Notes on Maxwell & Delaney

PSYCH 710

October 12, 2021

7 Factorial Designs

In previous chapters we analyzed data from so-called one-way experimental designs, in which subjects were randomly assigned to groups that differed on a single treatment or grouping variable. In this chapter we will analyze data from **factorial experiments**. Factorial experiments contain two or more experimental variables. In a completely crossed factorial experiment, each level of every variable is combined with each level of every other variable. So, if there are two variables with a and b levels, then there will be ab combinations of treatments, which are sometimes called cells. Subjects are then assigned randomly to the various combinations of treatments. Experiments that have equal n per cell are called **balanced** designs. Designs with unequal n are unbalanced. We will start by analyzing data from balanced designs, and then consider unbalanced data near the end of this chapter.

Our analyses will focus on factorial experiments with two experimental factors, A and B. (Higher-order factorial designs are examined in chapter 8.) The design of a factorial experiment that has two variables can be represented as a two-dimensional table, with the rows representing different levels of variable A and the columns representing different levels of variable B. Each cell within the table corresponds to a particular combination of treatments. For example, a_1b_3, which is the cell in row 1 and column 3, refers to the combination of level 1 of A and level 3 in B. I will often use the terms "cell" or group to refer to a particular combination of treatments. Cell jk refers to the cell in row j and column k.

7.1 Linear Model

The linear model for data collected in a two factor factorial experiment can be expressed as

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \tag{1}$$

where Y_{ijk} is the score of individual i in cell jk, μ is a constant, α_j is the effect of treatment a_j, β_k is the effect of b_k, $(\alpha\beta)_{jk}$ is the effect of the *combination* a_jb_k, and ϵ_{ijk} is the error term. Notice that α_j is the effect of a particular level of A without reference to B. In other words, α_j represents the effect of being in row a_j regardless of the level of B. Similarly, β_k represents the effect of being in column b_k regardless of the level of A. As we shall see, the α's and β's constitute the **main effects** of A and B, respectively. The effect $(\alpha\beta)_{jk}$, on the other hand, refers to the effect of begin in a particular *cell*, and collectively (i.e., across all cells) constitute the **interaction** between A and B, which often is represented as $A \times B$. Interactions are discussed in more detail in section 7.4.1.

The main effects are defined as

$$\alpha_j = \mu_{j.} - \mu_{..} \tag{2}$$

$$\beta_k = \mu_{.k} - \mu_{..} \tag{3}$$

where $\mu_{j.}$ is the marginal mean in row j, $\mu_{.k}$ is the marginal mean in column k, and $\mu_{..}$ is the population grand mean (i.e., the mean of the row and column marginal means). There is one coefficient for each row and column, and the effects are constrained so that they sum to zero:

$$\sum_{j=1}^{a} \alpha_j = 0 \tag{4}$$

$$\sum_{k=1}^{b} \beta_k = 0 \tag{5}$$

$$\tag{6}$$

The marginal row and column means are defined as

$$\mu_{j.} = \sum_{k=1}^{b} \mu_{jk}/b \tag{7}$$

$$\mu_{.k} = \sum_{j=1}^{a} \mu_{jk}/a \tag{8}$$

which are estimated from our sample as

$$\bar{Y}_{j.} = \sum_{k=1}^{b} \bar{Y}_{jk}/b \tag{9}$$

$$\bar{Y}_{.k} = \sum_{j=1}^{a} \bar{Y}_{jk}/a \tag{10}$$

For balanced designs, the formulae for marginal means can be rewritten as

$$\bar{Y}_{j.} = \left(\sum_{k=1}^{b} \sum_{i=1}^{n} Y_{ijk} \right) /nb \tag{11}$$

$$\bar{Y}_{.k} = \left(\sum_{j=1}^{a} \sum_{i=1}^{n} Y_{ijk} \right) /na \tag{12}$$

Note that the marginal row means are the means calculated from *all* the scores in each row. Similarly, the column marginal means are the means obtained from all of the scores in column. Hence, the row marginal means ignore the column effects, and the column marginal means ignore the row effects.

The interaction terms are defined as

$$(\alpha\beta)_{jk} = \mu_{jk} - (\mu_{..} + \alpha_j + \beta_k) \tag{13}$$

which represents the difference between the cell mean, μ_{jk}, and the sum of the grand mean ($\mu_{..}$) and the row and column main effects. There is one interaction coefficient for each cell. Finally, the interaction terms in each row and each column sum to zero:

$$\sum_{j=1}^{a} (\alpha\beta)_{jk} = 0 \text{ for each value of } k \tag{14}$$

$$\sum_{k=1}^{b} (\alpha\beta)_{jk} = 0 \text{ for each value of } j \tag{15}$$

The least squares estimates of the coefficients are

$$\begin{aligned} \alpha_j &= \bar{Y}_{j.} - \bar{Y}_{..} \\ \beta_k &= \bar{Y}_{.k} - \bar{Y}_{..} \\ (\alpha\beta)_{jk} &= \bar{Y}_{jk} - (\bar{Y}_{..} + (\bar{Y}_{j.} - \bar{Y}_{..}) + (\bar{Y}_{.k} - \bar{Y}_{..})) \\ &= \bar{Y}_{jk} - \bar{Y}_{j.} - \bar{Y}_{.k} + \bar{Y}_{..} \end{aligned}$$

The main effects have $(a-1)$ and $(b-1)$ degrees of freedom, the interaction has $(a-1) \times (b-1)$ degrees of freedom, and the model in Equation 1 has $1 + (a-1) + (b-1) + (a-1)(b-1)$ parameters.

7.1.1 simple interpretation of linear model

If one cuts through the mathematical notation, the model defined by Equation 1 has a very simple interpretation. In our discussion of the linear model for one-way designs, we noted that the sum of the intercept and group effects constituted a *predicted* score for individual i in group j (i.e., $\hat{Y}_{ij}$). We can extend this idea to our two-way factorial design:

$$Y_{ijk} = \hat{Y}_{ijk} + \epsilon_{ijk} \tag{16}$$

where $\hat{Y}_{ijk}$ is the predicted score. Equation 16 says that the score of individual i in cell jk is the sum of a predicted score and an error (or residual) term. Now, the predicted score is defined as the sum of the effects in the model:

$$\hat{Y}_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

It turns out that the sum of the effects equals the mean of the scores in cell ij. In other words, $\hat{Y}_{ijk} = \bar{Y}_{ij}$. $\bar{Y}_{ij}$ is an estimate of μ_{ij}, so the model can be re-written as

$$Y_{ijk} = \hat{Y}_{ijk} + \epsilon_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

So, Equation 1 is equivalent to saying that the observed score is equal to the sum of the cell mean and an error term, but it describes the cell mean as a sum of three separate effects.

7.2 main effect of A

Generally, the linear model in Equation 1 is used to evaluate three null hypotheses corresponding to the two main effects and the $A \times B$ interaction. The null hypothesis for the main effect of A is

$$\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \tag{17}$$

which is equivalent to

$$\bar{Y}_{1.} = \bar{Y}_{2.} = \cdots = \bar{Y}_{j.} \tag{18}$$

A significant main effect of A means that the row marginal means — i.e., the average of scores within a row and across all levels of B — are not all equal

Your book points out that the sum of squares for factor A, or SS_A, can be obtained by calculating the difference between $SS_{Residual}$ obtained with the restricted model

$$Y_{ijk} = \mu + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \tag{19}$$

to $SS_{Residual}$ obtained with the full model (i.e. Equation 1). The model in Equation 19 includes the $A \times B$ interaction but not the main effect of A, and therefore violates the **principle of marginality**.[1] One undesirable feature of such models is that SS_A depends crucially on the definition of α. Our definition of an effect is given by Equation 2, but other definitions are possible. For example, we could define effect α_j as the difference between the marginal mean of row j and the marginal row mean of a baseline group (e.g., $j = 1$), giving $\alpha_j = \bar{Y}_j - \bar{Y}_1$. This definition of an effect, which is sometimes called treatment coding, or reference-cell coding, is entirely reasonable. Ideally, we would not want our results to depend on arbitrary definitions of an effect. Unfortunately, in some situations, the SS_A derived from a comparison of Equations 1 and 19 *does* depend on our definition of α. This is one reason why some statisticians argue strongly against the use of models that violate the marginality principle (Venables and Ripley, 2002). In any case, it is important that you realize that comparison of Equations 1 and 19 yields SS_A *only* when α's are defined as in Equation 2.

Are there other model comparisons that can be used to evaluate the null hypothesis in Equation 17? The answer is, yes. For example we can compare the fits provided by the models

$$Y_{ijk} = \mu + \epsilon_{ijk} \quad (20)$$

$$Y_{ijk} = \mu + \alpha_j + \epsilon_{ijk} \quad (21)$$

Alternatively, we can compare the following two models

$$Y_{ijk} = \mu + \beta_k + \epsilon_{ijk} \quad (22)$$

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk} \quad (23)$$

Notice that both comparisons involve models that differ only in the presence or absence of α coefficients, and so therefore should provide estimates of SS_A. In fact, comparisons of models 1 and 19, 20 and 21, and 22 and 23 yield *identical* values for SS_A **provided that there are equal** n **per group and we use sum-to-zero coding.**

7.2.1 calculating F

The null hypothesis (Equation 17) is evaluated by comparing SS_A to sum of squared residuals ($SS_{Residuals}$) for the full model (i.e., Eq. 1). It can be shown that

$$SS_A = E_R - E_F = \sum_{j=1}^{a} \sum_{i=1}^{n} \alpha_j^2 \quad (24)$$

The sum of squared residuals is

$$SS_{Residuals} = E_F = \sum_{j=1}^{a} \sum_{k=1}^{b} \sum_{i=1}^{n} (Y_{ijk} - \bar{Y}_{jk})^2 \quad (25)$$

Finally, the F test for the main effect of A is

$$F_A = \frac{SS_A/(a-1)}{SS_{Residuals}/(ab(n-1))} \quad (26)$$

with numerator and denominator degrees of freedom of $(a-1)$ and $ab(n-1)$, respectively.

[1] In a quadratic model that contains linear (x) and quadratic (x^2) terms, the linear term is said to be *marginal* to the quadratic term. In the case of factors, a main effect, A, is said to be marginal to interactions that contain it (e.g., $A \times B$, $A \times B \times C$, etc.). According to the principle of marginality, if a model contains a higher-order term (e.g., AB), then it must also contain the terms marginal to it (e.g., A and B).

