation = Variation due to \( M_A \) + Variation unexplained by \( M_A \).

(5.7)

The explanatory power of the alternative model \( M_A \) over the null model \( M_0 \) can be measured in a number of ways, e.g., by comparing the variation due to the two models, or by comparing the variation unexplained by the two models. The most common measures start with the variation unexplained by the null model, and look at how much of that is explained by the alternative. That is, we subtract the variation due to the null model from the equations (5.5) and (5.7):

\[
\| y \|^2 - \| \hat{y}_0 \|^2 = \| y - \hat{y}_0 \|^2 \text{ and }
\]

\[
\| y \|^2 - \| \hat{y}_A \|^2 = \| \hat{y}_A \|^2 - \| \hat{y}_0 \|^2 + \| y - \hat{y}_A \|^2 ,
\]

(5.8)
yielding

\[
\text{Variation unexplained by } M_0 = \text{Variation explained by } M_A \text{ but not by } M_0 \\
+ \text{Variation unexplained by } M_A.
\]

The larger the “Variation explained by \( M_A \) but not by \( M_0 \)”, and the smaller the “Variation unexplained by \( M_A \)”, the more evidence there is that the more complicated model \( M_A \) is better than the simpler model \( M_0 \). These quantities need to be normalized somehow. One popular way is to take the ratio

\[
R^2 = \frac{\text{Variation explained by } M_A \text{ but not by } M_0}{\text{Variation unexplained by } M_0} = \frac{\|\hat{y}_A\|^2 - \|\hat{y}_0\|^2}{\|y - \hat{y}_0\|^2}.
\]

This quantity is sometimes called the coefficient of determination or the square of the multiple correlation coefficient. Usually it is called \( R \)-squared.

The squaredness suggests that \( R^2 \) must be nonnegative, and the “correlation” in the term suggests it must be no larger than 1. Both suggestions are true. The next section looks more closely at these sums of squares.

### 5.1.1 Note on calculating sums of squares

The sums of squares as in (5.5) for a generic model \( y = X\beta + \varepsilon \) can be obtained by finding the \( \hat{y} \) explicitly, then squaring the elements and summing. When \( X'X \) is invertible, there are more efficient ways, although they may not be as stable numerically. That is, once one has \( \hat{\beta} \) and \( X'X \) calculated, it is simple to use

\[
\|\hat{y}\|^2 = \hat{y}'\hat{y} = (X\hat{\beta})'X\hat{\beta} = \hat{\beta}'X'X\hat{\beta}.
\]

Then

\[
\|y - \hat{y}\|^2 = \|y\|^2 - \hat{\beta}'X'X\hat{\beta}.
\]

These formulas are especially useful if \( p \), the dimension of the \( \beta \) vector, is small relative to \( n \).

As a special case, suppose \( X = \mathbf{1}_n \). Then \( \hat{\beta} = \bar{y} \) and \( X'X = n \), so that (5.12) is

\[
\|y - \hat{y}\|^2 = \|y\|^2 - \bar{y}(n)\bar{y}.
\]

i.e.,

\[
\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2,
\]

the familiar “machine formula” used for calculating the sample standard deviation.

These days, typical statistical programs use efficient and accurate routines for calculating linear model quantities, so that the efficiency of formula (5.12) is of minor importance to us. Conceptually, it comes in handy, though.
5.2 Orthogonal complements

We are interested in the part of $\hat{\mathbf{y}}_A$ that is not part of $\mathcal{M}_0$, e.g., the difference in projections $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0$. It turns out that this is also a projection of $\hat{\mathbf{y}}_A$ onto some subspace.

**Definition 18** Suppose that $\mathcal{M}_0 \subset \mathcal{M}_A$. Then the orthogonal complement of $\mathcal{M}_0$ relative to $\mathcal{M}_A$, denoted $\mathcal{M}_{A:0}$, is the set of vectors in $\mathcal{M}_A$ that are orthogonal to $\mathcal{M}_0$, that is,

$$\mathcal{M}_{A:0} = \{ \mathbf{x} \in \mathcal{M}_A \mid \mathbf{x} \perp \mathcal{M}_0 \} = \mathcal{M}_A \cap \mathcal{M}_0^\perp. \quad (5.15)$$

It is not hard to prove that $\mathcal{M}_{A:0}$ is a vector space. In fact, any intersection of vector spaces is also a vector space. The projection onto $\mathcal{M}_{A:0}$ is just the difference of the projections on the two spaces.

**Proposition 16** If $\mathcal{M}_0 \subset \mathcal{M}_A$, then the projection of $\hat{\mathbf{y}}_A$ onto $\mathcal{M}_{A:0}$ is $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0$, hence the projection matrix $\mathbf{M}_{A:0}$ is $\mathbf{M}_A - \mathbf{M}_0$, where $\hat{\mathbf{y}}_A$ and $\mathbf{M}_0$ ($\hat{\mathbf{y}}_A$ and $\mathbf{M}_A$) are the projection onto $\mathcal{M}_0$ ($\mathcal{M}_A$) and corresponding projection matrix.

**Proof.** To see that $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0 \in \mathcal{M}_{A:0}$, note first that $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0 \in \mathcal{M}_A$ because both individual projections are in $\mathcal{M}_A$. Second, to see that $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0 \in \mathcal{M}_0^\perp$, note that $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0 = (\mathbf{y} - \hat{\mathbf{y}}_0) - (\mathbf{y} - \hat{\mathbf{y}}_A)$. Then by the definition of projection, $(\mathbf{y} - \hat{\mathbf{y}}_0) \in \mathcal{M}_0^\perp$ and $(\mathbf{y} - \hat{\mathbf{y}}_A) \in \mathcal{M}_A^\perp$. Because $\mathcal{M}_0 \subset \mathcal{M}_A$, $\mathcal{M}_A \subset \mathcal{M}_0^\perp$ (Why?), hence $(\mathbf{y} - \hat{\mathbf{y}}_A) \in \mathcal{M}_0^\perp$, so that $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0 \in \mathcal{M}_0^\perp$.

Next, we need to show $\mathbf{y} - (\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0) \in \mathcal{M}_{A:0}$. Write $\mathbf{y} - (\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0) = (\mathbf{y} - \hat{\mathbf{y}}_A) - \hat{\mathbf{y}}_0$. $(\mathbf{y} - \hat{\mathbf{y}}_A) \in \mathcal{M}_A \subset \mathcal{M}_{A:0}$, because $\mathcal{M}_{A:0} \subset \mathcal{M}_A$. Also, if $\mathbf{x} \in \mathcal{M}_{A:0}$, then $\mathbf{x} \perp \mathcal{M}_0$, hence $\mathbf{x} \perp \hat{\mathbf{y}}_0$, which means that $\hat{\mathbf{y}}_0 \in \mathcal{M}_{A:0}$. Thus $\mathbf{y} - (\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0) \in \mathcal{M}_{A:0}$. \hfill \qed

Write

$$\mathbf{y} - \hat{\mathbf{y}}_0 = (\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0) + (\mathbf{y} - \hat{\mathbf{y}}_A). \quad (5.16)$$

Because $\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0 \in \mathcal{M}_{A:0}$ and $\mathbf{y} - \hat{\mathbf{y}}_A \perp \mathcal{M}_A$, they are orthogonal, hence

$$\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 = \|\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}_A\|^2. \quad (5.17)$$

Thus by (5.8), $\|\hat{\mathbf{y}}_A\|^2 - \|\hat{\mathbf{y}}_0\|^2 = \|\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0\|^2$, and by (5.10),

$$R^2 = \frac{\|\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2}, \quad (5.18)$$

Now (5.17) also shows that $\|\hat{\mathbf{y}}_A - \hat{\mathbf{y}}_0\|^2 \leq \|\mathbf{y} - \hat{\mathbf{y}}_0\|^2$, hence indeed we have that $0 \leq R^2 \leq 1$. 
5.2. ORTHOGONAL COMPLEMENTS

5.2.1 Example

Consider the Leprosy example, with the covariate, so that the model is as in (2.61),

\[
\begin{pmatrix}
  y_{11} \\
  y_{12} \\
  \vdots \\
  y_{1,10} \\
  y_{21} \\
  y_{22} \\
  \vdots \\
  y_{2,10} \\
  y_{31} \\
  y_{32} \\
  \vdots \\
  y_{3,10}
\end{pmatrix}
= X_A \hat{\beta}_A + \varepsilon =
\begin{pmatrix}
  1 & 1 & 0 & 0 & z_{11} \\
  1 & 1 & 0 & 0 & z_{12} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 1 & 0 & 0 & z_{1,10} \\
  1 & 0 & 1 & 0 & z_{21} \\
  1 & 0 & 1 & 0 & z_{22} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 1 & 0 & z_{2,10} \\
  1 & 0 & 0 & 1 & z_{31} \\
  1 & 0 & 0 & 1 & z_{32} \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 0 & 1 & z_{3,10}
\end{pmatrix}
\begin{pmatrix}
  \mu \\
  \alpha_1 \\
  \alpha_2 \\
  \alpha_3 \\
  \gamma
\end{pmatrix}
+ \varepsilon
\]  

(5.19)

The large model is then \( \mathcal{M}_A = C(X_A) \). Consider the smaller model to be that without treatment effect. It can be obtained by setting \( \alpha_1 = \alpha_2 = \alpha_3 = 0 \) (or equal to any constant), so that \( \mathcal{M}_0 = span\{1_{10}, \hat{z}_i\} \). From (2.67) we know that \( \hat{\gamma}_A \) has elements

\[
\hat{\gamma}_{Aij} = \gamma_i + 0.987 (z_{ij} - \bar{z}_i).
\]  

(5.20)

Notice that model \( \mathcal{M}_0 \) is just a simple linear regression model, \( y_{ij} = \alpha + \beta z_{ij} + e_{ij} \), so we know how to estimate the coefficients. They turn out to be \( \hat{\alpha} = -3.886 \) and \( \hat{\beta} = 1.098 \), so

\[
\hat{\gamma}_{0ij} = -3.886 + 1.098 z_{ij}.
\]  

(5.21)

To find the decompositions, we first calculate \( \|y\|^2 = 3161 \). For model \( \mathcal{M}_A \), we would like to use formula (5.12), but need \( X'X \) invertible, which we do by dropping the \( 1_{10} \) vector from \( X_A \), so that we have the model (3.38). From (2.65) and (2.68), we can calculate the \( \hat{\beta}^* \) (without the \( \mu \)) to be

\[
\hat{\beta}^* = \begin{pmatrix}
  -3.881 \\
  -3.772 \\
  -0.435 \\
  0.987
\end{pmatrix}
\]  

(5.22)

The \( X'X^* \) matrix is given in (3.39), hence

\[
\|\hat{\gamma}_A\|^2 = \begin{pmatrix}
  -3.881 \\
  -3.772 \\
  -0.435 \\
  0.987
\end{pmatrix}' \begin{pmatrix}
  10 & 0 & 0 & 93 \\
  0 & 10 & 0 & 100 \\
  0 & 0 & 10 & 129 \\
  93 & 100 & 129 & 4122
\end{pmatrix} \begin{pmatrix}
  -3.881 \\
  -3.772 \\
  -0.435 \\
  0.987
\end{pmatrix} = 2743.80.
\]  

(5.23)
Then \( \|y - \hat{y}_A\|^2 \) is the difference, \( 3161 - 2743.80 \):

\[
\|y\|^2 = \|\hat{y}_A\|^2 + \|y - \hat{y}_A\|^2; \\
3161 = 2743.80 + 417.20. \tag{5.24}
\]

The decomposition for \( \mathcal{M}_0 \) is similar. We can find \( X_0'X_0 \), where \( X_0 = (1 30 z) \), from \( X^*X^* \) by adding the first three diagonals, and adding the first three elements of the first column:

\[
X_0'X_0 = \begin{pmatrix} 30 & 322 \\ 322 & 4122 \end{pmatrix}. \tag{5.25}
\]

Then

\[
\|\hat{y}_0\|^2 = \begin{pmatrix} -3.886 \\ 1.098 \end{pmatrix}' \begin{pmatrix} 30 & 322 \\ 322 & 4122 \end{pmatrix} \begin{pmatrix} -3.886 \\ 1.098 \end{pmatrix} = 2675.24, \tag{5.26}
\]

hence

\[
\|y\|^2 = \|\hat{y}_0\|^2 + \|y - \hat{y}_0\|^2; \\
3161 = 2675.24 + 485.76. \tag{5.27}
\]

The decomposition of interest then follows easily by subtraction:

\[
\|y - \hat{y}_0\|^2 = \|\hat{y}_A - \hat{y}_0\|^2 + \|y - \hat{y}_A\|^2; \\
485.76 = 68.56 + 417.20. \tag{5.28}
\]

The \( R^2 \) is then

\[
R^2 = \frac{\|\hat{y}_A - \hat{y}_0\|^2}{\|y - \hat{y}_0\|^2} = \frac{68.56}{485.76} = 0.141. \tag{5.29}
\]

That says that about 14% of the variation that the before measurements fail to explain is explained by the treatments. It is fairly small, which means there is still a lot of variation in the data not explained by the difference between the treatments. It may be that there are other variables that would be relevant, such as age, sex, weight, etc., or that bacterial counts are inherently variable.

### 5.3 Mean squares

Another popular ratio is motivated by looking at the expected values of the sums of squares under the two models. We will concentrate on the two difference vectors that decompose \( y - \hat{y}_0 \) in (5.17), \( \hat{y}_A - \hat{y}_0 \), and \( y - \hat{y}_A \). There are two models under consideration, the null \( \mathcal{M}_0 \) and the alternative \( \mathcal{M}_A \). Thus there are two possible distributions for \( y \):

\[
\mathcal{M}_0 : y \sim N_n(\mu_0, \sigma^2_0 I_n) \text{ for some } \mu_0 \in \mathcal{M}_0; \\
\mathcal{M}_A : y \sim N_n(\mu_A, \sigma^2_0 I_n) \text{ for some } \mu_A \in \mathcal{M}_A. \tag{5.30}
\]

The two difference vectors are independent under either model by the next proposition, because they are the projections on two orthogonal spaces, \( \mathcal{M}_{A,0} \) and \( \mathcal{M}_{A} \).
The projections are thus independent because they are uncorrelated and multivariate normal.

\[ Cov \left[ \left( \hat{y}_1, \hat{y}_2 \right) \right] = Cov \left[ \left( \frac{\mathbf{M}_1}{\mathbf{M}_2} \right) \mathbf{y} \right] \]
\[ = \sigma^2 \left( \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{pmatrix} \right) \left( \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{pmatrix} \right)' \]
\[ = \sigma^2 \left( \begin{pmatrix} \mathbf{M}_1 & 0 \\ 0 & \mathbf{M}_2 \end{pmatrix} \right). \] (5.31)

The projections are thus independent because they are uncorrelated and multivariate normal.

The means of the vectors may depend on the model. Below are the distributions under the two models:

\[
\begin{array}{c|c|c}
\mathcal{M}_0 & \mathcal{M}_A \\
\allowstackrel{\mathbf{y} - \hat{y}_A}{\mathbf{y} - \hat{y}_0} & \begin{array}{c}
\mathcal{M}_0 \\
\mathcal{M}_A
\end{array}
\end{array}
\]
\[
\begin{array}{c|c|c}
N_n((\mathbf{M}_A - \mathbf{M}_0)\mu_0, \sigma^2(\mathbf{M}_A - \mathbf{M}_0)) & N_n((\mathbf{M}_A - \mathbf{M}_0)\mu_A, \sigma^2(\mathbf{M}_A - \mathbf{M}_0)) \\
N_n((\mathbf{I}_n - \mathbf{M}_A)\mu_0, \sigma^2(\mathbf{I}_n - \mathbf{M}_A)) & N_n((\mathbf{I}_n - \mathbf{M}_A)\mu_A, \sigma^2(\mathbf{I}_n - \mathbf{M}_A))
\end{array}
\] (5.32)

Three of those means are zero: Because \( \mu_0 \) is in \( \mathcal{M}_0 \), it is also in \( \mathcal{M}_A \), hence \( \mathbf{M}_0 \mu_0 = \mu_0 \) and \( \mathbf{M}_A \mu_0 = \mu_0 \). Thus \( (\mathbf{M}_A - \mathbf{M}_0)\mu_0 = \mathbf{0} \) and \( (\mathbf{I}_n - \mathbf{M}_A)\mu_0 = \mathbf{0} \). Also, \( \mathbf{M}_A \mu_A = \mu_A \), so that \( (\mathbf{I}_n - \mathbf{M}_A)\mu_A = \mathbf{0} \). Thus all but the upper right have zero mean:

\[
\begin{array}{c|c|c}
\mathcal{M}_0 & \mathcal{M}_A \\
\allowstackrel{\mathbf{y} - \hat{y}_A}{\mathbf{y} - \hat{y}_0} & \begin{array}{c}
\mathcal{M}_0 \\
\mathcal{M}_A
\end{array}
\end{array}
\]
\[
\begin{array}{c|c|c}
N_n(\mathbf{0}_n, \sigma^2(\mathbf{M}_A - \mathbf{M}_0)) & N_n((\mathbf{M}_A - \mathbf{M}_0)\mu_A, \sigma^2(\mathbf{M}_A - \mathbf{M}_0)) \\
N_n(\mathbf{0}_n, \sigma^2(\mathbf{I}_n - \mathbf{M}_A)) & N_n((\mathbf{I}_n - \mathbf{M}_A)\mu_A, \sigma^2(\mathbf{I}_n - \mathbf{M}_A))
\end{array}
\] (5.33)

Notice that no matter which model is true, \( \mathbf{y} - \hat{y}_A \) has the same distribution, with mean \( \mathbf{0}_n \). On the other hand, \( \hat{y}_A - \hat{y}_0 \) has zero mean if the null model is true, but (potentially) nonzero mean if the alternative model is true. Thus this vector contains the information to help decide whether the null hypothesis is true.

We are actually interested in the sums of squares, so consider the expected values of the squared lengths of those vectors. For any random variable \( W_i \), \( E[W_i^2] = Var[W_i] + E[W_i]^2 \) (Why?), hence for the vector \( \mathbf{W} \),

\[
E[||\mathbf{W}||^2] = \sum_{i=1}^n E[W_i^2] \]
\[
= \sum_{i=1}^n Var[W_i] + \sum_{i=1}^n E[W_i]^2 \]
\[
= \text{trace}(\text{Cov}(\mathbf{W})) + E[\mathbf{W}]'E[\mathbf{W}]. \] (5.34)
For the covariance matrices in the table (5.33), we know \( \text{trace}(M_A) = p_A \) and \( \text{trace}(M_0) = p_0 \), where \( p_A \) and \( p_0 \) are the ranks of the respective vector spaces. Thus we have

\[
\begin{array}{c|cc}
\text{Expected sum of squares} & M_0 & M_A \\
\hline
\text{ESS}_{A0} & E[\|\hat{y}_A - \hat{y}_0\|^2] & \sigma_e^2(p_A - p_0) \\
\text{ESSE} & E[\|\hat{y} - \hat{y}_A\|^2] & \sigma_e^2(n - p_A)
\end{array}
\]

The “ESS” means expected sum of squares, and the “ESSE” means expected sum of squares of errors.

Now for the question: If the null hypothesis is not true, then \( E[\|\hat{y}_A - \hat{y}_0\|^2] \) will be relatively large. How large is large? One approach is the \( R^2 \) idea from the previous section. Another is to notice that if the null hypothesis is true, how large \( E[\|\hat{y}_A - \hat{y}_0\|^2] \) is depends on \( \sigma_e^2 \) (and \( n - p_A \)). Thus we could try comparing those. The key is to look at expected mean squares, which are obtained from table (5.35) by dividing by the degrees of freedom:

\[
\begin{array}{c|cc}
\text{Expected mean squares} & M_0 & M_A \\
\hline
\text{EMS}_{A0} & E[\|\hat{y}_A - \hat{y}_0\|^2]/(p_A - p_0) & \sigma_e^2 + \mu_A'(M_A - M_0)\mu_A/(p_A - p_0) \\
\text{EMSE} & E[\|y - \hat{y}_A\|^2]/(n - p_A) & \sigma_e^2
\end{array}
\]

The \( \text{EMSE} \) means “expected mean square error”. One further step simplifies even more: Take the ratio of the expected mean squares:

\[
\text{Ratio of expected mean squares} \quad M_0 \quad M_A \\
\hline
\text{EMS}_{A0}/\text{EMSE} & 1 & 1 + \frac{\mu_A'(M_A - M_0)\mu_A}{\sigma_e^2(p_A - p_0)}
\]

Now we have (sort of) answered the question: How large is large? Larger than 1. That is, this ratio of expected mean squares is 1 if the null hypothesis is true, and larger than 1 if the null hypothesis is not true. How much larger is semi-complicated to say, but it depends on \( \mu_A \) and \( \sigma_e^2 \).

That is fine, but we need to estimate this ratio. We will use the analogous ratio of mean squares, that is, just remove the “\( E[ \cdot \cdot \cdot ] \)”’s. This ratio is called the \( F \) ratio, named after R. A. Fisher:

**Definition 19** Given the above set up, the \( F \) ratio is

\[
F = \frac{MS_{A0}}{MSE} = \frac{\|\hat{y}_A - \hat{y}_0\|^2/(p_A - p_0)}{\|\hat{y} - \hat{y}_A\|^2/(n - p_A)}.
\]

Notice that \( \text{EMSE} \) is actually \( \hat{\sigma}_e^2 \) for the model \( M_A \).

The larger \( F \), the more evidence we have for rejecting the null hypothesis in favor of the alternative. The next section will deal with exactly how large is large, again. But first we continue with the example in Section 5.2.1. From (5.28) we obtain the sums of squares. The \( n = 30, p_0 = 2 \) and \( p_A = 4 \), hence

\[
MS_{A0} = \frac{68.56}{4 - 2} = 34.28, \quad \text{MSE} = \frac{417.20}{30 - 4} = 16.05, \quad F = \frac{34.28}{16.05} = 2.14.
\]
That $F$ is not much larger than 1, so there does not seem to be a very significant treatment effect. The next section shows how to calculate the significance level.

For both measures $R^2$ and $F$, the larger, the more one favors the alternative model. These measures are in fact equivalent in the sense of being monotone functions of each other:

$$R^2 = \frac{\|\hat{y}_A - \hat{y}_0\|^2}{\|y - \hat{y}_0\|^2} = \frac{\|\hat{y}_A - \hat{y}_0\|^2}{\|\hat{y}_A - \hat{y}_0\|^2 + \|y - \hat{y}_A\|^2},$$

so that

$$\frac{R^2}{1 - R^2} = \frac{\|\hat{y}_A - \hat{y}_0\|^2}{\|y - \hat{y}_A\|^2},$$

and

$$\frac{n - p_A}{p_A - p_0} \frac{R^2}{1 - R^2} = \frac{\|\hat{y}_A - \hat{y}_0\|^2/(p_A - p_0)}{\|y - \hat{y}_A\|^2/(n - p_A)} = F.$$

### 5.4 The $F$ distribution

Consider the $F$ statistic in (5.42). We know that the numerator and denominator are independent, by (5.31). Furthermore, under the null model $M_0$ in (5.33), from Proposition 14, Equation (4.44), we have as in Section 4.4.1 that

$$\|\hat{y}_A - \hat{y}_0\|^2 \sim \sigma_e^2 \chi_{p_A - p_0}^2 \quad \text{and} \quad \|y - \hat{y}_A\|^2 \sim \sigma_e^2 \chi_{n-p_A}^2,$$

so that the distribution of $F$ can be given as

$$F \sim \frac{\sigma_e^2 \chi_{p_A - p_0}^2/(p_A - p_0)}{\sigma_e^2 \chi_{n-p_A}^2/(n - p_A)} = \frac{\chi_{p_A - p_0}^2/(p_A - p_0)}{\chi_{n-p_A}^2/(n - p_A)},$$

where the $\chi^2$'s are independent. In fact, that is the definition of the $F$ distribution.

**Definition 20** If $U_1 \sim \chi_{\nu_1}^2$ and $U_2 \sim \chi_{\nu_2}^2$, and $U_1$ and $U_2$ are independent, then

$$F \equiv \frac{U_1/\nu_1}{U_2/\nu_2}$$

has an $F$ distribution with degrees of freedom $\nu_1$ and $\nu_2$. It is written

$$F \sim F_{\nu_1,\nu_2}.$$ 

Then, according to the definition, when $M_0$ is the true model,

$$F = \frac{MS_{A0}}{\text{MSE}} = \frac{\|\hat{y}_A - \hat{y}_0\|^2/(p_A - p_0)}{\|y - \hat{y}_A\|^2/(n - p_A)} \sim F_{p_A - p_0, n - p_A}.$$
Note. When $M_A$ is true, then the numerator and denominator are still independent, and the denominator is still $\chi^2_{n-p_A} / (n-p_A)$, but the $SS_{A_0}$ is no longer $\chi^2$. In fact, it is noncentral chi-squared, and the $F$ is noncentral $F$. We will not deal with these distributions, except to say that they are “larger” than their regular (“central”) cousins.

We can now formally test the hypotheses

$$H_0 : \mu \in M_0 \quad versus \quad H_A : \mu \in M_A,$$

based on $y \sim N_n(\mu, \sigma^2 \mathbf{I}_n)$. (We assume $\sigma^2 > 0$. Otherwise, $y = \bar{y}$, so it is easy to test the hypotheses with no error.) For level $\alpha$, reject the null hypothesis when

$$F > F_{\nu_1,\nu_2,\alpha},$$

where $F$ is in (5.47), and $F_{\nu_1,\nu_2,\alpha}$ is the upper $\alpha$ cutoff point of the $F$ distribution:

$$P[F_{p_A-p_0,n-p_A} > F_{\nu_1,\nu_2,\alpha}] = \alpha.$$

(5.50)

There are tables of these cutoff points, and most statistical software will produce them.

Example. Continuing the leprosy example, from (5.39) we have that $F = 2.16$. Also, $p_A - p_0 = 2$ and $n - p_A = 26$, hence we reject the null hypothesis ($M_0$) of no treatment effect at the $\alpha = 0.05$ level if

$$F > F_{2,26,0.05} = 3.369.$$

(5.51)

Because 2.16 is less than 3.369, we cannot reject the null hypothesis, which means there is not enough evidence to say that there is a treatment effect.

Does this conclusion contradict that from the confidence interval in (4.62) for $(\alpha_1 + \alpha_2)/2 - \alpha_3$, which shows a significant difference between the average of the two drugs and the placebo? Yes and no. The $F$ test is a less focussed test, in that it is looking for any difference among the three treatments. Thus it is combining inferences for the drug versus placebo contrast, which is barely significant on its own, and the drug A versus drug D contrast, which is very insignificant. The combining drowns out the first contrast, so that overall there does not appear to be anything significant. More on this phenomenon when we get to simultaneous inference.

