Notes on Maxwell & Delaney

PSYCH 710

October 3, 2021

3 Chapter 3

3.1 Linear Models

Whenever we do an analysis of variance we are determining which one of several linear models best fits our data. Linear models have the form:

$$Y_i = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e_i \tag{1}$$

where Y_i is the dependent variable for subject i, and the X's and β's are, respectively, predictors and parameters. The last element of the model, e_i, is called the error or residual term. Equation 1 represents a *linear* model because it is a weighted sum of predictor variables (i.e., the X's). It is important to note that only the parameters must be linear, not the predictor variables. So, the equation

$$Y_i = \beta_0 X_0 + \beta_1 \exp(X_1) + \beta_2 X_2^2 + \cdots + \beta_p X_p + e_i \tag{2}$$

is a linear equation because it is possible to rewrite it as

$$Y_i = \beta_0 X_0 + \beta_1 X_1' + \beta_2 X_2' + \cdots + \beta_p X_p + e_i \tag{3}$$

where $X_1' = \exp(X_1)$ and $X_2' = X_2^2$. However,

$$Y_i = \beta_0 X_0 + \exp(\beta_1 X_1) + (\beta_2 X_2)^2 + \cdots + \beta_p X_p + e_i \tag{4}$$

is not a linear model.

The predictor variables usually represent things that differ among subjects. For example, the X's might represent subject's age, gender, years of education, etc. In designed experiments, the X's also represent the treatments subjects receive, or the

experimental group to which they belong. One exception to this general rule is the first term in the model, $\beta_0 X_0$, which usually represents things that are constant across subjects. In the construction of the model, X_0 is typically set to 1, yielding a slightly simpler model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e_i \tag{5}$$

The parameter β_0 is often referred to as the model's intercept. All of the parameters and predictors form the predicted value, or estimate, of Y_i, which is designated at $\hat{Y}_i$. The residual term is the difference between the observed and estimated values of Y

$$e_i = Y_i - \hat{Y}_i \tag{6}$$

3.2 Least Squares

Much of statistics involves fitting models like Equation 5 to data. A common technique for accomplishing this task is to adjust the parameters of a model to find the set of parameters that minimizes the sum of the squared residuals

$$\sum e_i^2 \tag{7}$$

where the sum is taken across all i subjects. The parameters that minimize Equation 7 are said to provide the best-fitting model according to the *least squares* criterion of goodness-of-fit. Least squares is the method of estimating parameters that is used in the analysis of variance and multiple regression.

3.3 Comparing Models

A very wide range of linear models could be fit to any given data set. What criteria should we use to decide which model is the "best"? One obvious criterion would be to find the one that minimizes Equation 7. Unfortunately, such a criterion would not (by itself) be very useful because it is always possible to fit the data *exactly* (i.e., have Equation 7 go to zero) with a model that has one parameter, β, for each observation, Y_i. So, if we have 25 observations, we can always reproduce the observations exactly with a model that has 25 parameters.

What is lacking here is some notion of simplicity, or parsimony: We want to use a model that provides a good fit to the data but is, at the same time, as "simple" as possible. A common measure of model complexity is the number of parameters: a model with 4 parameters is judged to be more complex than a model with 3

parameters. With this notion of model complexity, we can restate the problem of model selection as choosing the model with the fewest parameters that provides a good fit to the data.

To illustrate how this is done, we will analyze the data presented in Table 3.2 of your textbook. The data are from a mood induction study by Pruitt (1988). Subjects had to view videoclips that were designed to induce a pleasant, unpleasant, or neutral mood. After viewing a videoclip, each subject rated his/her mood on several scales. In addition, each subject was videotaped, and an assistant (who did not know which videoclip was watched by the subject) later watched the videotape and rated each subject's mood on a 7-point scale. The assistant's ratings are presented in Table 3.2 in the textbook. There were 10 subjects per group. Our task is to determine if mood ratings were associated with videoclip condition (pleasant, neutral, and unpleasant). We fit the following two models to the data:

$$Y_{ij} = \mu + \alpha_j + e_{ij} \tag{8}$$

$$Y_{ij} = \mu + e_{ij} \tag{9}$$

Here, Y_{ij} represents the score (i.e., mood rating) for subject i in group j. In Equation 8, the observed score, Y_{ij} is the sum of a constant (μ), a group-specific effect (α_j), and a residual term (e_{ij}); the predicted score, $\hat{Y}_{ij}$, equals $\mu + \alpha_j$. In Equation 9, the observed score is the sum of a constant and a residual term, and the predicted score consists only of a constant. The effects are defined as $\alpha_j = \mu_j - \mu$, and satisfy the constraint that the sum of all effects is zero:

$$\sum_{j=1}^{a} \alpha_j = 0 \tag{10}$$

where a is the number of groups. Note that the models specified by Equations 8 and 9 are *nested* versions of each other because Equation 9 can be obtained by setting $\alpha_j = 0$. Equations 8 and 9 represent the full and reduced models, respectively. The question of interest is whether the full model provides a better fit to the data than the reduced model, even after taking into account its greater complexity.

Now that we've specified the models, we need to estimate the best-fitting (least squares) parameters. For the full model, it can be shown that the sum of squared residuals for Equation 8 is minimized by setting $\mu = \bar{Y}_u$ and $\alpha_j = \hat{\alpha}_j$, where

$$\bar{Y}_u = \sum_{j=1}^{a} \bar{Y}_j / a \tag{11}$$

$$\hat{\alpha}_j = \bar{Y}_j - \bar{Y}_u \ . \tag{12}$$

$\bar{Y}_u$ is simply the mean of the group means; differences in the size of the groups (if they exist) are ignored, and so $\bar{Y}_u$ is said to be the unweighted mean of the group means. $\hat{\alpha}_j$ is simply the difference between the mean of group j and $\bar{Y}_u$. For the reduced model, setting the one free parameter, μ, to the grand average, $\bar{Y}$, minimizes the sum of squared residuals.