7.3 main effect of B

The null hypothesis for the main effect of B is

$$\beta_{.1} = \beta_{2.} = \cdots = \beta_{.b} = 0 \tag{27}$$

which is equivalent to

$$\bar{Y}_{.1} = \bar{Y}_{.2} = \cdots = \bar{Y}_{.b} = 0 \tag{28}$$

A significant main effect of B means that the column marginal means — i.e., the average of scores within a column and across all levels of A — are not all equal

The null hypothesis can be evaluated by comparing the sum of squared residuals from the full model (i.e., Equation 1) to the sum of squared residuals for the restricted model

$$Y_{ijk} = \mu + \alpha_j + (\alpha\beta)_{jk} + \epsilon_{ijk} \tag{29}$$

Alternatively, we could compare

$$\begin{aligned} Y_{ijk} &= \mu + \epsilon_{ijk} && (30)\\ Y_{ijk} &= \mu + \beta_k + \epsilon_{ijk} && (31) \end{aligned}$$

or

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_j + \epsilon_{ijk} && (32)\\ Y_{ijk} &= \mu + \alpha_j + \beta_k + \epsilon_{ijk} && (33) \end{aligned}$$

All of the comments made about the various comparisons that can be done to calculate SS_A also apply to calculations of SS_B. It can be shown that

$$SS_B = \sum_{k=1}^{b} \sum_{i=1}^{n} \beta_k^2 \tag{34}$$

The F test for the main effect of B is

$$F_B = \frac{SS_B/(b-1)}{SS_{Residual}/(ab(n-1))} \tag{35}$$

with numerator and denominator degrees of freedom of $(b-1)$ and $ab(n-1)$, respectively.

7.4 $A \times B$ interaction

The null hypothesis for the AB interaction is

$$(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0 \tag{36}$$

A significant $A \times B$ interaction means that the at least one cell mean, $\bar{Y}_{jk}$, differs from the predicted value of $\mu + \alpha_j + \beta_k$.

The significance of the interaction is evaluated by comparing Equation 1 to

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk} \tag{37}$$

SS_{AB} is given by

$$SS_{AB} = \sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{n}(\alpha\beta)_{jk}^2 \tag{38}$$

The F test for the AB interaction is

$$F_{AB} = \frac{SS_{AB}/((a-1)(b-1))}{SS_{Residual}/(ab(n-1))} \tag{39}$$

with numerator and denominator degrees of freedom of $(a-1)(b-1)$ and $ab(n-1)$, respectively.

7.4.1 interpreting interactions

The linear model in Equation 1 is equivalent to

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \tag{40}$$

and that

$$\mu_{ij} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \tag{41}$$

Notice that Equation 41 implies that the interaction term, $(\alpha\beta)_{jk}$ is the difference between the mean of cell jk and $\mu + \alpha_j + \beta_k$. In other words, **the interaction term is the difference between the cell mean and the predictions of a model that includes only the intercept and main effects**.

An interaction is depicted graphically in Figure 1, which shows the average scores for four groups in a 2x2 factorial experiment. There is a main effect of A: on average, scores are higher in A_2 than A_1. There is a main effect of B: on average, scores are higher in groups that receive treatment B_1 than in groups that received treatment B_2. Notice, however, that the difference between B_1 and B_2 is smaller in the two groups that received treatment A_1 than in the two groups that received treatment A_2:

$$\bar{Y}_{11} - \bar{Y}_{12} < \bar{Y}_{21} - \bar{Y}_{22}$$

This inequality is represented graphically as a deviation from parallelism. A significant interaction means that the two lines representing how scores vary with treatment A — one line for subjects receiving treatment B_1 and another for subjects receiving treatment B_2 — are not parallel.

After looking at Figure 1, you might be tempted to conclude that the $A \times B$ interaction means that the effect of A was significant for subjects receiving treatment B_1 but not for subjects receiving treatment B_2. Alternatively, you might conclude that the effect of B is significant only in subjects who received treatment A_1, but not in subjects who received treatment B_2. **Such conclusions would be premature and (possibly) incorrect.** A significant interaction only means that $\bar{Y}_{11} - \bar{Y}_{12} \neq \bar{Y}_{21} - \bar{Y}_{22}$, or (equivalently) that $\bar{Y}_{12} - \bar{Y}_{22} \neq \bar{Y}_{11} - \bar{Y}_{21}$. Further tests are necessary to determine if, for example, A is or is not significant in subjects who received treatments B_1 or B_2.

A significant $A \times B$ interaction means that the effect of A *depends on the level of* B, and/or that the effect of B depends on the level of A. In this situation – when the effect of one variable depends on the level of the other variable – it often does not make sense to talk about, or further analyze, the main effects of A and B.

7.5 familywise Type I error rates

Testing for main effects of A and B and an $A \times B$ interaction will inflate the *experiment-wise* Type I error rate. Nevertheless, standard practice is to not correct for these multiple tests.

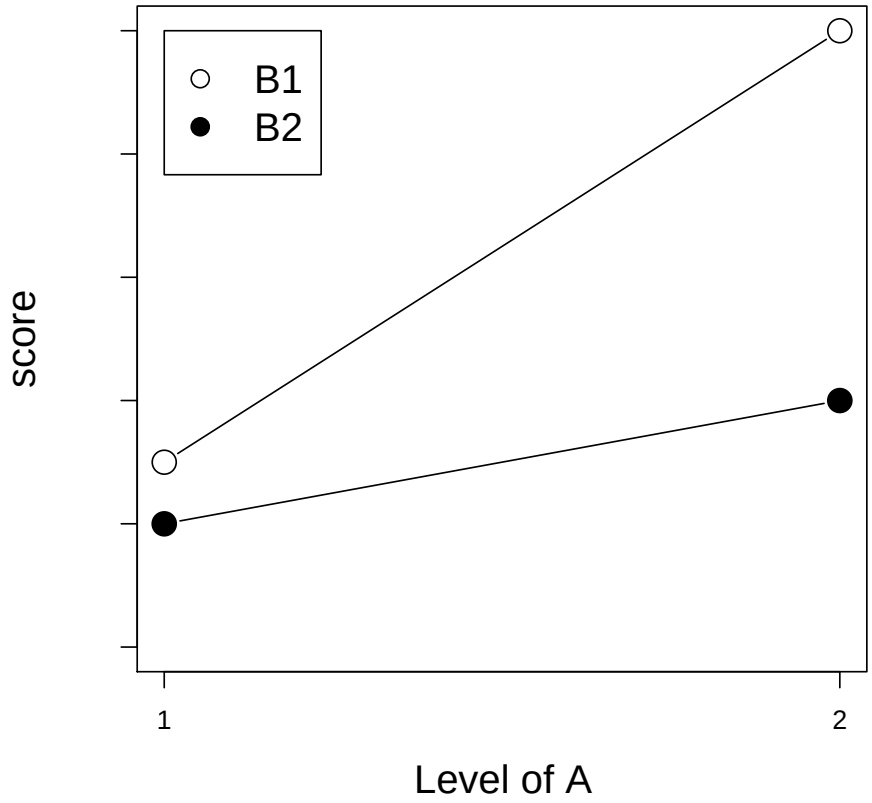

Figure 1: Illustration of an $A \times B$ interaction.

7.6 an example

In this section I will illustrate how to analyze data collected in an experiment that used a balanced factorial design. There are two *factors*, A and B, and each factor has two levels. The dependent variable is denoted by `y`. Finally, there are 6 scores per cell. Here is how I initialized R and created the fake data:

```
options(contrasts=c("contr.sum","contr.poly"))
a<-rep(c(-1,-1,1,1),each=6)
b<-rep(c(-1,1,-1,1),each=6)
ab<-rep(c(-1,1,1,-1),each=6)
y<-10+2*a+1*b+0.5*ab
set.seed(123456);
nz<-rnorm(y)
y<-y+nz;
af<-factor(a,labels=c("a1","a2"),ordered=F)
bf<-factor(b,labels=c("b1","b2"),ordered=F)
myData <- data.frame(y,af,bf)
names(myData) <- c("y","A","B")
```

As was the case for a one-way ANOVA, data from factorial designs can be analyzed with the `lm()` or `aov()` commands. However, the formula that defines the linear model is slightly more complicated that the one used to analyze data collected with one-way designs. In this factorial experiment, the linear model is defined with the the formula $y \sim 1 + A + B + A:B$, where `1` represents the intercept, `A` and `B` represent the main effects, and `A:B` is the A x B interaction. The formula can be interpreted as saying that `y` *is modeled as* the sum of an intercept term, a main effect of `A`, a main effect of `B`, and an A x B interaction. Here, I use the `lm()` and `anova()` commands to do the analysis and print the summary table:

```
lm.full.model <- lm(y ~ 1 + A + B + A:B, data=myData)
anova(lm.full.model)
```

The output is listed in Table 1. In R, the ANOVA table shows *sequential* sums-of-squares. So SS_A is derived from a comparison between models 20 and 21, SS_B comes from a comparison of models 21 and 23, and SS_{AB} comes from a comparison of models 23 and 1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	94.25	94.25	84.59	0.0000
B	1	23.21	23.21	20.83	0.0002
A:B	1	4.63	4.63	4.15	0.0550
Residuals	20	22.28	1.11		

Table 1: ANOVA table for full model.

Normally, the anova table for the full model is all you need. However, I want to show you that the terms in the full model really do come from a comparison of different reduced models. For example, we can compute SS_A by doing a direct comparison of models 20 and 21:

```
lm.01 <- lm(y ~ 1, data=myData)
lm.02.a <- lm(y ~ 1 + A, data=myData)
anova(lm.01, lm.02.a)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ 1 + A
##   Res.Df   RSS Df Sum of Sq    F  Pr(>F)
## 1     23 144.4
## 2     22  50.1  1      94.2 41.4 1.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the difference between the $SS_{Residual}$ for the two models is the same as SS_A for the full model. Next, we compare models 22 and 23. Again, the change in $SS_{Residual}$ is the same as SS_A from the full model.

```
lm.02.b <- lm(y ~ 1 + B, data=myData)
lm.03 <- lm(y ~ 1 + A + B, data=myData)
anova(lm.02.b, lm.03)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1 + B
## Model 2: y ~ 1 + A + B
##   Res.Df   RSS Df Sum of Sq    F  Pr(>F)
## 1     22 121.2
## 2     21  26.9  1      94.2 73.5 2.7e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And, finally, we calculate SS_A using Equation 24:

```
row.means <- with(myData,tapply(y, A, mean))
grand.mean <- mean(row.means)
alphas <- row.means - grand.mean
row.n <- with(myData,tapply(y, A, length))
levels(myData$A)
```

```
## [1] "a1" "a2"
```

```
a.levels <- 2
(SS.a <- sum((alphas^2) * row.n))
```

```
## [1] 94.25
```

All of these methods yield the same value for SS_A. Likewise, the values of SS_B and SS_{AB} can be calculated in several different ways. Trust me.