5.5 The ANOVA table

The ANOVA table is based on a systematic method for arranging the important quantities in comparing two nested model, taking off from the decomposition of the sums of squares and degrees of freedom. That is, write

$$\|y - \bar{y}_0\|^2 = \|\bar{y}_A - \bar{y}_0\|^2 + \|y - \bar{y}_A\|^2$$

$$n - p_0 = p_A - p_0 + n - p_A$$

(5.52)

in table form. The more generic form is
5.5. THE ANOVA TABLE

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_{A0} )</td>
<td>((M_A - M_0)y)^2</td>
<td>trace ((M_A - M_0))</td>
<td>(SS_{A0}/df_{A0})</td>
<td>(MS_{A0}/MSE)</td>
</tr>
<tr>
<td>(M_A )</td>
<td>((I_n - M_A)y)^2</td>
<td>trace ((I_n - M_A))</td>
<td>(SSE/dfE)</td>
<td>—</td>
</tr>
<tr>
<td>(M_0 )</td>
<td>((I_n - M_0)y)^2</td>
<td>trace ((I_n - M_0))</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The “source” refers to a vector space, the “sum of squares” is the squared length of the projection of \(y\) on that space, the “degrees of freedom” is the rank of the space, and the “mean square” is the sum of squares divided by the rank. To be a valid ANOVA table, the first two rows of the sum of squares and degrees of freedom columns add to the third row.

Writing the table more explicitly, we have

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_{A0} )</td>
<td>(\hat{y}_A - \hat{y}_0)^2</td>
<td>(p_A - p_0)</td>
<td>(SS_{A0}/df_{A0})</td>
<td>(MS_{A0}/MSE)</td>
</tr>
<tr>
<td>(M_A )</td>
<td>(y - \hat{y}_A)^2</td>
<td>(n - p_A)</td>
<td>(\hat{\sigma}^2 = SSE/dfE)</td>
<td>—</td>
</tr>
<tr>
<td>(M_0 )</td>
<td>(y - \hat{y}_0)^2</td>
<td>(n - p_0)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

It is also common to add the \(R^2\),

\[ R^2 = \frac{||\hat{y}_A - \hat{y}_0||^2}{||y - \hat{y}_0||^2}. \]  

In practice, more evocative names are given to the sources. Typically, the \(M_A\) space is called “Error” and the bottom space \(M_0\) is called “Total.” In a one-way ANOVA, the \(M_{A0}\) space may be called “Group effect”, or may refer to the actual groups. For example, the ANOVA table for the leprosy example would be

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>68.56</td>
<td>2</td>
<td>34.28</td>
<td>2.14</td>
</tr>
<tr>
<td>Error</td>
<td>417.20</td>
<td>26</td>
<td>16.05</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>485.76</td>
<td>28</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

\(R^2 = 0.141\)

For another example, consider simple linear regression \(y_i = \alpha + \beta x_i + e_i\), and let \(M_A = C(X)\) and \(M_0 = span\{1_n\}\), so that we are testing whether \(\beta = 0\), that is, whether the \(x_i\)'s are related to the \(y_i\)'s. Now (from Question 6 of Homework #5, for example),

\[ ||y - \hat{y}_0||^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2, \]  

\[ ||\hat{y}_A - \hat{y}_0||^2 = \beta^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}; \]  

\(p_0 = 1\) and \(p_A = 2\). Data on 132 male athletes, with \(x = height\) and \(y = weight\), has \(\hat{\alpha} = 657.30\) and \(\hat{\beta} = -5.003\) (which may seem strange, being negative). The ANOVA table is
$R^2 = 0.082$

The $F_{1,130,0.05} = 3.91$, so that the $\beta$ is very significant. On the other hand, $R^2$ is quite small, suggesting there is substantial variation in the data. It is partly because this model does not take into account the important factor that there are actually two sports represented in the data.
Chapter 6

One-way ANOVA

This chapter will look more closely at the one-way ANOVA model. The model has \( g \) groups, and \( N_i \) observations in group \( i \), so that there are \( n = N_1 + \cdots + N_g \) observations overall. Formally, the model is

\[
y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, g; \quad j = 1, \ldots, N_i,
\]

where the \( e_{ij} \)'s are independent \( N(0, \sigma^2_e) \)'s. Written in matrix form, we have

\[
y = X\beta + \epsilon = \begin{pmatrix} \mathbb{1}_{N_1} & \mathbb{1}_{N_1} & 0_{N_1} & \cdots & 0_{N_1} \\ \mathbb{1}_{N_2} & 0_{N_2} & \mathbb{1}_{N_2} & \cdots & 0_{N_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}_{N_g} & 0_{N_g} & 0_{N_g} & \cdots & 1_{N_g} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_g \end{pmatrix} + \epsilon, \quad \epsilon \sim N_n(\mathbb{0}_n, \sigma^2_e \mathbb{I}_n). \tag{6.2}
\]

The ANOVA is called \textbf{balanced} if there is the same number of observations in each group, that is, \( N_i = N \), so that \( n = Ng \). Otherwise, it is \textbf{unbalanced}. The balanced case is somewhat easier to analyze than the unbalanced one, but the difference is more evident in higher-way ANOVA’s. See the next chapter.

The next section gives the ANOVA table. Section 6.2 shows how to further decompose the group sum of squares into components based on orthogonal contrasts. Section 6.3 looks at “effects” and gives constraints on the parameters to make them estimable. Later, in Chapter 8, we deal with the thorny problem of multiple comparisons: A single confidence interval may have a 5\% chance of missing the parameter, but with many confidence intervals, each at 95\%, the chance that at least one misses its parameter can be quite high. E.g., with 100 95\% confidence intervals, you’d expect about 5 to miss. How can you adjust so that the chance is 95\% that they are all ok?

6.1 The ANOVA table

From all the work so far, it is easy to find the ANOVA table for testing whether there are any group effects, that is, testing whether the group means are equal. Here, \( \mathcal{M}_A = C(X) \)
for the $X$ in (6.2), and $M_0 = \text{span}\{1_n\}$. The ranks of these spaces are, respectively, $p_A = g$ and $p_0 = 1$. The projections have

$$\hat{y}_{Aij} = y_i \quad \text{and} \quad \hat{y}_{bij} = \bar{y},$$

(6.3)

where

$$\bar{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$$

is the sample mean of the observations in group $i$. Then the sums of squares are immediate:

$$SS_A = \| \hat{y}_A - \bar{y} \|^2 = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{g} N_i (\bar{y}_i - \bar{y})^2$$

$$SSE = \| y - \hat{y}_A \|^2 = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2$$

$$SST = \| y - \bar{y} \|^2 = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y})^2.$$  

(6.5)

(The $SST$ means “sum of squares total”. Often, the $SS_{A:0}$ is called the **between sum of squares** because it measures the differences between the group means and the overall mean, and the $SSE$ is called the **within sum of squares**, because it adds up the sums of squares of the deviations from each group’s mean. The table is then

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>$\sum_{i=1}^{g} N_i (\bar{y}_i - \bar{y})^2$</td>
<td>$g - 1$</td>
<td>$MSB$</td>
<td>$MSB/MSW$</td>
</tr>
<tr>
<td>Within</td>
<td>$\sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2$</td>
<td>$n - g$</td>
<td>$MSW$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$\sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y})^2$</td>
<td>$n - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$R^2 = \frac{\sum_{i=1}^{g} \sum_{j=1}^{N_i} (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y})^2}.$$  

### 6.2 Decomposing the between sum of squares

When there are more than two groups, the between sum of squares measure a combination of all possible differences among the groups. It is usually informative to be more specific about differences. One approach is to further decompose the between sum of squares using **orthogonal contrasts**. A **contrast** of a vector of parameters is a linear combination in which the coefficients sum to zero. For example, in the leprosy data we looked at contrasts of the group means $(\mu_1, \mu_2, \mu_3)'$, or, equivalently, the $(\alpha_1, \alpha_2, \alpha_3)'$:

$$\gamma_1 = \frac{1}{2} (\alpha_1 + \alpha_2) - \alpha_3 = \left( \frac{1}{2}, \frac{1}{2}, -1 \right) (\alpha_1, \alpha_2, \alpha_3)' \quad \text{and} \quad \gamma_2 = \alpha_1 - \alpha_2 = (1, -1, 0) (\alpha_1, \alpha_2, \alpha_3)'.$$  

(6.6)
6.2. DECOMPOSING THE BETWEEN SUM OF SQUARES

These two parameters are enough to describe any differences between the \( \alpha_i \)'s, which is to say that they can be used as parameters in place of the \( \alpha_i \)'s in the ANOVA model:

\[
\hat{y} = X^* \hat{\beta}^* + \epsilon = \begin{pmatrix} \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{2}{10} & \frac{1}{10} \\ \frac{1}{10} & -\frac{3}{10} & \frac{1}{10} \end{pmatrix} \begin{pmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{pmatrix} + \epsilon. \quad (6.7)
\]

[The model comes from the model (5.1) by solving for the \( \alpha_i \)'s in terms of the \( \gamma_i \)'s in the equations (6.6) plus \( \alpha_1 + \alpha_2 + \alpha_3 = 0 \).] This model (6.7) is indeed the regular one-way ANOVA model (5.1), that is, \( M_A = C(X^*) = C(X) \). Note also that the columns of \( X^* \) are orthogonal, which in particular means that

\[
M_{A,0} = \text{span} \left\{ \begin{pmatrix} \frac{1}{10} \\ \frac{1}{10} \\ \frac{1}{10} \end{pmatrix}, \begin{pmatrix} \frac{1}{10} \\ \frac{2}{10} \\ \frac{1}{10} \end{pmatrix}, \begin{pmatrix} \frac{1}{10} \\ -\frac{3}{10} \\ \frac{1}{10} \end{pmatrix} \right\}. \quad (6.8)
\]

More generally, suppose that there are orthogonal and nonzero vectors \( x_1, \ldots, x_q \) such that

\[
M_{A,0} = \text{span} \{ x_1, \ldots, x_q \}. \quad (6.9)
\]

(So that in the one-way ANOVA case, \( q = g - 1 \).) The projection onto \( M_{A,0} \) can be decomposed into the projections on the one-dimensional spaces \( M_1, \ldots, M_q \), where

\[
M_k = \text{span} \{ x_k \}. \quad (6.10)
\]

That is, if \( \hat{y}_{A,0} \) is the projection of \( \hat{y} \) onto \( M_{A,0} \), and \( \hat{y}_k \) is the projection onto \( M_k \), \( k = 1, \ldots, q \), then

\[
\hat{y}_{A,0} = \hat{y}_1 + \cdots + \hat{y}_q, \quad (6.11)
\]

and, because the \( M_k \)'s are mutually orthogonal,

\[
\|\hat{y}_{A,0}\|^2 = \|\hat{y}_1\|^2 + \cdots + \|\hat{y}_q\|^2. \quad (6.12)
\]

**Proof of (6.11).** We will show that \( \hat{z} \equiv \hat{y}_1 + \cdots + \hat{y}_q \) is the projection onto \( M_{A,0} \). First, \( \hat{z} \in M_A \), since it is a linear combination of the \( x_k \)'s, which are all in \( M_{A,0} \). Next, consider \( \hat{y} - \hat{z} \). To see that it is orthogonal to \( M_{A,0} \), it is enough to show that it is orthogonal to each \( x_k \), by (6.9). To that end,

\[
(y - \hat{z})'x_k = -\hat{y}_j'x_k - \cdots - \hat{y}_{k-1}'x_k + (y - \hat{y}_k)'x_k - \hat{y}_{k+1}'x_k - \cdots - \hat{y}_q'x_k. \quad (6.13)
\]

But \( \hat{y}_j'x_k = 0 \) if \( j \neq k \), because the \( x_j \)'s are orthogonal, and \( (y - \hat{y}_k)'x_k = 0 \), because \( \hat{y}_k \) is the projection of \( y \) onto \( M_k \). Thus \( (y - \hat{z})'x_k \), which means \( \hat{z} = \hat{y}_{A,0} \). \( \square \)

We know that the projection matrix for \( M_k \) is

\[
M_k = x_k(x_k'x_k)^{-1}x_k' = \frac{x_kx_k'}{\|x_k\|^2}, \quad (6.14)
\]
hence
\[ \|\hat{\gamma}_k\|^2 = \|M_k\hat{y}\|^2 = y'M_ky = \frac{(y'x_k)^2}{\|x_k\|^2}. \]  
(6.15)

The decomposition (6.12) then leads to an expanded ANOVA table, inserting a row for each \( M_k \):

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_1 )</td>
<td>( \frac{(y'x_1)^2}{|x_1|^2} )</td>
<td>1</td>
<td>( MS_1 )</td>
<td>( MS_1/MSW )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( M_{g-1} )</td>
<td>( \frac{(y'x_{g-1})^2}{|x_{g-1}|^2} )</td>
<td>1</td>
<td>( MS_{g-1} )</td>
<td>( MS_{g-1}/MSW )</td>
</tr>
<tr>
<td>Within</td>
<td>( \sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2 )</td>
<td>( n - g )</td>
<td>( MSW )</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>( \sum_{i=1}^{g} \sum_{j=1}^{N_i} (y_{ij} - \bar{y})^2 )</td>
<td>( n - 1 )</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Because the \( \text{rank}(M_k) = 1 \) for these vector spaces, the degrees of freedom are all 1, and the \( SS_k = MS_k \). Proposition 17 shows that the \( \hat{\gamma}_k \)'s are independent (because \( M_kM_j = 0 \) if \( k \neq j \)), and

\[ \hat{\gamma}_k \sim N_n(M_k(\mu_A - \mu_0), \sigma_e^2 M_k). \]  
(6.16)

Thus under the null hypothesis, since \( \text{trace}(M_k) = 1 \),

\[ \|\hat{\gamma}_1\|^2, \ldots, \|\hat{\gamma}_{g-1}\|^2 \]  
are independent \( \sigma_e^2 \chi^2_1 \)'s.

(6.17)

Thus the \( F_k = MS_k/MSW \) are indeed \( F_{1,n-g} \)'s, although they are not quite independent due to the common denominator \( MSW \).

These sums of squares can also be written in terms of the estimates of parameters in the generalization of the model (6.7),

\[ y = X^*\beta^* + \varepsilon = (\mathbf{1}_n, \bar{x}_1, \ldots, \bar{x}_q) \begin{pmatrix} \mu \\ \gamma_1 \\ \vdots \\ \gamma_q \end{pmatrix} + \varepsilon. \]  
(6.18)

By orthogonality of the columns of \( X^* \), we have that

\[ \beta^* = (X^*X^*)^{-1}X^*y = \begin{pmatrix} \bar{y} \\ \bar{x}'y/\|\bar{x}_1\|^2 \\ \vdots \\ \bar{x}'y/\|\bar{x}_q\|^2 \end{pmatrix}. \]  
(6.19)

Then with \( \gamma_k = \bar{x}_ky/\|\bar{x}_k\|^2 \)

\[ \|\hat{\gamma}_k\|^2 = \frac{(y'x_k)^2}{\|x_k\|^2} = \|x_k\|^2 \gamma_k^2. \]  
(6.20)
6.2. Decomposing the Between Sum of Squares

6.2.1 Example

Return to the one-way ANOVA model without covariate, (5.1), for the leprosy data. The table below has the basic statistics:

<table>
<thead>
<tr>
<th>Group</th>
<th>$\bar{y}_i$</th>
<th>$\sum_{j=1}^{10}(y_{ij} - \bar{y}_i)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Drug A</td>
<td>5.3</td>
<td>194.1</td>
</tr>
<tr>
<td>2. Drug D</td>
<td>6.1</td>
<td>340.9</td>
</tr>
<tr>
<td>3. Placebo</td>
<td>12.3</td>
<td>460.1</td>
</tr>
</tbody>
</table>

Then the $SSW = 194.1 + 340.9 + 460.1 = 995.1$, and with $\bar{y} = 7.9$,

$$SSB = \sum_{i=1}^{3} \sum_{i=1}^{10} (\bar{y}_i - \bar{y})^2 = 10[(5.3 - 7.9)^2 + (6.1 - 7.9)^2 + (12.3 - 7.9)^2] = 293.6.$$  \hspace{1cm} (6.22)

Hence the regular ANOVA table, with $g = 3$, is

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>293.6</td>
<td>2</td>
<td>146.8</td>
<td>3.98</td>
</tr>
<tr>
<td>Within</td>
<td>995.1</td>
<td>27</td>
<td>36.86</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>1288.7</td>
<td>29</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

$R^2 = 0.2278$.

The $F_{2,27,0.05} = 3.354$, so that there is a significant group effect. Recall from Section 5.5 that when using the covariate, the group differences are not significant.

Now we decompose the $SSB$, using the $\mathbf{x}_1$ and $\mathbf{x}_2$ as in (6.8). From (6.6),

$$\hat{\gamma}_1 = (\alpha_1 + \alpha_2)/2 - \alpha_3 = (5.3 + 6.1)/2 - 12.3 = -6.6, \text{ and}$$
$$\hat{\gamma}_2 = \alpha_1 - \alpha_2 = 5.3 - 6.1 = -0.8.$$ \hspace{1cm} (6.23)

Thus

$$SS_1 = \|\mathbf{x}_1\|^2 \hat{\gamma}_1^2 = (10(1/3)^2 + 10(1/3)^2 + 10(-2/3)^2)(-6.6)^2 = 290.4, \text{ and}$$
$$SS_2 = \|\mathbf{x}_2\|^2 \hat{\gamma}_2^2 = (10(1/2)^2 + 10(-1/2)^2)(-0.8)^2 = 3.2.$$ \hspace{1cm} (6.24)

Note that indeed, $SSB = 293.6 = SS_1 + SS_2 = 290.4 + 3.2$. The expanded ANOVA table is then

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug vs Placebo</td>
<td>290.4</td>
<td>1</td>
<td>290.4</td>
<td>7.88</td>
</tr>
<tr>
<td>Drug A vs Drug D</td>
<td>3.2</td>
<td>1</td>
<td>3.2</td>
<td>0.09</td>
</tr>
<tr>
<td>Within</td>
<td>995.1</td>
<td>27</td>
<td>36.86</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>1288.7</td>
<td>29</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Now $F_{1,27,0.05} = 4.21$, so that the Drug vs Placebo contrast is quite significant, but the contrast between the two drugs is completely nonsignificant.
6.2.2 Limitations of the decomposition

Every subspace has an orthogonal basis, many essentially different such bases if the rank is more than one. Unfortunately, the resulting subspaces $\mathcal{M}_k$ need not be ones of interest. In the leprosy example, the two subspaces nicely corresponded to two interesting contrasts. Such nice results will occur in the balanced one-way ANOVA if one is interested in a set of orthogonal contrasts of the $\alpha_i$’s. A contrast of $\underline{\alpha} = (\alpha_1, \ldots, \alpha_g)'$ is a linear combination $c'\underline{\alpha}$, where $c$ is any nonzero $g \times 1$ vector such that $c_1 + \cdots + c_g = 0$. Two contrasts $c_1'\underline{\alpha}$ and $c_2'\underline{\alpha}$ are orthogonal if their vectors $c_1$ and $c_2$ are orthogonal. For example, the $\gamma_i$’s in (6.6) are orthogonal contrasts of $\underline{\alpha}$ with

$$c_1 = \left(\frac{1}{2}, \frac{1}{2}, -1\right)' \quad \text{and} \quad c_2 = (1, -1, 0)'.$$

(6.25)

It is easy to see that these two vectors are orthogonal, and their components sum to 0.

If the contrasts of interest are not orthogonal, then their respective sums of squares do not sum to the between sum of squares. E.g., we may be interested in comparing each drug to the placebo, so that the contrast vectors are $(1, 0, -1)'$ and $(0, 1, -1)'$. Although these are contrasts, the two vectors are not orthogonal. Worse, if the model is unbalanced, even orthogonal contrasts will not translate back to an orthogonal basis for $\mathcal{M}_{A_0}$. Also, the model with covariates will not allow the decomposition, even with a balanced design.

On the positive side, balanced higher-way ANOVA models do allow nice decompositions.

6.3 Effects

We know that the general one-way ANOVA model (6.1), (6.2), is often parametrized in such a way that the parameters are not estimable. That usually will not present a problem, because interest is one contrasts of the $\alpha_i$’s, which are estimable, or (equivalently) testing whether there are any differences among the groups. Alternatively, one can place constraints on the parameters so that they are estimable. E.g., one could set $\mu = 0$, or $\alpha_G = 0$, or $\alpha_1 + \cdots + \alpha_G = 0$. One method for setting constraints is to define effects for the groups. Let $\mu_i$ be the population mean of group $i$, i.e., $\mu_i = \mu + \alpha_i$. The idea is to have a benchmark value $\mu^*$, being some weighted average of the $\mu_i$’s. Then the effect of a group is the amount that group’s mean exceeds $\mu^*$, that is, the group $i$ effect is defined to be

$$\alpha_i = \mu_i - \mu^*, \quad \mu^* = \sum_{i=1}^g w_i \mu_i,$$

(6.26)

where the $w_i$’s are nonnegative and sum to 1. Backing up, we see that (6.26) implies the constraint that $\mu = \mu^*$, or equivalently, that

$$\sum_{i=1}^g w_i \alpha_i = 0.$$

(6.27)
6.3. EFFECTS

One choice for $\mu^*$ is the unweighted average of the $\mu_i$'s, that is, $w_i = 1/g$, so that $\mu^* = \sum \mu_i / g$. The resulting constraint (6.27) is that the unweighted average of the $\alpha_i$'s is 0, which is the same as saying the sum of the $\alpha_i$'s is 0. The least squares estimate of $\mu_i$ is $\bar{y}_i$, so that this constraint leads to the estimates

$$\hat{\mu} = \frac{\sum_{i=1}^{g} \bar{y}_i}{g} \text{ and } \hat{\alpha}_i = \bar{y}_i - \hat{\mu}. \tag{6.28}$$

In the balanced case, (6.28) becomes

$$\hat{\mu} = \frac{\sum_{i=1}^{g} (\sum_{j=1}^{N} y_{ij} / N)}{g} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{N} y_{ij}}{n} = \bar{y}_r \text{ and } \hat{\alpha}_i = \bar{y}_i - \bar{y}_r. \tag{6.29}$$

That is, $\hat{\mu}$ is the straight average of the $y_{ij}$'s. In this case, the $SSB$ in (6.5) can be easily written as a function of the estimated effects,

$$SSB = \sum_{i=1}^{g} \sum_{j=1}^{N} (\bar{y}_i - \bar{y}_r)^2 = N \sum_{i=1}^{g} \hat{\alpha}_i^2. \tag{6.30}$$

In the unbalanced case, (6.29) is not true, hence neither is (6.30). An alternative is to weight the $\mu_i$'s by the numbers of observations in the groups, i.e., $w_i = N_i/n$. Then

$$\mu = \frac{\sum_{i=1}^{g} N_i \mu_i}{n} \text{ (hence } \sum_{i=1}^{g} N_i \alpha_i = 0), \tag{6.31}$$

and

$$\hat{\mu} = \frac{\sum_{i=1}^{G} N_i \bar{y}_i}{n} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{N_i} y_{ij}}{n} = \bar{y}_{..}, \text{ } \hat{\alpha}_i = \bar{y}_i - \bar{y}_{..}, \tag{6.32}$$

and (6.30) does hold. One objection to using the weighting (6.31) is that groups that happen to have more observations are more strongly represented in the benchmark, hence their effect would seem smaller. E.g., in an extreme case with $g = 2$, suppose $N_1 = 99$ and $N_2 = 1$. Then $\alpha_2 = -99\alpha_1$, so that the effect of the second group will always look much larger than that for group 1. Using the unweighted mean (6.28), $\alpha_2 = -\alpha_1$, so the groups are treated equally.

Another approach is to weight the groups according to their representation in the population, so that $\mu$ would be the population mean of $y$. This approach assumes there is an existing population, e.g., the groups are men and women, or based on age, etc. It would not be relevant in the leprosy example, since the population is not taking Drug A, Drug D, and a placebo.
Chapter 7

Two-way ANOVA

Now observations are categorized according to two variables, e.g., amount of sun/shade and type of fruit for the data in (1.23). The table contains the leaf area/dry weight for the citrus trees:

<table>
<thead>
<tr>
<th></th>
<th>Orange</th>
<th>Grapefruit</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>112</td>
<td>90</td>
<td>123</td>
</tr>
<tr>
<td>Half – shade</td>
<td>86</td>
<td>73</td>
<td>89</td>
</tr>
<tr>
<td>Shade</td>
<td>80</td>
<td>62</td>
<td>81</td>
</tr>
</tbody>
</table>

(From Table 11.2.1 in Statistical Methods by Snedecor and Cochran.) Writing it in terms of $y_{ij}$.

<table>
<thead>
<tr>
<th></th>
<th>Orange</th>
<th>Grapefruit</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>$y_{11}$</td>
<td>$y_{12}$</td>
<td>$y_{13}$</td>
</tr>
<tr>
<td>Half – shade</td>
<td>$y_{21}$</td>
<td>$y_{22}$</td>
<td>$y_{23}$</td>
</tr>
<tr>
<td>Shade</td>
<td>$y_{31}$</td>
<td>$y_{32}$</td>
<td>$y_{33}$</td>
</tr>
</tbody>
</table>

This is a $3 \times 3$ layout: three rows and three columns. More generally, one has $r$ rows and $c$ columns, for an “$r \times c$ layout,” and there could be more than one observation in each cell. For example, here is a $2 \times 3$ layout with varying numbers of observations per cell:

<table>
<thead>
<tr>
<th></th>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>$y_{111}$, $y_{112}$, $y_{113}$</td>
<td>$y_{121}$, $y_{122}$</td>
<td>$y_{131}$, $y_{132}$, $y_{133}$, $y_{134}$</td>
</tr>
<tr>
<td>Row 2</td>
<td>$y_{211}$</td>
<td>$y_{221}$, $y_{222}$, $y_{223}$</td>
<td>$y_{231}$, $y_{232}$, $y_{233}$</td>
</tr>
</tbody>
</table>

The $y_{ijk}$ is the $k^{th}$ observation in the $i^{th}$ row and $j^{th}$ column. The most general model sets $\mu_{ij} = E[y_{ijk}]$, the population mean of the $ij^{th}$ cell, so that

$$y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, c, \quad k = 1, \ldots, N_{ij}. \quad (7.4)$$

The $N_{ij}$ is the number of observations in cell $ij$. E.g., in (7.3), $N_{11} = 3$, $N_{12} = 2$, $N_{13} = 4$, etc. As in one-way ANOVA, the design is called balanced if the $N_{ij}$ are equal to, say, $N$. E.g., (7.2) is balanced, with $N = 1$ (hence we do not bother with the third subscript), but (7.3) is unbalanced.

