ST 518 Homework 6

Eric Warren

October 22, 2023

Contents

1 Problem 1

A completely randomized experiment investigates the effects of increasing nitrogen (N) and copper (Cu) in the diet of chickens. Feed conversion ratio (FCR) is observed on $n = 4$ chickens for each of four treatment combinations(diets), with output below. Data are available online as “fcr.dat” so you can check your answers, but you should be able to complete these problems without software.

1.1 Part A

Write a factorial effects model for the 16 observed FCR measurements which assumes that, for a given diet, FCR is normally distributed, with variance σ^2 that is constant across diets.

We can say that this factorial effects model is $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$ where for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ and E_{ijk} are i.i.d $N(0, \sigma^2)$ errors.

1.2 Part B

Estimate the simple effect of increasing copper when $N = 25$.

We can estimate the simple effect by only looking at the output for when $N = 25$. This occurs when $Cu = 10$ and gives us a mean value of 133 and when $Cu = 100$ and gives us a mean value of 146. Therefore, the simple effect of increasing copper when $n = 25$ is 146 - 133 = 13. For future purposes we will call this answer $\hat{\theta_1} = \bar{y_3} - \bar{y_1} = 13$ where $\bar{y_3} = 146$ and $\bar{y_1} = 133$.

1.3 Part C

Estimate the simple effect of increasing copper when $N = 45$.

We can estimate the simple effect by only looking at the output for when $N = 45$. This occurs when $Cu = 10$ and gives us a mean value of 130 and when $Cu = 100$ and gives us a mean value of 127. Therefore, the simple effect of increasing copper when $n = 45$ is subtracting these two values of 127 - 130 = -3. For future purposes we will call this answer $\hat{\theta_2} = \bar{y_4} - \bar{y_2} = -3$ where $\bar{y_4} = 127$ and $\bar{y_2} = 130$.

1.4 Part D

Estimate the difference in the simple effects of increasing copper across levels of Nitrogen.

This is just finding the difference between **Part B** and **Part C**. For purposes of this problem, we will call this estimate $\hat{\theta_3} = \hat{\theta_1} - \hat{\theta_2} = \bar{y_3} - \bar{y_1} - (\bar{y_4} - \bar{y_2}) = \bar{y_3} - \bar{y_1} - \bar{y_4} + \bar{y_2} = -13 - (-3) = 16$. Therefore, this difference shown by $\hat{\theta_3} = 16$.

1.5 Part E

Using significance level $\alpha = .05$, test the hypothesis that the simple effects of copper are constant across levels of nitrogen.

Here we are going to do a F-test to see if this simple effects of copper are consistent across the different levels of nitrogen. Our null hypothesis is that the effects of copper are consistent across all levels of nitrogen. The alternative hypothesis is that they are not. We are going to use our result from **Part D** to help us get our appropriate F-value. In this case, we have shown that $\hat{\theta_3} = 16$. Now let us note the contrast which is $(-1, 1, 1, -1)'$ which will get us our c_i values later. To get our F-value, we must use the formula $F = \frac{SS(\hat{\theta_3})/(a-1)(b-1)}{MS(E)}$ where $a = 2$ because there are only 2 nitrogen levels, $b = 2$ because there are only

2 copper levels, $SS(\hat{\theta_3}) = \frac{(\hat{\theta_3})^2}{\sum_{i=1}^{4} \frac{c_i^2}{n_i}} = \frac{16^2}{\frac{-1^2}{4}+\frac{1^2}{4}+\frac{1^2}{4}+\frac{-1^2}{4}} = \frac{256}{4*\frac{1}{4}} = \frac{256}{1} = 256$, $MS(E) = \frac{1}{4} * \sum_{i=1}^{4} s_i^2 = 36 + 16 + 58.6666667 + 57.3333333 = \frac{1}{4} * 168 = 42$. Now that we have found all the values for our F-test we can get our F-value to be $F = \frac{SS(\hat{\theta_3})/(a-1)(b-1)}{MS(E)} = \frac{256/(2-1)(2-1)}{42} = \frac{256}{42} = 6.095238$ on degrees of freedom $df = (a-1)*(b-1), ab(n-1) = (2-1)(2-1), 2*2(4-1) = 1*1, 4*3 = 1, 12$. We can get our F-critical value (or F^*) by using the `qf()` function in R. Our F^* value is `qf(1-alpha, df1, df2)` $=$ `qf(1-.05, 1, 12)` $= 4.7472253$. Since the F-value we obtained is greater the F-critical value ($6.095238 > 4.7472253$) then we reject our null hypothesis and say that the simple effects of copper are **not** consistent across the different levels of nitrogen.

1.6 Part F

Report the smallest level of significance at which the difference between simple copper effects across levels of nitrogen may be declared significant.

I am assuming for this problem that we are essentially finding the p-value we obtain from our F-value and use that as the smallest significance level we can use for the difference between copper effects still be declared significant. Remember for hypothesis testing, we declare significance when the p-value we obtain is less than our significance level (α). Note in **Part E** we showed that the F-value we obtained was 6.095238. Now we can use the `pf()` function to get the smallest level of significance at which we can still declare significance. To do this, we will put into the function `pf(F-value, df1, df2, lower.tail = F)` $=$ `pf(6.095238, 1, 12, lower.tail = F)` $= 0.0295555$. Therefore, the smallest level of significance at which the difference between simple copper effects across levels of nitrogen may be declared significant is 0.0295555.

1.7 Part G

Report a contrast sum of squares associated with the contrast tested in part (d).