7.7 R and marginality

In the previous section, I did not calculate SS_A by comparing models 1 and 19. Let's do so here:

```
lm.04 <- lm(y ~ 1 + B + A:B,data=myData)
anova(lm.04, lm.full.model)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1 + B + A:B
## Model 2: y ~ 1 + A + B + A:B
##   Res.Df  RSS Df Sum of Sq F Pr(>F)
## 1     20 22.3
## 2     20 22.3  0  3.55e-15
```

Notice that the results do not look like the ones obtained in the previous section. Why not? If you examine the anova table carefully you will see that the residual degrees of freedom equals 20 in both models. This result is surprising because we removed the A main effect in the reduced model, so the residuals degrees of freedom should be greater in that model. The problem here is that R does not look kindly upon models that violate the marginality principle, and so the effects of A have been incorporated surreptitiously into the reduced model. In fact, the command `anova(lm.04)` would show that the degrees of freedom for A and SS_A have been added to the values for the `A:B` interaction.

```
anova(lm.04)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## B          1   23.2    23.2    20.8 0.00019 ***
## B:A        2   98.9    49.4    44.4 4.4e-08 ***
## Residuals 20   22.3     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, a comparison of the effects for A and the $A \times B$ interaction calculated in the full model

```
dummy.coef(lm.full.model)
```

```
## Full coefficients are
##
## (Intercept):      10.47
## A:                   a1      a2
##                  -1.982   1.982
## B:                   b1      b2
##                 -0.9835  0.9835
## A:B:              a1:b1   a2:b1  a1:b2  a2:b2
##                  -0.439   0.439  0.439 -0.439
```

are added together to form the $A \times B$ interaction effects in the reduced model

```
dummy.coef(lm.04)

## Full coefficients are
##
## (Intercept):      10.47
## B:                   b1      b2
##                  -0.9835  0.9835
## B:A:               b1:a1   b2:a1  b1:a2  b2:a2
##                   -2.421  -1.543  2.421  1.543
```

The bottom line is that R intentionally makes it difficult for you to construct and compare models that violate the marginality principle.

7.8 measures of association strength & effect size

We are considering a factorial experiment in which the factors are *fixed* (i.e., all of the factors about which inferences are to be drawn are included in the experiment). In this case, the appropriate (Kirk, 1995) measure of association strength is **partial omega squared** ($\omega^2_{partial}$). $\omega^2_{partial}$ expresses the variance of each treatment effect relative to the sum of the treatment effect variance and error variance. Note that "error" refers to the error calculated from the full model: it is the unexplained variance. Hence, $\omega^2_{A,partial}$ expresses the variance among α's relative to the sum of the error variance and the α variance, and it *ignores* the variation in the dependent variable that is due to the effects of B or $A \times B$. Here are the definitions of $\omega^2_{partial}$ in terms of the treatment effects:

$$\omega^2_{A,partial} = \frac{\sum_{j=1}^{a}(\alpha_j^2/a)}{\sigma_e^2 + \sum_{j=1}^{a}(\alpha_j^2/a)}$$

$$\omega^2_{B,partial} = \frac{\sum_{k=1}^{b}(\beta_k^2/b)}{\sigma_e^2 + \sum_{k=1}^{b}(\beta_k^2/b)}$$

$$\omega^2_{AB,partial} = \frac{\sum_{j=1}^{a}\sum_{k=1}^{b}\left[(\alpha\beta)_{jk}^2/(ab)\right]}{\sigma_e^2 + \sum_{j=1}^{a}\sum_{k=1}^{b}\left[(\alpha\beta)_{jk}^2/(ab)\right]}$$

Usually it is easier to calculate $\omega^2_{partial}$ using the values listed in the ANOVA table for the full model. Here is how $\omega^2_{partial}$ is defined in terms of $SS_{treatment}$ and $MS_{Residuals}$

$$\omega^2_{A,partial} = \frac{SS_A - df_A MS_{Residuals}}{SS_A + (N - df_A)MS_{Residuals}}$$

$$\omega^2_{B,partial} = \frac{SS_B - df_B MS_{Residuals}}{SS_B + (N - df_B)MS_{Residuals}}$$

$$\omega^2_{AB,partial} = \frac{SS_{AB} - df_{AB} MS_{Residuals}}{SS_{AB} + (N - df_{AB})MS_{Residuals}}$$

Here is how $\omega^2_{partial}$ is defined in terms of F values:

$$\begin{aligned}
\omega^2_{A,partial} &= \frac{df_A(F_A - 1)}{df_A(F_A - 1) + N} \\
\omega^2_{B,partial} &= \frac{df_B(F_B - 1)}{df_B(F_B - 1) + N} \\
\omega^2_{AB,partial} &= \frac{df_{AB}(F_{AB} - 1)}{df_{AB}(F_{AB} - 1) + N}
\end{aligned}$$

If $\omega^2_{partial} < 0$, then it is set to 0.

For the data shown in Table 1, calculating $\omega^2_{partial}$ from F's yields

```
(omega.a <- (1 * (84.59 - 1))/(1 * (84.59 - 1) + 24))

## [1] 0.7769

(omega.b <- (1 * (20.83 - 1))/(1 * (20.83 - 1) + 24))

## [1] 0.4524

(omega.ab <- (1 * (4.15 - 1))/(1 * (4.15 - 1) + 24))

## [1] 0.116
```

According to Cohen's 1988 guidelines,

$$\begin{aligned}
\omega^2_{partial} &= 0.010 \text{ is a small association} \\
\omega^2_{partial} &= 0.059 \text{ is a medium association} \\
\omega^2_{partial} &\geq 0.138 \text{ is a large association}
\end{aligned}$$

Partial omega squared can be used to calculate Cohen's measure of effect size, f

$$f_{treatment} = \sqrt{\frac{\omega^2_{partial}}{1 - \omega^2_{partial}}} \tag{42}$$

According to Cohen, f's of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes, respectively.

7.8.1 effect size & association strength using `effectsize`

Measures of effect size and association strength can be calculated easily using commands in the `effectsize` package. Careful examination of the following code will show that the values returned by `cohens_f` do not follow Eq 42. The reason for the apparent error is that `cohens_f` calculates effect size using η^2, not ω^2. My reading of the literature is that it is more common to use ω^2.

```
library(effectsize)
cohens_f(lm.full.model) # uses eta-squared, not omega-squared

## # Effect Size for ANOVA (Type I)
##
## Parameter | Cohen's f (partial) |       90% CI
## ------------------------------------------------
## A         |                2.06 | [1.39, 2.69]
## B         |                1.02 | [0.56, 1.46]
## A:B       |                0.46 | [0.00, 0.84]

eta_squared(lm.full.model)

## # Effect Size for ANOVA (Type I)
##
## Parameter | Eta2 (partial) |       90% CI
## --------------------------------------
## A         |           0.81 | [0.66, 0.88]
## B         |           0.51 | [0.24, 0.68]
## A:B       |           0.17 | [0.00, 0.41]

omega_squared(lm.full.model)

## # Effect Size for ANOVA (Type I)
##
## Parameter | Omega2 (partial) |       90% CI
## ----------------------------------------
## A         |             0.78 | [0.61, 0.86]
## B         |             0.45 | [0.18, 0.64]
## A:B       |             0.12 | [0.00, 0.35]
```

7.9 power

In this section I will show you how to calculate power using the `pwr` package for `R`. If you have not yet installed the package on your computer, do so now using the following command:

```
install.packages("pwr")
```

After installing the package, you must load it into memory with the command `library(pwr)` command. Note that you only need to use this command once per R session. Now we can use `pwr.f2.test` to calculate the power for *any* F test. Suppose we have a two-factor design: factors A and B have three and two levels, respectively. Also, let's assume that the effect sizes for the main effects of A and B are 0.1 (i.e., small) and 0.25 (i.e., medium), respectively. Finally, we assume that we have six subjects per group, and therefore that the degrees of freedom for $MS_{residuals}$ in our ANOVA will be $3 \times 2 \times (6 - 1) = 30$. The following commands calculates the power of the test for a main effect of A when $\alpha = .05$:

```
library(pwr)
pwr.f2.test(u=3-1,v=30,f2=(0.1^2),sig.level=.05)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 30
##              f2 = 0.01
##       sig.level = 0.05
##           power = 0.07319
```

And here is the power of the test for a main effect of B:

```
pwr.f2.test(u=2-1,v=30,f2=(0.25^2),sig.level=.05)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 30
##              f2 = 0.0625
##       sig.level = 0.05
##           power = 0.2777
```

Note that `u` and `v` correspond to the degrees of freedom in the numerator and denominator, respectively, and `f2` is effect size *squared* (i.e., Cohen's f^2). The results of `pwr.f2.test` indicated that power is quite low for the tests of both main effects. `pwr.f2.test` also can be used to estimate the number of subjects that we would need to attain a power that was at least 0.8:

```
pwr.f2.test(u=3-1,f2=(0.1^2),sig.level=.05,power=.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 963.5
##              f2 = 0.01
##       sig.level = 0.05
##           power = 0.8
```

```
pwr.f2.test(u=2-1,f2=(0.25^2),sig.level=.05,power=.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 125.5
##              f2 = 0.0625
##       sig.level = 0.05
##           power = 0.8
```

To attain a power of 0.8 for our main effect of A, we would need a 963 degrees of freedom in the denominator of our F test, which means we would need $1 + [963/(3 \times 2)] = 162$ subjects per cell in our 3×2 design. To attain the same power for B, we would need 22 subjects per cell. Of course, we cannot satisfy both of these requirements simultaneously, so we would need approximately 162 subjects per cell to attain a power of *at least* 0.8 for both tests of main effects:

```
pwr.f2.test(u=3-1,v=(3*2*(162-1)),f2=(0.1^2),sig.level=.05)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 966
##              f2 = 0.01
##       sig.level = 0.05
##           power = 0.8011
```

```
pwr.f2.test(u=2-1,v=(3*2*(162-1)),f2=(0.25^2),sig.level=.05)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 966
##              f2 = 0.0625
##       sig.level = 0.05
##           power = 1
```

Obviously, it would be very difficult, if not impossible, to conduct such a study. Hopefully this example will highlight why you should think about the power of your experiments prior to doing them.

We'll conclude calculating power for a 3×4 factorial design (i.e. 3 levels on factor A and 4 levels on factor B). If the effect size for A is $f = 0.25$, how many subjects do we need to attain a power of 0.8? The function `pwr.f2.test` can be used to calculate the degrees of freedom that we need in the denominator of our F test:

```
pwr.f2.test(u=3-1,f2=(0.25^2),sig.level=.05,power=.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 154.2
##              f2 = 0.0625
##       sig.level = 0.05
##           power = 0.8
```

The denominator degrees of freedom is given by the formula

$$df_{residuals} = a \times b \times (n - 1)$$

so we would need $n = 14$ subjects per cell to achieve $df_{residuals} = 154$ and a power of 0.8.