Questions to ask include:
• Are there row effects? That is, are the means equal for the \( r \) rows?

• Are there column effects?

• Is there interaction? That is, are the row effects different for different columns, or vice versa?

More detailed questions will also be important, such as, “If there are differences, what are they?”

The two basic models are additive, such as in (1.27), and nonadditive, as in (1.28). Additive means that there is no interaction between rows and columns, that is, the difference between the means in any two rows is the same for each column. In the \( 3 \times 3 \) case, that requirement is

\[
\begin{align*}
\mu_{11} - \mu_{21} &= \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}; \\
\mu_{11} - \mu_{31} &= \mu_{12} - \mu_{32} = \mu_{13} - \mu_{33}; \\
\mu_{21} - \mu_{31} &= \mu_{22} - \mu_{32} = \mu_{23} - \mu_{33};
\end{align*}
\]

(7.5)

and in general,

\[
\mu_{ij} - \mu_{i'j'} = \mu_{i'j'} - \mu_{ij'} \quad \text{for all } i, i', j, j',
\]

(7.6)

e.i.,

\[
\mu_{ij} - \mu_{i'j'} = \mu_{ij'} + \mu_{i'j'} = 0 \quad \text{for all } i, i', j, j'.
\]

(7.7)

The no interaction requirement (7.7) is equivalent to the restriction that the group means satisfy

\[
\mu_{ij} = \mu + \alpha_i + \beta_j
\]

(7.8)

for some \( \mu, \alpha_i's \) and \( \beta_j's \). The model (7.8) is called additive because the row effect and column effect are added. The nonadditive model allows interaction, and is often called the saturated model, because it makes no restrictions on the \( \mu_{ij}'s \). The interaction terms are \( \gamma_{ij} \), so that the means have the form

\[
\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.
\]

(7.9)

As is, the formulation (7.9) is extremely overparametrized. E.g., one could get rid of \( \mu \) and the \( \alpha_i's \) and \( \beta_j's \), and still have the same model. Typically one places restrictions on the parameters so that the parameters are effects as in Section 6.3. At least in the balanced case, the common definitions are

\[
\begin{align*}
\mu &= \mu. \\
\alpha_i &= \mu_i - \mu. \\
\beta_j &= \mu.j - \mu. \\
\gamma_{ij} &= \mu_{ij} - \mu.i - \mu.j + \mu..
\end{align*}
\]

(7.10)
One can check that these definitions imply the constraints

\[
\begin{align*}
\sum_{i=1}^{r} \alpha_i &= 0, \\
\sum_{j=1}^{c} \beta_j &= 0, \\
\sum_{i=1}^{r} \gamma_{ij} &= 0 \text{ for each } j, \\
\sum_{j=1}^{c} \gamma_{ij} &= 0 \text{ for each } i.
\end{align*}
\] (7.11)

In the unbalanced case, one may wish to use weighted means.

There are five basic models to consider, four of which are special cases of the additive model. They are

\[
\begin{align*}
\text{Saturated} : \quad \mu_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij} ; \\
\text{Additive} : \quad \mu_{ij} &= \mu + \alpha_i + \beta_j ; \\
\text{Just row effect (no column effect)} : \quad \mu_{ij} &= \mu + \alpha_i ; \\
\text{Just column effect (no row effect)} : \quad \mu_{ij} &= \mu + \beta_j ; \\
\text{No row or column effect} : \quad \mu_{ij} &= \mu .
\end{align*}
\] (7.12)

Notice that if the interaction terms are in the model, then both the row and column effects are, too. The reason is that “no row effects” means that the row variable does not effect the \( y \), so that there cannot be any interactions. If there are, then the effect of the rows is different for the different columns, meaning it cannot be uniformly 0. Thus \( \mu_{ij} = \mu + \beta_j + \gamma_{ij} \) is not a sensible model. (Of course, there may be rare situations when such a model makes sense. I’ve never seen one.)

The next section looks at the subspaces corresponding to these models, without worrying about constraints too much, and fitting and testing the models.

### 7.1 The subspaces and projections

We start by writing the saturated model as \( y = X\beta + \epsilon \). We write out the vectors and matrices in the \( 2 \times 3 \) layout below. See (1.36) for another example. We put the observations into the vector \( y \) by going across the rows in (7.3). The matrices are then
\[ \begin{pmatrix} \frac{1}{N_{11}} & 0 & 0 & 0 \\ \frac{1}{N_{12}} & 0 & 0 & 0 \\ \frac{1}{N_{13}} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{N_{21}} & 0 & 0 & 0 \\ \frac{1}{N_{22}} & 0 & 0 & 0 \\ \frac{1}{N_{23}} & 0 & 0 & 0 \end{pmatrix} \]

\[ X = \begin{pmatrix} \frac{1}{N} & x^{(R)}_1 & x^{(R)}_2 & x^{(C)}_1 & x^{(C)}_2 & x^{(C)}_3 & x^{(I)}_{11} & x^{(I)}_{12} & x^{(I)}_{13} & x^{(I)}_{21} & x^{(I)}_{22} & x^{(I)}_{23} \end{pmatrix} \]

\[ y = \begin{pmatrix} y_{111} \\ y_{112} \\ \vdots \\ y_{11N_{11}} \\ y_{121} \\ y_{122} \\ \vdots \\ y_{12N_{12}} \\ y_{131} \\ y_{132} \\ \vdots \\ y_{13N_{13}} \\ y_{211} \\ y_{212} \\ \vdots \\ y_{21N_{21}} \\ y_{221} \\ y_{222} \\ \vdots \\ y_{22N_{22}} \\ y_{231} \\ y_{232} \\ \vdots \\ y_{23N_{23}} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \end{pmatrix} \]

\[(7.13)\]

The \( x^{(R)}_i \) is the vector indicating observations in row \( i \), the \( x^{(C)}_j \) indicates row \( j \), and \( x^{(I)}_{ij} \) indicates the \( ij^{th} \) cell. The subspaces for the various models are next:
7.1. THE SUBSPACES AND PROJECTIONS

Saturated: \( \mathcal{M}_{R \times C} = C(X) \);
Additive: \( \mathcal{M}_{R+C} = \text{span}\{ \mathbf{1}_n, \mathbf{z}_1^{(R)}, \ldots, \mathbf{z}_r^{(R)}, \mathbf{z}_1^{(C)}, \ldots, \mathbf{z}_c^{(C)} \} \);
Just row effect: \( \mathcal{M}_R = \text{span}\{ \mathbf{1}_n, \mathbf{z}_1^{(R)}, \ldots, \mathbf{z}_r^{(R)} \} \);
Just column effect: \( \mathcal{M}_C = \text{span}\{ \mathbf{1}_n, \mathbf{z}_1^{(C)}, \ldots, \mathbf{z}_c^{(C)} \} \);
No row or column effect: \( \mathcal{M}_\emptyset = \text{span}\{ \mathbf{1}_n \} \).

\( (7.14) \)

The \( R \) and \( C \) notations are self-explanatory. The \( R \times C \) indicates that the row, column and interaction effects are in the model, whereas \( R + C \) means just the row and column effects, i.e., no interactions.

The projections and ranks for these models, except the additive model, are easy to obtain using ideas from one-way ANOVA. The saturated model is really a large one-way ANOVA with \( G = r \times c \) groups, so the projection has elements \( \mathbf{y}_{ij} \) and the rank is \( r \times c \), at least if all \( N_{ij} > 0 \). The rank in general is the number of cells with \( N_{ij} > 0 \). The model with just column effects is a one-way ANOVA with the rows as groups, hence the projection has elements \( \mathbf{y}_{i.} \) and rank \( c \). Similarly for the model with just column effects. The model without any row or column effects is also familiar, \( \hat{y}_{\emptyset} = \mathbf{y}_{..} \), with rank 1. Here is a summary:

<table>
<thead>
<tr>
<th>Model</th>
<th>Subspace</th>
<th>Projection ( \hat{y}_{ijk} )</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturated</td>
<td>( \mathcal{M}_{R \times C} )</td>
<td>( \mathbf{y}_{ij}. )</td>
<td>( r \times c )</td>
</tr>
<tr>
<td>Additive</td>
<td>( \mathcal{M}_{R+C} )</td>
<td>( \mathbf{y}<em>{i.} + \mathbf{y}</em>{.j} - \mathbf{y}_{..} \ast )</td>
<td>( r + c - 1 )</td>
</tr>
<tr>
<td>Just row effect</td>
<td>( \mathcal{M}_R )</td>
<td>( \mathbf{y}_{i.} )</td>
<td>( r )</td>
</tr>
<tr>
<td>Just column effect</td>
<td>( \mathcal{M}_C )</td>
<td>( \mathbf{y}_{.j} )</td>
<td>( c )</td>
</tr>
<tr>
<td>No row or column effect</td>
<td>( \mathcal{M}_\emptyset )</td>
<td>( \mathbf{y}_{..} )</td>
<td>1</td>
</tr>
</tbody>
</table>

\( \ast \) Only for the balanced case

The results for the additive model needs some explanation. From (7.14) we see that \( \mathcal{M}_{R+C} \) is the span of \( r + c + 1 \) vectors. It is easy to see that \( \mathbf{1}_n = \mathbf{z}_1^{(R)} + \cdots + \mathbf{z}_r^{(R)} \), so that we can eliminate \( \mathbf{1}_n \) from the set of spanning vectors. The remaining \( r + c \) vectors are still not linearly independent, because

\[
(\mathbf{z}_1^{(R)} + \cdots + \mathbf{z}_r^{(R)}) - (\mathbf{z}_1^{(C)} + \cdots + \mathbf{z}_c^{(C)}) = \mathbf{1}_n - \mathbf{1}_n = \mathbf{0}_n.
\]

\( (7.16) \)

Let us drop one of those, say \( \mathbf{z}_r^{(C)} \), and suppose that

\[
\tilde{z} \equiv a_1 \mathbf{z}_1^{(R)} + \cdots + a_r \mathbf{z}_r^{(R)} + a_{r+1} \mathbf{z}_1^{(C)} + \cdots + a_{r+c-1} \mathbf{z}_{c-1}^{(C)} = \mathbf{0}_n.
\]

\( (7.17) \)

In the 2 \times 3 case of (7.14),

\[
\tilde{z} = \begin{pmatrix}
(a_1 + a_3) \mathbf{1}_{N_{11}} \\
(a_1 + a_4) \mathbf{1}_{N_{12}} \\
(a_2 + a_3) \mathbf{1}_{N_{21}} \\
(a_2 + a_4) \mathbf{1}_{N_{22}} \\
(\ast)
\end{pmatrix}.
\]

\( (7.18) \)
We now assume that $N_{ij} > 0$ for all cells, i.e., there are no empty cells. If there are, the rank can be a bit more challenging to figure out. Consider the elements in the $C^{th}$ column, i.e., the $z_{iCk}$’s. Because that column’s vector is missing, only the coefficient multiplying the appropriate row vector shows up, so that $z_{iCk} = a_i$. Thus if \( \hat{a} = \emptyset \), it must be that $a_1 = \cdots = a_r = 0$. But then from (7.17) it is easy to see that the $a_{r+1} = \cdots = a_{r+c-1} = 0$ as well. Thus those vectors are linearly independent, showing that the rank of $M_{R+C}$ is indeed $r + c - 1$.

In the unbalanced case, the projection for the additive model has no simple form. One way to obtain the projection matrix is to use the $X^* = (x_{11}^{(R)}, \ldots, x_{1r}^{(R)}, x_{21}^{(C)}, \ldots, x_{2c}^{(C)}$, so that $M_{R+C} = X^*(X^*X^*)^{-1}X^*$.

Turn to the balanced case, so that all $N_{ij} = N$. The table (7.15) has that the projection is

\[
\hat{y}_{R+C} = -\overline{y}_c \mathbb{1}_n + \overline{y}_1 \cdot x_{11}^{(R)} + \cdots + \overline{y}_r \cdot x_{r1}^{(R)} + \overline{y}_1 \cdot x_{11}^{(C)} + \cdots + \overline{y}_c \cdot x_{c1}^{(C)}.
\]

That vector is clearly in $M_{R+C}$, being a linear combination of the spanning vectors. We also need that $\overline{y} - \hat{y}_{R+C} \perp M_{R+C}$. Start by looking at $x_{i1}^{(R)}$:

\[
(\overline{y} - \hat{y}_{R+C})' x_{i1}^{(R)} = \overline{y}' x_{i1}^{(R)} + \sum_{j=1}^r \sum_{k=1}^N y_{ijk} - \overline{y}_1 \cdot x_{11}^{(R)} - \cdots - \overline{y}_r \cdot x_{r1}^{(R)} - \overline{y}_1 \cdot x_{11}^{(C)} - \cdots - \overline{y}_c \cdot x_{c1}^{(C)}.
\]

By inspecting (7.14) with $N_{ij} = N$, one can see that

\[
\begin{align*}
\overline{y}' x_{i1}^{(R)} &= \sum_{j=1}^N \sum_{k=1}^N y_{ijk} = \overline{y}_i \cdot Nc \\
\overline{y}_1 x_{11}^{(R)} &= Nc \\
x_{11}^{(R)} x_{11}^{(R)'} &= Nc \\
x_{i1}^{(R)'} x_{i1}^{(R)} &= 0 \text{ if } i \neq j \\
x_{i1}^{(C)'} x_{i1}^{(R)} &= N,
\end{align*}
\]

hence

\[
(\overline{y} - \hat{y}_{R+C})' x_{i1}^{(R)} = \overline{y}_i \cdot Nc + \sum_{j=1}^N \sum_{k=1}^N y_{ijk} Nc - \overline{y}_1 \cdot Nc - \cdots - \overline{y}_c \cdot Nc
\]

\[
= \overline{y}_i \cdot Nc - (\overline{y}_1 + \cdots + \overline{y}_c ) N
\]

\[
= \left( \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N y_{ijk} Nc/(Ncr) - \sum_{i=1}^r \sum_{k=1}^N y_{i1k}/(Nc) \right) + \cdots + \sum_{r=1}^r \sum_{k=1}^N y_{r1k}/(Nc) N
\]

\[
= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N y_{ijk}/r - \sum_{i=1}^r \sum_{k=1}^N y_{i1k} + \cdots + \sum_{r=1}^r \sum_{k=1}^N y_{r1k}/r
\]

\[
= 0.
\]

Likewise, $(\overline{y}' x_{i1}^{(R)} x_{i1}^{(R)'} = 0$ for each $j$. Thus $\overline{y} - \hat{y}_{R+C} \perp M_{R+C}$, showing that (7.19) is indeed the projection.
7.2 Testing hypotheses

There are many possible hypotheses testing problems based on nested subspaces among those in (7.12), e.g., to test whether there are interactions, we would have

\[ H_0 : \mu \in \mathcal{M}_{R+C} \quad \text{(additive model)} \quad \text{vs.} \quad H_A : \mu \in \mathcal{M}_{R \times C} \quad \text{(saturated model)}, \]

(7.23)

or to test whether there are row effects in the additive model,

\[ H_0 : \mu \in \mathcal{M}_C \quad \text{(just column effect, no row effect)} \quad \text{vs.} \quad H_A : \mu \in \mathcal{M}_{R+C} \quad \text{(additive model)}. \]

(7.24)

The picture below shows the various possible nestings, where the arrow points from the smaller to the larger subspace, that is, \( \mathcal{M}_{R+C} \rightarrow \mathcal{M}_{R \times C} \) means that \( \mathcal{M}_{R+C} \subset \mathcal{M}_{R \times C} \).

\[
\begin{align*}
\mathcal{M}_{R \times C} & \quad \rightarrow \quad \mathcal{M}_{R+C} \\
& \quad \leftarrow \quad \mathcal{M}_R \\
& \quad \leftarrow \quad \mathcal{M}_C \\
& \quad \leftarrow \quad \mathcal{M}_0
\end{align*}
\]

Notice that \( \mathcal{M}_R \) and \( \mathcal{M}_C \) are not nested, but every other pair is (one way or the other). That does not mean one cannot test \( \mathcal{M}_R \) versus \( \mathcal{M}_C \), just that the approach we are taking with \( F \) tests and ANOVA tables does not work.

Testing for interaction, testing for row effects, and testing for column effects, are the main goals. Testing for interaction is unambiguous, but there are several ways to set up testing row or column effects. Here are some possible testing problems:

- Testing for interaction – The additive versus the saturated model:

\[ H_0 : \mu \in \mathcal{M}_{R+C} \quad \text{versus} \quad H_A : \mu \in \mathcal{M}_{R \times C}. \]

(7.25)

- Testing for row effects:

  - With the understanding that if there are interactions, there automatically are row effects, one defines total row effects as either nonzero interactions, or if interactions are zero, nonzero row effects. The null hypothesis is then \( \mathcal{M}_C \), and the alternative is the saturated model:

\[ H_0 : \mu \in \mathcal{M}_C \quad \text{versus} \quad H_A : \mu \in \mathcal{M}_{R \times C}. \]

(7.26)

  - One may have already decided that there are no interactions, but there may be column effects. In that case, the testing problem is like in (7.26), but the alternative is the additive model:

\[ H_0 : \mu \in \mathcal{M}_C \quad \text{versus} \quad H_A : \mu \in \mathcal{M}_{R+C}. \]

(7.27)
It may be that one has already decided there are no column effects, either, so that the problem is really a one-way ANOVA:

\[ H_0 : \mu \in \mathcal{M}_\emptyset \quad versus \quad H_A : \mu \in \mathcal{M}_R. \]  

(7.28)

- Testing for column effects, analogous to testing for row effects:
  
  - Testing total column effects
    
    \[ H_0 : \mu \in \mathcal{M}_R \quad versus \quad H_A : \mu \in \mathcal{M}_{R \times C}. \]  
    
    (7.29)

  - Assuming there are no interactions:
    
    \[ H_0 : \mu \in \mathcal{M}_R \quad versus \quad H_A : \mu \in \mathcal{M}_{R+C}. \]  
    
    (7.30)

  - Assuming there are no row effects:
    
    \[ H_0 : \mu \in \mathcal{M}_\emptyset \quad versus \quad H_A : \mu \in \mathcal{M}_C. \]  
    
    (7.31)

- Simultaneously testing for row and column effects:
  
  - Testing total effects
    
    \[ H_0 : \mu \in \mathcal{M}_\emptyset \quad versus \quad H_A : \mu \in \mathcal{M}_{R \times C}. \]  
    
    (7.32)

  - Assuming there are no interactions:
    
    \[ H_0 : \mu \in \mathcal{M}_\emptyset \quad versus \quad H_A : \mu \in \mathcal{M}_{R+C}. \]  
    
    (7.33)

The ANOVA table for any of these testing problems is straightforward to calculate if all \( N_{ij} > 0 \) (see table (7.15)), although if it is unbalanced and involves the additive model, the \( \hat{y}_{R+C} \) may require a matrix inversion.

### 7.2.1 Example

The text (Table 7.4) repeats some data from Scheffé, *The Analysis of Variance*, problem 4.8, about the weights of female rats. The experiment had litters of rats born to one mother but raised by another. The row factor is the genotype of the litter, and the column factor is the genotype of the foster mother. There were \( n = 61 \) litters, and the \( y \)'s are the average weight of the litters in grams at 28 days. The design is unbalanced. The tables below contains the cell means, \( \overline{y}_{ij} \), and the \( N_{ij} \)'s, respectively. See the book for the individual data points.
7.2. TESTING HYPOTHESES

<table>
<thead>
<tr>
<th>Foster mother →</th>
<th>( \bar{y}_{ij} )</th>
<th>A</th>
<th>F</th>
<th>I</th>
<th>J</th>
<th>( \bar{y}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>63.680</td>
<td>52.400</td>
<td>54.125</td>
<td>48.960</td>
<td>55.112</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>52.325</td>
<td>60.640</td>
<td>53.925</td>
<td>45.900</td>
<td>54.667</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>47.100</td>
<td>64.367</td>
<td>51.600</td>
<td>49.433</td>
<td>52.907</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>54.350</td>
<td>56.100</td>
<td>54.533</td>
<td>49.060</td>
<td>52.973</td>
<td></td>
</tr>
<tr>
<td>( \bar{y}_j )</td>
<td>55.400</td>
<td>58.700</td>
<td>53.362</td>
<td>48.680</td>
<td>53.970</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Foster mother →</th>
<th>( N_{ij} )</th>
<th>A</th>
<th>F</th>
<th>I</th>
<th>J</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>14</td>
<td>16</td>
<td>15</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

These two table give enough information to easily find the sums of squares for all but the additive subspace. To find the projection onto the additive subspace \( \mathcal{M}_{R+C} \), we need to set up an \( X \) matrix with linearly independent columns. There are a number of possibilities. We will place the restrictions on the parameters:

\[
\sum_{i=1}^{4} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{4} \beta_j = 0,
\]

which means that \( \alpha_4 = -\alpha_1 - \alpha_2 - \alpha_3 \) (and likewise for the \( \beta_j \)'), giving

\[
X = (I_{41}, \; \xi^{(R)} - \xi^{(R)}; \xi^{(R)} - \xi^{(R)}; \xi^{(R)} - \xi^{(R)}; \xi^{(R)}; \xi^{(R)} - \xi^{(R)}; \xi^{(C)} - \xi^{(C)}; \xi^{(C)} - \xi^{(C)}; \xi^{(C)}), \]

and the \( \beta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)' \). Note that there are \( r + c - 1 = 7 \) columns, which is the correct number. The next table gives the estimates and the projections for each \( ij \):

<table>
<thead>
<tr>
<th>Additive model</th>
<th>( \hat{y}_{(R+C)ij} )</th>
<th>A</th>
<th>F</th>
<th>I</th>
<th>J</th>
<th>( \hat{\alpha}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foster mother →</td>
<td>( \hat{y}_{(R+C)ij} )</td>
<td>A</td>
<td>F</td>
<td>I</td>
<td>J</td>
<td>( \hat{\alpha}_i )</td>
</tr>
<tr>
<td>A</td>
<td>56.909</td>
<td>60.425</td>
<td>55.077</td>
<td>50.154</td>
<td>1.675</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>54.884</td>
<td>58.400</td>
<td>53.052</td>
<td>48.129</td>
<td>-0.350</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>54.255</td>
<td>57.771</td>
<td>52.423</td>
<td>47.501</td>
<td>-0.979</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>54.888</td>
<td>58.404</td>
<td>53.056</td>
<td>48.133</td>
<td>-0.346</td>
<td></td>
</tr>
</tbody>
</table>
| \( \beta_j \)  | 1.268           | 4.784  | -0.564 | -5.487 | \( \hat{\mu} = 53.966 \)

The next table gives the sums of squares, and ranks, for the subspaces.
CHAPTER 7. TWO-WAY ANOVA

Subspace | SS | rank |
---|---|---|
$\mathcal{M}_\emptyset$ | $\|\hat{y}_n\|^2 = 177681.7$ | 1 |
$\mathcal{M}_R$ | $\|\hat{y}_R\|^2 = 177741.8$ | 4 |
$\mathcal{M}_C$ | $\|\hat{y}_C\|^2 = 178453.3$ | 4 |
$\mathcal{M}_{R+C}$ | $\|\hat{y}_{R+C}\|^2 = 178516.9$ | 7 |
$\mathcal{M}_{R\times C}$ | $\|\hat{y}_{R\times C}\|^2 = 179341.0$ | 16 |
$\mathbb{R}^n$ | $\|y\|^2 = 181781.8$ | 61 |

Interaction. The first test is usually to see if there is any interaction. The next plots illustrate the interactions. The left-hand plot has plots the $\overline{y}_{ij}$’s versus column number $j$, and the means for each row are connected. If there is exact additivity, then these lines should be parallel. One of the lines looks particularly nonparallel to the others. The right-hand plot has the fits for the additive model, the $\hat{y}_{(R+C)ij1} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$’s. Because the model enforces additivity, the lines are parallel.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactions</td>
<td>824.1</td>
<td>9</td>
<td>91.567</td>
<td>1.689</td>
</tr>
<tr>
<td>Error</td>
<td>2440.8</td>
<td>45</td>
<td>54.240</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>3264.9</td>
<td>54</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The ANOVA table for testing additivity, $\mathcal{M}_{R+C}$ versus $\mathcal{M}_{R\times C}$, is now easy. From table (7.39),

$$\|\hat{y} - \hat{y}_{R\times C}\|^2 = 181781.8 - 179341.0 = 2440.8,$$

and

$$\|\hat{y} - \hat{y}_{R+C}\|^2 = 181781.8 - 178516.9 = 3264.9.$$

The degrees of freedom for those two sums of squares are, respectively, $n-rc$ and $n-(r+c-1)$, hence the ANOVA table is

$$\|\hat{y} - \hat{y}_{R\times C}\|^2 = 181781.8 - 179341.0 = 2440.8,$$
The $F_{9,45,0.05} = 2.096$, so that the interactions are not significantly different from 0. Thus, though the plot may suggest interactions, according to this test they are not statistically significant, so that there is not enough evidence to reject additivity.

**Note** that the degrees of freedom for interaction are $n-rc-(n-(r+c-1)) = (r-1) \times (c-1)$.

**Testing for row effects.** Presuming no interactions, we have two choices for testing whether there are row effects: Assuming there are no column effects, or allowing column effects. We will do both.

First, let us allow column effects, so that $M_0 = M_C$ and $M_A = M_{R+C}$. Then $\|\mathbf{y} - \hat{\mathbf{y}}_{R+C}\|^2 = 3264.9$ from (7.40), and

$$\|\mathbf{y} - \hat{\mathbf{y}}_C\|^2 = 181781.8 - 178453.3 = 3328.5. \quad (7.41)$$

The ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>63.6</td>
<td>3</td>
<td>21.200</td>
<td>0.531</td>
</tr>
<tr>
<td>Error</td>
<td>3264.9</td>
<td>54</td>
<td>60.461</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>3328.5</td>
<td>57</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The rows are not at all significant.

Now suppose there are no column effects, so that $M_0 = M_0$ and $M_A = M_R$. This time, $\|\mathbf{y} - \hat{\mathbf{y}}_R\|^2 = 181781.8 - 177741.8 = 4040.0$, $\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 = 181781.8 - 177681.7 = 4100.1$, (7.42) and

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>60.1</td>
<td>3</td>
<td>20.033</td>
<td>0.283</td>
</tr>
<tr>
<td>Error</td>
<td>4040.0</td>
<td>57</td>
<td>70.877</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>4100.1</td>
<td>60</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Again, the conclusion is no evidence of row effect, that is, effect of the litters’ genotypes. Note that though the conclusion is the same, the calculations are different depending on what one assumes about the column effects.