Remember in **Part E**, we said this contrast was $(-1, 1, 1, -1)'$. We can use this for the c_i values in our sum of squares calculation. We are going to call this sum of squares calculation $SS(\hat{\theta_3})$. Therefore, $SS(\hat{\theta_3}) = \frac{(\hat{\theta_3})^2}{\sum_{i=1}^{4} \frac{c_i^2}{n_i}} = \frac{16^2}{\frac{-1^2}{4}+\frac{1^2}{4}+\frac{1^2}{4}+\frac{-1^2}{4}} = \frac{256}{4*\frac{1}{4}} = \frac{256}{1} = 256$. This was also calculated in **Part E** for our F-test, but this is now reinforcing what we did to get this value.

1.8 Part H

Estimate the simple effect of increasing N when $Cu = 100$. Report a standard error and a 95% confidence interval for the effect. In light of this interval, can you declare the observed effect "significant" at level of significance $\alpha = .05$?

First let us find the simple effect of increasing N when $Cu = 100$. We are going to call this effect $\hat{\theta_4} = \bar{y_4} - \bar{y_3} = 127 - 146 = -19$.

Next we are going to report a standard error which is $\hat{SE}(\sum c_i \bar{y_{i.}}) = \sqrt{MS(E) * \sum \frac{c_i^2}{n_i}}$. Now, we know this contrast is just $(0, 0, -1, 1)'$ because we said that $\hat{\theta_4} = \bar{y_4} - \bar{y_3}$. We also said in **Part E** that the $MS(E) = \frac{1}{4} * \sum_{i=1}^{4} s_i^2 = 36+16+58.6666667+57.3333333 = \frac{1}{4}*168 = 42$. So we can say that $\hat{SE}(\sum c_i \bar{y_{i.}}) = \sqrt{MS(E) * \sum \frac{c_i^2}{n_i}} = \sqrt{42 * (\frac{0^2}{4} + \frac{0^2}{4} + \frac{-1^2}{4} + \frac{1^2}{4})} = \sqrt{42 * (\frac{2}{4})} = \sqrt{21} = 4.5825757$. So, $\hat{SE}(\sum c_i \bar{y_{i.}}) = 4.5825757$.

Now we can get the confidence interval for these two effects, which is just saying $\hat{\theta_4} \pm t(\alpha/2, N-t) * \sqrt{MS(E) * \sum \frac{c_i^2}{n_i}}$ Now we know that $\hat{\theta_4} = -19$, $\alpha = 0.05$ since it is a 95% confidence interval, $N - t =$

$ab(n-1) = 2*2(4-1)4*3 = 12$, and $\sqrt{MS(E)*\sum\frac{c_i^2}{n_i}} = \hat{SE}(\sum c_i\bar{y_i.}) = \sqrt{21}$. Therefore, $\hat{\theta_4} \pm t(\alpha/2, N-t)*\sqrt{MS(E)*\sum\frac{c_i^2}{n_i}} = -19 \pm t(0.05/2, 12)*\sqrt{21} = -19 \pm t(0.025, 12)*\sqrt{21}$. So our 95% confidence interval is $(-19 - t(0.025, 12)*\sqrt{21}, -19 + t(0.025, 12)*\sqrt{21})$. Now $t(0.025, 12) =$ `qt(0.025, 12, lower.tail = F)` $= 2.1788128$. Therefore our confidence interval when plugging the t-value of $t(0.025, 12)$ into it which was $(-19 - t(0.025, 12)*\sqrt{21}, -19 + t(0.025, 12)*\sqrt{21})$ is (-28.9845747, -9.0154253).

Lastly, since we can see that 0 is not contained in our 95% confidence interval we have statistically significant evidence to conclude (or declare) that the observed effect is "significant" at level of significance of $\alpha = .05$.

1.9 Part I

Estimate the main effect of increasing Cu. Give the F-ratio for a test of no effect, along with degrees of freedom.

We can estimate the main effects of increasing Cu by finding the mean of the FCR scores when $Cu = 100$ and then subtract the mean scores when $Cu = 10$. Therefore, we can say that this estimation which we will call $\hat{\theta_5} = \frac{1}{2}*(\bar{y_3} + \bar{y_4}) - \frac{1}{2}*(\bar{y_1} + \bar{y_2}) = \frac{1}{2}*(\bar{y_3} + \bar{y_4} - \bar{y_1} - \bar{y_2}) = \frac{1}{2}*(146 + 127 - 133 - 130) = 5$. Therefore the estimate on the main effect of increasing Cu is $\hat{\theta_5} = 5$.

Now we are going to get a F-ratio for an F-test for a test of no effect. We know that we can get the F-value from the equation $F = \frac{SS(\hat{\theta_5})/df_1}{MS(E)}$ where $SS(\hat{\theta_5}) = \frac{\hat{\theta_5}^2}{\frac{1}{n_{group_1}} + \frac{1}{n_{group_2}}} = \frac{5^2}{\frac{1}{4+4} + \frac{1}{4+4}} = \frac{25}{\frac{2}{8}} = 100$, $df_1 = 1$ since we are only 2 different options (and then subtract 1 to get 1 numerator degree of freedom), and from **Part E** we showed the $MS(E) = \frac{1}{4}*\sum_{i=1}^{4} s_i^2 = 36 + 16 + 58.6666667 + 57.3333333 = \frac{1}{4}*168 = 42$. Therefore our F-value is $F = \frac{SS(\hat{\theta_5})/df_1}{MS(E)} = \frac{100/1}{42} = 2.3809524$ which has 1 numerator degree of freedom and $ab(n-1) = 2*2(4-1) = 12$ denominator degrees of freedom. Therefore our F-ratio is $F = 2.3809524$ with 1, 12 degrees of freedom. Note we cannot make a conclusion since we are not given a significance level.