```
pwr.f2.test(u=3-1,v=3*4*(14-1),f2=(0.25^2),sig.level=.05)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 156
##              f2 = 0.0625
##       sig.level = 0.05
##           power = 0.8049
```

	Bio.drugX	Bio.drugY	Bio.drugZ	None.drugX	None.drugY	None.drugZ
1	170	186	180	173	189	202
2	175	194	187	194	194	228
3	165	201	199	217	217	190
4	180	215	170	206	206	206
5	160	219	204	199	199	224

Table 2: Data from biofeedback experiment.

7.10 linear contrasts

A significant main effect implies that the marginal means are not all equal, but it does not tell us how the means differ. To get a clearer picture of the differences among means, we often will do one or more linear contrasts. In this section I will illustrate the calculation and evaluation of linear contrasts using the data from Table 7.5 in Maxwell and Delaney (2004, see Table 2). The data come from a 3(drug) x 2(biofeedback) factorial experiment. Five subjects were randomly assigned to each of the six experimental conditions. The results of a 2 (biofeedback) $\times$ 3 (drug) ANOVA are presented in Table 3. The main effects of drug and biofeedback were significant, as was the drug $\times$ biofeedback interaction.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
drug	2.00	1882.07	941.03	5.21	0.01
biofeedback	1.00	1904.03	1904.03	10.54	0.00
drug:biofeedback	2.00	1248.07	624.03	3.46	0.05
Residuals	24.00	4334.80	180.62		

Table 3: ANOVA for biofeedback experiment.

Imagine that I want to compare drugs x and y to each other. If the levels of `drug` are x, y, and z, then I would perform a linear contrast with contrast weights $c1 = [-1,\ 1,\ 0]$. The following example shows how to perform such a contrast using `aov()`:

```
# levels(mw75$drug) # "drugX""drugY""drugZ"
c1 <- c(-1,1,0)
myC <- cbind(c1)
contrasts(mw75$drug) <- myC
mw.aov.02<-aov(score~drug*biofeedback,data=mw75)
summary(mw.aov.02,split=list(drug=list(1)))

##                          Df Sum Sq Mean Sq F value Pr(>F)
## drug                      2   1882     941    5.21 0.0132 *
##   drug: C1                1   1638    1638    9.07 0.0060 **
## biofeedback               1   1904    1904   10.54 0.0034 **
## drug:biofeedback          2   1248     624    3.46 0.0480 *
##   drug:biofeedback: C1    1   1110    1110    6.15 0.0206 *
## Residuals                24   4335     181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The contrast was significant, so we reject the null hypothesis of no difference between the *marginal means* of drugs x and y. Notice that the output also shows how the value of the linear contrast varies with, or depends upon, the level of `biofeedback`: In other words, the output lists the contrast $\times$ biofeedback interaction, which in this case is significant, $F(1, 24) = 6.15$, $p = .02$. Next, let's compute the contrast directly from the group means and information in the omnibus ANOVA table:

```
(drug.means <- with(mw75,tapply(score,drug,mean)) )

## drugX drugY drugZ
## 183.9 202.0 199.0

with(mw75,tapply(score,list(drug,biofeedback),length))

##       Bio None
## drugX   5    5
## drugY   5    5
## drugZ   5    5
```

```
n <- 5
a <- 2
MS.resid <- 181
df.resid <- 24
psi <- sum(c1*drug.means)
( SScontrast <- (n*a)*(psi^2)/ (sum(c1^2)) )

## [1] 1638

(F <- SScontrast/MS.resid)

## [1] 9.05

( p.value <- 1-pf(F,df1=1,df2=df.resid) )

## [1] 0.006084
```

Note that the calculation of $SS_{contrast}$ uses $n \times a = 10$, rather than $n = 5$, because we are considering marginal means, not cell means, and there are 10 subjects who receive each drug. Also note that the denominator degrees of freedom for the p-value equals the residuals degrees of freedom. Finally, let's consider the case where we wanted to compare the different drugs with two orthogonal contrasts:

```
# levels(mw75$drug) # "drugX""drugY""drugZ"
c1 <- c(-1,1,0)
c2 <- c(-1,-1,2)
myC <- cbind(c1,c2)
contrasts(mw75$drug) <- myC
mw.aov.02<-aov(score~drug*biofeedback,data=mw75)
summary(mw.aov.02,split=list(drug=list(1,2)))

##                          Df Sum Sq Mean Sq F value Pr(>F)
## drug                      2   1882     941    5.21 0.0132 *
##   drug: C1                1   1638    1638    9.07 0.0060 **
##   drug: C2                1    244     244    1.35 0.2565
## biofeedback               1   1904    1904   10.54 0.0034 **
## drug:biofeedback          2   1248     624    3.46 0.0480 *
##   drug:biofeedback: C1    1   1110    1110    6.15 0.0206 *
##   drug:biofeedback: C2    1    138     138    0.76 0.3907
## Residuals                24   4335     181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the second contrast, which compares the marginal mean of drug z to the mean of the other two drugs, is not significant (and does not interact with biofeedback).

7.10.1 linear contrasts with `emmeans`

Linear contrasts can be performed easily with the `R` package, `emmeans`. The following block of code shows how to analyze the main effect of `drug` with contrasts `c1` and `c2`, which were defined in the previous section.

```
library(emmeans)
mw.em <- emmeans(mw.aov.02,specs="drug")
con <- contrast(mw.em, method=list(c1,c2))
summary(con,adjust="none")
```

```
##  contrast     estimate    SE df t.ratio p.value
##  c(-1, 1, 0)      18.1  6.01 24   3.011  0.0060
##  c(-1, -1, 2)     12.1 10.41 24   1.162  0.2565
##
## Results are averaged over the levels of: biofeedback
```

The t test results are equivalent to the F tests in the previous section. However, this analysis ignores any potential interaction between `drug` and `biofeedback`. The next block of code performs the contrasts separately at each level of `biofeedback`. The values of each contrast in each biofeedback are listed in the `estimate` column. Each t test evaluates the null hypothesis that the value of `estimate` equals zero. Only contrast `c1` in the biofeedback condition is significant.

```
mw.em2 <- emmeans(mw.aov.02,specs="drug",by="biofeedback")
contrast(mw.em2,method=list(drug=list(c1,c2)), by="biofeedback")
```

```
## biofeedback = Bio:
##  contrast          estimate   SE df t.ratio p.value
##  drug.c(-1, 1, 0)      33.0  8.5 24   3.882  0.0007
##  drug.c(-1, -1, 2)      3.0 14.7 24   0.204  0.8402
##
## biofeedback = None:
##  contrast          estimate   SE df t.ratio p.value
##  drug.c(-1, 1, 0)       3.2  8.5 24   0.376  0.7099
##  drug.c(-1, -1, 2)     21.2 14.7 24   1.440  0.1628
```

The values of the contrasts differ across the biofeedback conditions. For example, the values of contrast `c1` were 33 and 3.2 in the two biofeedback conditions, a difference of 29.8. However, that information is insufficient to determine if the contrasts differ *significantly* across biofeedback conditions. That inference requires us to perform a significance test on the difference itself, an operation that we perform in the next block of code.

```
contrast(mw.em2,
         interaction=list(drug=list(c1,c2),biofeedback=list(c(1,-1)) ),
         by=NULL)
```

```
##  drug_custom  biofeedback_custom estimate   SE df t.ratio p.value
##  c(-1, 1, 0)  c(1, -1)              29.8 12.0 24   2.479  0.0206
##  c(-1, -1, 2) c(1, -1)             -18.2 20.8 24  -0.874  0.3907
```

The first line in the output evaluates the null hypothesis that the difference between the values of contrast `c1` calculated in the two biofeedback conditions, $33 - 3.2 = 29.8$, is zero. According to our analysis, this difference is significant, so the difference between drugs x and y differs significantly between the two biofeedback conditions. The second line in the output evaluates the null hypothesis that the difference between contrast `c2` in the two conditions, $3 - 21.2 = -18.2$, is zero. The t test indicates that the difference is not significant, so we do not reject the hypothesis that contrast `c2` did not differ across biofeedback conditions.

7.10.2 Tukey HSD

In some situations we want to evaluate all pairwise comparisons among marginal means. In such cases, a Tukey HSD task is appropriate. Using the `which` option in the `TukeyHSD` command allows you to specify which set of marginal means should be tested:

```
TukeyHSD(mw.aov.02,which="drug")

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = score ~ drug * biofeedback, data = mw75)
##
## $drug
##              diff        lwr   upr  p adj
## drugY-drugX  18.1   3.09063 33.11 0.0160
## drugZ-drugX  15.1   0.09063 30.11 0.0484
## drugZ-drugY -3.0 -18.00937 12.01 0.8724

TukeyHSD(mw.aov.02,which="biofeedback")

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = score ~ drug * biofeedback, data = mw75)
##
## $biofeedback
##           diff   lwr   upr  p adj
## None-Bio 15.93 5.805 26.06 0.0034
```

7.11 simple main effects

A significant interaction means that the effect of A *depends* on the level of B or, equivalently, that the effect of B depends on the level of A. Hence, a significant $A \times B$ interaction means that tests of main effects may be give a very misleading picture of the effects of A and B. Usually, when the $A \times B$ interaction is significant we should *not* go on to test for main effects. Instead, we should determine if the effect of one variable (e.g., A) is significant within each individual level of the other variable (e.g., B_1, $B_2, \cdots$, B_k). Tests for the significance of one variable within the level of another variable are known as **simple main effects**.

7.11.1 an example

I will give an example of how to compute simple main effects with the data shown in Table 7.11 in your textbook, which is saved as a data file on the CD that came with your textbook. That file does not contain variable names, so R provides default names V1, V2, and V3. I'm going to read the data file, change the variable names to something more useful, and then convert them to factors. If you do not have the textbook's CD, you can get download the data file from the web:

```
# mw11 <- read.table(file = "chapter_7_table_11.dat")
file_path <- "http://pnb.mcmaster.ca/bennett/psy710/datasets/mw/chapter_7_table_11.dat"
mw11 <- read.table(file=url(file_path ))
names(mw11)