3.3.1 F formula

Next, we need to derive a quantitative measure of the relative goodness-of-fit of the two models. We denote the **sum of squared residuals** for the best-fitting full and reduced models as E_F and E_R, respectively. Associated with E_F and E_R are degrees-of-freedom $\mathrm{df}_F = N - a$ and $\mathrm{df}_R = N - 1$, respectively, where N is the total number of observations and a is the number of groups. Note that $\mathrm{df}_R - \mathrm{df}_F = a - 1$ is the difference between the number of parameters estimated in the full model (3 α's and 1 intercept) and the reduced model (1 intercept). The formula for computing the difference between the two models is

$$F = \frac{(E_R - E_F)/(\mathrm{df}_R - \mathrm{df}_F)}{E_F/\mathrm{df}_F} \tag{13}$$

Equation 13 can be used to compare *all* nested linear models. All tests in ANOVA, analysis of covariance, and multiple regression can be computed using this formula.

3.3.2 Null Hypothesis Testing

Finally, we are in a position to evaluate the hypothesis of no difference between the goodness-of-fit of the full and reduced models. Note that this comparison is equivalent to evaluating the hypothesis that all of the groups have the same mean; or (equivalently) that all α_j's are zero. More formally, we are comparing the hypotheses

$$\begin{aligned} H0: &\quad \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \\ H1: &\quad \alpha_j \neq 0 \end{aligned}$$

The null hypothesis is that all of the effects are zero, and therefore that all group means are equal. The alternative hypothesis is that at least one effect is not zero, and therefore that not all group means are equal. When the residuals, e_{ij}, are distributed as independent, normal random variables, with mean of zero and a constant variance, then F in Equation 13 follows an F distribution with $(\mathrm{df}_R - \mathrm{df}_F)$ and df_F degrees of freedom in the numerator and denominator, repsectively (Figure 1). Under the null hypothesis, therefore, large values of F should be relatively rare (Figure 2). Using

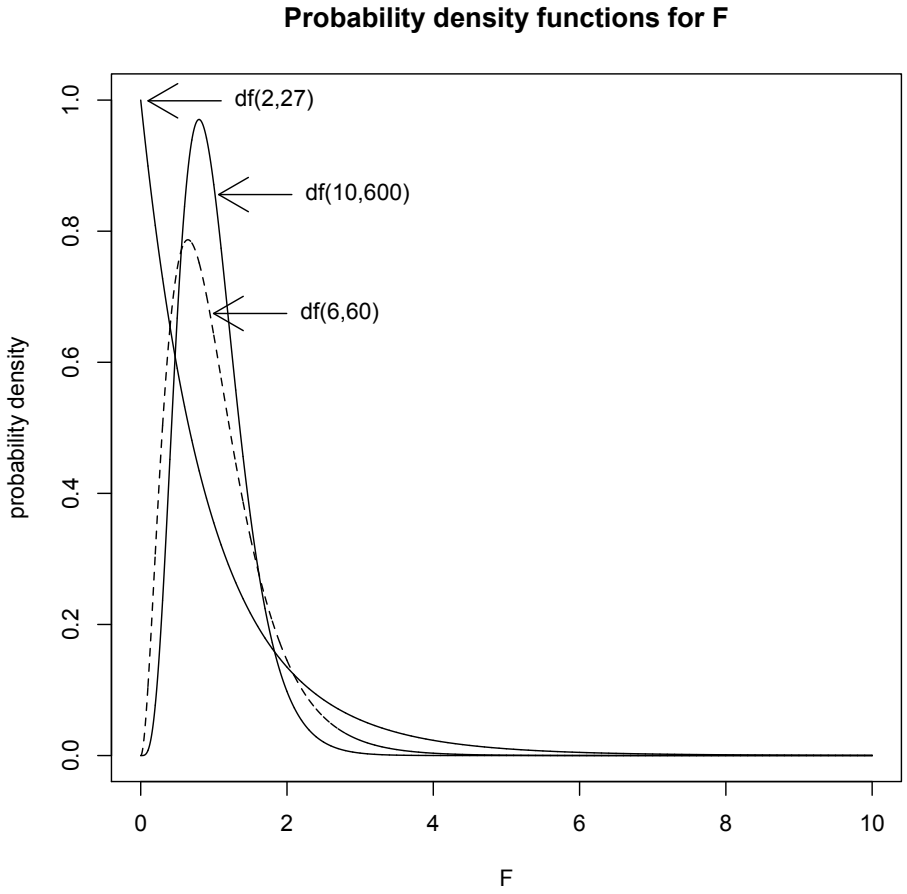

Figure 1: The F distribution is determined by two parameters which correspond to the degrees of freedom of the numerator and denominator of the F ratio. This figure shows the probability density functions of three different F distributions.

the F distribution, we can calculate the probability of obtaining a value of F that is as large or larger than the observed value of F under the assumption that the null hypothesis is true. If the p-value is smaller than our criterion value, typically .05 or .01, then we reject the null hypothesis in favour of the alternative. If the p-value is not smaller than our criterion, then we do not reject the null hypothesis.

3.3.3 Relation to ANOVA

Imagine an experiment in which we measure some aspect of behaviour on N subjects who were assigned randomly to a groups with the constraint that each group has the same number (n) of subjects (i.e., $N = an$). A standard ANOVA table is shown in Table 1. The independent variable, Group, has $a - 1$ degrees of freedom; the other item listed in the Source column, Residuals, has $a(n - 1) = N - a$ degrees of freedom. The total degrees of freedom is equal to one less than the total number of subjects (i.e., $N - 1$). Each source also has a Sum-of-Squared Error (SS) and a Mean Squared Error (MS). The SS and MS values for Group are referred to as *between-group* Sum-of-Squares and Mean Squared Error (i.e., SS_B and MS_B), whereas the values for Residuals often are referred to as *within-group* Sum-of-Squares and Mean Squared Error (i.e., SS_W and MS_W).