1.10 Part J

Given the analysis you've done so far, is it appropriate to say the the observed effect of copper in this experiment is not significant (using level $\alpha = .05$)? Explain.

From **Part I** we showed the F-ratio we are using to do this test is $F = 2.3809524$ with 1, 12 degrees of freedom. To compare this to our F-critical value, we can get the F-critical value for a significance level of 0.05 by using the `qf()` function and inputting `qf(alpha, df1, df2, lower.tail = F)` which in this case is `qf(0.05, 1, 12, lower.tail = F)` which gives us the corresponding F-critical value of 4.7472253. Since this F-critical value is larger than the F-value we calculated from our test in **Part I** we fail to reject the null hypothesis that there is no effect of increasing copper. Therefore it is appropriate to say the the observed effect of copper in this experiment is not significant.

1.11 Part K

Report the contrast sums of squares for the main effect of copper and the main effect of nitrogen.

From **Part J** we found the contrast sums of squares for the main effect of copper which was done by $SS(\hat{\theta_5}) = \frac{\hat{\theta_5}^2}{\frac{1}{n_{group_1}} + \frac{1}{n_{group_2}}} = \frac{5^2}{\frac{1}{4+4} + \frac{1}{4+4}} = \frac{25}{\frac{2}{8}} = 100$. So the contrast sums of squares for the main effect of copper is $SS(\hat{\theta_5}) = 100$

Now let us find the estimate of the main effect of increasing nitrogen by finding the mean of the FCR scores when $N = 45$ and then subtract the mean scores when $N = 25$. Therefore, we can say that this estimation which we will call $\hat{\theta_6} = \frac{1}{2}*(\bar{y_2} + \bar{y_4}) - \frac{1}{2}*(\bar{y_1} + \bar{y_3}) = \frac{1}{2}*(\bar{y_2} + \bar{y_4} - \bar{y_1} - \bar{y_3}) = \frac{1}{2}*(130 + 127 - 133 - 146) = 5$.

Therefore the estimate on the main effect of increasing Cu is $\hat{\theta_5} = -11$. Now we can find the contrast sums of squares for the main effect of nitrogen which can be done by $SS(\hat{\theta_6}) = \frac{\hat{\theta_6}^2}{\frac{1}{n_{group_1}} + \frac{1}{n_{group_2}}} = \frac{-11^2}{\frac{1}{4+4} + \frac{1}{4+4}} = \frac{121}{\frac{2}{8}} = 484$. So the contrast sums of squares for the main effect of nitrogen is $SS(\hat{\theta_6}) = 484$.

1.12 Part L

Obtain an ANOVA table which partitions the variability between the four treatments into meaningful components.

We found the sum of squares for Copper, Nitrogen, and Nitrogen:Copper (interaction). Keep in mind the degrees of freedom are all 1 since there are only 2 different values of Copper and 2 different values of Nitrogen. So $df_{Nitrogen} = 2 - 1 = 1$, $df_{Copper} = 2 - 1 = 1$, $df_{Nitrogen:Cooper} = (2 - 1)(2 - 1) = 1$. Moreover since the degrees of freedom are all 1 for these parts then the sum of squares for these values equals their respective mean squared values too (because of dividing the sum of squares values by 1 degree of freedom). Next, we know that the $MS(E) = 42$ from **Part E**. Going backwards, we have shown in earlier parts that the error degrees of freedom is 12 (from $ab(n-1) = 4(2*2-1) = 12$). So the $SS(Error) = MS(E)*df_{Error} = 42*12 = 504$. Lastly we can get the corrected total values by saying that degrees of freedom is $a*b-1 = 4*4-1 = 15$ and the sum of squares value is adding up the previous 4 sum of squares which is $100+484+256+504 = 1344$.

Next we need to show F-values. The F-value is shown by doing $F = \frac{SS(\hat{\theta}_i)/df}{MS(E)}$. We are going to show the different F-values below. - For Copper: $F = \frac{100/1}{42} = 2.3809524$ on degrees of freedom 1, 12 - For Nitrogen: $F = \frac{484/1}{42} = 11.5238095$ on degrees of freedom 1, 12 - For Copper and Nitrogen Interaction: $F = \frac{256/1}{42} = 6.0952381$ on degrees of freedom 1, 12

Lastly, we need to get the P-values for the respective things we are modeling. Thus, we are going to use the `pf()` function to do this by inputting `pf(F-value, df1, df2, lower.tail = F)`. We will show this below. - For Copper: p-value is `pf(100/42, 1, 12, lower.tail = F)` = 0.1487723 - For Nitrogen: p-value is `pf(484/42, 1, 12, lower.tail = F)` = 0.0053218 - For Copper and Nitrogen Interaction: p-value is `pf(256/42, 1, 12, lower.tail = F)` = 0.0295555

Now we can make our ANOVA table.