## [1] "V1" "V2" "V3"

names(mw11) <- c("group", "task", "score")
names(mw11)

## [1] "group" "task"  "score"

sapply(mw11,class)

##     group      task     score
## "integer" "integer" "integer"
```

The `group` and `task` variables are `integer` variables because the different levels were defined with numbers rather than letters. I want my ANOVA to treat group and task as `factors`, so I will create two new variables `gf` and `tf`:

```
mw11$gf <- factor(mw11$group,
                  labels = c("amnesic", "huntingtons", "control"),
                  ordered = FALSE)
mw11$tf <- factor(mw11$task,
                  labels = c("grammar", "classification", "recognition"),
                  ordered = FALSE)
names(mw11)

## [1] "group" "task"  "score" "gf"    "tf"

sapply(mw11,class)

##     group      task     score        gf        tf
## "integer" "integer" "integer"  "factor"  "factor"
```

I use the `lm()` and `anova()` commands to do the analysis and print the summary table. In this factorial experiment, the linear model is defined with the formula

$$score \sim 1 + gf + tf + gf : tf$$

where `1` represents the intercept, `gf` and `tf` represent the group and task main effects, and `gf:tf` is the group x task interaction. The formula can be interpreted as saying that `score` *is modeled as* the sum of an intercept term, a group main effect, a task main effect, and a group x task interaction. Here is the ANOVA:

```
mw11.lm.01 <- lm(score ~ 1+ gf + tf + gf:tf, data = mw11)
anova(mw11.lm.01)
```

```
## Analysis of Variance Table
##
## Response: score
##           Df Sum Sq Mean Sq F value  Pr(>F)
## gf         2   5250    2625   16.64 7.6e-06 ***
## tf         2   5250    2625   16.64 7.6e-06 ***
## gf:tf      4   5000    1250    7.92 0.00011 ***
## Residuals 36   5680     158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MS.w <- 157.8
df.w <- 36
```

The interaction between group and task is significant. I save the $MS_{Resdiuals}$, or MS_{Within}, and the degrees of freedom because we will need them later. The command `interaction.plot` can be used to get a better sense of what the interaction means. The following code chunk creates Figure 2.

```
interaction.plot(x.factor=mw11$gf,
                 trace.factor=mw11$tf,
                 response=mw11$score,
                 trace.label="task")
```

It looks as though the effect of task may be significant for the group of Huntington's patients, but not the other groups. I will evaluate this idea by testing the effect of task at each level of group. (N.B. I am not saying that this test is the most interesting one. I am simply going to use it to illustrate how to calculate simple main effects).

The key part of the analysis is to extract the data from each level of group using R's `subset` command. Here is an example of how to use the command:

```
# subset(mw11, gf == "control")
# subset(mw11, tf == "classification")
subset(mw11, tf == "classification" & gf == "control")
```

```
##    group task score      gf             tf
## 36     3    2    92 control classification
## 37     3    2    65 control classification
## 38     3    2    86 control classification
## 39     3    2    67 control classification
## 40     3    2    90 control classification
```

Next we construct the appropriate linear models and display the anova tables for the effect of task for each group:

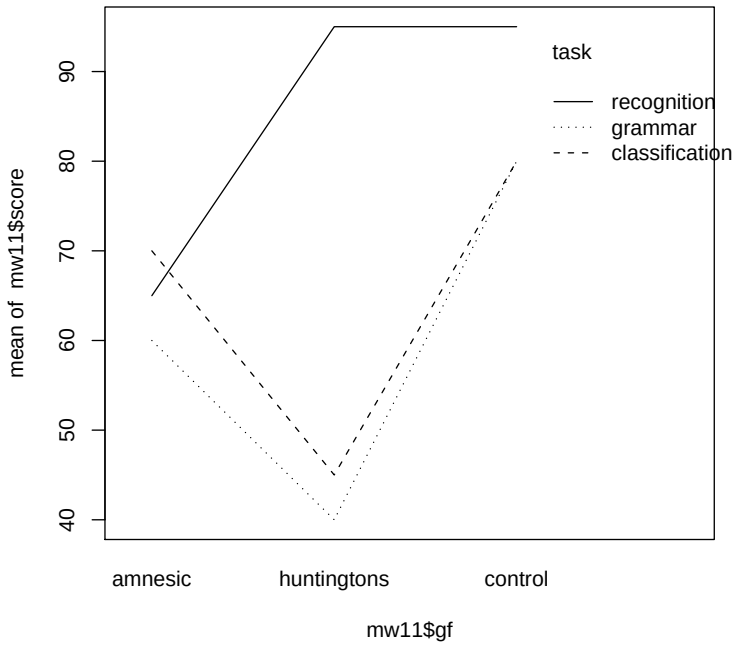

Figure 2: Interaction plot.

```
lm.task.hunt <- lm(score ~ 1 + tf, data = subset(mw11, gf == "huntingtons"))
lm.task.amnesic <- lm(score ~ 1 + tf, data = subset(mw11, gf == "amnesic"))
lm.task.control <- lm(score ~ 1 + tf, data = subset(mw11, gf == "control"))
```

Finally, we print the anova table for each model, extract SS_{task}, and then compute F and p values using MS_{Within} and df_{Within} from our original analysis. Next I evaluate the simple main effect of task for the Huntington's group. Notice how I recalculate F:

```
anova(lm.task.hunt)

## Analysis of Variance Table
##
## Response: score
##           Df Sum Sq Mean Sq F value  Pr(>F)
## tf         2   9250    4625    29.4 2.4e-05 ***
## Residuals 12   1890     157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(F.task.hunt <- 4625/MS.w)

## [1] 29.31

(p.task.hunt <- 1 - pf(F.task.hunt, df1 = 2, df2 = df.w))

## [1] 2.792e-08
```

The simple main effect is significant. We could now proceed to do contrasts or pairwise comparisons among the three tasks to see which ones differ. Below, I compute pairwise differences using Tukey's HSD method.

You might think that performing a Tukey HSD test on a simple main effect is done by passing the `aov` object to `TukeyHSD`:

```
aov.task.hunt <- aov(score ~ tf, data = subset(mw11, gf == "huntingtons"))
TukeyHSD(aov.task.hunt)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = score ~ tf, data = subset(mw11, gf == "huntingtons"))
##
## $tf
##                              diff    lwr   upr  p adj
## classification-grammar          5 -16.18 26.18 0.8068
## recognition-grammar            55  33.82 76.18 0.0000
## recognition-classification     50  28.82 71.18 0.0001
```

However, this test is not optimal because it is based on an error term calculated only from a portion of the data rather than all of the data. The following code shows how to calculate F from the means of the three tasks in the `huntingtons` groups:

```
with(subset(mw11,gf=="huntingtons"),tapply(score,tf,length))
```

```
##        grammar classification    recognition
##              5              5              5
```

```
n <- 5
b <- 3
(tf.means.hunt<-with(subset(mw11,gf=="huntingtons"),tapply(score,tf,mean)) )
```

```
##        grammar classification    recognition
##             40             45             95
```

```
pw1<-c(-1,1,0)
pw2<-c(-1,0,1)
pw3<-c(0,-1,1)
psi.pw1 <- sum(tf.means.hunt * pw1)
psi.pw2 <- sum(tf.means.hunt * pw2)
psi.pw3 <- sum(tf.means.hunt * pw3)
SS.pw1 <- n*b*(psi.pw1^2) / sum(pw1^2)
SS.pw2 <- n*b*(psi.pw2^2) / sum(pw2^2)
SS.pw3 <- n*b*(psi.pw3^2) / sum(pw3^2)
MS.w
```

```
## [1] 157.8
```

```
(F.pw1 <- SS.pw1/MS.w)

## [1] 1.188

(F.pw2 <- SS.pw2/MS.w)

## [1] 143.8

(F.pw3 <- SS.pw3/MS.w)

## [1] 118.8
```

Now that we have our F scores, we need to compare them to a critical F calculated using formula that is appropriate for the HSD test. In R, you can calculate this critical F using the `qtukey` command:

```
alpha.fw <- .05
df.w

## [1] 36

(F.tukey.critical <- ( qtukey(1-alpha.fw,nmeans=b,df=df.w)^2) / 2 )

## [1] 5.975
```

Note that $\alpha_{FW} = .05$, that the number of means equals the number of different tasks, and that the degrees of freedom equals the df for the error term that was used to calculate the F's. Comparing the observed F's to the critical F indicates that performance in the recognition task differed from performance in the other two tasks, but that performance in the grammar and classification tasks did not differ from each other. Here is the simple main effect of task for the amnesic group:

```
anova(lm.task.amnesic)

## Analysis of Variance Table
##
## Response: score
##           Df Sum Sq Mean Sq F value Pr(>F)
## tf         2    250     125    0.79   0.48
## Residuals 12   1896     158

(F.task.amnesic <- 125/MS.w)

## [1] 0.7921

(p.task.amnesic <- 1 - pf(F.task.amnesic, df1 = 2, df2 = df.w))

## [1] 0.4606
```

The simple main effect of task among amnesic subjects is not significant. Finally, we evaluate the simple main effect of task for the control group. The simple main effect of task among control subjects is not significant:

```
anova(lm.task.control)
```

```
## Analysis of Variance Table
##
## Response: score
##           Df Sum Sq Mean Sq F value Pr(>F)
## tf         2    750     375    2.38   0.14
## Residuals 12   1894     158
```

```
(F.task.control <- 375/MS.w)
```

```
## [1] 2.376
```

```
(p.task.control <- 1 - pf(F.task.control, df1 = 2, df2 = df.w))
```

```
## [1] 0.1073
```

In summary, our analysis suggests that task has a significant effect on performance among Huntington's patients but not among subjects in the other groups. Moreover, the effect of task in Huntington's patients appears to reflect the fact that recognition scores are significantly higher than scores on the grammar and classification tasks .

7.11.2 simple main effects using `emmeans`

The analyses performed in the previous sections can be done simply and efficiently with commands in the `R` package `emmeans`. The following code illustrates how to calculate the simple main effect of `task` in each group:

```
library(emmeans)
mw11.em <- emmeans(mw11.lm.01,specs=~tf|gf) # note the formula!
joint_tests(mw11.em,by="gf") # simple main effect of task for each group
```

```
## gf = amnesic:
##  model term df1 df2 F.ratio p.value
##  tf           2  36   0.792  0.4606
##
## gf = huntingtons:
##  model term df1 df2 F.ratio p.value
##  tf           2  36  29.313  <.0001
##
## gf = control:
##  model term df1 df2 F.ratio p.value
##  tf           2  36   2.377  0.1073
```

The `joint_tests` command shown above evaluates the null hypothesis that the two α's corresponding to the effects of task (at each level of `gf`) are zero. The simple main effect of task is significant only in the `huntingtons` group. Next, we can evaluate differences among tasks within each group using Tukey's HSD.