Source	df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	$a - 1$	SS_B	MS_B	$\mathrm{MS}_B/\mathrm{MS}_W$	p
Residuals	$a(n - 1)$	SS_W	MS_W		

Table 1: A standard ANOVA table for a one-way design.

The elements of Equation 13 are closely tied to various components of a standard ANOVA table (e.g., Table 1). For example, it can be shown that

$$F = \frac{(E_R - E_F)/(\mathrm{df}_R - \mathrm{df}_F)}{E_F/\mathrm{df}_F} = \frac{\mathrm{MS}_B}{\mathrm{MS}_W} \tag{14}$$

From Equation 14 it is possible to show that

$$\begin{aligned} E_F &= \mathrm{SS}_W \\ E_R &= \mathrm{SS}_B + \mathrm{SS}_W = \mathrm{SS}_{Total} \\ E_R - E_F &= \mathrm{SS}_{Total} - \mathrm{SS}_W = \mathrm{SS}_B \end{aligned}$$

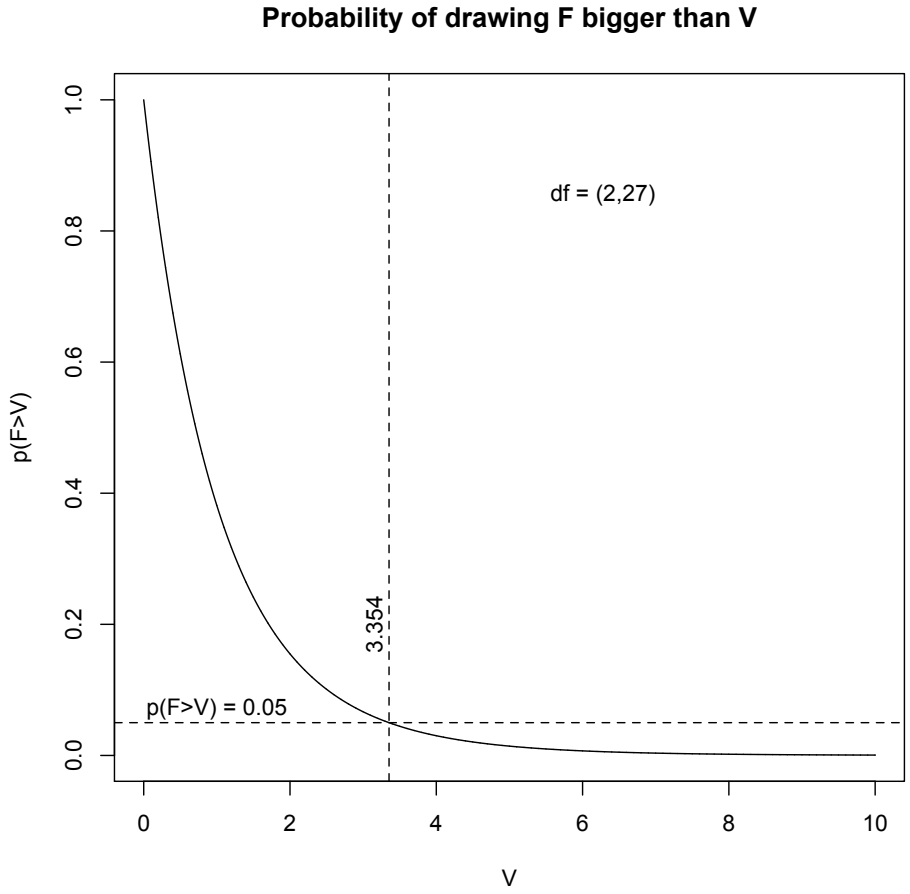

Figure 2: This figure plots the probability of drawing a random number from an $F(2, 27)$ distribution that is greater than some criterion, denoted as V. For example, the dotted lines indicate that the probability of drawing an F value that is ≥ 3.354 is 0.05. Note that the probability of drawing values greater than 4 is very low.

MS_W and E_F/df_F are estimates of the population error variance, σ_e^2. If the null hypothesis is true, MS_B and $(E_R - E_F)/(\mathrm{df}_R - \mathrm{df}_F)$ also are estimates of σ_e^2. However, if the null hypothesis is false, then

$$E(\mathrm{MS}_B) = \sigma_e^2 + \frac{\sum_j n_j \alpha_j^2}{a-1} \tag{15}$$

where n_j is the number of subjects in group j. So, when $\alpha_j \neq 0$ (for at least one group, j), $E(\mathrm{MS}_B)$ will tend to be larger than $E(\mathrm{MS}_W)$, and the F ratio (see Table 1) ought to be greater than one.

Finally, "Residual Standard Error", a measure of goodness-of-fit that is provided by many statistical software packages, can be shown to be equal to $\sqrt{E_F/\mathrm{df}_F}$.

3.3.4 Numerical Example

The data from the mood-induction experiment are shown in Table 2. You can load the data into R with the following command:

```
load(url("http://www.pnb.mcmaster.ca/bennett/psy710/datasets/mood_data.Rdata"))
```

Next, we want to set up R so that it defines effects as in Equation 10:

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

The `contr.sum` parameter in the `options` command tells R to use the sum-to-zero definition (Eq 10) of the α's when the grouping variable is a `factor`. If you do not use this command, then R will use a different definition for the α's, and your results will not match those shown here or in the textbook.

In R, we fit the full and reduced models to the data with the following commands:

```
names(mood.data)

## [1] "group" "mood"

mood.full <- lm(mood~1+group,data=mood.data)
mood.restricted <- lm(mood~1,data=mood.data)
```

The sum of the square residuals and degrees of freedom for each model are given by

```
(E.full<-sum(residuals(mood.full)^2))
```

```
## [1] 26
```

```
(E.restricted<-sum(residuals(mood.restricted)^2))
```

```
## [1] 72.67
```

```
(df.full<-mood.full$df.residual)
```

```
## [1] 27
```

```
(df.restricted<-mood.restricted$df.residual)
```

```
## [1] 29
```

Finally, we can calculate F and a p-value:

```
(F <- ( (E.restricted-E.full)/(df.restricted-df.full)/(E.full/df.full) ) )
```

```
## [1] 24.23
```

```
(p.value <- 1-pf(F,df1=(df.restricted-df.full),df2=df.full) )
```

```
## [1] 9.421e-07
```

Because $F = 24.23$ and $p \ll .001$, we reject the null hypothesis that all $\alpha_j = 0$ and accept the alternative hypothesis that $\alpha_j \neq 0$ for at least one group j.