Source	DF	Sum of Squares	Mean Square	F-value	P-value
Copper	1	100	100	2.3809524	0.1487723
Nitrogen	1	484	484	11.5238095	0.0053218
Copper and Nitrogen Interaction	1	256	256	6.0952381	0.0295555
Error	12	504	42		
Corrected Total	15	1344			

Similarly we could make an ANOVA table combining all treatments (and interaction in on thing called "Model"). Here we would add the sum of squares for copper, nitrogen, and its interaction which would be $100 + 484 + 256 = 840$ with degrees of freedom $1 + 1 + 1 = 3$. The mean square value would be the sum of squares divided by the degrees of freedom which is $840/3 = 280$. The F-value would be this mean square value divided by the MS(E) which is $280/42 = 6.6666667$ on 3, 12 degrees of freedom with the p-value being `pf(280/42, 3, 12, lower.tail = F)` = 0.0067142.

Now we can make this version of the ANOVA table.

Source	DF	Sum of Squares	Mean Square	F-value	P-value
Model	3	840	280	6.6666667	0.0067142
Error	12	504	42		
Corrected Total	15	1344			

And now we have made appropriate ANOVA tables showing the effects of copper and nitrogen (along with its interaction effect).

1.13 Part M

Briefly characterize the observed effects of copper and nitrogen on FCR, reporting appropriate p-values along the way.

Here we are going to look at three things: if copper alone has an effect on FCR, if nitrogen alone has an effect on FCR, and if the interaction between copper and nitrogen has an effect on FCR. Please note the ANOVA table we used for **Part L** which is attached below.

Source	DF	Sum of Squares	Mean Square	F-value	P-value
Copper	1	100	100	2.3809524	0.1487723
Nitrogen	1	484	484	11.5238095	0.0053218
Copper and Nitrogen Interaction	1	256	256	6.0952381	0.0295555
Error	12	504	42		
Corrected Total	15	1344			

- Does copper have an effect on FCR? What we are going to test here is if copper does have this effect. In this case, we called copper α_i so we are going to test using the null hypothesis that $H_0 : \alpha_i = 0$ with the alternative hypothesis being $H_A : \alpha_i \neq 0$. From **Part L** and what is attached to this part, we made an ANOVA table that went through and did this test by an appropriate F-test. Remember our F-value is calculated by $F = \frac{SS(Copper)/df_{Copper}}{MS(E)} = \frac{100/1}{42} = 2.3809524$. We then find the appropriate p-value using the F-value and the appropriate degrees of freedom which is 1, 12 (1 from the degrees of freedom for copper and 12 being the error degrees of freedom). We use the function `pf()` with inputs of the F-value we got along with the first and second degrees of freedom. Thus, we should be getting `pf(100/42, 1, 12, lower.tail = F)` which gives us the p-value 0.1487723. Most of the time, we use a significance level of $\alpha = 0.05$. Since our p-value is larger than our alpha level, we do not have significantly significant evidence to say that the copper amount has an effect on FCR. Therefore, it is plausible to say that copper might not have an effect on FCR.
- Does nitrogen have an effect on FCR? What we are going to test here is if nitrogen does have this effect. In this case, we called nitrogen β_j so we are going to test using the null hypothesis that $H_0 : \beta_j = 0$ with the alternative hypothesis being $H_A : \beta_j \neq 0$. From **Part L** and what is attached to this part, we made an ANOVA table that went through and did this test by an appropriate F-test. Remember our F-value is calculated by $F = \frac{SS(Nitrogen)/df_{Nitrogen}}{MS(E)} = \frac{484/1}{42} = 11.5238095$. We then find the appropriate p-value using the F-value and the appropriate degrees of freedom which is 1, 12 (1 from the degrees of freedom for nitrogen and 12 being the error degrees of freedom). We use the function `pf()` with inputs of the F-value we got along with the first and second degrees of freedom. Thus, we should be getting `pf(484/42, 1, 12, lower.tail = F)` which gives us the p-value 0.0053218. Most of the time, we use a significance level of $\alpha = 0.05$. Since our p-value is much smaller than our alpha level (and most significance levels people tend to use), we have significantly significant evidence to say

that the nitrogen amount has an effect on FCR. Therefore, it is plausible to say that nitrogen has an effect on FCR.

- Does the copper and nitrogen interaction have an effect on FCR? What we are going to test here is if the copper and nitrogen interaction has this effect. In this case, we called the copper and nitrogen interaction $(\alpha\beta)_{ij}$ so we are going to test using the null hypothesis that $H_0 : (\alpha\beta)_{ij} = 0$ with the alternative hypothesis being $H_A : (\alpha\beta)_{ij} \neq 0$. From **Part L** and what is attached to this part, we made an ANOVA table that went through and did this test by an appropriate F-test. Remember our F-value is calculated by $F = \frac{SS(Copper*Nitrogen)/df_{Copper*Nitrogen}}{MS(E)} = \frac{256/1}{42} = 6.0952381$. We then find the appropriate p-value using the F-value and the appropriate degrees of freedom which is 1, 12 (1 from the degrees of freedom for the copper and nitrogen interaction and 12 being the error degrees of freedom). We use the function `pf()` with inputs of the F-value we got along with the first and second degrees of freedom. Thus, we should be getting `pf(256/42, 1, 12, lower.tail = F)` which gives us the p-value 0.0295555. Most of the time, we use a significance level of $\alpha = 0.05$. Since our p-value is smaller than our alpha level, we have significantly significant evidence to say that the copper and nitrogen interaction amount has an effect on FCR. Therefore, it is plausible to say that the copper and nitrogen interaction has an effect on FCR. Now please note we would be saying the exact opposite if our significance level (α) was less than the p-value. In this case, we would say that we do not have significantly significant evidence to say that the copper and nitrogen interaction amount has an effect on FCR; moreover, it would be plausible to say that the copper and nitrogen interaction might not have an effect on FCR. So it all depends with the interaction term of what significance level you want to use.