```
library(emmeans)
pairs(mw11.em) # tukey adjustment used by default
```

```
## gf = amnesic:
##  contrast                    estimate   SE df t.ratio p.value
##  grammar - classification         -10 7.94 36  -1.259  0.4273
##  grammar - recognition             -5 7.94 36  -0.629  0.8050
##  classification - recognition       5 7.94 36   0.629  0.8050
##
## gf = huntingtons:
##  contrast                    estimate   SE df t.ratio p.value
##  grammar - classification          -5 7.94 36  -0.629  0.8050
##  grammar - recognition            -55 7.94 36  -6.923  <.0001
##  classification - recognition     -50 7.94 36  -6.294  <.0001
##
## gf = control:
##  contrast                    estimate   SE df t.ratio p.value
##  grammar - classification           0 7.94 36   0.000  1.0000
##  grammar - recognition            -15 7.94 36  -1.888  0.1567
##  classification - recognition     -15 7.94 36  -1.888  0.1567
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

In the `huntingtons` group, the `recognition` task differs significantly from the `grammar` and `classification` task. Notice that each Tukey test has 36 degrees of freedom, which is the same as the residuals degrees of freedom in our original ANOVA. Also, the value of `SE` equals the value you get when the error term is based on $MS_{residuals}$ in the original ANOVA. Thus, the comparisons calculated by `emmeans` are using an error term that is calculated from all of the data rather than subsets of the data.

7.11.3 Type I Error Rates

Analyzing the simple main effect of task required us to do three ANOVAs (one for each level of group). Obviously, performing multiple tests of simple main effects will increase the familywise Type I error rate. There is some disagreement on the issue of whether you should use the Bonferroni method to control the Type I error rate. One point of view is that you should not do so because you would examine simple main effects only if the omnibus interaction was significant. Another point of view is that the familywise Type I error rate should be kept at the level used in the omnibus ANOVA. In other words, if $\alpha = .05$ was used to evaluate significant interaction and main effects, then you should set $\alpha_{FW} = .05$ when you are analyzing simple main effects. In the example in the preceding section, we would set the per-comparison alpha to $\alpha_{PC} = .05/3 = .0167$.

7.12 Unbalanced Data

So far our analyses of factorial experiments have assumed that the designs are **balanced**: they have equal n in each combination of the experimental factors. There are occasions, however, when our

data are **unbalanced**. In the following sections we describe how unbalanced data complicate the analysis of variance, and suggest ways of addressing these complications.

In the following sections, I assume that the full model for a two factor experiment is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{43}$$

and that

$$\Sigma_i \alpha_i = \Sigma_i \beta_j = \Sigma_i (\alpha\beta)_{ij} = 0 \tag{44}$$

The sum-to-zero constraints in Equation 44 are especially important for interpreting Type III sums of squares (Section 7.12.5).

7.12.1 An example

The following example is taken directly from Chapter 13 in Howell (2002).

An experiment was conducted to study the effect of alcohol on vigilance. Subjects were placed in a driving simulator and told to avoid cars when they emerged suddenly from driveways, and pedestrians who stepped suddenly into the street. Subjects were assigned randomly to two groups: the no-alcohol group drank a non-alcoholic beverage before the driving test, whereas the alcohol group drank an alcoholic beverage. The dependent measure was the number of errors each driver made in one half-hour session. The experiment was conducted in two labs at universities in Michigan and Arizona. The data are presented in Table 4. There are unequal numbers of subjects in each cell of Table 4, so the data are unbalanced. Inspection of the cell means indicates that the results were what you would expect: subjects in the alcoholic group made more errors than subjects in the no-alcohol group. Furthermore, the cell means indicate that the pattern of results was, as expected, the same for subjects tested in Michigan and Arizona.

Table 4: Data from drinking study.

	no alcohol	alcohol	Row Means
Michigan	13 15 14 16 12 $\bar{Y}_{11} =$ 14	18 20 22 19 21 23 17 18 22 20 $\bar{Y}_{12} = 20$	$\bar{Y}_{1\cdot} = 18$
Arizona	13 15 18 14 10 12 16 17 15 10 14 $\bar{Y}_{21} = 14$	24 25 17 16 18 $\bar{Y}_{22} =$ 20	$\bar{Y}_{2\cdot} = 15.9$
Column Means	$\bar{Y}_{\cdot 1} = 14$	$\bar{Y}_{\cdot 2} = 20$	

So, what's the problem? Well, if you look at the row means, it appears that subjects from Michigan made more errors than subjects from Arizona: Despite the fact that the cell means in the two rows are exactly the same, the marginal row means differ. Why? Because 10 out of the 15 subjects from Michigan were in the alcohol condition, whereas 11 out of the 16 subjects from Arizona were in the no-alcohol condition. In other words, the effect of `alcohol` is masquerading as an effect of `state`. This mixing up of row and column effects is a general property of unbalanced factorial designs: the row, column, and interaction effects are no longer orthogonal to each other.

The fact that the row, column and interaction effects are no longer orthogonal greatly complicates the analysis of variance. To see why, consider the following two linear models:

$$score \sim 1 + alcohol + state + alcohol : state \quad (45)$$

$$score \sim 1 + state + alcohol + state : alcohol \quad (46)$$

The ANOVA tables produced by R for Models 45 and 46 are presented in Table 5. Although the models differ only in the order of terms, the sums of squares assigned by the models to the main effects differ significantly.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
state	1	34.96	34.96	5.13	0.0318
alcohol	1	243.75	243.75	35.77	0.0000
state:alcohol	1	0.00	0.00	0.00	1.0000
Residuals	27	184.00	6.81		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
alcohol	1	278.71	278.71	40.90	0.0000
state	1	0.00	0.00	0.00	1.0000
alcohol:state	1	0.00	0.00	0.00	1.0000
Residuals	27	184.00	6.81		

Table 5: ANOVA tables for Model 1 (top) and Model 2 (bottom).

7.12.2 Proportional Cell Frequencies

Before I continue to discuss the problems associated with analyzing unbalanced data, I want to describe a case where unbalanced data are *not* hard to analyze. Suppose we had 36 subjects from Michigan, with 24 in the alcohol condition and 12 in the no-alcohol condition. Also, let's suppose that there were 24 subjects from Arizona, with 16 in the alcohol condition and 8 in the no-alcohol condition. In this case, the ratio of subjects in the alcohol and no-alcohol conditions is 2:1 at both levels of the other variable, `state`. If the proportion of cases in one variable is the same at all levels of the other variables, then the design exhibits **proportionality** and can be analyzed as though the data were balanced. So everything in the following conditions applies to situations where the data are unbalanced and cell sizes are not proportional.

7.12.3 Type I (Sequential) Sums of Squares

R prints so-called Type I, or sequential, sums of squares, which are computed by comparing a series of nested models. The sums of squares in the top of Table 5 are derived by comparing the following four nested models:

$$1.1)\ score \sim 1$$

$$1.2)\ score \sim 1 + state$$

$$1.3)\ score \sim 1 + state + alcohol$$

$$1.4)\ score \sim 1 + state + alcohol + state : alcohol$$

SS_{state} is derived by comparing models 1.1 and 1.2, $SS_{alcohol}$ is derived by comparing models 1.2 and 1.3, and $SS_{state:alcohol}$ is derived by comparing models 1.3 and 1.4. In the first step of this sequential comparison process, SS_{state} represents *all* of the variation in the dependent variable that is associated with `state`. In the second step, $SS_{alcohol}$ represents the variation in the dependent variable that is associated with `alcohol` *after* removing the variation assigned to `state`. The third step assigns all of the variation in the dependent variable that is associated with the `state` X `alcohol` interaction after removing variation assigned to the two main effects.

The sums of squares in the bottom of Table 5 are derived by comparing the following four nested models:

$$\begin{aligned}
&2.1)\ score &&\sim 1\\
&2.2)\ score &&\sim 1 + alcohol\\
&2.3)\ score &&\sim 1 + alcohol + state\\
&2.4)\ score &&\sim 1 + alcohol + state + alcohol : state
\end{aligned}$$

$SS_{alcohol}$ is derived by comparing models 2.1 and 2.2, SS_{state} is derived by comparing models 2.2 and 2.3, and $SS_{alcohol:score}$ is derived by comparing models 2.3 and 2.4. In this case, $SS_{alcohol}$ represents *all* of the variation in the dependent variable that is associated with `alcohol`, but SS_{state} represents the variation that is associated with `state` *after* removing the variation assigned to `alcohol`. The last step is the same as before: $SS_{alcohol:state}$ represents the variation in the dependent variable that is associated with the `state` X `alcohol` interaction after removing the variation that assigned to the two main effects.

In Section 7.12, I noted that the fact that the design is unbalanced means that the main effects of `state` and `alcohol` are not orthogonal. The lack of orthogonality means that some of the variation in the dependent variable that is associated with `state` is also associated with `alcohol`; and some of the variation in the dependent variable that is associated with `alcohol` is also associated with `state`. This state of affairs is often summarized by saying that `state` and `alcohol` account for "overlapping" portions of the variation of the dependent variable. In such cases, the variation assigned to each variable depends on the order in which the assignment takes place. When we compare models 1.2 and 1.1, some variation in the dependent variable is assigned to SS_{state}. However, our example was constructed in such a way that the "effect" of `state` actually was due entirely to the effect of alcohol. So, after comparing models 2.2 and 2.1, and assigning alcohol-associated variation in the dependent variable to $SS_{alocohol}$, there is nothing left to assign to SS_{state}.

One last comment. This discussion should make it clear that each main effect has *two* values for Type I sum of squares: one is calculated by comparing models that ignore the other effects (e.g., comparing models 1.1 and 1.2, or 2.1 and 2.2) and another that is calculated by comparing models that control for the other main effect (e.g., comparing 1.2 and 1.3, or 2.2 and 2.3). *Both* values are Type I sums of squares, so you need to be clear which value you are referring to when you discuss the "Type I sum of squares for factor A". In your textbook, Type I sums of squares for A are calculated while ignoring both the other main effect, B, and the $A \times B$ interaction.