Fortunately, we do not have to calculate F this way every time we want to compare models. A short-cut would be to use the two-model version of R's anova command:

	Pleasant	Neutral	Unpleasant
	6	5	3
	5	4	3
	4	4	4
	7	3	4
	7	4	4
	5	3	3
	5	4	1
	7	4	2
	7	4	2
	7	5	4
mean	6	4	3

Table 2: Data from mood-induction experiment.

```
anova(mood.restricted,mood.full)

## Analysis of Variance Table
##
## Model 1: mood ~ 1
## Model 2: mood ~ 1 + group
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     29 72.7
## 2     27 26.0  2      46.7 24.2 9.4e-07 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `anova` command lists the formula for each model as well as a table listing the residual degrees of freedom, the residual sums-of-squares (RSS), the difference $\mathrm{df}_R - \mathrm{df}_F$, the difference between the sums of squared residuals for the two models (46.7), F, and p. Note that the F and p values are the same as the ones we calculated by hand.

There is even a shorter shortcut that can be used in this case, namely the single-model version of the anova command:

```
anova(mood.full)
```

```
## Analysis of Variance Table
##
## Response: mood
##           Df Sum Sq Mean Sq F value  Pr(>F)
## group      2   46.7   23.33    24.2 9.4e-07 ***
## Residuals 27   26.0    0.96
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This version of the command lists an anova table. Note that the values of F and p are the same as those obtained before. Also note that the "Sum Sq" value of 46.7 is the same one calculated before as the difference between the sum of squared residuals for the full and restricted model. In fact, it *is* the same value. The SS_{Group} term represents the change in the sum of squared residuals that occurs when the factor "group" is dropped from the analysis. Or, in other words, it is the effect of setting all α_j's to zero.

3.3.5 Confidence intervals for α_j

Rejecting the null hypothesis implies that $\alpha_j \neq 0$ for at least one group. However, often we want to know more about the α's : In some situations it is crucial to know what the α's *are*, not just whether or not they differ from zero. The values of the model parameters are given by R's summary function:

```
summary(mood.full)
```

```
##
## Call:
## lm(formula = mood ~ 1 + group, data = mood.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -2     -1      0      1      1
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.333      0.179   24.19  < 2e-16
## group1         1.667      0.253    6.58  4.7e-07
## group2        -0.333      0.253   -1.32      0.2
##
## (Intercept) ***
## group1      ***
## group2
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.981 on 27 degrees of freedom
## Multiple R-squared:  0.642,  Adjusted R-squared:  0.616
## F-statistic: 24.2 on 2 and 27 DF,  p-value: 9.42e-07
```

You should confirm for yourself that the Intercept is $\bar{Y}_u$, and that the values listed as group1 and group2 correspond to α_1 and α_2, respectively. Where is α_3? It is not shown in the table, but it can be calculate using Equation 10. The t and p values beside each parameter evaluate the null hypothesis that the parameter is zero. In this case, the Intercept and group1 parameters differ significantly from zero.

Next to each estimated parameter is a standard error, which can be used to calculate a confidence interval. To calculate a confidence interval, we first need to find the so-called critical value of t, from the t distribution, corresponding to our residual degrees of freedom and the α level (N.B. Here α refers to our criterion for rejecting the null hypothesis (e.g., $\alpha = .05$), not the effects in our linear model). The residual degrees of freedom for the full model is 27 (you can see it listed in the summary). Let us set $\alpha = .05$ to compute the $100(1-\alpha)\%$, or 95%, confidence interval. We need to calculate the critical value of t, $t_{critical,\alpha/2}$, for which the probability of drawing a random value (from a t distribution with $df = 27$) that is greater than or equal to $t_{critical,\alpha/2}$ is $\alpha/2$.

```
t.alpha <- .05/2
(t.critical <- qt(1-t.alpha,df=27) )

## [1] 2.052
```

In the previous calculation, note that α was divided by 2.

The 95% confidence interval is calculated with the formula

$$\alpha_j \pm t_{critical,\alpha/2}\ \mathrm{se}_j \tag{16}$$

where se_j is the standard error of parameter j. For group1, the 95% confidence interval is $1.667 \pm 2.052 \times 0.253 = (1.148, 2.186)$. The interval (1.148, 2.186) is called the 95% confidence interval because an interval constructed this way will contain the true population mean 95% of the time.

A simpler way of calculating confidence intervals for parameters in a linear model is to use R's `confint()` function:

```
confint(mood.full)

##                2.5 % 97.5 %
## (Intercept)  3.9657 4.7009
## group1       1.1468 2.1865
## group2      -0.8532 0.1865

confint(mood.full,level=.99)

##                0.5 % 99.5 %
## (Intercept)  3.8369 4.8297
## group1       0.9647 2.3687
## group2      -1.0353 0.3687
```

The first call lists the 95% confidence intervals; the second lists the 99% confidence intervals.

3.3.6 Measures of association & effect size

One common measure of *association strength* between between the dependent variable and the predictors (i.e., between Y_i and $\hat{Y}_i$) is R^2, which is known as Multiple-R squared, the coefficient of determination, and eta squared (η^2). R^2 represents the amount of variance in the dependent variable that is accounted for, or explained by, the linear model. One problem with R^2 is that it is biased: the value estimated from the data is higher than the value in the population, and the bias increases as sample size decreases. Adjusted-R^2, denoted by $\tilde{R}^2$, is an unbiased, or at least a less biased, estimate of the population R^2. Both R^2 and $\tilde{R}^2$ are printed by R's `summary()` function.