In conclusion, from doing our appropriate statistical tests we have found that it is plausible to say that copper has an effect on FCR, nitrogen does not have an effect on FCR, and the copper and nitrogen interaction has an effect on FCR (as long as the significance level is the standard α level is 0.05 – note if the significance level is less than 0.0295555 then we would be saying that is plausible to say that the copper and nitrogen interaction does not have an effect on FCR).

1.14 Part N

It turns out that a control was also run (with $n = 4$), without any added Cu or N. The mean FCR was $\bar{y_0} = 131$. The observed contrast of this mean with the average of the others is $\hat{\theta} = 131 - \frac{1}{4}(133 + 130 + 146 + 127) = -3$. Compute $SS(diet)$, the diet sum of squares (on $df = 5 - 1 = 4$) from a one-way analysis of variance using all five diets.

We know that the sum of squares for the control (or this contrast which is $(1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4})'$) is equal to $SS(control) = SS(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum_{i=1}^{t} \frac{c_i^2}{n_i}}$. We know in this case that $\hat{\theta} = -3$ and $\sum_{i=1}^{t} \frac{c_i^2}{n_i} = \frac{1^2}{4} + \frac{\frac{-1}{4}^2}{4} + \frac{\frac{-1}{4}^2}{4} + \frac{\frac{-1}{4}^2}{4} + \frac{\frac{-1}{4}^2}{4} = \frac{1^2}{4} + 4 * \frac{\frac{-1}{4}^2}{4} = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}$. Now $SS(control) = SS(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum_{i=1}^{t} \frac{c_i^2}{n_i}} = \frac{-3^2}{\frac{5}{16}} = 28.8$. Therefore the Sum of Squares of just the control is 28.8.

Now to get the sum of squares for diet we need to add $SS(Diet) = SS(copper) + SS(nitrogen) + SS(copper : nitrogen) + SS(control) = 100 + 484 + 256 + 28.8 = 868.8$. Therefore, $SS(Diet) = 868.8$.

2 Problem 2

First we are going to read in the data.

```
library(tidyverse)
(barley <- read_table("barley.dat"))
```

```
## # A tibble: 30 x 3
##    seedage   h20     y
##      <dbl> <dbl> <dbl>
##  1       1     1    11
##  2       1     2     8
##  3       2     1     7
##  4       2     2     1
##  5       3     1     9
##  6       3     2     5
##  7       4     1    13
##  8       4     2     1
##  9       5     1    20
## 10       5     2    11
## # i 20 more rows
```

2.1 Part A

Posit a factorial effects model for these data. Why might the homogeneity of variance assumption might be questionable?

We can say that this factorial effects model is $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$ where for $i = 1, 2, 3, 4, 5$ and $j = 1, 2$ and E_{ijk} are i.i.d $N(0, \sigma^2)$ errors. We can store this model below.

```
barleyModel <- lm(y ~ as.factor(seedage) + as.factor(h20) + as.factor(seedage)*as.factor(h20), barley)
```

The homogeneity of variance assumption might be questionable because where you plant the seeds (the land plot and its soil) could cause a difference in the seed production and thus these confounding variables could cause unequal variance.

2.2 Part B

Report the p-value for a test of interaction between water and seed age.

Here we can make an ANOVA test for our model we made before in **Part A**.

```
anova(barleyModel)
```

```
## Analysis of Variance Table
##
## Response: y
##                                   Df  Sum Sq Mean Sq F value   Pr(>F)
## as.factor(seedage)                 4 1321.13  330.28  5.5293 0.003645 **
## as.factor(h20)                     1 1178.13 1178.13 19.7232 0.000251 ***
## as.factor(seedage):as.factor(h20)  4  208.87   52.22  0.8742 0.496726
## Residuals                         20 1194.67   59.73
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

As we can see from our ANOVA table, the p-value is 0.4967256, which is a fairly high value that with most significance levels we use cause us to say there is not enough evidence of statistical significance for the interaction term.

If we wanted, we could have done this without the ANOVA function if we knew the MS(E) and the SS(Interaction). From the table, we could see that the $SS(Interaction) = 208.87$ and our $MS(E) = 52.22$. We also know that the interaction degrees of freedom is $(a-1)(b-1) = (5-1)(2-1) = 4*1 = 4$. We also know our error degrees of freedom is $a*b(n-1) = 5*2(3-1) = 10*2 = 20$. Now our F-value is $F = \frac{SS(interaction)/df1}{MS(E)} = \frac{208.87/4}{59.73} = 0.8742$ with degrees of freedom 4, 20. We can use the `pf()` function with inputs `pf(F-value, df1, df2, lower.tail=F)` = `pf(0.8742, 4, 20, lower.tail=F)` = 0.496705 which rounds to the same p-value we got from our ANOVA table. Therefore our p-value is about 0.4967.

2.3 Part C

What does it mean for the effects of water and seed age to be plausibly additive?

If we are saying the effects to be plausibly additive, we are saying that these effects of water and seed age do not interact and that each effect has an additive effect on the response variable which in this case is y (or the number of seeds that sprout). We know that additivity means that the effect of one independent variable(s) on the response variable does NOT depend on the value of another independent variable(s). So in this case the effect of seed age on the response variable (seeds sprouting) does not depend on the value of our other indpendent variable (water) and vice versa where the effect of water on the response variable (seeds sprouting) does not depend on the value of our other indpendent variable (seed age).

2.4 Part D

Obtain a 95% confidence interval for the main effect of water. Does more water lead to more sprouting seeds?

Here we are going to find the means of our different groupings of water and seed age.