7.12.4 Type II Sums of Squares

Type I sums of squares are derived from sequential comparisons of nested models that can result in an asymmetry between the sums of squares for the two main effects. In the top of Table 5, for example, SS_{state} was calculated while ignoring the effects of `alcohol`, whereas $SS_{alcohol}$ was calculated after controlling for the effects of `state`. In the bottom of Table 5, on the other hand $SS_{alcohol}$ was calculated while ignoring the effects of `state`, whereas SS_{state} was calculated after controlling for

the effects of `alcohol`. An alternative would be to calculate the sum of squares for each main effect after controlling for the other main effect. For example, we could estimate $SS_{alcohol}$ by comparing the residuals in the following nested models

$$\begin{aligned} score &\sim 1 + state \\ score &\sim 1 + state + alcohol \end{aligned}$$

and estimate SS_{state} by comparing these nested models

$$\begin{aligned} score &\sim 1 + alcohol \\ score &\sim 1 + alcohol + state \end{aligned}$$

These are Type II sums of squares: They are the sum of squares calculated after controlling for the other main effect but ignoring the interaction.

More generally, the Type II sum of squares for an effect is calculated by comparing nested models that lack all interactions that include that effect. For example, suppose an experiment contained factors A, B, and C. The Type II sum of squares for C is obtained by comparing the nested models

$$\begin{aligned} score &\sim 1 + A + B + A:B \\ score &\sim 1 + A + B + A:B + C \end{aligned}$$

Note how the models do not contain any interactions that include C (i.e., A:C, B:C, and A:B:C). According to this definition, the Type I and Type II sum of squares for the State X Alcohol interaction are equivalent.

The second line in each part of Table 5 lists the sum of squares for one main effect after controlling for the other main effect (but ignoring the interaction). These are the Type II sums of squares for `state` and `alcohol`. The Type II sum of squares for the interaction is equivalent to the Type I sum of squares for the interaction, and is listed in the third line of both ANOVA tables.

7.12.5 Type III Sums of Squares

Finally, we consider the sums of squares for an effect that are calculated after accounting for all other effects in the model. SS_{state} is calculated by comparing

$$\begin{aligned} score &\sim 1 + alcohol + alcohol:state \\ score &\sim 1 + alcohol + alcohol:state + state \end{aligned}$$

$SS_{alcohol}$ is calculated by comparing

$$\begin{aligned} score &\sim 1 + state + alcohol:state \\ score &\sim 1 + state + alcohol:state + alcohol \end{aligned}$$

and $SS_{state:alcohol}$ is calculated by comparing

$$\begin{aligned} score &\sim 1 + state + alcohol \\ score &\sim 1 + state + alcohol + alcohol:state \end{aligned}$$

These are Type III sums of squares. They are sums of squares that are obtained by comparing nested models that control for all other effects. One advantage of Type III sums of squares is that they are *not* dependent on the order in which the factors are specified in the linear model.

Note that the Type III sum of squares for the interaction is the same as the Type I and Type II sums of squares. In general, the Type I, II and III sums of squares for the highest-order interaction in the experimental design will be the same. In a 2-factorial design, the Type II sum of squares for each main effect control for the other main effect but ignore the interaction. Type III sum of squares control for the other main effect *and* the interaction. Therefore, Type II and III sums of squares for the main effects will be the same when $SS_{interaction}$ is zero.

7.12.6 principle of marginality

One might think that Type III sums of squares could be computed in R by comparing the two `aov` models:

```
aov.full <- aov(y ~ 1 + state + alcohol + state:alcohol, data=myData)
aov.reduced <- aov(y ~ 1 + state + state:alcohol, data=myData)
anova(aov.reduced, aov.full)
```

Although this sequence of commands looks correct, it will unfortunately not produce the Type III sums of squares for `alcohol`. The reason for this failure is that the formula in `aov.reduced` violates the **principle of marginality**, which holds that a model that contains an A:B interaction must also contain main effects of A and B. Some (most?) statisticians argue that, in general, such models should not be used. Here is one reason why: Suppose we construct a linear model that relates a dependent variable to a covariate

$$y \sim 1 + \alpha x + \beta x^2$$

and we find that α does not differ significantly from zero. We might be tempted to simplify our model by dropping x:

$$y \sim 1 + \beta x^2$$

This model also violates the principle of marginality because it contains the higher-order term, x^2, but not the lower-order term x. What's the problem? Well, suppose we change the units of x by adding and/or multiplying it by a constant: $x' = ax + b$. This type of linear transformation of x should not change the results of our statistical analyses, right? Well, if the model obeys the principle of marginality, then a linear transformation of x does not change the results in any significant way: the sum-of-squares associated with x and x^2 remain the same. However, when the model violates the principle of marginality, then a linear transformation of x *does* change the sum-of-squares associated with x^2. And we are not just talking about tiny changes: the effects of linearly transforming x can be huge.

Consider another example. We have used the sum-to-zero constraint (Equation 44) when defining the effects in the linear model in Equation 43. It is possible, however, to define effects in other ways. For example, we could designate one condition in our experimental design as the *baseline* in which all of the effects are zero, and the effects for all other conditions would be defined relative to this baseline. Using this definition, it is possible that all of the effects are greater than zero, so the sum-to-zero constraint does not apply. Nevertheless, this definition – which is known as a `treatment` definition of an effect, and actually is the default definition used by R – is perfectly reasonable. The results of our ANOVA should not depend on which definition we use: we should get the same results if we use the sum-to-zero constraint or define the effects relative to a baseline condition. And we do get the same results – the sums of squares assigned to the various factors in our model remain the same, *so long as the model obeys the principle of marginality*. When the model violates the principle of marginality, then the results obtained by an ANOVA will depend on the (seemingly arbitrary) decision of how we define the effects in our model.

All of this is to say that R tries hard to enforce the principle of marginality. Therefore, the code listed at the start of this section, which includes a model that violates the principle of marginality, will not yield the Type III sum of squares for `alcohol`. The reason for the failure is that R surreptitiously introduces the `alcohol` term into the reduced model, and so the difference in $SS_{residual}$ in the two models will be zero. Think of it this way: R it trying to save you from yourself. Nevertheless, Type III sums of squares can be calculated easily in R if you know the secret handshake.

7.12.7 calculating Type III sums of squares in R

The command `drop1(aovModel,. ~ .,test="F")` calculates the change in $SS_{residuals}$ that is caused by removing each term in `aovModel`, one at a time, and evaluates each change with an F test. (The second part of the command, . ~ ., tells R that it should drop *all* terms in the model, one at a time.) Notice that this procedure is exactly how Type III sums of squares are computed, so `drop1` calculates Type III sums of squares.

The following code shows how to compute the Type III sum of squares for our example:

```
howell<-read.csv("howell.csv")
aov.01<-aov(score~1+state+alcohol+state:alcohol,data= howell)
drop1(aov.01,.~.,test="F")


## Single term deletions
##
## Model:
## score ~ 1 + state + alcohol + state:alcohol
##               Df Sum of Sq RSS  AIC F value  Pr(>F)
## <none>                     184 63.2
## state          1         0 184 61.2     0.0       1
## alcohol        1       244 428 87.4    35.8 2.2e-06 ***
## state:alcohol  1         0 184 61.2     0.0       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that the results are not dependent on the order of the terms in the model (to within rounding error):

```
aov.02<-aov(score~1+alcohol+state+state:alcohol,data=howell)
drop1(aov.02,.~.,test="F")


## Single term deletions
##
## Model:
## score ~ 1 + alcohol + state + state:alcohol
##               Df Sum of Sq RSS  AIC F value  Pr(>F)
## <none>                     184 63.2
## alcohol        1       244 428 87.4    35.8 2.2e-06 ***
## state          1         0 184 61.2     0.0       1
## alcohol:state  1         0 184 61.2     0.0       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.12.8 linking everything to hypotheses about means

The previous sections described Type I, II, and III sums of squares in terms of comparisons among nested models. Normally, however, we express experimental hypotheses in terms of group means: we want our analyses of main effects, for example, to evaluate null hypotheses of no difference among row and column means. What hypotheses about means are being evaluated when we compute Type I, II, and III sums of squares?

7.12.9 the interaction term

In the case of the `A:B` interaction, the answer is simple: Type I, II, and III sums of squares are all equal and all test the same hypothesis, namely that every cell mean is simply the sum of the intercept and two main effects:

$$\bar{Y}_{jk} = \mu + \alpha_j + \beta_k \tag{47}$$

or, in other words, that $(\alpha\beta)_{jk} = 0$ for all j and k.

7.12.10 Type I: weighted means

Type I sums of squares evaluate null hypotheses about *weighted* row and column marginal means. A weighted mean takes into account the different n in each group. So, for an $a_{row} \times b_{column}$ design, the weighted row and column marginal means are

$$\bar{Y}_{r.(W)} = \frac{\Sigma_{k=1}^{b} n_k \bar{Y}_{rk}}{(n_1 + n_2 \cdots + n_b)} \tag{48}$$

$$\bar{Y}_{.c(W)} = \frac{\Sigma_{j=1}^{a} n_j \bar{Y}_{jc}}{(n_1 + n_2 \cdots + n_a)} \tag{49}$$

Type I sums of squares evaluate the null hypotheses that the weighted marginal row and column means are equal:

$$\bar{Y}_{1.(W)} = \bar{Y}_{2.(W)} = \cdots = \bar{Y}_{a.(W)} \tag{50}$$

$$\bar{Y}_{.1(W)} = \bar{Y}_{.2(W)} = \cdots = \bar{Y}_{.b(W)} \tag{51}$$

Are these hypotheses interesting? Usually, no, because differences in weighted means depend on the cell n's. However, in some cases it is important to incorporate information about cell n into our analyses (see Howell and McConaughy (1982) for an example), and in those cases testing hypotheses about weighted marginal means is appropriate.

7.12.11 Type II: crazy!

Type II sums of squares can be used to test the following crazy null hypothesis:

$$\Sigma_{k=1}^{b} \left(n_{jk} - (n_{jk}^2 / n_{.k}) \right) \mu_{jk} = \Sigma_{j \neq j'} \Sigma \left(n_{jk} n_{j'k} / n_{.k} \right) \mu_{j'k} \tag{52}$$

where $j = 1,\ 2,\ 3, \cdots, (a-1)$. Nobody knows what this hypothesis means. So, Type II sums of squares generally do not test anything that you or I can understand. An exception to this general rule is when when $SS_{interaction} \approx 0$. I expand on this point in see Section 7.12.12.

7.12.12 Type III: unweighted means

Type III sums of squares evaluate null hypotheses about unweighted marginal means. In an $a_{row} \times b_{column}$ design, the unweighted mean for row r, $\bar{Y}_{r.(U)}$, is simply the mean of the b cell means in row r, and the unweighted mean for column c, $\bar{Y}_{.c(U)}$, is the mean of the a cell means in column c:

$$\bar{Y}_{r.(U)} = \frac{\Sigma_{j=1}^{b} \bar{Y}_{rj}}{b} \tag{53}$$

$$\bar{Y}_{.c(U)} = \frac{\Sigma_{i=1}^{a} \bar{Y}_{ic}}{a} \tag{54}$$

Unweighted marginal means do not take into account the differences in n within each cell.