Another common measure of association is omega-squared (ω^2), which is the variance of the treatment, or group, effects (i.e., the α's) divided by the sum of the sum of the variance of the treatment effects and error variance:

$$\omega^2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2} \tag{17}$$

The values of η^2 and ω^2 can be calculated from quantities listed in ANOVA tables. The formula for eta-squared is

$$\eta^2 = \frac{SS_{group}}{SS_{total}} \tag{18}$$

and the formula for ω^2 is

$$\hat{\omega}^2 = \frac{SS_{Group} - (a-1)MS_{Resid}}{SS_{total} + MS_{Resid}} \tag{19}$$

where (a-1) is the degrees of freedom for Group. For one-way designs, the value of adjusted-R^2, which usually is listed in regression summary tables, is very similar to ω^2. Cohen (1988) suggested that ω^2 values of 0.01, 0.06, and 0.14 corresponded to weak, moderate, and strong associations.

Cohen's f is a measure of *effect size*. It is the ratio of the standard deviation of the α's divided by the standard deviation of the residuals. For this one-way design, f is well-approximated by the equation

$$\hat{f} \approx \sqrt{\frac{\tilde{R}^2}{1 - \tilde{R}^2}} \tag{20}$$

Alternatively, if you have access to the degrees of freedom and F value for the effect in question, and the total sample size, you can use the formula

$$\hat{f} = \sqrt{\left(\frac{\text{df.effect}}{\text{N.total}}\right)(\text{F.effect} - 1)} \tag{21}$$

Cohen's f expresses the standard deviation among effects relative to the standard deviation of residuals. According to Cohen (1988), small, medium, and large effects correspond to f's of 0.1, 0.25, and 0.4, respectively. Among other things, f is useful for calculations of power.

3.3.7 Example: Association Strength & Effect Size

In this section we illustrate how to calculate association strength and effect size using the results from the mood-induction study.

```
anova(mood.full)

Analysis of Variance Table

Response: mood
          Df Sum Sq Mean Sq F value   Pr(>F)
group      2   46.7   23.33    24.2 9.4e-07 ***
Residuals 27   26.0    0.96
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using Equation 18 and 19 we see that η^2 is $46.7/(46.7 + 26)$, or 0.64. and $\hat{\omega}^2$ is

$$\frac{46.7 - 2 \times 0.96}{46.7 + 26 + 0.96} = 0.608$$

For 1-way designs ω^2 is approximately equal to adjusted R-squared ($\tilde{R}^2$):

```
( adj.R2 <- summary(mood.full)$adj.r.squared ) # omega-squared

## [1] 0.6157
```

Cohen's f can be estimated from adjusted R-squared and from the quantities in the ANOVA table using Equations 20 and 21:

```
sqrt( adj.R2/(1-adj.R2) ) # cohen's f

## [1] 1.266

df.effect <- 2
N.total <- 27+2+1
F.effect <- 24.2
sqrt( (df.effect/N.total)*(F.effect-1) ) # cohen's f

## [1] 1.244
```

These measures of effect size and association strength can be calculated using several commands in the `effectsize` package. One nice feature of these commands is that they return confidence intervals. By default, the 90% intervals are returned, but we can ask for the 95% interval instead:

```
# install.packages("effectsize")
library(effectsize)
eta_squared(mood.full,ci=0.95)

## # Effect Size for ANOVA
##
## Parameter | Eta2 |       95% CI
## ------------------------------
## group     | 0.64 | [0.39, 0.77]

omega_squared(mood.full,ci=0.95)

## # Effect Size for ANOVA
##
## Parameter | Omega2 |       95% CI
## --------------------------------
## group     |   0.61 | [0.34, 0.75]

cohens_f(mood.full,ci=0.95)

## # Effect Size for ANOVA
##
## Parameter | Cohen's f |       95% CI
## ----------------------------------
## group     |      1.34 | [0.79, 1.84]
```

Note that the value of Cohen's f returned by `cohens_f` differs from the one calculated using values in the ANOVA table. The reason for this difference is that `cohens_f` calculates f using an algorithm that is equivalent to Equation 20, but uses R^2 (or, equivalently, η^2) instead instead of $\tilde{R}^2$:

$$\hat{f} = \sqrt{\frac{\eta^2}{1-\eta^2}} = \sqrt{\frac{0.64}{1-0.64}} = 1.33$$

3.4 Power

The power of a test refers to the probability of rejecting the null hypothesis when it is false. In the case of a oneway ANOVA, power depends on the number of groups, number of subjects per group, α level, and effect size.

3.4.1 estimating sample size from a pilot study

As an example of how to use power to plan your experiments, consider the case where you are planning to measure reaction time (RT) on three groups. Based on previous experiments, or perhaps your own pilot data, you think that the average RTs in each group will be 400, 450, 500, and the within-group standard deviation will be 100. Cohen's f is given by

$$\frac{\sigma_m}{\sigma_e} = \frac{\sqrt{\frac{(400-450)^2+(450-450)^2+(500-450)^2}{3}}}{100} = \frac{50}{100} = .408 \tag{22}$$

According to Cohen, this is a large effect. Now, the question is how many subjects should you test in each group in the actual experiment to achieve a power of 0.8? To answer this question, you need to specify the number of groups (3), the significance level that you will use ($\alpha = .05$), and the effect size ($f = 0.41$). So, given these assumptions, how many subjects do we need to attain a power of 0.8? To find out, we should use R's `pwr.anova.test()` function in the `pwr` package:

```
library(pwr)
pwr.anova.test(k=3,f=.41,sig.level=.05,power=.8,n=NULL)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 20.14
##               f = 0.41
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

It turns out that I will need nearly 20 subjects per group to attain the desired power. If I had set $\alpha = .01$ instead of .05, then I would need ≈ 30 subjects per group.