```r
(barleyGroupMeans <- barley %>%
  group_by(h20, seedage) %>%
  summarize(mean = mean(y), count = length(y)))
```

```
## # A tibble: 10 x 4
## # Groups:   h20 [2]
##      h20 seedage  mean count
##    <dbl>   <dbl> <dbl> <int>
##  1     1       1  8.67     3
##  2     1       2 13.3      3
##  3     1       3 21        3
##  4     1       4 25.3      3
##  5     1       5 34        3
##  6     2       1  4.67     3
##  7     2       2  3.67     3
##  8     2       3  7.67     3
##  9     2       4  6.67     3
## 10     2       5 17        3
```

We have now shown the group means for each amount of water along with seed age for each plot that has seeds sprout. The next thing we are going to do is find the mean effect of water difference by subtracting the mean value of `h20 = 1` from the mean value of `h20 = 2` (or another way of saying this is mean(h20 = 2) - mean(h20 = 1)) and call this $\hat{\theta}$. Then we are going to make a 95% confidence interval by the formula: $\hat{\theta} \pm t(\alpha/2, a*b(n-1))\sqrt{MS(E)\sum \frac{c_i}{n_i}} = \hat{\theta} \pm t(\alpha/2, N-t)\sqrt{MS(E)\sum \frac{c_i}{n_i}}$ where we know that $\alpha = 0.05$, $MS(E) \approx 59.73$ from **Part B** ANOVA table, and $a*b(n-1) = 5*2(3-1) = 20$. We also know that the contrast is $(-1, 1)'$ with $n_i = 15$ (since there are 5 groups being looked at with 3 in each group so 15 total observations in each part of the contrast. Thus, $\sum \frac{c_i}{n_i} = \frac{-1^2}{15} + \frac{1^2}{15} = \frac{1}{15} + \frac{1}{15} = \frac{2}{15}$. We are going to show this confidence interval below with the values we have made.

```r
# Get the main effect (theta_hat)
theta_hat <- mean(barleyGroupMeans$mean[1:5]) - mean(barleyGroupMeans$mean[6:10])

# Make the confidence interval
mse <- anova(barleyModel)$"Mean Sq"[[4]]
df2 <- anova(barleyModel)$Df[[4]]
alpha <- 0.05
contrast_sum <- 2/15
multiplier <- qt(alpha/2, df2, lower.tail = F)

# Lower Bound
lower_bound <- theta_hat - (multiplier * sqrt(mse * contrast_sum))

# Upper bound
upper_bound <- theta_hat + (multiplier * sqrt(mse * contrast_sum))

# Final confidence interval
(confidence_interval <- c(lower_bound, upper_bound))
```

```
## [1]  6.646463 18.420203
```

As we can see our 95% confidence interval gives us a lower bound of 6.6464634 and an upper bound of 18.4202032. This means we are 95% confident that the main effect of water is between 6.6464634 and 18.4202032. Since 0 is not in the 95% confidence interval (and all values are positive since we found the effect by doing the mean of more water minus the mean of less water), we have statistically significant evidence that adding more water leads to more sprouting seeds. Lastly, I would say that these results should have some caution to them since in our model used to test this has questionable homogeneous variance (as said in **Part A**).

2.5 Part E

Estimate the *linear* effect of seed age. Are older seeds more likely to sprout?

We know that since our degrees of freedom for seed age is 4 we can have up to a 4-degree model (from linear, quadratic, cubic, and quartic). Also we know the contrast for the linear effect for seed age is $\hat{\theta_L} = -2\bar{y_{.1.}} - \bar{y_{.2.}} + 0\bar{y_{.3.}} + \bar{y_{.4.}} + 2\bar{y_{.5.}}$. Since we are analyzing a balanced design, we can get the following terms from **Part D**.

- $\bar{y_{.1.}} = \frac{\bar{y_{11.}} + \bar{y_{21.}}}{2}$ = (8.6666667 + 4.6666667) / 2 = 6.6666667
- $\bar{y_{.2.}} = \frac{\bar{y_{12.}} + \bar{y_{22.}}}{2}$ = (13.3333333 + 3.6666667) / 2 = 8.5
- $\bar{y_{.3.}} = \frac{\bar{y_{13.}} + \bar{y_{23.}}}{2}$ = (21 + 7.6666667) / 2 = 14.3333333
- $\bar{y_{.4.}} = \frac{\bar{y_{14.}} + \bar{y_{24.}}}{2}$ = (25.3333333 + 6.6666667) / 2 = 16
- $\bar{y_{.5.}} = \frac{\bar{y_{15.}} + \bar{y_{25.}}}{2}$ = (34 + 17) / 2 = 25.5

Now we can plug those values into our formula to find the linear effect of seed age. This is $\hat{\theta_L} = -2\bar{y_{.1.}} - \bar{y_{.2.}} + 0\bar{y_{.3.}} + \bar{y_{.4.}} + 2\bar{y_{.5.}}$ = -2 x 6.6666667 - 8.5 + 0 x 14.3333333 + 16 + 2 x 25.5 = 45.1666667. Therefore, the linear effect of seed age is 45.1666667.

2.6 Part F

After averaging over to two levels of water, test for lack of fit of a model in which the effect of seed age on sprouting frequency is linear.

Here we are going to do a lack of fit test. Again since the degrees of freedom for the seed age is 4 we can have up to a 4^{th} level polynomial model. We are going to do a lack of fit test where we see if the effect of seed age on sprouting frequency is plausibly linear. Our H_0: The effect of seed age on sprouting frequency is plausibly linear and H_A: The effect of seed age on sprouting frequency is NOT plausibly linear.

We are going to do this by saying that $F = \frac{(SS(Trt)-SS(\hat{\theta_L}))/t-1-p}{MS(E)}$ where t is the number of treatment levels, p is the reduced model's degree polynomial, and $SS(\hat{\theta_L}) = \frac{\hat{\theta_L}^2}{\sum_{i=1}^{5}\frac{c_i^2}{n_i}}$. First we know that $t = 5$ for 5 levels of treatment for seed age. We also know $p = 1$ since in linear regression it only uses the one degree in the polynomial model. We found $SS(Trt) = 1321.13$ in our anova table in **Part B** and $MS(E) = 59.73$ also found in **Part B**. Lastly we know that $\hat{\theta_L} = 45.16667$ from **Part E** which we need to find $SS(\hat{\theta_L}) = \frac{\hat{\theta_L}^2}{\sum_{i=1}^{5}\frac{c_i^2}{n_i}} = \frac{45.16667^2}{\frac{-2^2}{6}+\frac{-1^2}{6}+\frac{0^2}{6}+\frac{1^2}{6}+\frac{2^2}{6}} = \frac{2040.028}{\frac{4}{6}+\frac{1}{6}+\frac{0}{6}+\frac{1}{6}+\frac{4}{6}} = \frac{2040.028}{\frac{10}{6}} = 1224.017$. Now that we have everything we can find $F = \frac{(SS(Trt)-SS(\hat{\theta_L}))/t-1-p}{MS(E)} = \frac{(1321.13-1224.017)/5-1-1}{59.73} = \frac{97.113/3}{59.73} = \frac{32.371}{59.73} = 0.54196$ on degrees of freedom $t-1-p, ab(n-1) = 5-1-1, 5*2(3-1) = 3, 20$. Now we want to compare our F-value of 0.54196 with 3, 20 degrees of freedom. We know that since our F-value < 1, it will be less than our critical value. Therefore, we fail to reject our H_0 and can say that the effect of seed age on sprouting frequency is plausibly linear as it does not suffer from a lack of fit.

3 Problem 3

First we are going to read in the data.