**Testing for column effects.** From above, it seems perfectly reasonable to assume no row effects, so we will do just that. Then to test for column effects, we have $M_0 = M_0$ and $M_A = M_C$. Equations (7.41) and (7.42) already have the necessary calculations, so the ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns</td>
<td>771.6</td>
<td>3</td>
<td>257.2</td>
<td>3.629</td>
</tr>
<tr>
<td>Error</td>
<td>3328.5</td>
<td>57</td>
<td>70.877</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>4100.1</td>
<td>60</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Now $F_{3,57,0.05} = 2.766$, so there does appear a significant column effect, that is, the foster mother’s genotype does appear to affect the weights of the litters.

Taking everything together, it looks like the best model has just the column effects, $y_{ijk} = \mu + \beta_j + e_{ijk}$. 
7.3 The balanced case

The example in the previous section required specifying whether or not there are row effects before knowing how to test for column effects, and vice versa. That is,

\[ \| \hat{y}_{R+C} - \hat{y}_C \|^2 \neq \| \hat{y}_R - \hat{y}_\emptyset \|^2 \quad \text{and} \quad \| \hat{y}_{R+C} - \hat{y}_R \|^2 \neq \| \hat{y}_C - \hat{y}_\emptyset \|^2. \] (7.43)

It turns out that in the balanced case, the row sums of squares are the same whether or not the column effects are in the model, so that there are equalities in (7.43). The main reason is the next lemma.

**Lemma 3** In the balanced case, \( M_{R \emptyset} \perp M_{C \emptyset}. \)

**Proof.** Elements \( x \in M_{R \emptyset} \) are in \( M_R \), which means \( x_{ijk} = a_i \), and are orthogonal to \( M_\emptyset \), which means that the elements sum to zero. Because the design is balanced,

\[ \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N x_{ijk} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N a_i = (cN) \sum_{i=1}^r a_i, \] (7.44)

hence

\[ \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N x_{ijk} \Rightarrow \sum_{i=1}^r a_i = 0. \] (7.45)

Similarly, if \( z \in M_{C \emptyset} \), \( z_{ijk} = b_j \), and

\[ \sum_{j=1}^c b_j = 0. \] (7.46)

Then

\[ x'z = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N x_{ijk}z_{ijk} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^N a_i b_j = N(\sum_{i=1}^r a_i)(\sum_{j=1}^c b_j) = 0. \] (7.47)

Thus \( x \perp z \), i.e., \( M_{R \emptyset} \perp M_{C \emptyset} \). \( \square \)

Now the main result shows that the projection onto the additive-effect space can be decomposed into the projections onto the row- and column-effect spaces.

**Proposition 18** In the balanced case,

\[ \hat{y}_{(R+C) \emptyset} = \hat{y}_{R \emptyset} + \hat{y}_{C \emptyset}. \] (7.48)

“Proof.” The proof is basically the same as for the decomposition (6.11) of the between sum of squares. That is, one has to show that \( \hat{y}_{R \emptyset} + \hat{y}_{C \emptyset} \in M_{(R+C) \emptyset} \) and \( y - (\hat{y}_{R \emptyset} + \hat{y}_{C \emptyset}) \perp M_{(R+C) \emptyset}. \)

The proof of these will be left to the reader. (Or more accurately, to the homework-doers.) \( \square \)
7.3. THE BALANCED CASE

Even in the unbalanced case, \( \hat{y}_{(R+C)} \) can be written as the sum of something in \( \mathcal{M}_{R} \) plus something in \( \mathcal{M}_{C} \). It is just that in the unbalanced case, those two somethings might not be the projections on their respective spaces.

Now we can expand on the decomposition (7.48). Clearly,

\[
y - \hat{y}_0 = (y - \hat{y}_{R+C}) + (\hat{y}_{R+C} - \hat{y}_R) + (\hat{y}_R - \hat{y}_0),
\]

but then (7.48) yields

\[
y - \hat{y}_0 = (y - \hat{y}_{R+C}) + (\hat{y}_{R+C} - \hat{y}_{R+C}) + (\hat{y}_R - \hat{y}_R) + (\hat{y}_C - \hat{y}_0).
\]

The four terms on the right-hand side are projections onto, respectively, \( \mathcal{M}_{R \times C}, \mathcal{M}_{(R 	imes C) - (R+C)}, \mathcal{M}_{R}, \) and \( \mathcal{M}_{C} \). It is always true (in balanced or unbalanced designs) that the first one is orthogonal to the other three, and the second is orthogonal to the last two. In the balanced case, the last two are also orthogonal (Lemma 3), so that the sums of squares add up:

\[
\begin{align*}
\|y - \hat{y}_0\|^2 &= \|y - \hat{y}_{R+C}\|^2 + \|\hat{y}_{R+C} - \hat{y}_C\|^2 + \|\hat{y}_R - \hat{y}_0\|^2 + \|\hat{y}_C - \hat{y}_0\|^2; \\
\text{Total SS} &= \text{Error SS} + \text{Interaction SS} + \text{Row SS} + \text{Column SS}.
\end{align*}
\]

The balance also allows easy calculation of the sums of squares using the estimated effects (7.10):

\[
\begin{align*}
\|y - \hat{y}_{R+C}\|^2 &= \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{N} (y_{ijk} - \hat{y}_{ijk})^2 \\
\|\hat{y}_{R+C} - \hat{y}_C\|^2 &= \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{N} (\hat{y}_{ijk} - (\hat{y}_{i\cdot} + \hat{y}_{j\cdot} - \hat{y}_{\cdot\cdot}))^2 = N \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\beta}_{ij}^2 \\
\|\hat{y}_R - \hat{y}_0\|^2 &= \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{N} (\hat{y}_{i\cdot} - \hat{y}_{\cdot\cdot})^2 = Nc \sum_{i=1}^{r} \hat{\alpha}_{i}^2 \\
\|\hat{y}_C - \hat{y}_0\|^2 &= \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{N} (\hat{y}_{j\cdot} - \hat{y}_{\cdot\cdot})^2 = Nr \sum_{j=1}^{c} \hat{\beta}_{j}^2.
\end{align*}
\]

Another benefit is that the sum of squares for testing row effects is the same whether or not the column effects are included in the model. Starting with the Row SS in the presence of column effects, we have

\[
\begin{align*}
\|\hat{y}_{R+C} - \hat{y}_C\|^2 &= \|\hat{y}_{R+C} - \hat{y}_R + \hat{y}_R - \hat{y}_0\|^2 \\
&= \|\hat{y}_R - \hat{y}_R + \hat{y}_C - \hat{y}_0\|^2 \\
&= \|\hat{y}_R - \hat{y}_0\|^2,
\end{align*}
\]

which is the Row SS without the column effects.

The decomposition (7.51) leads to the expanded ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>( Nc \sum_{i=1}^{r} \hat{\alpha}_{i}^2 )</td>
<td>( r - 1 )</td>
<td>( SSR/\text{df}R )</td>
<td>( \text{MSR/\text{MSE}} )</td>
</tr>
<tr>
<td>Columns</td>
<td>( Nr \sum_{j=1}^{c} \hat{\beta}_{j}^2 )</td>
<td>( c - 1 )</td>
<td>( SSC/\text{df}C )</td>
<td>( \text{MSC/\text{MSE}} )</td>
</tr>
<tr>
<td>Interactions</td>
<td>( N \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\gamma}_{ij}^2 )</td>
<td>( (r - 1)(c - 1) )</td>
<td>( SSI/\text{df}I )</td>
<td>( \text{MSI/\text{MSE}} )</td>
</tr>
<tr>
<td>Error</td>
<td>( \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{N} (y_{ijk} - \hat{y}_{ijk})^2 )</td>
<td>( n - rc )</td>
<td>( SSE/\text{df}E )</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>( \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{N} (y_{ijk} - \hat{y}_{ijk})^2 )</td>
<td>( n - 1 )</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

It is then easy to perform the \( F \) tests.
### 7.3.1 Example

Below are data on 30 hyperactive boys, measuring “out-of-seat behavior.” Each boy was given one of two types of therapy, Behavioral or Cognitive, and Ritalin in one of three dosages, Low, Medium, or High. The design is balanced, so that here are \( N = 5 \) boys receiving each therapy/dose combination.

<table>
<thead>
<tr>
<th>Dose ( \rightarrow )</th>
<th>( L )</th>
<th>( M )</th>
<th>( H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral</td>
<td>54</td>
<td>56</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>56</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>55</td>
<td>56</td>
</tr>
<tr>
<td>Cognitive</td>
<td>52</td>
<td>50</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>54</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>57</td>
<td>55</td>
</tr>
</tbody>
</table>

[From http://espse.ed.psu.edu/statistics/Chapters/Chapter12/.

The various means:

\[
\begin{array}{ccc}
\bar{y}_{ij} & \bar{y}_{i..} & \bar{y}_{..} = 54.767 \\
55.000 & 54.000 & 54.000 \\
52.000 & 56.000 & 58.000 \\
\end{array}
\]

Plugging those numbers into (7.52) yields

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapy (Rows)</td>
<td>5.64</td>
<td>1</td>
<td>5.640</td>
<td>1.627</td>
</tr>
<tr>
<td>Dosage (Columns)</td>
<td>31.27</td>
<td>2</td>
<td>15.635</td>
<td>4.508</td>
</tr>
<tr>
<td>Interactions</td>
<td>63.26</td>
<td>2</td>
<td>31.630</td>
<td>9.123</td>
</tr>
<tr>
<td>Error</td>
<td>83.20</td>
<td>24</td>
<td>3.467</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>183.37</td>
<td>29</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Now \( F_{2,24,0.001} = 9.339 \), so that the interaction \( F \) is almost significant at the \( \alpha = 0.001 \) level, and certainly at the 0.005 level, meaning there is interaction. Thus there are also row and column effects. The \( F_{1,24,0.05} = 4.260 \), which means it looks as though the row effects are not significant, but it really means the row effect averaged over columns is not significant.

Consider the difference between the therapies for each dosage. The estimate of the difference for column \( j \), \( \mu_{1j} - \mu_{2j} \), is \( \bar{y}_{1j} - \bar{y}_{2j} \), which has variance \((2/5)\sigma_e^2\), because it is a difference of two independent means of 5 observations. Hence \( se = \sqrt{(2/5) \times 3.467} = 1.178 \), and

\[
\begin{align*}
\hat{\mu}_{11} - \hat{\mu}_{21} &= 55 - 52 = 3.0, \quad t = 2.55 \\
\hat{\mu}_{12} - \hat{\mu}_{22} &= 54 - 55.6 = -1.6, \quad t = -1.36 \\
\hat{\mu}_{13} - \hat{\mu}_{23} &= 54 - 58 = -4.0, \quad t = -3.40
\end{align*}
\]

Now we see that for low dosage, cognitive therapy has a significantly lower score than behavioral therapy, while for high dosage, the behavioral therapy has a significantly lower score than cognitive. (The difference is not significant for the medium dosage.) These differences are canceled when averaging, so it looks like therapy has no effect.

Another way to look at the data is to notice that for behavioral therapy, Ritalin dosage seems to have no effect, whereas for cognitive therapy, the scores increase with dosage.
7.4 Balance is good

Balanced designs are preferable to unbalanced ones for a number of reasons:

1. **Interpretation is easier.** In balanced two-way designs, the row and column effects are not confounded. That is, the row effects are the same whether or not the column effects are in the model, and vice versa. In unbalanced designs, there is confounding. E.g., suppose these are the cell means and \( N_{ij} \)'s for a small ANOVA:

<table>
<thead>
<tr>
<th>Fertilizer ↓; Soil →</th>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>115.3</td>
<td>52.7</td>
</tr>
<tr>
<td>B</td>
<td>114.0</td>
<td>52.8</td>
</tr>
</tbody>
</table>

If \( y_{ijk} \) measures yield of corn, then you would expect a strong column effect, with the first column having better yield than the second. With those data, \( \overline{y}_{1.} = 115.04 \) and \( \overline{y}_{2.} = 52.77 \). Now if you ignored the columns, and just tested for row effect, you see that Fertilizer A is better than Fertilizer B: \( \overline{y}_{1.} = 102.78 \) and \( \overline{y}_{2.} = 73.2 \). Would that be because A is better, or because A happens to have more observations in the good soil? That question arises because quality of the soil and fertilizer are **confounded**. We can see that for each type of soil, A and B are about equal. If all \( N_{ij} = 10 \), then there would not be such confounding, and the two fertilizers would look essentially equal: \( \overline{y}_{1.} = 84 \) and \( \overline{y}_{2.} = 83.4 \).

2. **Computations are easier.** With a balanced design, no matrix inversions are necessary. Everything is based on the various means.

3. **The design is more efficient.** By efficiency we mean that the variances of the estimates tend to be low for a given number of observations. That is, consider the \( N_{ij} \)'s above, where the total \( n = 40 \). Then the overall variance of the estimates of the cell means is

\[
\text{Var}(\overline{y}_{11}.) + \text{Var}(\overline{y}_{12}.) + \text{Var}(\overline{y}_{21}.) + \text{Var}(\overline{y}_{22}.) = \sigma_e^2 \left( \frac{1}{20} + \frac{1}{5} + \frac{1}{5} + \frac{1}{10} \right) = (0.55) \sigma_e^2. \tag{7.59}
\]

If, on the other hand, all \( N_{ij} = 10 \), which still gives \( n = 40 \),

\[
\text{Var}(\overline{y}_{11}.) + \text{Var}(\overline{y}_{12}.) + \text{Var}(\overline{y}_{21}.) + \text{Var}(\overline{y}_{22}.) = \sigma_e^2 \left( \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} \right) = (0.40) \sigma_e^2. \tag{7.60}
\]

That is, the overall variance in the balanced case is 27% smaller than in the unbalanced case, with the same number of observations.

The above benefits are true in general in statistics: balance is good.
Chapter 8

Multiple comparisons

When you preform a hypothesis test, or find a confidence interval, you often try to control the chance of an error. In the testing situation, the Type I error is the chance you reject the null hypothesis when the null hypothesis is true, that is, it is the chance of a false positive. One often wants that chance to be no larger than a specified level \( \alpha \), e.g., 5%. With confidence intervals, you wish that the chance the parameter of interest is covered by the interval to be at least \( 1 - \alpha \), e.g., 95%.

Suppose there are several tests or confidence intervals begin considered simultaneously. For example, in the leprosy example, one may want confidence intervals for \( \gamma_1 = (\alpha_1 + \alpha_2)/2 - \alpha_3 \) and \( \gamma_2 = \alpha_1 - \alpha_2 \). If each confidence interval has a 95% chance of covering its true value, what is the chance that both cover their true values simultaneously? It is somewhere between 90% and 95%. Thus we can at best guarantee that there is a 90% chance both are correct. The difference between 90% and 95% may not be a big deal, but consider more than two intervals, say 5, or 10, or 20. The chance that ten 95% intervals are all correct is bounded from below by 50%: For twenty, the bound is 0%!

To adjust the intervals so that the chance is 95% (or whatever is desired) that all are correct, one must widen them. Suppose there are \( J \) intervals, and \( \alpha_O \) is the overall error rate, that is, we want the chance that all intervals cover their parameters to be at least \( 1 - \alpha_O \). Then instead of using \( \hat{\gamma}_i \pm t_{\nu,\alpha/2} se(\hat{\gamma}_i) \), the intervals would be

\[
\hat{\gamma}_i \pm C_{\alpha_O} se(\hat{\gamma}_i),
\]

where the \( C_{\alpha_O} \) is some constant that is larger than the \( t_{\nu,\alpha} \). This constant must satisfy

\[
P[\gamma_i \in (\hat{\gamma}_i \pm C_{\alpha_O} se(\hat{\gamma}_i)) \ , i = 1, \ldots, J] \geq 1 - \alpha_O.
\]

There are many methods for choosing the \( C_{\alpha_O} \), some depending on the type of parameters being considered. We will consider three methods:

- **Bonferroni**, for \( J \) not too large;
- **Tukey**, for all pairwise comparisons, i.e., all \( \mu_i - \mu_j \)'s;
- **Scheffé**, for all contrasts of group means (so that \( J = \infty \)).
8.1 Bonferroni Bounds

Bonferroni bounds are the simplest and most generally applicable, although they can be quite conservative. The idea is that if each interval has individually a chance of $1 - \alpha$ of covering its parameter, then the chance all $J$ simultaneously cover their parameters is $1 - J\alpha$. To see this fact, let $A_i$ be the event that the $i$th interval is ok, and $\alpha_i$ be its individual coverage probability. That is,

$$A_i \text{ is true iff } \gamma_i \in (\hat{\gamma}_i \pm t_{\nu,\alpha_i/2} se(\hat{\gamma}_i)); \quad P[A_i] = 1 - \alpha_i. \quad (8.3)$$

Then if $1 - \alpha_O$ is the desired chance all intervals cover their parameters, we need that

$$P[A_1 \cap A_2 \cap \cdots \cap A_J] \geq 1 - \alpha_O. \quad (8.4)$$

Looking at complements instead, we have that $P[A_i^c] = \alpha_i$, and wish to have

$$P[A_1^c \cup A_2^c \cup \cdots \cup A_J^c] \leq \alpha_O. \quad (8.5)$$

The probability of a union is less than or equal to the sum of probabilities, hence

$$P[A_1^c \cup A_2^c \cup \cdots \cup A_J^c] \leq P[A_1^c] + P[A_2^c] + \cdots + P[A_J^c] = \alpha_1 + \cdots + \alpha_J. \quad (8.6)$$

So how can we choose the $\alpha_i$’s so that

$$\alpha_1 + \cdots + \alpha_J \leq \alpha_O? \quad (8.7)$$

Just take each $\alpha_i = \alpha_O/J$. (Of course, there are many other choices that will work, too.) In any case, we have proved the next proposition:

**Proposition 19** If each of $J$ confidence intervals has a coverage probability of $1 - \alpha_O/J$, then the chance that all intervals cover their parameters is at least $1 - \alpha_O$.

In the leprosy example, with the covariate, we have from (4.62) that the 95% confidence interval for $(\alpha_1 + \alpha_2)/2 - \alpha_3$ is

$$(-3.392 \pm 2.056 \times 1.641) = (-6.77, -0.02), \quad (8.8)$$

where the $t_{26,0.025} = 2.056$. For $\alpha_1 - \alpha_2$, it is

$$(-1.09 \pm 2.056 \times 1.796) = (-4.78, 2.60). \quad (8.9)$$

Now if we want simultaneous coverage to be 95%, we would take the individual $\alpha_i = 0.05/2 = 0.025$, so that the $t$ in the intervals would be

$$t_{26,0.025/2} = t_{26,0.0125} = 2.379, \quad (8.10)$$
which is a bit larger than the 2.056. The intervals are then

\[
\begin{align*}
(\alpha_1 + \alpha_2)/2 - \alpha_3 : & \quad (-3.392 \pm 2.379 \times 1.641) = (-7.30, 0.51); \\
\alpha_1 - \alpha_2 : & \quad (-1.09 \pm 2.379 \times 1.796) = (-5.36, 3.18). \\
\end{align*}
\]

(8.11)

Notice that these are wider, and in fact now the first one includes zero, suggesting that the first contrast is not significant. Recall the \( F \) test for the treatment effects around (5.51). We found the effect not significant, but the Drugs versus Placebo contrast was significant. Part of the disconnect was that the \( F \) test takes into account all possible differences. When we adjust for the fact that we are considering two contrasts, neither contrast is not significant, agreeing with the \( F \) test.

The big advantage of the Bonferroni approach is that there are no special assumptions on the relationships between the intervals. They can be any type \( (t, z, \text{etc.}) \); they can be from the same experiment or different experiments; they can be on any combination of parameters. They main drawback is the conservativeness. That is, because there is a \( \geq \) in (8.4), one is likely to be understating the overall coverage probability, maybe by a great deal if \( J \) is large. More accurate bounds in special cases lead to smaller intervals (which is good), without violating the coverage probability bound. The next two sections deal with such cases.

### 8.1.1 Testing

Testing works the same way. That is, if one has \( J \) hypothesis tests and wishes to have an overall Type I error rate of \( \alpha_O \), then each individual test is performed at the \( \alpha_O/J \) level. This result can be shown using (8.6) again, where now \( A_i \) is the event that the \( i^{th} \) test accepts the null hypothesis (so that \( A_i^c \) is that it rejects), and the probabilities are calculated assuming all null hypotheses are true.

To illustrate on the example above, in order to achieve an overall \( \alpha_O = 0.05 \), we use individual \( t \)-tests of level \( \alpha_O/2 = 0.025 \). That is, we reject the \( i^{th} \) null hypothesis, \( H_0 : \gamma_i = 0 \), if

\[
\frac{|\hat{\gamma}_i|}{se(\hat{\gamma}_i)} \geq t_{26.0.025/2} = 2.379. 
\]

(8.12)

The two statistics are \(| -3.392|/1.641 = 2.067\) and \(| -1.09|/1.796 = 0.607\), neither of which are significant.

### 8.2 Tukey’s method for comparing means pairwise

In the balanced one-way ANOVA, one may wish to make all possible pairwise comparisons between the means, that is, test all the null hypotheses \( H_0 : \mu_i = \mu_j \), or equivalently, the \( H_0 : \alpha_i = \alpha_j \)'s. Or, finding the confidence intervals for all the \( \alpha_i - \alpha_j \)'s. In this case, we have that

\[
\alpha_i - \alpha_j = \bar{y}_i - \bar{y}_j, \quad \text{and} \quad se(\alpha_i - \alpha_j) = \hat{\sigma}_e \sqrt{2/N}.
\]

(8.13)
Assuming that there are \( g \) means to compare, we have \( J = \binom{g}{2} \) confidence intervals.

The goal here is to find the \( C \) so that

\[
P[\mu_i - \mu_j \in (\overline{y}_i - \overline{y}_j, \pm C_{\alpha_o}\hat{\sigma}_e \sqrt{2/N}) \text{ for all } 1 \leq i < j \leq g] \geq 1 - \alpha_o. \tag{8.14}
\]

First,

\[
\mu_i - \mu_j \in (\overline{y}_i - \overline{y}_j, \pm C_{\alpha_o}\hat{\sigma}_e \sqrt{2/N}) \quad \Leftrightarrow \quad \frac{|(\overline{y}_i - \overline{y}_j) - (\mu_i - \mu_j)|}{\hat{\sigma}_e / \sqrt{N}} \leq C_{\alpha_o} \sqrt{2}
\]

\[
\Leftrightarrow \frac{|(\overline{y}_i - \overline{y}_j)^2 - (\mu_i - \mu_j)^2|}{\hat{\sigma}_e / \sqrt{N}} \leq C_{\alpha_o} \sqrt{2}. \tag{8.15}
\]

Next,

\[
\frac{|(\overline{y}_i - \mu_i) - (\overline{y}_j - \mu_j)|}{\hat{\sigma}_e / \sqrt{N}} \leq C_{\alpha_o} \sqrt{2} \quad \text{for all } 1 \leq i < j \leq g \quad \Leftrightarrow \quad T \equiv \max_{1 \leq i \leq g} (\overline{y}_i - \mu_i) - \min_{1 \leq i \leq g} (\overline{y}_i - \mu_i) \leq C_{\alpha_o} \sqrt{2}. \tag{8.16}
\]

This equation follows from the fact that all the elements in a set are less than or equal to a given number if and only if the largest is. The distribution of this \( T \) is defined next.

**Definition 21** Suppose \( Z_1, \ldots, Z_g \) are independent \( N(0, 1) \)'s, and \( U \sim \chi^2_\nu \), where \( U \) is independent of the \( Z_i \)'s. Then

\[
T = \frac{\max_{1 \leq i \leq g} Z_i - \min_{1 \leq i \leq g} Z_i}{\sqrt{U/\nu}} \tag{8.17}
\]

has the **studentized range** distribution, with parameters \( g \) and \( \nu \). It is denoted

\[
T \sim T_{g,\nu}. \tag{8.18}
\]

To apply this definition to \( T \) in (8.16), set \( Z_i = \sqrt{N}(\overline{y}_i - \mu_i)/\sigma_e \)'s, so that they are independent \( N(0, 1) \)'s, and set \( \nu \hat{\sigma}_e^2 / \sigma_e^2 = U \) which is independent of the \( Z_i \)'s and distributed \( \chi^2_\nu \), where here \( \nu = n - g \). Then from (8.16)

\[
T = \frac{\max_{1 \leq i \leq g} \sqrt{N}(\overline{y}_i - \mu_i)/\sigma_e - \min_{1 \leq i \leq g} \sqrt{N}(\overline{y}_i - \mu_i)/\sigma_e}{\sqrt{(\nu \hat{\sigma}_e^2 / \sigma_e^2)/\nu}} = \frac{\max_{1 \leq i \leq g} Z_i - \min_{1 \leq i \leq g} Z_i}{\sqrt{U/\nu}} \sim T_{g,\nu}. \tag{8.19}
\]
8.2. **Tukey’s Method for Comparing Means Pairwise**

Letting $T_{g,\nu,\alpha_O}$ be the upper $\alpha_O$th cutoff point of the $T_g,\nu$ distribution, the confidence intervals for the difference $\mu_i - \mu_j$ is

$$\overline{y}_i - \overline{y}_j \pm T_{g,\nu,\alpha_O} \hat{\sigma}_e / \sqrt{N},$$

(8.20)

and the chance that all intervals cover their difference is $1 - \alpha_O$.

**Note.** Be careful about the $\sqrt{2}$. Comparing (8.14) to (8.20), we see that the constant $C_{\alpha_O} = T_{g,\nu,\alpha_O} / \sqrt{2}$, so that the intervals are not $\pm T_{g,\nu,\alpha_O} \times \hat{\sigma}_e$, but $\pm T_{g,\nu,\alpha_O} \times \hat{\sigma}_e / \sqrt{2}$.

**Example.** Look again at the leprosy data, without the covariates in this case. The $g = 3$, $\overline{y}_1 = 5.3, \overline{y}_2 = 6.1, \overline{y}_3 = 12.3, \hat{\sigma}_e^2 = 36.86, \nu = 27$, and $N = 10$. Then $T_{3,27,0.05} = 3.506$. (Tables for the studentized range are not as easy to find as for the usual $t$’s and $F$’s, but there are some online, e.g., [http://cse.niaes.affrc.go.jp/miwa/probcalc/s-range/](http://cse.niaes.affrc.go.jp/miwa/probcalc/s-range/), or you can use the functions `ptukey` and `qtukey` in R.)