Suppose we can test only 10 subjects per group due to the constraints imposed by time and money. What is the power of such a study?

```
pwr.anova.test(k=3,f=.41,sig.level=.05,power=NULL,n=10)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 10
##               f = 0.41
##       sig.level = 0.05
##           power = 0.4614
##
## NOTE: n is number in each group
```

The power is 0.46, so I have only a 46% chance of rejecting the null hypothesis if it is in fact false.

The built-in function `power.anova.test()` is an alternative to `pwr.anova.test`:

```
groupMeans <- c(400,450,500);
power.anova.test(groups=3,
                 between.var=var(groupMeans),
                 within.var=100^{2},
                 sig.level=.05,
                 power=.8,n=NULL)


##
##      Balanced one-way analysis of variance power calculation
##
##           groups = 3
##                n = 20.3
##      between.var = 2500
##       within.var = 10000
##        sig.level = 0.05
##            power = 0.8
##
## NOTE: n is number in each group
```

`pwr.anova.test()` and `power.anova.test()` are very useful functions that can be used in a variety of ways when planning experiments. It is important to keep in mind, however, that power calculations, like all other statistical calculations, are valid only if the assumptions behind the calculations are correct. In the case of power calculations, the assumptions are that the data are distributed as independent, normal random variables with constant variance. If those assumptions are false, then the results of the power calculations can be misleading, sometimes seriously so.

3.5 Statistical Assumptions

The interpretation of the p-value calculated for the F statistic (Equation 13) rests on the assumption that the observed F does indeed follow an F distribution. Three assumptions must be met for this to be true:

1. The population distribution of scores on the dependent variable, Y, must be normal *within each group*.

2. The population variances of scores Y must be equal for all a groups. This is often called the homogeneity of variance assumption.

3. The scores must be statistically independent of each other. Typically, this assumption is met by assigning subjects randomly to groups/treatments.

3.5.1 robustness

In practice, the above assumptions are almost never met exactly, and so it is important to consider the effects that violations of the assumptions have on our analyses.

In general, ANOVA is robust to violations of the normality assumption. Specifically, if the distribution of scores in *all* groups deviate from normality in the *same* way, the actual Type I error rate often will be reasonably close to the expected Type I error rate. Of course, deviations from normality do not have an all-or-none effect on our analyses: the greater the deviation from normality, then the stronger the effect on our actual Type I error rate. Also, the effect of non-normality on Type I error rate increases in cases where sample sizes are unequal across groups. Finally, deviations from non-normality that *differ* across groups - for example, if scores are skewed positively in one group and negative in another group - can significantly affect power. Finally, you should be aware that Wilcox and Keselman (2003) have argued that even apparently small deviations from normality can dramatically reduce power.

It is generally assumed that ANOVA is robust to violations of the homogeneity of variance assumption, *as long as samples sizes in each group are equal*. If sample sizes are equal, the F test performs well if the ratio of the largest and smallest variances among the groups is ≈ 3 or less. If sample sizes are unequal, however, then even moderate heterogeneity of variance can inflate Type I error rates significantly. Also, if the scores within each group are distributed non-normally, then moderate heterogeneity of variance will inflate Type I error rates even if group n's are equal (Wilcox and Keselman, 2003).

The F test is not robust to violations of the independence-of-errors assumption: Violations of this assumption will result in very poor control of Type I error rates. Kenny and Judd (1986) discuss how non-independence can affect your data analyses.

3.5.2 tests for non-normality

There are many statistical methods that can be used to formally test whether your data are distributed non-normally. Two that are implemented in R are the Kolmogorov-Smirnov test (`ks.test()`) and the Shapiro-Wilk's test (`shapiro.test()`). Of the two, the Shapiro-Wilk's test is preferred because it has higher power. However, the power of both of these tests is not very good, so they will not be sensitive to small deviations from normality. One way of solving this problem is to use a more liberal decision criterion (e.g., $\alpha = 0.1$).

```
shapiro.test(residuals(mood.full) )
```

```
##
##       Shapiro-Wilk normality test
##
## data:  residuals(mood.full)
## W = 0.85, p-value = 5e-04
```

As you can see, the p-value is quite small, so we reject the null hypothesis that the residuals are distributed normally. We'll return to the cause of the non-normality in a moment.

A very useful graphical method for searching for non-normality is to plot the residuals of your analysis in a qqplot using R's `qqnorm` function. If the scores are distributed normally, then they will fall along a straight line in a qqplot. The following commands were used to create Figure 3:

```
qqnorm(residuals(mood.full),main="residuals from mood.full");
qqline(residuals(mood.full) );
```

The second function, `qqline`, adds a reference line. You can see that the residuals do not fall along the line: although there follow a linear trend, there is a strange scalloping, or staircase, effect. This effect is due to the nature of the relatively coarse nature of the dependent variable (i.e., an integer on a 7-point mood scale). We can test this idea by adding a very small amount of noise to the residuals. Adding the noise jitter makes the data look much more normally distributed (see Figure 4).

```
tmp<-residuals(mood.full)+rnorm(residuals(mood.full),mean=0,sd=0.333)
shapiro.test(tmp)
```

```
##
##       Shapiro-Wilk normality test
##
## data:  tmp
## W = 0.95, p-value = 0.2
```