Provided that the effects in the linear model satisfy the sum-to-zero constraint (as in Equation 44), Type III sums of squares can be used to evaluate the null hypotheses that the unweighted marginal row and column means are equal:

$$\bar{Y}_{1.(U)} = \bar{Y}_{2.(U)} = \cdots = \bar{Y}_{a.(U)} \tag{55}$$

$$\bar{Y}_{.1(U)} = \bar{Y}_{.2(U)} = \cdots = \bar{Y}_{.b(U)} \tag{56}$$

I will say it again: Type III sums of squares can be used to test hypotheses about unweighted marginal means *provided that the effects satisfy the sum-to-zero constraint*[2]. To illustrate this point, I am going to use `drop1` to calculate $SS_{alcohol}$ when we define an effect relative to a baseline condition:

```
options(contrasts=c("contr.treatment","contr.poly")) # treatment/baseline definition
aov.03 <- aov(score~alcohol+state+alcohol:state,data=howell)
drop1(aov.03,.~.,test="F")
```

```
## Single term deletions
##
## Model:
## score ~ alcohol + state + alcohol:state
##               Df Sum of Sq RSS  AIC F value  Pr(>F)
## <none>                     184 63.2
## alcohol        1       124 308 77.2    18.2 0.00022 ***
## state          1         0 184 61.2     0.0 1.00000
## alcohol:state  1         0 184 61.2     0.0 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
options(contrasts=c("contr.sum","contr.poly")) # reset
```

Now $SS_{alcohol} = 123.75$, which is considerably less than the value of 184 calculated in Section 7.12.7.

[2]Actually, the requirement is that the contrasts used to define the effects are orthogonal in the row-basis of the model matrix. This means that several definitions of effects can be used to test hypotheses about marginal means using Type III sums of squares, but the only one considered here is the sum-to-zero definition.

7.12.13 When $SS_{interaction} \approx 0$

Type II and III sums of squares are equal when $SS_{interaction} \approx 0$. In that situation, Type II sums of squares can be used to evaluate null hypotheses about marginal unweighted means. Moreover, in this case Type II sums of squares will provide a more sensitive test of the null hypothesis because the degrees of freedom associated with the interaction effects will be incorporated into $df_{Residuals}$. Therefore, some researchers recommend that you first evaluate the interaction term and if it is not significant – with $p > 0.20$ or $p > 0.25$ – proceed to use Type II sums of squares to evaluate hypotheses about marginal unweighted means. Others recommend against this procedure, arguing instead that the interaction term should remain in the model even if it is not significant, and that you should use Type III tests to evaluate the main effects (Faraway, 2005).

7.12.14 A Numerical Example

To illustrate the ideas described in the previous section, I will analyze the data shown in Table 7.23 (page 339) in your textbook. The data are from an experiment that evaluates the effectiveness of three types of therapy on alleviating depression. The factors are type of `therapy` (cognitive-behavioural, Rogerian, and Assertiveness Training) and severity of `depression` (mild, moderate, and severe)[3]. The dependent variable is the score on the depression scale of the MMPI.

The data are stored in the file `chapter_7_table_23.dat` on the CD that came with your textbook. I copied the file to my working directory and then read it into R with the following command:

```
# mwtable23<-read.table("chapter_7_table_23.dat")
file_path <- "http://pnb.mcmaster.ca/bennett/psy710/datasets/mw/chapter_7_table_23.dat"
mwtable23 <- read.table(file=url(file_path))
names(mwtable23)<-c("therapy","depression","mmpi")
mwtable23$therapy<-factor(mwtable23$therapy,labels=c("cognitive","rogerian","assertive") )
mwtable23$depression<-factor(mwtable23$depression,labels=c("mild","moderate","severe"))
class(mwtable23$therapy)

## [1] "factor"

class(mwtable23$depression)

## [1] "factor"
```

In the next block of code, I first check that the data are unbalanced. Then I construct an `aov` object and use `drop1` to estimate the Type III sums of squares for the two main effects and interaction:

```
with(mwtable23,tapply(mmpi,list(therapy,depression),length) )

##           mild moderate severe
## cognitive    3        5      7
## rogerian     5        6      4
## assertive    5        4      6
```

[3]Subjects are not assigned to severity of depression, so `depression` is a blocking variable, not a true experimental factor. Nevertheless, we will use a fixed-effects model to analyze these data.

```
options(contrasts=c("contr.sum","contr.poly") ) # use sum-to-zero coding
mw23.aov.01<-aov(mmpi~therapy+depression+therapy:depression,data=mwtable23)
drop1(mw23.aov.01,.~.,test="F")

## Single term deletions
##
## Model:
## mmpi   therapy + depression + therapy:depression
##                     Df Sum of Sq  RSS AIC F value  Pr(>F)
## <none>                           1005 158
## therapy              2       205 1210 162    3.67   0.036
## depression           2      1181 2187 189   21.15 8.4e-07
## therapy:depression   4        14 1020 150    0.13   0.972
## ---
## Signif. codes:  0 '   ' 0.001 '  ' 0.01 ' ' 0.05 '.' 0.1 ' ' 1
```

The first thing to note is that the interaction is not significant, so it makes sense to examine the main effects. The Type III sums of squares for `therapy` and `depression` are 205 and 1181, respectively. Both are significant, so we reject the null hypotheses that the unweighted marginal means (of mmpi scores) do not vary across types of `therapy` or level of `depression`. (N.B. This second main effect is not interesting.)

In this example, $SS_{therapy:depression}$ was very small and the p value was very large ($p = 0.971$). In other words, if interaction effects do exist then they are likely to be very small. Therefore, we might consider evaluating the main effects using Type II sums of squares. The following commands list two ANOVA tables for models that contain two main effects but no interaction. The second line in each table is the sum of squares for that main effect after controlling for the other main effect while ignoring the interaction (which we think is approximately zero). In other words, those lines contain the Type II sum of squares and, because the interaction is assumed to be zero, test the same null hypotheses as the Type III sums of squares:

```
mw23.aov.02<-aov(mmpi~therapy+depression,data=mwtable23)
summary(mw23.aov.02)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## therapy      2    101      51    1.98    0.15
## depression   2   1253     627   24.58 1.1e-07 ***
## Residuals   40   1020      25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mw23.aov.03<-aov(mmpi~depression+therapy,data=mwtable23)
summary(mw23.aov.03)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## depression   2   1116     558   21.89 3.8e-07 ***
## therapy      2    238     119    4.68   0.015 *
## Residuals   40   1020      25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Type II sums of squares for `therapy` and `depression` are 238.48 and 1253.19, respectively. Both are significant, so we reject the null hypothesis of no difference between marginal row means and marginal column means. Again, it is important to remember that this null hypothesis is begin evaluated because the interaction is assumed – with good reason, in this case – to be zero. If the interaction was not zero, then Type II sums of squares would be evaluating the crazy null hypothesis listed in Section 7.12.11.

The Type I sums of squares for `therapy` and `depression` – the versions calculated by ignoring all other effects, which are listed in the first lines of the two ANOVA tables – are 101.11 and 1115.82, respectively. These sums of squares can be used to evaluate the null hypotheses that the weighted marginal means do not vary across levels of `depression` or types of `therapy`. The effect of `therapy` is not significant ($F(2, 40) = 1.98$, $p = 0.15$), so we do not reject the null hypothesis of no difference among weighted marginal means. The effect of `depression` is significant (though this finding is not interesting).

7.12.15 Using `Anova` in the `car` package

In this section I describe how to use the `Anova()` command (N.B., note the capital A) to compute Type II and III sums of squares. Note that you do not *need* to use `Anova()`; `drop1()`, along with the standard `anova()` or `summary()` commands, are adequate. However, `Anova()` has some features that will be useful for analyzing within-subjects designs, which will be covered in later chapters. Therefore, I thought it might be useful to introduce the command here.

The `Anova()` command is part of the `car` package, which is a suite of commands and data files for `R` written and maintained by Professor John Fox. You can read about the `car` package at `http://cran.r-project.org/web/packages/car/index.html`. To use `car`, you need to install the software on your computer. From within `R`, type the following command:

```
install.packages(pkgs="car")
```

This command installs `car` on your computer; you only have to use it once (plus whenever you periodically update your software). The command `library("car")` loads `car` into memory: you have to use it once per session. You can read about the contents of `car` by typing `?car` at R's command prompt. You can learn more about `Anova` by typing `?Anova` at the command prompt. Again, take note of the capital A in `Anova`.

Finally, we can calculate Type II and III sums of squares. Type III sums of squares only make sense if we are using sum-to-zero coding of our effects. Therefore, I am going to make sure the coding is correct by resetting the coding option.

```
library("car") # load car (just once per session)
options(contrasts=c("contr.sum","contr.poly")) # just in case
Anova(mw23.aov.01,type="II")

## Anova Table (Type II tests)
##
## Response: mmpi
##                    Sum Sq Df F value  Pr(>F)
## therapy               238  2    4.27   0.022 *
## depression           1253  2   22.44 4.7e-07 ***
## therapy:depression     14  4    0.13   0.972
## Residuals            1005 36
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(mw23.aov.01,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: mmpi
##                     Sum Sq Df F value  Pr(>F)
## (Intercept)         121174  1 4338.72 < 2e-16 ***
## therapy                205  2    3.67   0.036 *
## depression            1181  2   21.15 8.4e-07 ***
## therapy:depression      14  4    0.13   0.972
## Residuals             1005 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

Cohen, J. (1988). *Statistical power analysis for the behavioral science*. Erlbaum, Hillsdale, NJ, 2nd edition.

Faraway, J. J. (2005). *Linear models with R*. Chapman & Hall/CRC, Boca Raton.

Howell, D. C. (2002). *Statistical methods for psychology*. Duxbury/Thomson Learning, Pacific Grove, CA, 5th edition.

Howell, D. C. and McConaughy, S. H. (1982). Nonorthogonal analysis of variance: Putting the question before the answer. *Educational and Psychological Measurement*, 42(1):9–24.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Brooks/Cole, 3rd edition.

Maxwell, S. E. and Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective*. Lawrence Erlbaum Associates, Mahwah, N.J., 2nd ed edition.

Venables, W. and Ripley, B. (2002). *Modern applied statistics with S*. Statistics and Computing. Springer-Verlag, New York, 4th edition edition.