The intervals are then

$$\left( \overline{y}_i - \overline{y}_j \pm 3.506 \times \sqrt{\frac{36.86}{10}} \right) = (\overline{y}_i - \overline{y}_j \pm 6.731).$$

(8.21)

The intervals are in the “Tukey” column:

<table>
<thead>
<tr>
<th>Means</th>
<th>Tukey</th>
<th>Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1 - \mu_2$</td>
<td>(-7.531, 5.931)</td>
<td>(-7.729, 6.129)</td>
</tr>
<tr>
<td>$\mu_1 - \mu_3$</td>
<td>(-13.731, -0.269)</td>
<td>(-13.929, -0.071)</td>
</tr>
<tr>
<td>$\mu_2 - \mu_3$</td>
<td>(-12.931, 0.531)</td>
<td>(-13.129, 0.729)</td>
</tr>
</tbody>
</table>

We see that the second interval does not contain 0, so that Drug A can be declared better than the placebo here. The other two intervals contain 0.

For comparison, we also present the Bonferroni intervals, where here $J = \binom{3}{2} = 3$. They use $t_{27,0.05/6} = 2.552$, so that

$$\left( \overline{y}_i - \overline{y}_j \pm 2.552 \times \sqrt{\frac{2 \times 36.86}{10}} \right) = (\overline{y}_i - \overline{y}_j \pm 6.929).$$

(8.22)

These intervals are slightly wider than the Tukey intervals, because Bonferroni is conservative whereas Tukey is exact here.

In the unbalanced case, it is hard to obtain exact intervals, but conservative Tukey-type intervals (which are less conservative than Bonferroni) use $(N_i + N_j)/2$ in place of $N$. That is, the intervals are

$$\overline{y}_i - \overline{y}_j \pm T_{g,\nu,\alpha_O} \hat{\sigma}_e / \sqrt{(N_i + N_j)/2}.$$  

(8.23)
8.3 Scheffé’s method for all contrasts

What looks like a particularly difficult problem is to find a $C_{\alpha_O}$ so that the chance that confidence intervals for all possible contrasts contain their parameters is at least $1 - \alpha_O$. Let $\bar{c} = (\alpha_1, \ldots, \alpha_g)'$, so that a contrast of the $\alpha_i$’s is $c^t \bar{c}$, where $c = (c_1, \ldots, c_g)'$ and $c_1 + \cdots + c_g = 0$. Note that any contrast of the $\alpha_i$’s is the same contrast of the $\mu_i$’s. The estimate of $c^t \bar{c}$ is

$$
\hat{c}^t \bar{c} = c_1 \bar{y}_1 + \cdots + c_g \bar{y}_g.
$$

(8.24)

We wish to find the constant $C_{\alpha_O}$ so that

$$
P[\hat{c}^t \bar{c} \in (\hat{c}^t \bar{c} \pm C_{\alpha_O}se(\hat{c}^t \bar{c})) \text{ for all contrasts } \alpha] \geq 1 - \alpha_O.
$$

(8.25)

Because the variance of each mean is $\sigma^2/N_i$, and the means are independent, the estimated standard error is

$$
se(\hat{c}^t \bar{c}) = \hat{\sigma}_e \sqrt{c_1^2/N_1 + \cdots + c_g^2/N_g},
$$

(8.26)

It turns out to be useful to express the contrast using the $\hat{a}$ for which $\hat{c}^t \bar{c} = \hat{a}^t y$ and $\hat{c}^t \bar{A} = \hat{a}^t \mu$, that is,

$$
\hat{a} = \begin{pmatrix}
(c_1/N_1) \mathbf{1}_{N_1} \\
(c_2/N_2) \mathbf{1}_{N_2} \\
\vdots \\
(c_g/N_g) \mathbf{1}_{N_g}
\end{pmatrix},
$$

(8.27)

so that $se(\hat{c}^t \bar{c}) = \hat{\sigma}_e \|\hat{a}\|$. Notice that $\hat{a} \in M_{\bar{A}}$, and

$$
\mathbf{1}_n^t \hat{a} = N_1(c_1/N_1) + \cdots + N_g(c_g/N_g) = c_1 + \cdots + c_g = 0,
$$

(8.28)

so that $\hat{a} \not\perp M_{\emptyset}$. Which means that $\hat{a} \in M_{A\emptyset}$. Note also the reverse, that is, for any $\hat{a} \in M_{A\emptyset}$, there is a corresponding $c$. If $\hat{a} = (a_1 \mathbf{1}_{N_1}, \ldots, a_g \mathbf{1}_{N_g})'$, then $c = (N_1 a_1, \ldots, N_g a_g)'$, which sums to 0 because $\mathbf{1}_n^t \hat{a} = 0$. Thus the probability inequality in (8.25) is the same as

$$
P[\hat{a}^t \mu \in (\hat{a}^t y \pm C_{\alpha_O} \hat{\sigma}_e \|\hat{a}\|) \text{ for all } \hat{a} \in M_{A\emptyset}] \geq 1 - \alpha_O.
$$

(8.29)

Similar to the reasoning in (8.15) and (8.16), we have that

$$
\hat{a}^t \mu \in (\hat{a}^t y \pm C_{\alpha_O} \hat{\sigma}_e \|\hat{a}\|) \text{ for all } \hat{a} \in M_{A\emptyset} \iff \frac{(\hat{a}^t(y - \mu))^2}{\hat{\sigma}_e^2 \|\hat{a}\|^2} \leq C_{\alpha_O}^2 \text{ for all } \hat{a} \in M_{A\emptyset}
$$

$$
\iff \max_{\hat{a} \in M_{A\emptyset}} \frac{(\hat{a}^t(y - \mu))^2}{\hat{\sigma}_e^2 \|\hat{a}\|^2} \leq C_{\alpha_O}^2
$$

(8.30)

Because $\hat{a} \in M_{A\emptyset}$, $M_{A\emptyset} \hat{a} = \hat{a}$, hence

$$
\hat{a}^t(y - \mu) = \hat{a}^t M_{A\emptyset}(y - \mu) = \hat{a}^t (\hat{y}_{A\emptyset} - M_{A\emptyset} \mu).
$$

(8.31)
The Cauchy-Schwarz Inequality says that \((a^t b)^2 \leq \|a\|^2\|b\|^2\), hence
\[
\frac{\bar{a}^t (\bar{y}_{A\emptyset} - M_{A\emptyset}\mu)^2}{\|\bar{a}\|^2} \leq \frac{\|\bar{a}\|^2\|\bar{y}_{A\emptyset} - M_{A\emptyset}\mu\|^2}{\|\bar{a}\|^2} = \|\bar{y}_{A\emptyset} - M_{A\emptyset}\mu\|^2.
\] (8.32)

Taking \(\hat{a} = \|\bar{y}_{A\emptyset} - M_{A\emptyset}\mu\|^2 (\in M_{A\emptyset})\), the inequality in (8.32) is an equality. Thus,
\[
\max_{\hat{a} \in M_{A\emptyset}} \frac{(\hat{a}^t (Y - \mu))^2}{\|\hat{a}\|^2} = \max_{\hat{a} \in M_{A\emptyset}} \frac{(\hat{a}^t (\bar{y}_{A\emptyset} - M_{A\emptyset}\mu))^2}{\|\hat{a}\|^2} = \|\bar{y}_{A\emptyset} - M_{A\emptyset}\mu\|^2.
\] (8.33)

Because \(\bar{y}_{A\emptyset} - M_{A\emptyset}\mu \sim N_n(\bar{a}, \sigma_e^2 M_{A\emptyset})\), and \(\text{trace}(M_{A\emptyset}) = g - 1\),
\[
\|\bar{y}_{A\emptyset} - M_{A\emptyset}\mu\|^2 \sim \sigma_e^2 \chi_{g-1}^2.
\] (8.34)

Now \(\hat{\sigma}_e^2\) is independent of \(\bar{y}_{A\emptyset}\), and is distributed \((\sigma_e^2 / (n - g))\chi_{n-g}^2\), so that
\[
\max_{\hat{a} \in M_{A\emptyset}} \frac{(\hat{a}^t (Y - \mu))^2}{\hat{\sigma}_e^2 \|\hat{a}\|^2} = \frac{\sigma_e^2 \chi_{g-1}^2}{(\sigma_e^2 / (n - g))\chi_{n-g}^2} = \frac{\chi_{g-1}^2}{\chi_{n-g}^2 / (n - g)} = (g - 1) \frac{\chi_{g-1}^2 / (g - 1)}{\chi_{n-g}^2 / (n - g)} \sim (g - 1) F_{g-1,n-g}.
\] (8.35)

Looking at (8.30) and (8.35), we see that
\[
P[\bar{a}^t \mu \in (\hat{a}^t \mu \pm C_\alpha \hat{\sigma}_e \|\bar{a}\|)] \text{ for all } \hat{a} \in M_{A\emptyset} = P[(g - 1) F_{g-1,n-g} \leq C_{\alpha_0}^2],
\] (8.36)
which means that
\[
C_{\alpha_0} = \sqrt{(g - 1) F_{g-1,n-g,\alpha_0}}.
\] (8.37)

**Example.** Consider the four contrasts for the leprosy data: the Drugs vs. Placebo, plus the three pairwise comparisons. For \(\alpha_0 = 0.05\), \(F_{2,27,0.05} = 3.354\), so that \(C_{\alpha_0} = \sqrt{2 \times 3.354} = 2.590\). The table has the \(\hat{\gamma} \pm 2.590 \times \sigma_e(\hat{\gamma})\)’s:

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Estimate</th>
<th>se</th>
<th>Scheffé</th>
<th>Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\alpha_1 + \alpha_2)/2 - \alpha_3)</td>
<td>-6.6</td>
<td>2.351</td>
<td>(-12.689, -0.511)</td>
<td>(-12.891, -0.309)</td>
</tr>
<tr>
<td>(\alpha_1 - \alpha_2)</td>
<td>-0.8</td>
<td>2.715</td>
<td>(-7.832, 6.232)</td>
<td>(-8.065, 6.465)</td>
</tr>
<tr>
<td>(\alpha_1 - \alpha_3)</td>
<td>-7.0</td>
<td>2.715</td>
<td>(-14.032, 0.032)</td>
<td>(-14.265, 0.265)</td>
</tr>
<tr>
<td>(\alpha_2 - \alpha_3)</td>
<td>-6.2</td>
<td>2.715</td>
<td>(-13.232, 0.832)</td>
<td>(-13.465, 1.065)</td>
</tr>
</tbody>
</table>
The only significant contrast is the first. For comparison, we also have the Bonferroni intervals, which are a little wider. The $t_{27, \alpha/8} = 2.676$. Note that the Tukey intervals are not directly applicable here, because the first contrast is not a paired comparison.

A nice feature of the Scheffé intervals is that we can add any number of contrasts to our list without widening the intervals. With Bonferroni, each additional contrast would widen all of them. Of course, if one is only interested in a few contrasts, Scheffé can be quite conservative, since it protects against all. Bonferroni is conservative, too, but may be more or less so. The following table gives an idea in this example (with $\alpha_O = 0.05$, $g = 3$, $n - g = 27$):

<table>
<thead>
<tr>
<th># of contrasts</th>
<th>Scheffé cutoff</th>
<th>Bonferroni cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.59</td>
<td>2.052</td>
</tr>
<tr>
<td>2</td>
<td>2.59</td>
<td>2.373</td>
</tr>
<tr>
<td>3</td>
<td>2.59</td>
<td>2.552</td>
</tr>
<tr>
<td>4</td>
<td>2.59</td>
<td>2.676</td>
</tr>
<tr>
<td>5</td>
<td>2.59</td>
<td>2.771</td>
</tr>
<tr>
<td>10</td>
<td>2.59</td>
<td>3.057</td>
</tr>
<tr>
<td>20</td>
<td>2.59</td>
<td>3.333</td>
</tr>
<tr>
<td>100</td>
<td>2.59</td>
<td>3.954</td>
</tr>
</tbody>
</table>

Between 3 and 4 the advantage switches from Bonferroni to Scheffé. With only three means, one is not likely to want to look at more than 5 or so contrasts, so either approach is reasonable here.

### 8.3.1 Generalized Scheffé

The key to the derivation of the Scheffé bounds was the appearance of the vector space $\mathcal{M}_{A\emptyset}$. Similar bounds can be obtained for any set of linear combinations of the parameters as long as one can find an associated vector space. The model is the general one, $y = X \beta + \epsilon$, $\epsilon \sim N_n(0_n, \sigma^2_e I_n)$. Suppose one is interested in the set of linear combinations

$$\{ \lambda' \beta \mid \lambda \in \Lambda \}, \quad (8.38)$$

where $\Lambda$ is any set of $p \times 1$ vectors (presuming $\beta$ is $p \times 1$). For each $\lambda$, let $\hat{\lambda}$ be the vector so that $\hat{\lambda}' y$ is the least squares estimate of $\lambda' \beta$. Then define the vector space

$$\mathcal{M}_\Lambda = \text{span}\{ \hat{\lambda} \mid \lambda \in \Lambda \}. \quad (8.39)$$

Recall that the least squares estimates have $\hat{\lambda} \in \mathcal{M}(= C(X))$, hence $\mathcal{M}_\Lambda \subset \mathcal{M}$. The Scheffé intervals are then

$$\hat{\lambda}' \beta \pm \sqrt{q} F_{q, \nu, \alpha_O} \text{se}(\hat{\lambda}' \beta), \quad q = \text{rank}(\mathcal{M}_\Lambda), \nu = n - p. \quad (8.40)$$

For example, in the two-way additive ANOVA model, if one wishes all contrasts on the row effects $\alpha_i$, then $q = r - 1$ and $\nu = n - r - c + 1$. 
8.4. CONCLUSION

For another example, consider the leprosy data with covariate, Section 3.4.2. We are again interested in all contrasts of the $\alpha_i$’s, but with the covariate included in the model, the estimates are not based on just the group means. Now the $\nu = 26$, and $\hat{\sigma}^2_e = 16.05$. The dimension of the $M_\Lambda$ will be $g - 1 = 3$ again. To see this, note that any contrast vector $(c_1, c_2, c_3)'$ has associated $\lambda = (0, c_1, c_2, c_3, 0)'$, since $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3, \gamma)'$, which can be written as a linear combination of two special $\lambda$’s,

$$
(0, c_1, c_2, c_3, 0)' = c_1(0, 1, 0, -1, 0)' + c_2(0, 0, 1, -1, 0)',
$$

because $c_3 = -c_1 - c_2$. Then

$$
M_\Lambda = \text{span}\{(\bar{\alpha}(0,1,0,-1,0)', \bar{\alpha}(0,0,1,-1,0)')\}.
$$

Those two vectors are linearly independent, so $\text{rank}(M_\Lambda) = 2$. The $C_{\alpha_0} = \sqrt{2F_{26,0.05}} = 2.596$. Just for a change, instead of confidence intervals, the next table has the $t$-statistics, where the contrast is significant if the $|t|$ exceeds 2.596.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Estimate</th>
<th>se</th>
<th>$t=\text{estimate/se}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\alpha_1 + \alpha_2)/2 - \alpha_3$</td>
<td>-3.392</td>
<td>1.641</td>
<td>-2.067</td>
</tr>
<tr>
<td>$\alpha_1 - \alpha_2$</td>
<td>-0.109</td>
<td>1.796</td>
<td>-0.061</td>
</tr>
<tr>
<td>$\alpha_1 - \alpha_3$</td>
<td>-3.446</td>
<td>1.887</td>
<td>-1.826</td>
</tr>
<tr>
<td>$\alpha_2 - \alpha_3$</td>
<td>-3.337</td>
<td>1.854</td>
<td>-1.800</td>
</tr>
</tbody>
</table>

Now none of the contrasts is significant. Actually, we could have foreseen this fact, since the $F$-test for testing the $\alpha_i$’s are equal was not significant. For the Scheffé intervals, none are significant if and only if the $F$ test is not significant.

8.4 Conclusion

Which of these methods to use, or whether to use other methods floating around, depends on the situation. Bonferroni is the most widely useful, and although is conservative, as long as $J$ is not too large, is reasonable. Tukey’s is the best if you are comparing means pairwise. Scheffé is more general than Tukey, but is usually conservative because one rarely wants to look at all contrasts. Whether Bonferroni or Scheffé is less conservative depends on how many contrasts one is interested in. Fortunately, it is not too hard to do both, and pick the one with the smallest $C_{\alpha_0}$.

There are other approaches to the multiple comparison problem than trying to control the overall error rate. One that has arisen recently is called the False Discovery Rate (FDR), which may be relevant when testing many (like 100) similar hypotheses. For example, one may be testing many different compounds as carcinogens, or testing many different genes to see if they are associated with some disease. In these cases, trying to control the chance of any false positives may mean one does not see any significant differences. Of course, you can
use a fairly hefty $\alpha_O$, like 50%, or just not worry about the overall level, if you are willing to put up with a 5% of false positives.

The FDR approach is slightly different, wanting to control the number of false positives among the positives, that is,

$$FDR = E \left[ \frac{\text{# Rejections when null is true}}{\text{Total # of rejections}} \right].$$  \hspace{1cm} (8.43)

The idea is that if the $FDR \leq 0.05$, then among all your rejections, approximately only 5% were falsely rejected. In the examples, you are willing to put up with false rejections, as long as the rejections are predominantly correct. So, e.g., if you are testing 100 compounds, and 30 are found suspect, then you are fairly confident that only 1 or 2 ($30 \times 0.05 = 1.5$) have been falsely accused. There is some controversy about what you do with the ratio when there are 0 rejections, and people are still trying to find good ways to implement procedures to control FDR, but it is an interesting idea to be aware of.
Chapter 9

Random effects

So far, we have been concerned with fixed effects, even though that term has yet to be uttered. By fixed, we mean the actual levels in the row or columns are of interest in themselves: The types of fruit trees, the treatments for leprosy, the genotypes of rats. We are interested in the actual Drug A or Drug D, or the Shmouti Orange type of tree. An example where an effect is not of direct interest is in HW #7:

Here are data from an experiment in which 3 people tested four formulations of hot dogs. Each person tested three of each formulation. The people then rated each hot dog from 0 to 14 on its texture, where 0 is soft, 14 is hard. (From course notes of J. W. Sutherland at http://www.me.mtu.edu/~jwsuther/doe/)

<table>
<thead>
<tr>
<th>Formulation</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>7.6</td>
<td>6.5</td>
<td>11.4</td>
<td>9.5</td>
</tr>
<tr>
<td>Person 2</td>
<td>7.2</td>
<td>13.6</td>
<td>12.9</td>
<td>10.7</td>
</tr>
<tr>
<td>Person 3</td>
<td>7.0</td>
<td>10.2</td>
<td>10.2</td>
<td>8.1</td>
</tr>
</tbody>
</table>

The formulations of hot dogs are fixed, because you can imagine being interested in those actual effects. On the other hand, the people effects are probably not of interest in themselves, but rather as representatives of a population. That is, someone in another part of the country might use the same formulations for their hot dogs, but are unlikely to use the same people. For modeling the data, the people would be considered a random sample from a population, so would be considered random effects.

Agricultural experiments often have random effects represented by plots of land. (Or, e.g., litters of pigs.) An experiment comparing three fertilizers may be conducted by splitting the experimental land into ten plots, and applying each fertilizer to a subplot of each plot. This design is a two-way ANOVA, with plots as rows and fertilizers as columns. The fertilizers are fixed effects, because the same fertilizers would be of interest elsewhere. The plots would be random effects, thought of as a sample of all possible farm plots, because the outside world does not really care about those particular plots.

One person’s fixed effect might be another’s random effects, so if the plots are yours, and you are going to use them over and over, you very may well consider the plots as fixed effects.
In practice, whether to consider an effect fixed or random depends on whether you wish to make inferences about the actual effects, or about the population which they represent.

The next several sections look at some models with random effects: one random effect, two random effects, or one random and one fixed effect.

### 9.1 One way random effects

Consider a population \( I \) of individuals. E.g., suppose \( I \) is the student population of a large midwestern university. Think of \( I \) as a list of the students’ id numbers. Of interest in \( \mu \), the average lung capacity of the population. The typical approach would be to take a random sample, of twenty-five, say, from the population, measure the lung capacity of the sampled people, and average the measurements. There is likely to be measurement error, that is, the results for person \( I \) would not be exactly that person’s lung capacity, but the capacity \( \pm \) some error. To help better estimate each person’s capacity, several, say five, independent measurements are made on each individual. The data would then look like a one-way ANOVA, with \( g = 25 \) groups (people) and \( N = 5 \) observations per group. Let \( \mu_I \) be the true lung capacity of individual \( I \), and \( Y_{ij}, j = 1, \ldots, N \) be the measurements for that individual, so that the model is

\[
y_{ij} = \mu_I + e_{ij}, \quad j = 1, \ldots, N. \tag{9.2}
\]

The “\( I \)” is capitalized to suggest that it is random, hence that \( \mu_I \) is random. That is, as \( I \) runs over the population of id numbers, \( \mu_I \) runs over the true lung capacities of the students. We will model the distribution by

\[
\mu_I \sim N(\mu, \sigma^2_A), \tag{9.3}
\]

in particular, \( \mu \) is the average lung capacity of the population. (Because the population is finite, this assumption cannot hold exactly, of course. See Section 9.5.) The \( e_{ij} \)'s are then the measurement errors, which will be assumed to be independent \( N(0, \sigma^2_e) \), and independent of the \( \mu_I \)'s. Turning to effects, let \( \alpha_I = \mu_I - \mu \), so that, with (9.2) and (9.3), we have the model

\[
y_{ij} = \mu + \alpha_I + e_{ij}, \quad \alpha_I \sim N(0, \sigma^2_A), \quad e_{ij} \sim N(0, \sigma^2_e), \tag{9.4}
\]

with the \( e_{ij} \)'s and \( \alpha_I \)'s independent.

The data is assumed to arise by taking a sample \( I_1, \ldots, I_g \) from \( I \), then for each of those individuals, taking \( N \) measurements, so that the data are

\[
y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, g; \quad j = 1, \ldots, N. \tag{9.5}
\]

The double subscript is a little awkward, so we will use the notation

\[
y_{i,j} = y_{ij}, \quad e_{i,j} = e_{ij}, \quad \text{and} \quad \alpha_i = A_i : \quad y_{ij} = \mu + A_i + e_{ij}. \tag{9.6}
\]

Replacing the \( \alpha \) with \( A \) is supposed to emphasize its randomness.
Although the model looks like the usual (fixed effect) ANOVA model, the randomness of the $A_i$’s moves the group effects from the mean to the covariance. That is, the $y_{ij}$’s all have the same mean $\mu$, but they are not all independent: $y_{11}$ and $y_{12}$ are not independent, because they share $A_1$, but $y_{11}$ and $y_{21}$ are independent. To find the actual distribution, we write the model in the usual matrix formulation, but break out the $\mathbf{1}_n$ part:

$$
\mathbf{y} = \mu \mathbf{1}_n + \left( \begin{array}{c} x_1, \ldots, x_g \end{array} \right) \left( \begin{array}{c} A_1 \\ A_2 \\ \vdots \\ A_g \end{array} \right) + \boldsymbol{\varepsilon} = \mu \mathbf{1}_n + \mathbf{X}_G \mathbf{A} + \boldsymbol{\varepsilon}. \quad (9.7)
$$

By assumption on $\mathbf{A}$,

$$
\mathbf{A} \sim N_g(\mathbf{0}, \sigma^2_A \mathbf{I}_g), \quad (9.8)
$$

and is independent of $\boldsymbol{\varepsilon}$, hence

$$
E[\mathbf{y}] = \mu \mathbf{1}_n \quad \text{Cov}[\mathbf{y}] = \text{Cov}[\mathbf{X}_G \mathbf{A}] + \text{Cov}[\boldsymbol{\varepsilon}] = \sigma^2_A \mathbf{X}_G \mathbf{X}_G' + \sigma^2_e \mathbf{I}_n. \quad (9.9)
$$

If $g = 2$ and $N = 3$, this covariance is explicitly

$$
\text{Cov}[\mathbf{y}] = \sigma^2_A \left( \begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{array} \right) \left( \begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{array} \right) + \sigma^2_e \mathbf{I}_6. \quad (9.10)
$$

Thus the $y_{ij}$’s all have the same variance, but the covariances depend on whether they are from the same group.

Before proceeding, we introduce Kronecker products, which make the subsequent calculations easier.

### 9.2 Kronecker products

The relevant matrices for balanced layouts, especially with many factors, can be handled more easily using the Kronecker product notation. Next is the definition.
Definition 22 If \( A \) is a \( p \times q \) matrix and \( B \) is an \( n \times m \) matrix, then the Kronecker product is the \((np) \times (mq)\) matrix \( A \otimes B \) given by

\[
A \otimes B = \begin{pmatrix}
a_{11}B & a_{12}B & \cdots & a_{1q}B \\
a_{21}B & a_{22}B & \cdots & a_{2q}B \\
\vdots & \vdots & \ddots & \vdots \\
a_{p1}B & a_{p2}B & \cdots & a_{pq}B
\end{pmatrix}.
\quad (9.12)
\]

The \( X_G \) matrix for the balanced one-way ANOVA can then be written as

\[
X_g = I_g \otimes \mathbf{1}_N.
\quad (9.13)
\]

E.g., with \( g = 3 \) and \( N = 2 \),

\[
I_3 \otimes \mathbf{1}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}
\quad (9.14)
\]

Kronecker product are nice because many operations can be performed componentwise. Here is a collection of some properties.

Proposition 20 Presuming the operations (addition, multiplication trace, inverse) make sense,

\[
(A \otimes B)' = A' \otimes B'
\quad (9.15)
\]

\[
(A \otimes B)(C \otimes D) = (AC) \otimes (BD)
\quad (9.16)
\]

\[
(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}
\quad (9.17)
\]

\[
(A \otimes B) + (A \otimes D) = A \otimes (B + D)
\quad (9.18)
\]

\[
(A \otimes B) + (C \otimes B) = (A + C) \otimes B
\quad (9.19)
\]

\[
(A \otimes B) \otimes C = A \otimes (B \otimes C)
\quad (9.20)
\]

\[
\text{trace}(A \otimes B) = \text{trace}(A) \times \text{trace}(B).
\quad (9.21)
\]
These properties can be used to show that if \( X = X_1 \otimes X_2 \), and the columns of these matrices are linearly independent, the projection matrix onto \( C(X) \) is

\[
M = X(X'X)^{-1}X' \\
= (X_1 \otimes X_2)((X_1 \otimes X_2)'(X_1 \otimes X_2))^{-1}(X_1 \otimes X_2)' \\
= (X_1 \otimes X_2)((X'_1X_1) \otimes (X'_2X_2))^{-1}(X_1 \otimes X_2)' \\
= (X_1X'_1X_1)^{-1} \otimes (X'_2X_2)^{-1}(X_1 \otimes X_2') \\
= (X_1X'_1X_1)^{-1} \otimes (X_2(X'_2X_2)^{-1}X'_2). \quad (9.22)
\]

That is, it decomposes into the Kronecker product of two little projection matrices.