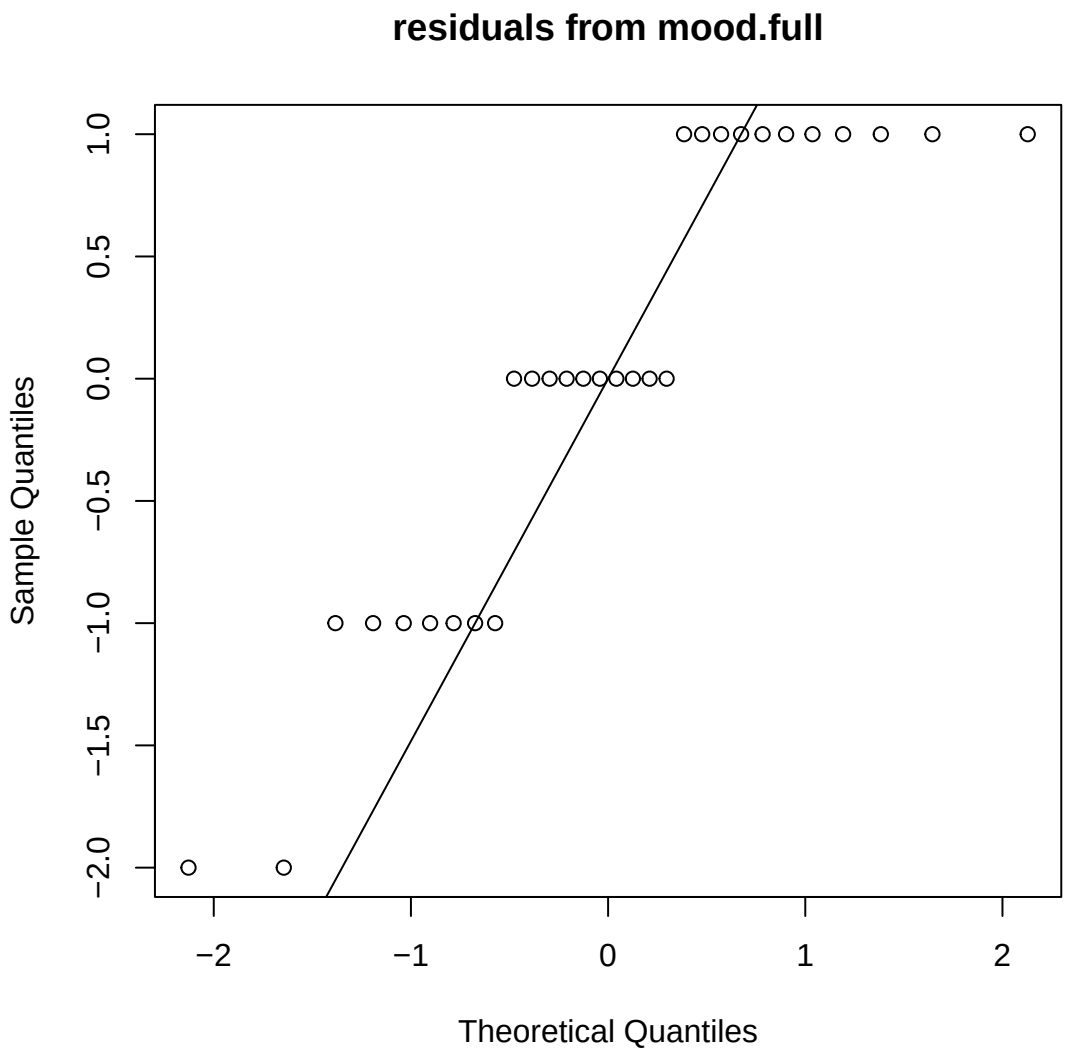

Figure 3: qqnorm plot of residuals of mood.full model.

```
qqnorm(tmp,main="noise-jittered residuals");
qqline(tmp);
```

3.5.3 tests for homogeneity of variance

There are many statistical methods that can be used to formally test for heterogeneity of variance. One common test that is implemented in R is the Bartlett test (`bartlett.test()`). As is the case for tests of non-normality, these tests often lack power, so you might consider using a liberal decision criterion (e.g., $\alpha = 0.1$).

```
bartlett.test(mood.data$mood,mood.data$group)


##
##      Bartlett test of homogeneity of variances
##
## data:  mood.data$mood and mood.data$group
## Bartlett's K-squared = 2.6, df = 2, p-value
## = 0.3
```

As you can see, $p = 0.2708$, so we do not reject the null hypothesis of equal variances among groups.

You might also consider using `boxplot()` to graphically inspect your data for differences in variance.

3.5.4 data transformations

What should you do if you find that your data violate the assumption of normality and/or homogeneity of variance? One possible remedy is to transform your data such that the transformed scores are more nearly normal and exhibit equal variances. Some common transformations (Kirk, 1995) are:

1. Square-root Transformation: $Y' = \sqrt{Y}$. This transformation is useful when $\sigma_e^2 = k\mu$. If some scores are less than 10, then consider using $Y' = \sqrt{Y + 0.5}$.

2. Log Transformation: $Y' = \log_{10}(Y)$. Useful when $\sigma_e = k\mu$ and/or cases where data are skewed positively. If any scores are zero, use $Y' = \log_{10}(Y + 1)$.