```
(shrimp <- read_table("shrimp.dat"))
```

```
## # A tibble: 36 x 4
##        a     b     c     y
##    <dbl> <dbl> <dbl> <dbl>
##  1    25    80    10    86
##  2    25    80    10    52
##  3    25    80    10    73
##  4    25    80    25   544
##  5    25    80    25   371
##  6    25    80    25   482
##  7    25    80    40   390
##  8    25    80    40   290
##  9    25    80    40   397
## 10    25   160    10    53
## # i 26 more rows
```

3.1 Part A

Estimate each of these contrasts.

- $\theta_1 = -\mu_{1.1} + 0\mu_{1.2} + \mu_{1.3}$
- $\theta_2 = -\mu_{2.1} + 0\mu_{2.2} + \mu_{2.3}$

- $\theta_3 = \mu_{1.1} - 2\mu_{1.2} + \mu_{1.3}$
- $\theta_4 = \mu_{2.1} - 2\mu_{2.2} + \mu_{3.3}$
- $\theta_5 = \theta_1 - \theta_2$
- $\theta_6 = \theta_3 - \theta_4$

The first thing we want to do is find the means of y (and standard deviations if we want) of all the groupings of variable a and variable c since for example $\mu_{1.1}$ is saying that we want the mean of our response variable when looking at the first (or lowest level) of our first variable (which is variable a in this case), averaging over second variable (variable b), and then finding the first (or lowest level) of our third variable (which is variable c in this case). I am going to do some `group_by()` and `mean()` along with the `sd()` and the `length()` or count of how many responses we have for each grouping.

```
(
  shrimpGroupMeans <- shrimp %>%
    group_by(a, c) %>%
    summarize(
      mean = mean(y),
      sd = sd(y),
      count = length(y)
    )
)
```