### 9.3 One-way random effects: Estimation and testing

As in (9.13), the balanced one-way random effects ANOVA model is

\[
y = \mu \mathbb{1}_n + X_G A = \mu (\mathbb{1}_g \otimes \mathbb{1}_N) + (I_g \otimes \mathbb{1}_N)A + e. \quad (9.23)
\]

The fixed effect version would have \( \alpha \) in place of \( A \). Writing \( \mathbb{1}_n \) as a Kronecker product may seem unnecessary here, but it does help to make the notation consistent.

Inferences that would be of interest include estimating \( \mu \) (and the standard error), \( \sigma_A^2 \) and \( \sigma_e^2 \). The null hypothesis for testing for group effects in this case would be that all \( \alpha_i \)'s in the population are equal to 0, not just the ones in the sample. That is, we would test whether the variance of the \( A_i \)'s is zero:

\[
H_0 : \sigma_A^2 = 0 \quad versus \quad H_A : \sigma_A^2 > 0. \quad (9.24)
\]

We start by finding the relevant projections and their distributions.

By (9.22), the projection matrix for \( M_G \) is

\[
M_G = I_g \otimes (\mathbb{1}_N(1/1_N)^{-1}1_N') = I_g \otimes ((1/N)1_N1_N') = I_g \otimes J_N, \quad (9.25)
\]

where we are defining

\[
J_k = (1/k) \mathbb{1}_k \otimes \mathbb{1}_k, \quad (9.26)
\]

the \( k \times k \) matrix consisting of all \( (1/k)'s. \) Note that it is the projection matrix for \( span\{\mathbb{1}_k\} \).

Similarly, we can write \( M_\theta \), the projection matrix for \( span\{\mathbb{1}_n\} \), as

\[
M_\theta = J_g \otimes J_N \quad (9.27)
\]

(which also equals \( J_n \)). A summary:

\[
\begin{array}{c|c|c|c|c|c}
\text{Constant:} & \mathbb{1}_n & I_g \otimes \mathbb{1}_N & M_\theta & = & J_g \otimes J_N \\
\text{Groups:} & X_G & I_g \otimes \mathbb{1}_N & M_G & = & I_g \otimes J_N \\
\end{array} \quad (9.28)
\]
Then we have
\[ M_{G\emptyset} = M_G - M_\emptyset = (I_g \otimes J_N) - (J_g \otimes J_N) = (I_g - J_g) \otimes J_N = H_g \otimes J_N, \] (9.29)
where now we are defining
\[ H_k = I_k - J_k. \] (9.30)
This \( H_k \) is the projection matrix for \( \text{span}(\{ J_k \})^\perp \), that is, it subtracts the mean from each element of a vector. Note that because they are projections on orthogonal spaces (or by easy calculation),
\[ J_k H_k = H_k J_k = 0. \] (9.31)
Also,
\[ I_n - M_G = (I_g \otimes I_N) - (I_g \otimes J_N) = I_g \otimes (I_N - J_N) = I_g \otimes H_N. \] (9.32)

The mean and covariance of \( y \) in (9.23) can be written
\[ E[y] = 1_n \mu, \] (9.33)
and
\[
Cov[y] = \sigma_A^2 (I_g \otimes 1_N)(I_g \otimes 1_N)' + \sigma_e^2 I_n \\
= \sigma_A^2 (I_g \otimes 1_N 1_N^\prime) + \sigma_e^2 I_n \\
= N \sigma_A^2 (I_g \otimes J_N) + \sigma_e^2 I_n \\
= N \sigma_A^2 M_G + \sigma_e^2 I_n.
\] (9.34)

We calculate the same projections as for the fixed-effect case, but they have different distributions than before. In particular, the projections onto \( M_{G\emptyset} \) and \( M_{\emptyset} \) are
\[ \tilde{y}_{G\emptyset} = M_{G\emptyset} y \quad \text{and} \quad y - \tilde{y}_{G\emptyset} = (I_n - M_G)y. \] (9.35)

The means of both projections are \( 0_n \), because \( 1_n \) is orthogonal to the two relevant spaces. The covariances can be found using (9.8):
\[ \text{Cov}[\tilde{y}_{G\emptyset}] = N \sigma_A^2 M_{G\emptyset} M_G M_{G\emptyset} + \sigma_e^2 M_G M_{G\emptyset} = (N \sigma_A^2 + \sigma_e^2) M_G M_{G\emptyset}. \] (9.36)

We are using the fact that \( M_{G\emptyset} M_G = M_{G\emptyset} \), because if \( M_1 \subset M_2 \), then \( M_1 M_2 = M_2 M_1 = M_1 \). Next,
\[ \text{Cov}[y - \tilde{y}_{G\emptyset}] = \sigma_A^2 (I_n - M_G) M_G (I_n - M_G) + \sigma_e^2 (I_n - M_G) = \sigma_e^2 (I_n - M_G). \] (9.37)

Finally, independence follows because
\[ \text{Cov}[\tilde{y}_{G\emptyset}, y - \tilde{y}_{G\emptyset}] = N \sigma_A^2 (I_n - M_G) M_G M_{G\emptyset} + \sigma_e^2 (I_n - M_G) M_{G\emptyset} = 0 + 0 = 0. \] (9.38)

Thus we have that \( \| \tilde{y}_{G\emptyset} \|^2 \) and \( \| y - \tilde{y}_{G\emptyset} \|^2 \) are independent, and (because \( \text{trace}(M_{G\emptyset}) = g - 1 \) and \( \text{trace}(I_n - M_G) = n - g \)),
\[ \| \tilde{y}_{G\emptyset} \|^2 \sim (N \sigma_A^2 + \sigma_e) \chi^2_{g-1} \quad \text{and} \quad \| y - \tilde{y}_{G\emptyset} \|^2 \sim \sigma_e^2 \chi^2_{n-g}. \] (9.39)
The expected means squares are then easy to obtain:

\[ E[||\hat{y}_{G,\emptyset}||^2/(g - 1)] = N\sigma_A^2 + \sigma_e \quad \text{and} \quad E[||y - \hat{y}_{G,\emptyset}||^2/(n - g)] = \sigma_e^2. \]  

(9.40)

Now we can write out the ANOVA table, where we put in a column for expected mean squares instead of sample mean squares:

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>E[Mean square]</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups ((\mathcal{M}_{G,:}))</td>
<td>(</td>
<td></td>
<td>\hat{y}_{G,\emptyset}</td>
<td></td>
</tr>
<tr>
<td>Error ((\mathcal{M}_{\emptyset}'))</td>
<td>(</td>
<td></td>
<td>y - \hat{y}_{G,\emptyset}</td>
<td></td>
</tr>
<tr>
<td>Total ((\mathcal{M}_{\emptyset}))</td>
<td>(</td>
<td></td>
<td>y - \hat{y}_{\emptyset}</td>
<td></td>
</tr>
</tbody>
</table>

Notice that when \(\sigma_A^2 = 0\), then the two expected mean squares are equal. That is, one can use the \(F\) statistic to test \(H_0: \sigma_A^2 = 0\). Is the \(F \sim F_{g - 1, n - g}\) under the null hypothesis? Yes, just as in the fixed effect case, from (9.39), the numerator and denominator are independent \(\chi^2\)'s divided by their degrees of freedom (the \(\sigma_e^2\)'s cancel).

Also, \(\hat{\sigma}_e^2 = MSE\) as before, and we can find an unbiased estimate of \(\sigma_A^2\):

\[ E[MSG,: - MSE] = N\sigma_A^2, \quad \text{hence} \quad E\left[\frac{MSG,: - MSE}{N}\right] = \sigma_A^2. \]  

(9.41)

Turning to estimating \(\mu\), the least squares estimate is \(\hat{\mu} = \overline{y}..\). Note that \(\hat{\mu} = (1/n)\overline{y}'.\overline{y}\), so that

\[ Var[\hat{\mu}] = \frac{1}{n^2} \text{Cov}[\overline{y}, (1/n)\overline{y}] = \frac{1}{n^2} (N\sigma_A^2\overline{y}', \overline{M}_{G,:}\overline{y} + \sigma_e^2\overline{y}', \overline{y}) = \frac{1}{n^2} (\sigma_A^2\overline{y}', \overline{y} + n\sigma_e^2) = \frac{1}{n^2} (N^2g\sigma_A^2 + n\sigma_e^2) = \frac{1}{g}\sigma_A^2 + \frac{1}{n^2}\sigma_e^2, \]  

(9.42)

because \(n = Ng\). [A direct way to calculate the variance is to note that \(\overline{y}.. = \mu + \overline{A} + \overline{e}..\), and \(\overline{A}\) and \(\overline{e}..\) are independent means of \(g\) and \(n\) iid elements, respectively.] To find the estimated standard error, note that \(MSG,:\) estimates \(n Var[\hat{\mu}]\), so that

\[ se(\hat{\mu}) = \sqrt{\frac{MSG,:}{n}}. \]  

(9.43)

**Example.** This is Example 10.13.1 from Snedecor and Cochran, *Statistical Methods*. Two boars from each of four litters of pigs were studied, and each one’s average weight gain (until they became 225 pounds or so) was recorded. So the groups are the litters, \(g = 4\), and \(N = 2\). The data:
The ANOVA table:

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Litters</td>
<td>0.3288</td>
<td>3</td>
<td>0.1096</td>
<td>7.3805</td>
</tr>
<tr>
<td>Error</td>
<td>0.0594</td>
<td>4</td>
<td>0.0148</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>0.3882</td>
<td>7</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The $F_{3,4,0.05} = 6.591$, which suggests that $\sigma_a^2 > 0$, that is, there are differences between the litters. The $\bar{y} = 1.255$, and $se = \sqrt{0.1096/8} = 0.1170$. Because the variance estimate in the standard error is based on the litter mean square, the degrees of freedom are 3, so the 95% confidence interval is

$$(\bar{y} \pm t_{3,0.025}se) = (1.255 \pm 3.182 \times 0.1170) = (0.883, 1.627). \quad (9.44)$$

**Warning.** One might be tempted to treat the 8 observations as independent, so that we have just a regular sample of $n = 8$, with the $se = s/\sqrt{8} = 0.2355/\sqrt{8} = 0.0833$, where $s$ is the sample standard deviation of the observations. This standard error is somewhat less than the correct one of 0.1170. The confidence interval for the mean would be

$$(\bar{y} \pm t_{n-1,0.025}se) = (1.255 \pm 2.365 \times 0.0833) = (1.058, 1.452), \quad (9.45)$$

which is just a bit more than half (53%) the width of the correct one. Thus treating these observations as independent would be cheating.

What is correct is to look at the group means $\bar{y}_1, \ldots, \bar{y}_4$, as a sample of 4 independent observations (which they are). Then $se = s^*/\sqrt{4}$, where $s^* = 0.2341$ is the sample standard deviation of the means. Indeed, this $se = 0.1170$, which is correct.

## 9.4 Two-way random effects ANOVA

Return to the lung capacity example, and imagine that in addition to choosing a number of individuals, there are a number of doctors administering the lung capacity measurements, and that each doctor tests each individual $N$ times. Then for individual $I$ and doctor $J$ (not Julius Erving), $y_{IJK}, k = 1, \ldots, N$, represent the $N$ measurements. We assume there are $r$ individuals chosen as a simple random sample from the population $I$ and $c$ doctors chosen as a simple random sample from the population $J$, and the individuals and doctors are chosen independently. Letting $\mu_{IJ}$ be the true lung capacity of individual $I$ measured by doctor $J$, the model is

$$y_{IJK} = \mu_{IJ} + e_{IJK}. \quad (9.46)$$
The average for person $I$, averaged over all doctors, is
\[ \mu_I = E_J[\mu_{IJ}], \] (9.47)
where “$E_J$” indicates taking expected value with $J \in J$ as the random variable. Similarly, the average for doctor $J$, averaged over all $I \in I$, is
\[ \mu_J = E_I[\mu_{IJ}]. \] (9.48)
Both $\mu_I$ and $\mu_J$ are random. The overall mean $\mu$ is the average over all individuals and doctors:
\[ \mu = E_{I,J}[\mu_{IJ}] = E_I[\mu_I] = E_J[\mu_J]. \] (9.49)
The effects are then defined as for the fixed effect model, except that they are random:
Row effect: $\alpha_I = \mu_I - \mu$
Column effect: $\beta_J = \mu_J - \mu$
Interaction effect: $\gamma_{IJ} = \mu_{IJ} - \mu_I - \mu_J + \mu$ (9.50)
From (9.47), (9.48) and (9.49), one can see that
\[ E_I[\alpha_I] = 0 = E_J[\beta_J] = E_I[\gamma_{IJ}] = E_J[\gamma_{IJ}], \] (9.51)
which are the analogs of the usual constraints in the fixed effect model. We denote the population variances of these effects by
\[ \text{Var}_I[\alpha_I] = \sigma_A^2, \text{Var}_J[\beta_J] = \sigma_B^2, \text{Var}_{I,J}[\gamma_{IJ}] = \sigma_C^2. \] (9.52)
Then using (9.51) in (9.46), the model is
\[ y_{IJK} = \mu + \alpha_I + \beta_J + \gamma_{IJ} + e_{IJK}, \quad k = 1, \ldots, N. \] (9.53)
The actual data are based on taking simple random samples of the $I$’s and $J$’s, that is, $I_1, \ldots, I_r$ are the individuals, and $J_1, \ldots, J_c$ are the doctors, so that the data are
\[ y_{IJK} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad i = 1, \ldots, r; \quad j = 1, \ldots, c; \quad k = 1, \ldots, N. \] (9.54)
As in the one-way case, there are too many double subscripts, so that we change notation to
\[ y_{ijk} = \mu + A_i + B_j + C_{ij} + e_{ijk}, \quad i = 1, \ldots, r; \quad j = 1, \ldots, c; \quad k = 1, \ldots, N. \] (9.55)
so that $A_i = \alpha_i, B_j = \beta_j$, and $C_{ij} = \gamma_{ij}$.
The effects are then modeled by assuming that they are all independent and normal, which can be written as
\[ A \sim N_r(0, \sigma_A^2 I_r), \quad B \sim N_c(0, \sigma_B^2 I_c), \quad C \sim N_{rc}(0_{rc}, \sigma_C^2 I_{rc}), \quad A, B, \text{ and } C \text{ independent}. \] (9.56)
A justification for the uncorrelatedness of the effects is found in Section 9.5. Finally, the model written in matrix form is
\[
\bar{y} = \mu \mathbb{1}_n + X_R A + X_C B + X_{R \times C} C + \epsilon,
\]
where in addition to (9.56), the \( \epsilon \sim N_n(0, \sigma^2 I_n) \) is independent of the \( A, B \) and \( C \). Also, \( X_R \) is the \( n \times r \) matrix with vectors indicating the rows, \( X_C \) is the \( n \times c \) matrix with vectors indicating the columns, and \( X_{R \times C} \) is the \( n \times (rc) \) matrix with vectors indicating the cells.

Now the distribution of \( \bar{y} \) is multivariate normal with mean \( \mu \) and
\[
\text{Cov}[\bar{y}] = \sigma^2_A X_R X_R' + \sigma^2_B X_C X_C' + \sigma^2_C X_{R \times C} X_{R \times C}' + \sigma^2_e I_n.
\]

Inferences that are of interest include estimating \( \mu \) and the variances, and testing for interactions (\( \sigma^2_A = 0 \)) or for row or column effects (\( \sigma^2_A = 0 \) or \( \sigma^2_B = 0 \)). We will approach the task by finding the usual sums of squares and their expected mean squares. Because we have a balanced situation, the matrices can be written using Kronecker products, which helps in finding the \( E[MS] \)'s. The matrices are in the left half:

\[
\begin{align*}
\text{Constant:} & \quad \mathbb{1}_n = \mathbb{1}_r \otimes \mathbb{1}_c \otimes \mathbb{1}_N \quad \text{M}_\emptyset = J_r \otimes J_c \otimes J_N \\
\text{Rows:} & \quad X_R = \mathbb{I}_r \otimes \mathbb{1}_c \otimes \mathbb{1}_N \quad \text{M}_R = J_r \otimes J_c \otimes J_N \\
\text{Columns:} & \quad X_C = \mathbb{1}_r \otimes I_c \otimes \mathbb{1}_N \quad \text{M}_C = J_r \otimes I_c \otimes J_N \\
\text{Rows x Columns:} & \quad X_{R \times C} = I_r \otimes I_c \otimes \mathbb{1}_N \quad \text{M}_{R \times C} = I_r \otimes I_c \otimes J_N.
\end{align*}
\]

The projection matrices are then found by finding the component project matrices, e.g.,
\[
\text{M}_R = (I_r (I'_r I_r)^{-1} I'_r) \otimes (1_c (1'_c 1_c)^{-1} 1'_c) \otimes (1_N (1'_N 1_N)^{-1} 1'_N) = I_r \otimes J_c \otimes J_N.
\]

From these, we can find the projection matrices on the orthogonal subspaces, starting with
\[
\text{M}_{R \emptyset} = \text{M}_R - \text{M}_\emptyset = (I_r \otimes J_c \otimes J_N) - (J_r \otimes J_c \otimes J_N) = (I_r - J_r) \otimes (J_c \otimes J_N) = H_r \otimes J_c \otimes J_N,
\]
and
\[
\text{M}_{C \emptyset} = \text{M}_C - \text{M}_\emptyset = (J_r \otimes I_c \otimes J_N) - (J_r \otimes J_c \otimes J_N) = J_r \otimes (I_c - J_c) \otimes J_N = J_r \otimes H_c \otimes J_N.
\]

Note that
\[
\text{M}_{R \emptyset} \text{M}_{C \emptyset} = (H_r \otimes J_c \otimes J_N) (J_r \otimes H_c \otimes J_N) = (H_r J_r) \otimes (J_c H_c) \otimes J_N = 0 \otimes 0 \otimes J_N = 0,
\]
so that \( \mathcal{M}_{R \emptyset} \perp \mathcal{M}_{C \emptyset} \), as we know. We also know that \( \text{M}_{(R+C) \emptyset} = \text{M}_{R \emptyset} + \mathcal{M}_{C \emptyset} \), so that
\[
\begin{align*}
\text{M}_{(R \times C)-(R+C)} &= \text{M}_{(R \times C) \emptyset} - \text{M}_{(R+C) \emptyset} \\
&= \text{M}_{(R \times C) \emptyset} - \text{M}_{R \emptyset} - \text{M}_{C \emptyset} \\
&= I_r \otimes I_c \otimes J_N - J_r \otimes J_c \otimes J_N - H_r \otimes J_c \otimes J_N - J_r \otimes H_c \otimes J_N \\
&= I_r \otimes I_c \otimes J_N - (J_r + H_r) \otimes J_c \otimes J_N - J_r \otimes H_c \otimes J_N \\
&= I_r \otimes I_c \otimes J_N - I_r \otimes J_c \otimes J_N - J_r \otimes H_c \otimes J_N \\
&= I_r \otimes H_c \otimes J_N - J_r \otimes H_c \otimes J_N \\
&= H_r \otimes H_c \otimes J_N.
\end{align*}
\]
Finally,
\[ I_n - M_{R \times C} = I_r \otimes I_c \otimes I_N - I_r \otimes I_c \otimes J_N = I_r \otimes I_c \otimes H_N. \] (9.65)

We can summarize these orthogonal projections:

\[
\begin{align*}
\text{Constant} & : & M_0 & = & J_r \otimes J_c \otimes J_N \\
\text{Row effects} & : & M_{R \otimes} & = & H_r \otimes J_c \otimes J_N \\
\text{Column effects} & : & M_{C \otimes} & = & J_r \otimes H_c \otimes J_N \\
\text{Interactions} & : & M_{(R \times C) \otimes (R + C)} & = & H_r \otimes H_c \otimes J_N \\
\text{Error} & : & M_{R \times C} & = & I_n - M_{R \times C} = I_r \otimes I_c \otimes H_N.
\end{align*}
\]

Now for the sums of squares. The idea is to find the distribution of the projections of \( \hat{y} \) on each of the spaces in (9.66). They are straightforward to obtain using the Kronecker products, although may get a bit tedious. The covariance in (9.58) has the \( XX' \) matrices, which can also be more easily written:

\[
\begin{align*}
X_R X'_R & = I_r \otimes (L_r' \otimes (L_N 1_N')) = I_r \otimes (cJ_c) \otimes (NJ_N) = (cN)M_R, \\
X_C X'_C & = (L_r' \otimes I_c) \otimes (L_N 1_N') = rJ_r \otimes I_c \otimes (NJ_N) = (rN)M_C, \\
X_{R \times C} X'_{R \times C} & = I_r \otimes I_c \otimes (L_N 1_N') = I_r \otimes I_c \otimes (NJ_N) = N M_{R \times C}.
\end{align*}
\] (9.67)

Then
\[ Cov[\hat{y}] = cN\sigma^2_A M_R + rN\sigma^2_B M_C + N\sigma^2_C M_{R \times C} + \sigma^2_e I_n. \] (9.68)

Start with the constant term, that is, \( \hat{y}_0 = M_0 \hat{y} \). The
\[ E[\hat{y}_0] = M_0(\mu_{\perp_n}) = \mu_{\perp_n}, \] (9.69)

and

\[
\begin{align*}
Cov[\hat{y}_0] & = M_0Cov[\hat{y}]M_0 \\
& = M_0(cN\sigma^2_A M_R + rN\sigma^2_B M_C + N\sigma^2_C M_{R \times C} + \sigma^2_e I_n)M_0 \\
& = cN\sigma^2_A M_0 + rN\sigma^2_B M_0 + N\sigma^2_C M_0 + \sigma^2_e M_0 \\
& = (cN\sigma^2_A + rN\sigma^2_B + N\sigma^2_C + \sigma^2_e)M_0.
\end{align*}
\] (9.70)

The projections on the other spaces have mean zero, so we need to worry about just the covariances. For the projection onto \( M_{R \otimes} \), we have
\[ Cov[\hat{y}_{R \otimes}] = M_{R \otimes}(cN\sigma^2_A M_R + rN\sigma^2_B M_C + N\sigma^2_C M_{R \times C} + \sigma^2_e I_n)M_{R \otimes}. \] (9.71)

We know that \( M_{R \otimes} M_R = M_{R \otimes} \) and \( M_{R \otimes} M_{R \times C} = M_{R \otimes} \) by the subset property. Also,
\[ M_{R \otimes} M_C = (H_r \otimes J_c \otimes J_N)(J_r \otimes I_c \otimes J_N) = 0, \] (9.72)

because \( H_r J_r = 0 \). Thus
\[ Cov[\hat{y}_{R \otimes}] = (cN\sigma^2_A + N\sigma^2_C + \sigma^2_e)M_{R \otimes}. \] (9.73)
Similarly,
\[
\text{Cov}[\hat{\mathbf{y}}_{C,\emptyset}] = (rN\sigma_B^2 + N\sigma_C^2 + \sigma^2_e)\mathbf{M}_{C,\emptyset}. \tag{9.74}
\]

For \(\hat{\mathbf{y}}_{(R\times C)-(R+C)}\), we have that \(\mathbf{M}_{(R\times C)-(R+C)}\mathbf{M}_{R} = \mathbf{M}_{(R\times C)-(R+C)}\mathbf{M}_{C} = \mathbf{0}\) because \(\mathcal{M}_R\) and \(\mathcal{M}_C\) are orthogonal to \(\mathcal{M}_{(R\times C)-(R+C)}\). Thus
\[
\text{Cov}[\hat{\mathbf{y}}_{(R\times C)-(R+C)}] = (N\sigma_C^2 + \sigma^2_e)\mathbf{M}_{(R\times C)-(R+C)}. \tag{9.75}
\]

Finally, because \(\mathbf{I}_n - \mathbf{M}_{R\times C}\) is orthogonal to all the other projection matrices,
\[
\text{Cov}[\mathbf{y} - \hat{\mathbf{y}}_{R\times C}] = \sigma^2_e(\mathbf{I}_n - \mathbf{M}_{R\times C}), \tag{9.76}
\]
as usual.

These projections can also be seen to be independent, e.g.,
\[
\text{Cov}[\hat{\mathbf{y}}_{R,\emptyset}, \hat{\mathbf{y}}_{C,\emptyset}] = \mathbf{M}_{R,\emptyset}(cN\sigma_A^2\mathbf{M}_R + rN\sigma_B^2\mathbf{M}_C + N\sigma_C^2\mathbf{M}_{R\times C} + \sigma^2_e\mathbf{I}_n)\mathbf{M}_{C,\emptyset}
\]
\[
= (cN\sigma_A^2\mathbf{M}_{R,\emptyset} + N\sigma_C^2\mathbf{M}_{R,\emptyset} + \sigma^2_e\mathbf{M}_{R,\emptyset})\mathbf{M}_{C,\emptyset}
\]
\[
= 0, \tag{9.77}
\]
because \(\mathbf{M}_{R,\emptyset}\mathbf{M}_{C,\emptyset} = 0\).