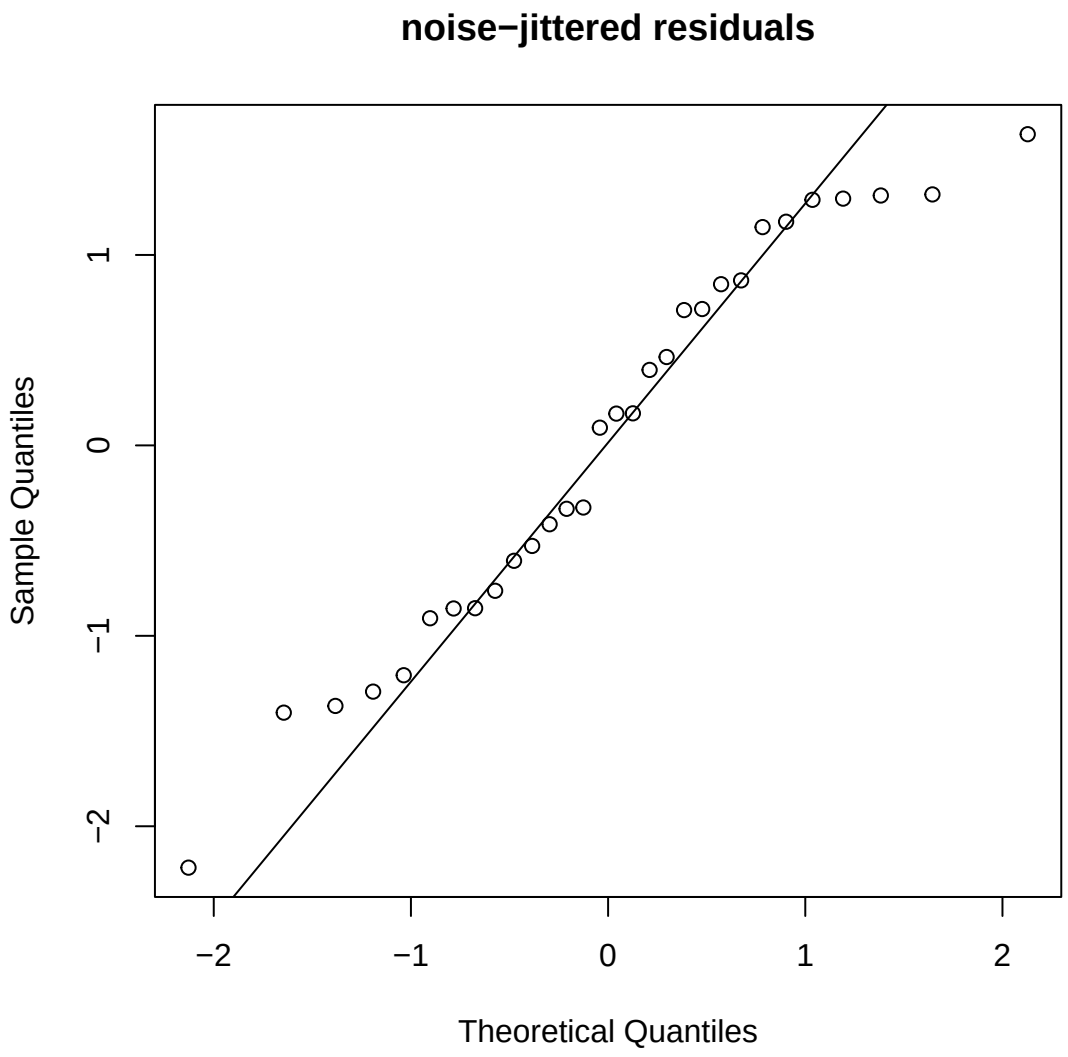

Figure 4: qqnorm plot of jittered residuals of mood.full model.

3. Reciprocal Transformation: $Y' = 1/Y$. Useful when $\mu^2 = k\sigma_e$ and/or cases where data are skewed positively. If any scores are zero, use $Y' = 1/(Y + 1)$.

4. Angular or Inverse Sine Transformation: $Y' = 2\arcsin(\sqrt{Y})$. Useful when Y is a proportion.

   Sometimes the inverse sine (or arcsin) transformation is given as $Y' = \arcsin(\sqrt{Y})$ (i.e., without the 2). The two version of the transformation will give exactly the same results. In fact, any linear transformation (e.g., $Y' = kY + b$) will have no effect on the results of your F test.

It is important to remember that any conclusions that you make as a result of statistical tests on transformed data *apply to the transformed data*, not the original data. So, if you use a log transformation, then you conclusions will apply to the log-transformed scores, not necessarily to the original scores.

3.5.5 alternative analyses

Finally, we consider what to do if assumptions of normality and/or homogeneity of variance are violated, and we choose not to do analyses on transformed data.

If the data within each group are distributed normally but have different variances, then you should consider using the Welch correction procedure described in your textbook (pages 131-135). The basic problem here is that, in cases where the groups have unequal variances, the F calculated by Equation 13 will not be distributed as an F statistics with $\mathrm{df}_R - \mathrm{df}_F$ and df_F degrees of freedom. However, it will be distributed *approximately* as an F variable with reduced degrees of freedom. The formula for correcting the degrees of freedom is given on page 134 in the textbook. You can use R's `oneway.test()` function evaluate group differences with this approach.

```
oneway.test(mood~group,data=mood.data)


##
##      One-way analysis of means (not assuming
##      equal variances)
##
## data:  mood and group
## F = 18, num df = 2, denom df = 17, p-value =
## 6e-05
```

In this case the effect of group is still significant.

The Welch correction for degrees of freedom still assumes that the data are distributed normally. If the data are not distributed normally, then you might consider using a non-parametric procedures. So-called non-parametric procedures make minimal assumptions about the data, and so are appropriate when the scores are non-normal and/or differ in variance. Your book describes the Kruskal-Wallis test, which is appropriate for the one-way designs we are considering here. In R, the Kruskal-Wallis test is used in the following way:

```
kruskal.test(mood~group,data=mood.data)


##
##      Kruskal-Wallis rank sum test
##
## data:  mood by group
## Kruskal-Wallis chi-squared = 19, df = 2,
## p-value = 7e-05
```

Again, the effect of group is significant.

You might think that you should always use non-parametric procedures because they make weaker assumptions about the data than do parametric procedures (like the F test). However, such a decision would be unwise: when the assumptions of normality and equal variance are approximately true, parametric tests are *much* more powerful than non-parametric tests. Some writers have argued that traditional non-parametric tests are so severely "underpowered" that they should almost never be used (Wilcox, 1992). A variety of modern, non-parametric tests have been developed that have considerably more power than traditional ones (Efron and Tibshirani, 1993; Wilcox, 2005). Such methods are beyond the scope of this course, but are covered in other courses. If you routinely analyze data that violate the assumptions of normality and homogeneity of variance, then you should seriously think about learning these methods.

References

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Kenny, D. A. and Judd, C. M. (1986). Consequences of violationg the independence assumption in the analysis of variance. *Psychological Bulletin*, 99:422–31.

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

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science*, 1(3):101–105.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Elsevier/Academic Press, Boston.

Wilcox, R. R. and Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8(3):254–74.