The sums of squares of these projections (except onto \(\mathcal{M}_\emptyset\)) are then chi-squared with degrees of freedom being the trace of the projection matrices, and constant out front being the \(\sigma^2\)-part of the covariances. Thus the ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>E[Mean square]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows ((\mathcal{M}_{R,\emptyset}))</td>
<td>(|\hat{\mathbf{y}}_{R,\emptyset}|^2)</td>
<td>(r - 1)</td>
<td>(cN\sigma_A^2 + N\sigma_C^2 + \sigma^2_e)</td>
</tr>
<tr>
<td>Columns ((\mathcal{M}_{C,\emptyset}))</td>
<td>(|\hat{\mathbf{y}}_{C,\emptyset}|^2)</td>
<td>(c - 1)</td>
<td>(rN\sigma_B^2 + N\sigma_C^2 + \sigma^2_e)</td>
</tr>
<tr>
<td>Interactions ((\mathcal{M}_{(R\times C)-(R+C)}))</td>
<td>(|\hat{\mathbf{y}}_{(R\times C)-(R+C)}|^2)</td>
<td>((r - 1)(c - 1))</td>
<td>(N\sigma_C^2 + \sigma^2_e)</td>
</tr>
<tr>
<td>Error ((\mathcal{M}_{R\times C}^\perp))</td>
<td>(|y - \hat{\mathbf{y}}_{R\times C}|^2)</td>
<td>(n - rc)</td>
<td>(\sigma^2_e)</td>
</tr>
<tr>
<td>Total ((\mathcal{M}_\emptyset^\perp))</td>
<td>(|\hat{\mathbf{y}}_\emptyset|^2)</td>
<td>(n - 1)</td>
<td>—</td>
</tr>
</tbody>
</table>

### 9.4.1 Estimation and testing

Unbiased estimates of the variance components are immediately obtainable from the expected mean square column of the ANOVA table:

\[
\hat{\sigma}_{A}^2 = \frac{MS_{\text{Rows}} - MS_{\text{Interactions}}}{cN},
\]
\[
\hat{\sigma}_{B}^2 = \frac{MS_{\text{Columns}} - MS_{\text{Interactions}}}{rN},
\]
\[
\hat{\sigma}_{C}^2 = \frac{MS_{\text{Interactions}} - MSE}{N},
\]
\[
\hat{\sigma}_{e}^2 = MSE. \tag{9.78}
\]
For $\hat{\mu} = \overline{y}$, note that because $\hat{y}_\emptyset = \hat{\mu}_\emptyset$, (9.70) shows that $Var[\hat{\mu}]$ is the upper-right (or any other) elements of $Cov[\hat{y}_\emptyset]$, so that

$$Var[\hat{\mu}] = \frac{cN\sigma_A^2 + rN\sigma_B^2 + N\sigma_C^2 + \sigma_e^2}{n} = \frac{1}{r} \sigma_A^2 + \frac{1}{c} \sigma_B^2 + \frac{1}{rc} \sigma_C^2 + \frac{1}{n} \sigma_e^2.$$  \hspace{1cm} (9.79)

The estimated standard error would then be

$$se(\hat{\mu}) = \sqrt{\frac{MS_{Rows} + MS_{Columns} - MS_{Interactions}}{n}}.$$  \hspace{1cm} (9.80)

For testing, use the expected mean squares column to decide what ratio to take for the $F$ test. For example, to test for average (over the population) column effects, i.e.,

$$H_0 : \sigma_A^2 = 0 \ \text{versus} \ \ H_A : \sigma_A^2 > 0,$$  \hspace{1cm} (9.81)

The $F = MS_{Rows}/MS_{Interactions}$, because under the null hypothesis, both means squares have the same constant $N\sigma_C^2 + \sigma_e^2$. The degrees of freedom for the $F$ are then $r - 1$ and $(r - 1)(c - 1)$. Generally, testing for row effects when there are interaction effects present is uninteresting, so this particular test has little application.

It is of interest to test for interaction,

$$H_0 : \sigma_C^2 = 0 \ \text{versus} \ \ H_A : \sigma_C^2 > 0,$$  \hspace{1cm} (9.82)

which would use $F = MS_{Interactions}/MSE$, just as in the fixed effect case, with degrees of freedom $(r - 1)(c - 1)$ and $n - rc$.

If one assumes there are no interactions, then $\sigma_C^2 = 0$, so that the ANOVA table simplifies to

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>E[Mean square]</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows ($\mathcal{M}_{R\emptyset}$)</td>
<td>$|\hat{y}_R\emptyset|^2$</td>
<td>$r - 1$</td>
<td>$cN\sigma_A^2 + \sigma_e^2$</td>
<td>$MS_{Rows}/MSE$</td>
</tr>
<tr>
<td>Columns ($\mathcal{M}_{C\emptyset}$)</td>
<td>$|\hat{y}_{C\emptyset}|^2$</td>
<td>$c - 1$</td>
<td>$rN\sigma_B^2 + \sigma_e^2$</td>
<td>$MS_{Columns}/MSE$</td>
</tr>
<tr>
<td>Error ($\mathcal{M}_{R+C\emptyset}$)</td>
<td>$|y - \hat{y}_{R+C}|^2$</td>
<td>$n - r - c + 1$</td>
<td>$\sigma_e^2$</td>
<td>—</td>
</tr>
<tr>
<td>Total ($\mathcal{M}_\emptyset$)</td>
<td>$|y - \overline{y}|^2$</td>
<td>$n - 1$</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The $F$ tests for row and column effects are then as for the fixed effect case.
9.5 The distribution of the effects

9.5.1 One-way ANOVA

Starting with the one-way random effects ANOVA, the justification of the distributional assumptions, at least as far as the means and covariances go, is based on \( I_1, \ldots, I_r \)'s being a simple random sample from the population \( \mathcal{I} \), where \( \mathcal{I} \) is a very large relative to \( r \). The population mean and variance of the \( \mu_I \) is then defined by the usual finite-population values,

\[
\mu = E_I[\mu_I] = \frac{\sum_{i \in I} \mu_i}{N_I} \quad \text{and} \quad \sigma_A^2 = \text{Var}_I[\mu_I] = \frac{\sum_{i \in I} (\mu_i - \mu)^2}{N_I}. \tag{9.83}
\]

(\( N_I \) is the number of individuals in the population.) It then follows that the effects \( \alpha_I = \mu_I - \mu \) have

\[
E_I[\alpha_I] = 0 \quad \text{and} \quad \text{Var}_I[\alpha_I] = \sigma_A^2. \tag{9.84}
\]

Because we are sampling without replacement, \( I_i \) and \( I_j \) are not independent, hence the \( \alpha_I \)'s are not independent. The covariances between pairs are equal, so it is enough to find the covariance between \( \alpha_{I_1} \) and \( \alpha_{I_2} \). Because the means are zero,

\[
\text{Cov}[\alpha_{I_1}, \alpha_{I_2}] = \frac{\sum_{i \neq j} \alpha_i \alpha_j}{N_I(N_I - 1)}. \tag{9.85}
\]

Then

\[
\text{Cov}[\alpha_{I_1}, \alpha_{I_2}] = -\frac{\sigma_A^2}{N_I - 1}. \tag{9.87}
\]

At this point we could stop and say, reasonably, that if \( N_I \) is very large, the correlation, \( 1/(N_I - 1) \), is negligible, so that assuming independence should be fine. But to be more precise, for \( \underline{\alpha} = (\alpha_{I_1}, \ldots, \alpha_{I_g})' \) as in (9.8), we have to modify the covariance matrix to account for the correlations, so that

\[
\text{Cov}[\underline{\alpha}] = \sigma_A^2 \left( \frac{N_I}{N_I - 1} \mathbf{I}_g - \frac{1}{N_I - 1} \underline{L}_g \underline{L}_g' \right). \tag{9.88}
\]

Then

\[
\text{Cov}[\mathbf{X}_G \underline{\alpha}] = \sigma_A^2 \left( \frac{N_I}{N_I - 1} \mathbf{X}_G \mathbf{X}_G' - \frac{1}{N_I - 1} \mathbf{X}_G \underline{L}_g \underline{L}_g' \mathbf{X}_G' \right). \tag{9.89}
\]

Now \( \mathbf{X}_G \underline{L}_g = (\mathbf{I}_g \otimes \underline{1}_N)(\underline{L}_g \otimes 1) = \underline{L}_g \otimes \underline{1}_N \), so that \( \mathbf{X}_G \underline{L}_g \underline{L}_g' \mathbf{X}_G' = gN(\mathbf{J}_g \otimes \mathbf{J}_N) = gn\mathbf{M}_\emptyset \), hence

\[
\text{Cov}[\mathbf{X}_G \underline{\alpha}] = \sigma_A^2 \left( \frac{N_I}{N_I - 1} nM_G - \frac{1}{N_I - 1} gn\mathbf{M}_\emptyset \right). \tag{9.90}
\]
For $y$, we then have that
\[
\text{Cov}[y] = \sigma_A^2 \left( \frac{N_T}{N_T - 1} N M_G - \frac{1}{N_T - 1} g n M_\emptyset \right) + \sigma_e^2 I_n. \tag{9.91}
\]

The covariances of the projections onto $M_{G\emptyset}$ and $M_G$ are
\[
\text{Cov}[\hat{y}_{G\emptyset}] = M_{G\emptyset} (\sigma_A^2 \left( \frac{N_T}{N_T - 1} N M_G - \frac{1}{N_T - 1} g n M_\emptyset \right) + \sigma_e^2 I_n) M_{G\emptyset}
\]
\[
= (N \sigma_A^2 \frac{N_T}{N_T - 1} + \sigma_e^2) M_{G\emptyset} \tag{9.92}
\]
and
\[
\text{Cov}[y - \hat{y}_G] = (I_n - M_G) (\sigma_A^2 \left( \frac{N_T}{N_T - 1} N M_G - \frac{1}{N_T - 1} g n M_\emptyset \right) + \sigma_e^2 I_n) (I_n - M_G)
\]
\[
= \sigma_e^2 (I_n - M_G). \tag{9.93}
\]

These two projections can also be shown to be independent. The ANOVA table now is

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>E[Mean square]</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups ($M_{G\emptyset}$)</td>
<td>$|\hat{y}_{G\emptyset}|^2$</td>
<td>$g - 1$</td>
<td>$N \sigma_A^2 N_T / (N_T - 1) + \sigma_e^2$</td>
<td>$MS_{G\emptyset} / MSE$</td>
</tr>
<tr>
<td>Error ($M_{\emptyset}$)</td>
<td>$|y - \hat{y}_G|^2$</td>
<td>$n - g$</td>
<td>$\sigma_e^2$</td>
<td>—</td>
</tr>
<tr>
<td>Total ($M_{G\emptyset}$)</td>
<td>$|y - \hat{y}|^2$</td>
<td>$n - 1$</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Comparing this table to that in Section 9.1, we see that the only difference is the factor $N_T / (N_T - 1)$, which is practically 1. Of course, typically one does not really know $N_T$, so ignoring the factor is reasonable. But note that the $F$ test for testing $\sigma_A^2 = 0$ is exactly the same as before.
Chapter 10

Mixed Models

A mixed model is one with some fixed effects and some random effects. The most basic is the randomized block design, exemplified by the example on hot dogs in (9.1), where the people are random “blocks”, and the hot dog formulations are the fixed effects. Mixed models can be described in general by the model

$$y = X_{\text{Random}}A + X_{\text{Fixed}}\beta + e,$$

(10.1)

where the $X$’s are fixed design matrices, $A$ is a vector $(a \times 1)$ of random effects, and $\beta$ is a vector of fixed effects. All the models we have considered so far are of this form, mostly with no fixed effects, and in Chapter 9, with random effects plus the simple fixed effect matrix $X_{\text{Fixed}} = 1_n$.

The distributional assumption on $A$ is key, and can make the analysis of the model quite challenging. Often one assumes $A \sim N_a(0_a, \Sigma_A)$, where $\Sigma_A$ may have a simple form, or may not. With the assumption that $e$ is independent of $A$, and $e \sim N_n(0_n, \sigma^2_e I_n)$, the distribution of $y$ is

$$y \sim N_n(X_{\text{Fixed}}\beta, X_{\text{Random}}\Sigma_A X_{\text{Random}}' + \sigma^2_e I_n).$$

(10.2)

We will not be dealing with the general case, but present the two-way balanced mixed model, which is fairly straightforward to analyze. More complex models need complex algorithms, found in packages like SPSS and SAS. The next section looks at the balanced randomized block case, without interactions. Section 10.2 adds potential interactions. In each case, we find the ANOVA table and $E[MS]$’s.

10.1 Randomized blocks

As in the hot dog example, the rows are randomly chosen from a large population $I$, and the columns represent $c$ fixed treatments. The parameter $\mu_{ij}$ is the average for individual $I$ and column treatment $j$, e.g., the average rating person $I$ would give to hot dog formulation $j$. The column means are found by averaging over the entire population,

$$\mu_j = E[I][\mu_{ij}].$$

(10.3)
It is a fixed parameter, because formulation \( j \) is a fixed treatment. The overall average and column effects are then defined as usual:

\[
\mu = \frac{\mu_1 + \cdots + \mu_c}{c}, \quad \beta_j = \mu_j - \mu,
\]

so that \( \beta_1 + \cdots + \beta_c = 0 \).

For the row effects, we look at the

\[
\alpha^*_{ij} = \mu_{ij} - \mu_j = \mu_{ij} - \mu - \beta_j
\]

for each \( j \), being the \( I \)th individual’s mean for treatment \( j \) relative to the population. Note that \( E_{I}[\alpha^*_{ij}] = 0 \). If these values are different for different \( j \)’s, then there is interaction. It is certainly possible there is interaction, e.g., some people may tend to rate soft hot dogs higher than average, and hard hot dogs softer than average. No interaction means each person is consistent over formulations, that is, is the same amount above average (or below average) for each \( j \). Calling that amount \( \alpha_I \), we than have

\[
\alpha^*_{ij} = \alpha_I
\]

for each \( j \), so that

\[
\mu_{ij} = \mu + \alpha_I + \beta_j
\]

which is of course the usual additive model.

In the balanced design, each individual has \( N \) independent measurements for each treatment, so that the model is

\[
y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad i = 1, \ldots, r; \quad j = 1, \ldots, c; \quad k = 1, \ldots, N.
\]

Rewriting, with \( A_i = \alpha_i \), we have

\[
y_{ijk} = \mu + A_i + \beta_j + e_{ijk}, \quad i = 1, \ldots, r; \quad j = 1, \ldots, c; \quad k = 1, \ldots, N,
\]

or with matrices,

\[
y = \mu \mathbf{1}_n + \mathbf{X}_R \mathbf{A} + \mathbf{X}_C \mathbf{\beta} + \mathbf{e}.
\]

We model the \( A_i \)’s here the same as for the one-way random effects model, so that

\[
\mathbf{A} \sim N_r(\mathbf{0}_r, \sigma^2_A \mathbf{I}_r), \quad \mathbf{e} \sim N_n(\mathbf{0}_n, \sigma^2_e \mathbf{I}_n),
\]

and \( \mathbf{A} \) and \( \mathbf{e} \) are independent.

The mean and covariance of \( y \) are

\[
E[y] = \mu \mathbf{1}_n + \mathbf{X}_C \mathbf{\beta}, \quad Cov[y] = \sigma^2_A \mathbf{X}_R \mathbf{X}_R' + \sigma^2_e \mathbf{I}_n
\]

or, using Kronecker products,

\[
E[y] = (\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\mu + (\mathbf{1}_r \otimes \mathbf{I}_c \otimes \mathbf{1}_N)\mathbf{\beta},
\]

\[
Cov[y] = cN\sigma^2_A \mathbf{M}_R + \sigma^2_e \mathbf{I}_n,
\]
as in (9.67). We obtain the usual projections, onto $\mathcal{M}_{R\emptyset}$, $\mathcal{M}_{C\emptyset}$, and $\mathcal{M}_{R+C}$. (The error is from the additive model.)

For the means, we have

$$ E[\hat{y}_{R\emptyset}] = M_{R\emptyset}((\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\mu + (\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\beta) $$

$$ = (H_r \otimes J_c \otimes J_N)(\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\mu + (H_r \otimes J_c \otimes J_N)(\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\beta $$

$$ = 0, \quad \text{(10.14)} $$

because $H_r \mathbf{1}_r = \mathbf{0}_r$. Similarly,

$$ E[\bar{y} - \hat{y}_{R+C}] = 0. \quad \text{(10.15)} $$

The other one is not zero,

$$ E[\hat{y}_{C\emptyset}] = M_{C\emptyset}((\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\mu + (\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\beta) $$

$$ = (J_r \otimes H_c \otimes J_N)(\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\mu + (J_r \otimes H_c \otimes J_N)(\mathbf{1}_r \otimes \mathbf{1}_c \otimes \mathbf{1}_N)\beta $$

$$ = (\mathbf{1}_r \otimes H_c \otimes \mathbf{1}_N)\beta. \quad \text{(10.16)} $$

Then

$$ ||E[\hat{y}_{C\emptyset}]||^2 = \beta'(\mathbf{1}_r \otimes H_c \otimes \mathbf{1}_N)'(\mathbf{1}_r \otimes H_c \otimes \mathbf{1}_N)\beta $$

$$ = \beta'(\mathbf{1}_r \otimes H_c H_c \otimes \mathbf{1}_N)\beta $$

$$ = rN\|\beta\|^2. \quad \text{(10.17)} $$

The last step follows from $H_c \beta = \beta$ because the $\beta_j$'s sum to 0.

The covariances are $M(cN\sigma^2_A M_R + \sigma^2_e I_n)M$ for the various projection matrices $M$. The $M_{C\emptyset}M_R = 0$, $M_{R\emptyset}M_R = M_{R\emptyset}$, and $(I_n - M_{R+C})M_R = 0$. One can also show that the three projections are independent. Putting these together, we have

| Projection | $||\text{Mean}||^2$ | Covariance | $E[MS]$ |
|------------|-------------------|------------|---------|
| $\hat{y}_{R\emptyset}$ | $0$ | $(cN\sigma^2_A + \sigma^2_e)M_{R\emptyset}$ | $cN\sigma^2_A + \sigma^2_e$ |
| $\hat{y}_{C\emptyset}$ | $rN\|\beta\|^2$ | $\sigma^2_e M_{C\emptyset}$ | $rN\|\beta\|^2 / (c - 1) + \sigma^2_e$ |
| $\hat{y} - \hat{y}_{R+C}$ | $0$ | $\sigma^2_e (I_n - M_{R+C})$ | $\sigma^2_e$ |

The expected mean squares are found by dividing the $||\text{Mean}||^2$ by the degrees of freedom, then adding that to the $\sigma^2$ part of the covariance. The ANOVA table is then

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>$E[\text{Mean square}]$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows ($M_{R\emptyset}$)</td>
<td>$</td>
<td></td>
<td>\hat{y}_{R\emptyset}</td>
<td></td>
</tr>
<tr>
<td>Columns ($M_{C\emptyset}$)</td>
<td>$</td>
<td></td>
<td>\hat{y}_{C\emptyset}</td>
<td></td>
</tr>
<tr>
<td>Error ($M_{R+C}$)</td>
<td>$</td>
<td></td>
<td>\hat{y} - \hat{y}_{R+C}</td>
<td></td>
</tr>
<tr>
<td>Total ($\hat{y}$)</td>
<td>$</td>
<td></td>
<td>\hat{y} - \hat{y}_0</td>
<td></td>
</tr>
</tbody>
</table>
10.1.1 Estimation and testing

Testing is the same as for the fixed effects additive case, except that for the rows, we are testing $H_0: \sigma_A^2 = 0$. The estimation of contrasts in the $\beta_j$'s (or $\mu_j$'s) is also the same. That is, if $\ell'\beta$ is a contrast, then its estimate is 

$$\ell'\hat{\beta} = c_1\bar{y}_1 + \cdots + c_c\bar{y}_c = \hat{\alpha}'\bar{y}$$  

(10.19)

where $\hat{\alpha} \in \mathcal{M}_{C,0}$. (Specifically, $\hat{\alpha} = (\mathbb{1}_r \otimes \mathcal{C} \otimes \mathbb{1}_N)/rN$.) Then

$$Cov[\hat{\alpha}'y] = \hat{\alpha}'Cov[y]\hat{\alpha} = \hat{\alpha}'(\sigma_A^2XRX_R' + \sigma_e^2I_n)\hat{\alpha} = ||\hat{\alpha}||^2\sigma_e^2,$$  

(10.20)

because $\hat{\alpha} \in \mathcal{M}_{C,0} \perp \mathcal{M}_R$, so that $\hat{\alpha}'X_R = \mathbb{0}_c'$. Note that $||\hat{\alpha}||^2 = r||\mathcal{C}||^2N/(rN)^2 = ||\mathcal{C}||^2/rN$, so that we have

$$se(\ell'\beta) = ||\ell||\sqrt{MSE/rN}.$$  

(10.21)

Example. Five soybean treatments (one a control) were tested on each of five plots of land. That is, each of five plots of land had five batches of soybean seeds, each batch receiving one of the five treatments. The $y$ is then the number of failures (out of 100 seeds) for that batch. Thus there are $r = 5$ plots, the random effects, and $c = 5$ treatments, the fixed effects. Each plot $\times$ treatment combination produced only one observation, so $N = 1$. (From Snedecor and Cochran, Statistical Methods, Sixth Edition, Table 11.2.1.)

Here is the ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plots</td>
<td>49.84</td>
<td>4</td>
<td>12.46</td>
<td>2.303</td>
</tr>
<tr>
<td>Treatments</td>
<td>83.84</td>
<td>4</td>
<td>20.96</td>
<td>3.874</td>
</tr>
<tr>
<td>Error</td>
<td>86.56</td>
<td>16</td>
<td>5.41</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>220.24</td>
<td>24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The $F_{4,16,0.05} = 3.007$. Thus there is a significant treatment effect, but we can accept $\sigma_A^2 = 0$. The latter means that the plots a reasonably homogeneous, so maybe next time it is not worth bothering to use randomized blocks, but just randomly place the 25 sets of seeds around the land.

10.2 Two-way mixed model with interactions

Again we have the rows begin randomly selected from $\mathcal{I}$, and the columns being fixed treatments. With $\mu_{Ij}$ being the mean of the $I^{th}$ individual receiving the $j^{th}$ treatment, we again define

$$\mu_j = E[\mu_{Ij}], \quad \mu = \frac{\mu_1 + \cdots + \mu_c}{c}, \quad \beta_j = \mu_j - \mu,$$  

(10.22)
10.2. TWO-WAY MIXED MODEL WITH INTERACTIONS

\[
\alpha_{Ij}^* = \mu_{Ij} - \mu - \beta_j. \quad (10.23)
\]

These definitions imply that

\[
E_I[\alpha_{Ij}^*] = 0 \text{ for each } j, \text{ and } \beta_1 + \cdots + \beta_c = 0. \quad (10.24)
\]

No interactions means the \(\alpha_{Ij}^*\) is the same for each \(j\). Here we are not making that assumption, although may wish to test that hypothesis. We can define the average effect for \(I\) to be

\[
\alpha_I = \frac{\alpha_{I1}^* + \cdots + \alpha_{Ic}^*}{c}, \quad (10.25)
\]

and the interactions to be

\[
\gamma_{Ij} = \alpha_{Ij}^* - \alpha_I. \quad (10.26)
\]

Note that

\[
E_I[\gamma_{Ij}] = 0 \text{ and } \gamma_{I1} + \cdots + \gamma_{Ic} = 0. \quad (10.27)
\]

Then the general model is

\[
y_{ijk} = \mu + \alpha_I + \beta_j + \gamma_{Ij} + e_{ijk}. \quad (10.28)
\]

Because \(I\) is random, \(\alpha_I\) and the \(\gamma_{Ij}\)'s are random. Unfortunately, we cannot justifiably assume they are independent. In particular, the \(\gamma_{Ij}\)'s cannot be independent (unless they are all 0) because they have to sum to 0 over \(j\). In Section 10.2.2 we present an approach to the joint distribution of the \(\alpha_I\) and \(\gamma_{Ij}\)'s that yields fairly simple results. Here, we will present just the outcome.

The data are then \(I_1, \ldots, I_r\), a simple random sample from \(I\). Making a similar change in notation as from (9.54) to (9.55), the model is

\[
y_{ijk} = \mu + A_i + \beta_j + C_{ij} + e_{ijk}, \quad i = 1, \ldots, r; \quad j = 1, \ldots, c; \quad k = 1, \ldots, N. \quad (10.29)
\]

The \(A_i\) and \(C_{ij}\)'s are random, the \(\beta_j\)'s are fixed. In matrix form,

\[
\underline{y} = \mu \underline{1}_n + X_R \underline{A} + X_C \underline{\beta} + X_{R \times C} \underline{C} + \underline{e}. \quad (10.30)
\]

The ANOVA table (under assumptions from Section 10.2.2) is

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>E[Mean square]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows (random)</td>
<td>(|\tilde{y}<em>{R \cdot} - \mu</em>{R \cdot}|^2)</td>
<td>(r - 1)</td>
<td>(cN\sigma_A^2 + \sigma_e^2)</td>
</tr>
<tr>
<td>Columns (fixed)</td>
<td>(|\tilde{y}<em>{C \cdot} - \mu</em>{C \cdot}|^2)</td>
<td>(c - 1)</td>
<td>(N\sigma_C^2 + \sigma_e^2)</td>
</tr>
<tr>
<td>Interactions (random)</td>
<td>(|\tilde{y}<em>{(R \times C) - (R+C)} - \mu</em>{(R \times C) - (R+C)}|^2)</td>
<td>((r - 1)(c - 1))</td>
<td>(N\sigma_{\beta}^2/(c - 1) + N\sigma_C^2 + \sigma_e^2)</td>
</tr>
<tr>
<td>Error</td>
<td>(|\tilde{y} - \tilde{\mu}|^2)</td>
<td>(n - rc)</td>
<td>(\sigma^2_e)</td>
</tr>
<tr>
<td>Total</td>
<td>(|\tilde{y} - \tilde{\mu}|^2)</td>
<td>(n - 1)</td>
<td>—</td>
</tr>
</tbody>
</table>
It is helpful to compare the expected mean squares for different models. Below we have them (all in the two-way balanced ANOVA with interactions), with zero, one or two of the effects random:

<table>
<thead>
<tr>
<th>Source</th>
<th>Rows fixed</th>
<th>Rows random</th>
<th>Rows random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>$cN|a|^2/(r-1) + \sigma^2_e$</td>
<td>$cN\sigma^2_A + N\sigma^2_C + \sigma^2_e$</td>
<td>$cN\sigma^2_A + \sigma^2_e$</td>
</tr>
<tr>
<td>Columns</td>
<td>$rN|\beta|^2/(c-1) + \sigma^2_e$</td>
<td>$rN\sigma^2_B + N\sigma^2_C + \sigma^2_e$</td>
<td>$rN|\beta|^2/(c-1) + N\sigma^2_C + \sigma^2_e$</td>
</tr>
<tr>
<td>Interactions</td>
<td>$N|\gamma|^2/((r-1)(c-1)) + \sigma^2_e$</td>
<td>$N\sigma^2_C + \sigma^2_e$</td>
<td>$N\sigma^2_C + \sigma^2_e$</td>
</tr>
<tr>
<td>Error</td>
<td>$\sigma^2_e$</td>
<td>$\sigma^2_e$</td>
<td>$\sigma^2_e$</td>
</tr>
</tbody>
</table>

10.2.1 Estimation and testing

To test for interactions ($\sigma^2_C = 0$) or for block effects ($\sigma^2_A = 0$), the usual $F$ tests apply, that is, $MS_{\text{Interactions}}/MSE$ and $MS_{\text{Rows}}/MSE$, respectively. For testing the fixed effects, which are usually the effects of most interest, the $F$ uses the interaction mean square in the denominator:

$$H_0 : \beta = 0 \quad , \quad F = \frac{MS_{\text{Columns}}}{MS_{\text{Interactions}}} \sim F_{c-1,(r-1)(c-1)} \text{ under the null hypothesis.} \quad (10.31)$$

Turn to contrasts in the column effects, $c^T\beta$ with $c_1 + \cdots + c_c = 0$. The estimate is the usual one,

$$\hat{c^T\beta} = c_1\overline{y}_1 + \cdots + c_c\overline{y}_c. \quad (10.32)$$

The standard error uses the interactions mean square (not the MSE), in the standard error:

$$se(\hat{c^T\beta}) = \|c\|\sqrt{MS_{\text{Interactions}}/rN}. \quad (10.33)$$

See (10.49).

10.2.2 Distributional assumptions on the random effects

One approach to modeling the joint distribution of the $\alpha^*_{ij}$'s from (10.23), $\alpha^*_{ij} = \mu + \mu - \beta_j$, is to assume that the variances are equal (to $\sigma^2$, say), and the correlations are equal (to $\rho$). That is, letting $\alpha^*_i = (\alpha^*_{i1}, \ldots, \alpha^*_{ic})'$, assume that

$$Cov[\alpha^*_i] = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{pmatrix}. \quad (10.34)$$
10.2. TWO-WAY MIXED MODEL WITH INTERACTIONS

Such matrices can be written as a linear combination of $\mathbf{I}$ and $\mathbf{J}$, in this case

$$Cov[\mathbf{\alpha}_i^*] = (1 - \rho)\sigma^2\mathbf{I}_c + c\rho\sigma^2\mathbf{J}_c.$$  \hspace{1cm} (10.35)

This model is restrictive, but captures the idea that the effects are exchangeable, that is, the distribution is the same when permuting the order of the $i$’s. It can make sense when there is no particular reason to believe that two treatments are any more or less related than any other two treatments. It may not be reasonable if the $i$’s represent time points, and one expects measurements made close together in time are more highly correlated.

Now (10.25) and (10.26) show that

$$\mathbf{\alpha}_i^* = \mathbf{\alpha}_I \mathbf{L}_c + \mathbf{\gamma}_I,$$  \hspace{1cm} (10.36)

where $\mathbf{\gamma}_I = (\gamma_{i1}, \ldots, \gamma_{ic})'$. We can write the quantities on the right as

$$\alpha_I \mathbf{L}_c = \mathbf{J}_c \mathbf{\alpha}_I^*$$  and $$\mathbf{\gamma}_I = \mathbf{H}_c \mathbf{\alpha}_I^*.$$  \hspace{1cm} (10.37)

Then from (10.35),

$$Cov[\alpha_I \mathbf{L}_c] = \mathbf{J}_c((1 - \rho)\sigma^2\mathbf{I}_c + c\rho\sigma^2\mathbf{J}_c)\mathbf{J}_c = ((1 - \rho)\sigma^2 + c\rho\sigma^2)\mathbf{J}_c = \sigma^2(1 + (c - 1)\rho)\mathbf{J}_c,$$

$$Cov[\mathbf{\gamma}_I] = \mathbf{H}_c((1 - \rho)\sigma^2\mathbf{I}_c + c\rho\sigma^2\mathbf{J}_c)\mathbf{H}_c = \sigma^2(1 - \rho)\mathbf{H}_c.$$  \hspace{1cm} (10.38)

Also, $Cov[\alpha_I \mathbf{L}_c, \mathbf{\gamma}_I] = \mathbf{0}$, so that the vectors are independent. For convenience, define $\sigma^2_\mathbf{A} = Var[\alpha_I] = \sigma^2(1 + (c - 1)\rho)/c$ and $\sigma^2_\mathbf{C} = \sigma^2(1 - \rho)$, so that

$$Cov[\alpha_I \mathbf{L}_c] = c\sigma^2_\mathbf{A}\mathbf{J}_c$$  and $$Cov[\mathbf{\gamma}_I] = \sigma^2_\mathbf{C}\mathbf{H}_c.$$  \hspace{1cm} (10.39)

Letting $\mathbf{y}_I$ be the $cN$ observations from individual $I$, we have that

$$\mathbf{y}_I = \begin{pmatrix} y_{i11} \\ y_{i12} \\ \vdots \\ y_{i1N} \\ \vdots \\ y_{ic1} \\ \vdots \\ y_{icN} \end{pmatrix} = (\mathbf{L}_c \otimes \mathbf{1}_N)\mu + (\mathbf{L}_c \otimes \mathbf{1}_N)\alpha_I + (\mathbf{I}_c \otimes \mathbf{1}_N)\beta + (\mathbf{I}_c \otimes \mathbf{1}_N)\mathbf{\gamma}_I + \mathbf{e}_I,$$  \hspace{1cm} (10.40)

$\mathbf{e}_I$ being the part of $\mathbf{e}$ for individual $I$. Then

$$Cov[\mathbf{y}_I] = Var[\alpha_I](\mathbf{L}_c \otimes \mathbf{1}_N)(\mathbf{L}_c \otimes \mathbf{1}_N)' + (\mathbf{I}_c \otimes \mathbf{1}_N)Cov[\mathbf{\gamma}_I](\mathbf{I}_c \otimes \mathbf{1}_N)' + \sigma^2_\mathbf{e}\mathbf{I}_{cN} = (Nc)\sigma^2_\mathbf{A}(\mathbf{J}_c \otimes \mathbf{J}_N) + \sigma^2_\mathbf{C}(\mathbf{I}_c \otimes \mathbf{1}_N)(\mathbf{H}_c \otimes 1)(\mathbf{I}_c \otimes \mathbf{1}_N)' + \sigma^2_\mathbf{e}\mathbf{I}_{cN} = (Nc)\sigma^2_\mathbf{A}(\mathbf{J}_c \otimes \mathbf{J}_N) + N\sigma^2_\mathbf{C}(\mathbf{H}_c \otimes \mathbf{J}_N) + \sigma^2_\mathbf{e}(\mathbf{I}_c \otimes \mathbf{I}_N).$$  \hspace{1cm} (10.41)
Turning to the data, we have $I_1, \ldots, I_r$ a random sample from $\mathcal{I}$. We again model the individuals as being independent, so that with $y = (y_{11}, \ldots, y_{rt})'$, the overall covariance is $I_r \otimes \text{Cov}[y_i]$ in (10.41):

$$\text{Cov}[y] = Nc\sigma_A^2(I_r \otimes J_c \otimes J_N) + N\sigma_C^2(I_r \otimes H_c \otimes J_N) + \sigma_e^2(I_r \otimes I_c \otimes I_N).$$

(10.42)

Also, for the mean we use $1_n \otimes E[y_i]$, so by (10.40),

$$E[y] = (1_r \otimes 1_c \otimes 1_N)\mu + (1_r \otimes I_c \otimes 1_N)\beta.$$

(10.43)

For the projections, we round up the usual suspects from (9.66):

$$
\begin{align*}
M_{R, \emptyset} &= H_r \otimes J_c \otimes J_N \\
M_{C, \emptyset} &= J_c \otimes H_c \otimes J_N \\
M_{(R \times C), (R+C)} &= H_r \otimes H_c \otimes J_N \\
I_n - M_{R \times C} &= I_r \otimes I_c \otimes H_N.
\end{align*}
$$

(10.44)

For these, all the means are zero except for the second one:

$$E[\tilde{y}_{C, \emptyset}] = M_{C, \emptyset} = (J_c \otimes H_c \otimes J_N)(1_r \otimes I_c \otimes 1_N)\beta = (1_r \otimes H_c \otimes 1_N)\beta,$$

(10.45)

hence

$$\|E[\tilde{y}_{C, \emptyset}]\|^2 = rN\|\beta\|^2.$$

(10.46)

Applying the projection matrices to the covariance in (10.42) yields

$$
\begin{align*}
\text{Cov}[\tilde{y}_{R, \emptyset}] &= Nc\sigma_A^2(H_r \otimes J_c \otimes J_N) + 0 + \sigma_e^2(H_r \otimes J_c \otimes J_N) = (Nc\sigma_A^2 + \sigma_e^2)M_{R, \emptyset} \\
\text{Cov}[\tilde{y}_{C, \emptyset}] &= 0 + N\sigma_C^2(J_r \otimes H_c \otimes J_N) + \sigma_e^2(J_r \otimes H_c \otimes J_N) = (N\sigma_C^2 + \sigma_e^2)M_{C, \emptyset}, \\
\text{Cov}[\tilde{y}_{(R \times C), (R+C)}] &= 0 + N\sigma_C^2(H_r \otimes H_c \otimes J_N) + \sigma_e^2(H_r \otimes H_c \otimes J_N) = (N\sigma_C^2 + \sigma_e^2)M_{(R \times C), (R+C)}, \\
\text{Cov}[\tilde{y} - \tilde{y}_{R \times C}] &= 0 + 0 + \sigma_e^2(I_r \otimes I_c \otimes H_N) = \sigma_e^2(I_n - M_{R \times C}).
\end{align*}
$$

(10.47)

The expected mean squares are then found by adding $\|E[\tilde{y}]\|^2/df$ to the $\sigma^2$ part of the covariances, i.e., (10.46) and (10.47) combine to show

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>$\text{E[Mean square]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows (random)</td>
<td>$r - 1$</td>
<td>$cN\sigma_A^2 + \sigma_e^2$</td>
</tr>
<tr>
<td>Columns (fixed)</td>
<td>$c - 1$</td>
<td>$rN|\beta|^2/(c - 1) + N\sigma_C^2 + \sigma_e^2$</td>
</tr>
<tr>
<td>Interactions (random)</td>
<td>$(r - 1)(c - 1)$</td>
<td>$N\sigma_C^2 + \sigma_e^2$</td>
</tr>
<tr>
<td>Error</td>
<td>$n - rc$</td>
<td>$\sigma_e^2$</td>
</tr>
</tbody>
</table>

It is also easy to show these projections are independent (because the projection matrices are orthogonal), hence we have the ANOVA table at the end of Section 10.2.
Finally, for contrast $\mathbf{c}' \beta$, where $\hat{\mathbf{a}} = (\mathbf{1}_r \otimes \mathbf{c} \otimes \mathbf{1}_N)/rN$, we have by (10.42),

$$\text{Var}[\hat{\mathbf{a}}' y] = \hat{\mathbf{a}}' \text{Cov}[y] \hat{\mathbf{a}} = (\mathbf{1}_r \otimes \mathbf{c} \otimes \mathbf{1}_N)/(rN)^2$$

$$= \left( N\sigma^2 \mathbf{I}_r \otimes \mathbf{J}_c \otimes \mathbf{J}_N + N\sigma^2 \mathbf{H}_c \otimes \mathbf{J}_N + \sigma^2 \mathbf{I}_r \otimes \mathbf{I}_c \otimes \mathbf{I}_N \right)$$

$$= \left( 0 + rN^2 \|\mathbf{c}\|^2 \sigma_C^2 + rN \|\mathbf{c}\|^2 \sigma_e^2 \right)/(rN)^2$$

$$= \|\mathbf{c}\|^2 \left( \frac{N\sigma_C^2 + \sigma_e^2}{rN} \right). \quad (10.48)$$

We used the fact that the elements of $\mathbf{c}$ sum to 0, hence $\mathbf{H}_c \mathbf{c} = \mathbf{c}$ and $\mathbf{J}_c \mathbf{c} = \mathbf{0}_r$. The estimated standard error then uses the interaction mean square, i.e.,

$$\text{se}(\hat{\mathbf{a}}' y) = \|\mathbf{c}\| \sqrt{\frac{MS_{\text{Interactions}}}{rN}}. \quad (10.49)$$

**Example.** Problem 8.1 in Scheffe’s *The Analysis of Variance* has data on flow rates of fuel through three types of nozzles, which are the fixed effects. Each of five operators (the random effects) tested each nozzle three times. Thus we have $r = 3$, $c = 5$, and $N = 3$. The ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$E[MS]$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nozzles (fixed)</td>
<td>1427.0</td>
<td>2</td>
<td>713.5</td>
<td>$cN|\mathbf{c}|^2/(r-1) + N\sigma_C^2 + \sigma_e^2$</td>
<td>3.13</td>
</tr>
<tr>
<td>Operators (random)</td>
<td>798.8</td>
<td>4</td>
<td>199.7</td>
<td>$rN\sigma_B^2 + \sigma_e^2$</td>
<td>1.97</td>
</tr>
<tr>
<td>Interactions (random)</td>
<td>1821.5</td>
<td>8</td>
<td>227.7</td>
<td>$N\sigma_C^2 + \sigma_e^2$</td>
<td>2.25</td>
</tr>
<tr>
<td>Error</td>
<td>3038.0</td>
<td>30</td>
<td>101.3</td>
<td>$\sigma_e^2$</td>
<td>$---$</td>
</tr>
<tr>
<td>Total</td>
<td>7085.2</td>
<td>44</td>
<td>$---$</td>
<td>$---$</td>
<td>$---$</td>
</tr>
</tbody>
</table>

To test for nozzle effects, we use $F = MS_{\text{Nozzle}}/MS_{\text{Interactions}} = 3.13$. The $F_{2,30,0.05} = 4.459$, so this effect is not significant. To test the operator effect, $F = MS_{\text{Operators}}/MSE = 1.97$, which does not look significant. ($F_{4,30,0.05} = 2.690$.) For interactions, $F_{8,30,0.05} = 2.266$, so it is just barely not significant. It is close enough to be leery of assuming the interactions are 0. If you could assume that, then the $F$ for nozzles would be $MS_{\text{Nozzle}}/MSE = 7.04$, which is very significant ($F_{2,30,0.05} = 3.316$).
Chapter 11

Three- (and higher-) way ANOVA

There is no problem handling models with more than two factors. If all factors are fixed, then the usual linear model theory applies. If some are, or all, factors are random, then things become more complicated, but in principle everything we have done so far carries over. We will indicate the factors by $A, B, C$, etc., rather than rows, columns, .... The number of levels will be the respective lower-case letters, that is $a$ levels of factor $A$, $b$ levels of factor $B$, etc.

The three-way ANOVA is given by

$$y_{ijkl} = \mu_{ijk} + e_{ijkl}, \quad i = 1, \ldots, a; \quad j = 1, \ldots, b; \quad k = 1, \ldots, c; \quad l = 1, \ldots, N_{ijk}. \tag{11.1}$$

The $\mu_{ijk}$ is the average for observations at level $i$ of factor $A$, level $j$ of factor $B$, and level $k$ of factor $C$. The $N_{ijk}$ is the number of observations at that combination of levels.

The simplest model in the additive one, which as in the two-way ANOVA, means that each factor adds the same amount no matter the levels of the other factors, so that

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k \tag{11.2}$$

for some parameters $\alpha_i, \beta_j, \gamma_k$. These parameters (effects) are not estimable without some constraints. Effects can be defined in terms of the means. The simplest is to take straight averages. A dot indicates averaging over that index, e.g,

$$\mu_{.j.} = \frac{\sum_{i=1}^{a} \sum_{k=1}^{c} \mu_{ijk}}{ac}. \tag{11.3}$$

Then the overall mean is $\mu = \mu_{.\ldots}$, and the main effects are

$$\alpha_i = \mu_{i..} - \mu \quad \beta_j = \mu_{.j.} - \mu \quad \gamma_k = \mu_{..k} - \mu. \tag{11.4}$$

These definition imply the constraints

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{k=1}^{c} \gamma_k = 0. \tag{11.5}$$
Violations of the additive model can take a number of different forms. Two-way interactions are the nonadditive parts of the $\mu_{ij}$, $\mu_{i.k}$ and $\mu_{.jk}$'s. That is, if the $\mu_{ij}$'s satisfied an additive model, then we would have $\mu_{ij} = \mu + \alpha_i + \beta_j$. The $AB$ interactions are the differences $\mu_{ij} - (\mu + \alpha_i + \beta_j)$. In order not to run out of Greek letters, we denote these interactions $(\alpha\beta)_{ij}$. This $(\alpha\beta)$ is to be considered a single symbol, not the product of two parameters. Then

$$
(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)
$$

$$
(\alpha\gamma)_{ik} = \mu_{i.k} - (\mu + \alpha_i + \gamma_k)
$$

$$
(\beta\gamma)_{jk} = \mu_{.jk} - (\mu + \beta_j + \gamma_k)
$$

These parameters sum to zero over either of their indices, e.g.,

$$
\sum_{i=1}^{a}(\alpha\beta)_{ij} = \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0,
$$

and similarly for the others. The model with all two-way interactions is then

$$
\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.
$$

The model (11.9) is not the saturated model, that is, it may not hold. The difference between $\mu_{ijk}$ and the sum of those parameters is called the three-way interaction:

$$
(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - (\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk})
$$

$$
= \mu_{ijk} - \mu_{ij} - \mu_{i.k} - \mu_{.jk} + \mu_{i..} + \mu_{i.j} + \mu_{i..} - \mu_{..k} - \mu_{..}.
$$

A non-zero two-way interaction, say between factors $A$ and $B$, means that the effect of factor $A$ can be different for different levels of factor $B$, and vice versa. A non-zero three-way interaction means that the effect of factor $A$ can be different for each combination of levels of $B$ and $C$; or that the effect of factor $B$ can be different for each combination of levels of $A$ and $C$; or that the effect of factor $C$ can be different for each combination of levels of $A$ and $B$. For purposes of interpretation, it is easiest if there are few such high-order interactions, although in some cases these interactions may be very important. For example, which drug is best may depend on a combination of sex, age, and race.

The saturated model, which puts no restrictions on the means, is then

$$
\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.
$$
Adding in the errors, it is, from (11.1),

\[ y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}. \tag{11.12} \]

Writing it out in matrix form, it is

\[ y = \mathbf{1}_n \mu + X_A \alpha + X_B \beta + X_C \gamma + X_{AB} (\alpha \beta) + X_{AC} (\alpha \gamma) + X_{BC} (\beta \gamma) + X_{ABC} (\alpha \beta \gamma) + e. \tag{11.13} \]

Rather than formally describe these matrices in general, we look at the balanced case, so that they can be given using Kronecker products. There are four components to each matrix, one for each factor, then the \( \mathbf{1}_N \) for the repetitions. For each matrix, the relevant factors' components receive identity matrices, the others receive vectors of ones:

\[ \begin{align*}
\mathbf{1}_a &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_N \\
X_A &= I_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c \otimes \mathbf{1}_N \\
X_B &= \mathbf{1}_a \otimes I_b \otimes \mathbf{1}_c \otimes \mathbf{1}_N \\
X_C &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes I_c \otimes \mathbf{1}_N \\
X_{AB} &= I_a \otimes I_b \otimes \mathbf{1}_c \otimes \mathbf{1}_N \\
X_{AC} &= I_a \otimes \mathbf{1}_b \otimes I_c \otimes \mathbf{1}_N \\
X_{BC} &= \mathbf{1}_a \otimes \mathbf{1}_b \otimes I_c \otimes \mathbf{1}_N \\
X_{ABC} &= I_a \otimes I_b \otimes I_c \otimes \mathbf{1}_N
\end{align*} \tag{11.14} \]

### 11.1 Hierarchy of models

In the two-way ANOVA, the models considered preserve a certain hierarchy: If any of the main effects is in the model, \( \mu \) is in the model, and if there is two-way interaction, then the two main effects are in the model. For three-way models, the same type of conditions are typically invoked:

- If a main effect is in the model, the \( \mu \) is in the model;

- If a two-way interaction is in the model, then the two corresponding main effects are in the model (e.g., if \( (\alpha\gamma)_{ik} \) is in the model, so are \( \alpha_i \) and \( \gamma_k \));

- If the three-way interaction is in the model, then all two-way interactions are in the model (so that it is the saturated model).

(How many such models are there?)
Each model can be described by giving which factors and interactions are present. Here are some examples (and notation):

<table>
<thead>
<tr>
<th>Model</th>
<th>Vector space</th>
<th>( \mu_{ijk} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturated</td>
<td>( M_{A \times B \times C} )</td>
<td>( \mu + \alpha_i + \beta_j + \gamma_k + (\alpha \beta)<em>{ij} + (\alpha \gamma)</em>{ik} + (\beta \gamma)<em>{jk} + (\alpha \beta \gamma)</em>{ijk} )</td>
</tr>
<tr>
<td>All two-way</td>
<td>( M_{A \times B + B \times C + A \times C} )</td>
<td>( \mu + \alpha_i + \beta_j + \gamma_k + (\alpha \beta)<em>{ij} + (\alpha \gamma)</em>{ik} + (\beta \gamma)_{jk} )</td>
</tr>
<tr>
<td>Additive</td>
<td>( M_{A + B + C} )</td>
<td>( \mu + \alpha_i + \beta_j + \gamma_k )</td>
</tr>
<tr>
<td>No ( A \times B )  interaction</td>
<td>( M_{B \times C + A \times C} )</td>
<td>( \mu + \alpha_i + \beta_j + \gamma_k + (\alpha \gamma)<em>{ik} + (\beta \gamma)</em>{jk} )</td>
</tr>
<tr>
<td>( A \times B ) interaction</td>
<td>( M_{A \times B + C} )</td>
<td>( \mu + \alpha_i + \beta_j + \gamma_k + (\alpha \beta)_{ij} )</td>
</tr>
<tr>
<td>No ( B ) effect</td>
<td>( M_{A \times C} )</td>
<td>( \mu + \alpha_i + \gamma_k + (\alpha \gamma)_{ik} )</td>
</tr>
<tr>
<td>No effects</td>
<td>( M_{\emptyset} )</td>
<td>( \mu )</td>
</tr>
</tbody>
</table>

(11.15)

There is a myriad of testing problems, with the null model nested in the alternative model. For example, testing for three-way interactions has the all two-way interaction model (11.9) as null, and the saturated model as alternative:

\[
H_0 : M_{A \times B + B \times C + A \times C} \text{ versus } H_A : M_{A \times B \times C}. \tag{11.16}
\]

Testing for additivity given no three-way interaction has the additive model (11.2) as null and the all two-way model as alternative:

\[
H_0 : M_{A + B + C} \text{ versus } H_A : M_{A \times B + B \times C + A \times C}. \tag{11.17}
\]

Generally, the idea is to find the simplest model that still fits. That is, if a parameter is found to be significant, then it stays in the model, along with the other parameters it implies. That is, if \((\alpha \beta)_{ij}\) is found significant, then the \(\alpha_i\)'s and \(\beta_j\)'s must be in the model, whether significant or not. Testing all the potential models against each other can be time consuming (although software these days can do it in seconds), and can also end up with more than one model. In the balanced case, it is much easier, because each parameter can be tested on its own.

### 11.1.1 Orthogonal spaces in the balanced case

In the balanced case, the saturated model space can be decomposed into orthogonal spaces, each one corresponding to a set of effects. In the two-way balanced ANOVA, we have

\[
M_{A \times B} = M_{\emptyset} + M_{A \emptyset} + M_{B \emptyset} + M_{(A \times B) \cdot (A + B)},
\tag{11.18}
\]

where all the spaces on the right-hand side are orthogonal. The corresponding equation for the three-way balanced model is

\[
M_{A \times B \times C} = M_{\emptyset} + M_{A \emptyset} + M_{B \emptyset} + M_{C \emptyset} + M_{(A \times B) \cdot (A + B)} + M_{(A \times C) \cdot (A + C)} + M_{(B \times C) \cdot (B + C)} + M_{(A \times B \times C) \cdot (A + B + B \times C + A \times C)}. \tag{11.19}
\]
11.2. Example

Again, the spaces on the right are all orthogonal. It is easiest to see by finding the projection matrices, which are given using Kronecker products with H’s in the relevant slots, and J’s elsewhere. That is,

\[
\begin{array}{l|l|l}
\text{Source} & \text{Sum of squares} & \text{Degrees of freedom} & \text{F} \\
\hline
\text{Main } A \text{ effect} & \|\hat{y}_{A,0}\|^2 & a - 1 & MS_A/MSE \\
\text{Main } B \text{ effect} & \|\hat{y}_{B,0}\|^2 & b - 1 & MS_B/MSE \\
\text{Main } C \text{ effect} & \|\hat{y}_{C,0}\|^2 & c - 1 & MS_C/MSE \\
\text{AB interaction} & \|\hat{y}_{(A\times B)-(A+B)}\|^2 & (a - 1)(b - 1) & MS_{AB int}/MSE \\
\text{AC interaction} & \|\hat{y}_{(A\times C)-(A+C)}\|^2 & (a - 1)(c - 1) & MS_{AC int}/MSE \\
\text{BC interaction} & \|\hat{y}_{(B\times C)-(B+C)}\|^2 & (b - 1)(c - 1) & MS_{BC int}/MSE \\
\text{ABC interaction} & \|\hat{y}_{(A\times B\times C)-(A\times B\times C+A\times C)}\|^2 & (a - 1)(b - 1)(c - 1) & MS_{ABC int}/MSE \\
\text{Error} & \|\hat{y} - \hat{y}_0\|^2 & n - abc & -- \\
\hline
\text{Total} & \|\hat{y} - \hat{y}_0\|^2 & n - 1 & -- \\
\end{array}
\]

These matrices can be found using the techniques for the two-way model. The degrees of freedom are easy to find by taking traces, recalling that traces of a Kronecker product multiply, and that \(\text{trace}(J_k) = 1\) and \(\text{trace}(H_k) = k - 1\).

From this table, it is easy to construct the ANOVA table. If all effects are fixed, then the F’s all use the \(MSE\) in the denominator. (The MS column is left out.)

11.2 Example

Snedecor and Cochran (Statistical Methods) has an example on food supplements for pigs’ corn. The factors are A: Lysine at \(a = 4\) levels, B: Methionine at \(b = 3\) levels, and C: Soybean Meal at \(c = 2\) levels. Here, \(N = 1\), so we will assume there is no three-way interaction, and use \(\|\hat{y} - \hat{y}_{A\times B\times C}\|^2\) for the error sum of squares. Measured is the average weight gain. Next is the ANOVA table, along with the relevant F cutoff point:
Looking at the two-way interactions, only the Lysine×Soy Meal interaction is significant. Of the main effects, only the Soy Meal is significant, and it is very significant. But, because Lysine does appear in a significant interaction, we have to include the Lysine main effect. Methionine does not appear in any significant effect, so we can drop it. Thus the best model appears to include Lysine, Soy Meal, and their interaction,

\[ y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + e_{ijk}. \]  

(11.21)