```
## # A tibble: 6 x 5
## # Groups:   a [2]
##       a     c  mean    sd count
##   <dbl> <dbl> <dbl> <dbl> <int>
## 1    25    10  70.5  15.1     6
## 2    25    25 399.  114.      6
## 3    25    40 306.   70.0     6
## 4    35    10 370.   56.5     6
## 5    35    25 293.   45.4     6
## 6    35    40 237.   38.1     6
```

Now that we have our means we can now estimate the contrasts. For purpose of this problem, variable a is temperature, variable b is density, and variable c is salinity.

- $\hat{\theta_1} = -\hat{\mu_{1.1}} + 0\hat{\mu_{1.2}} + \hat{\mu_{1.3}} = -\hat{\mu_{1.1}} + \hat{\mu_{1.3}}$. From looking at our means table, we can see that $\hat{\mu_{1.1}} =$ 70.5 because it is where we find our first (lowest) value of variable a and our first (lowest) value of variable c and $\hat{\mu_{1.3}} =$ 305.6666667 because it is where we find our first (lowest) value of variable a and our third (highest) value of variable c. Therefore, $\hat{\theta_1} = -\hat{\mu_{1.1}} + 0\hat{\mu_{1.2}} + \hat{\mu_{1.3}} = -\hat{\mu_{1.1}} + \hat{\mu_{1.3}} =$ -70.5 + 305.6666667 = 235.1666667
- $\hat{\theta_2} = -\hat{\mu_{2.1}} + 0\hat{\mu_{2.2}} + \hat{\mu_{2.3}} = -\hat{\mu_{2.1}} + \hat{\mu_{2.3}}$. From looking at our means table, we can see that $\hat{\mu_{2.1}} =$ 369.5 because it is where we find our second (highest) value of variable a and our first (lowest) value of variable c and $\hat{\mu_{2.3}} =$ 236.8333333 because it is where we find our second (highest) value of variable a and our third (highest) value of variable c. Therefore, $\hat{\theta_2} = -\hat{\mu_{2.1}} + 0\hat{\mu_{2.2}} + \hat{\mu_{2.3}} = -\hat{\mu_{2.1}} + \hat{\mu_{2.3}} =$ -369.5 + 236.8333333 = -132.6666667
- $\hat{\theta_3} = -\hat{\mu_{1.1}} + 2\hat{\mu_{1.2}} + \hat{\mu_{1.3}}$ From looking at our means table, we can see that $\hat{\mu_{1.1}} =$ 70.5 because it is where we find our first (lowest) value of variable a and our first (lowest) value of variable c, $\hat{\mu_{1.2}} =$ 399.3333333 because it is where we find our first (lowest) value of variable a and our second (middle) value of variable c, and $\hat{\mu_{1.3}} =$ 305.6666667 because it is where we find our first (lowest) value of variable a and our third (highest) value of variable c. Therefore, $\hat{\theta_3} = -\hat{\mu_{1.1}} + 2\hat{\mu_{1.2}} + \hat{\mu_{1.3}} =$ 70.5 - 2 x 399.3333333 + 305.6666667 = -422.5

- $\hat{\theta_4} = -\hat{\mu_{2.1}} + 2\hat{\mu_{2.2}} + \hat{\mu_{2.3}}$ From looking at our means table, we can see that $\hat{\mu_{2.1}} = 369.5$ because it is where we find our second (highest) value of variable a and our first (lowest) value of variable c, $\hat{\mu_{2.2}} = 293.1666667$ because it is where we find our second (highest) value of variable a and our second (middle) value of variable c, and $\hat{\mu_{2.3}} = 236.8333333$ because it is where we find our second (highest) value of variable a and our third (highest) value of variable c. Therefore, $\hat{\theta_4} = -\hat{\mu_{2.1}} + 2\hat{\mu_{2.2}} + \hat{\mu_{2.3}} =$ 369.5 - 2 x 293.1666667 + 236.8333333 = 20
- $\hat{\theta_5} = \hat{\theta_1} - \hat{\theta_2}$ We know that $\hat{\theta_1} =$ 235.1666667 and $\hat{\theta_2} =$ -132.6666667 so $\hat{\theta_5} = \hat{\theta_1} - \hat{\theta_2} =$ 235.1666667 - -132.6666667 = 367.8333333
- $\hat{\theta_6} = \hat{\theta_3} - \hat{\theta_4}$ We know that $\hat{\theta_3} =$ -422.5 and $\hat{\theta_4} = 20$ so $\hat{\theta_6} = \hat{\theta_3} - \hat{\theta_4} =$ -422.5 - 20 = -442.5

3.2 Part B

Identify which estimates differ significantly from 0. Use the results to characterize the interaction between temperature and salinity.

First we are going to make an ANOVA table of our full model to help get the MS(E) value that we will use to do statistically testing on our contrasts.

```
(anova_table3 <- anova(lm(y ~ factor(a)*factor(b)*factor(c), shrimp)))
```

```
## Analysis of Variance Table
##
## Response: y
##                               Df Sum Sq Mean Sq F value         Pr(>F)
## factor(a)                      1  15376   15376  5.2952        0.03038 *
## factor(b)                      1  21219   21219  7.3073        0.01242 *
## factor(c)                      2  96762   48381 16.6615 0.000029012870 ***
## factor(a):factor(b)            1   8711    8711  2.9999        0.09610 .
## factor(a):factor(c)            2 300855  150428 51.8041 0.000000001959 ***
## factor(b):factor(c)            2    674     337  0.1161        0.89086
## factor(a):factor(b):factor(c)  2  24038   12019  4.1392        0.02855 *
## Residuals                     24  69691    2904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this table we can pull our MS(E) to be 2903.7777778. We will use that value as we do our F-tests on our contrasts. Note $F = \frac{SS(\hat{\theta_i})/(a-1)(c-1)}{MS(E)}$ on degrees of freedom (a-1)(c-1), abc(n-1). Now, $(a-1)(c-1) = (2-1)(3-1) = 1 * 2 = 2$ and $abc(n-1) = 2 * 2 * 3(3-1) = 12(2) = 24$. Plugging everything in our F-tests will consist of $F = \frac{SS(\hat{\theta_i})/(a-1)(c-1)}{MS(E)} = \frac{SS(\hat{\theta_i})/2}{2903.778}$ on degrees of freedom 2, 24. For each θ_i we are testing, we want the null hypothesis to be that $H_0 : \theta_i = 0$ with the alternative hypothesis being that $H_A : \theta_i \neq 0$. Also note, we will be using a significance level of $\alpha = 0.05$ so we are comparing if our F-value we obtain is larger than the F-critical value of F(.05, 2, 24) = 3.4028261.

- θ_1: The only part we are missing from our F-test is the $SS(\hat{\theta_1})$. We can get that by saying that $SS(\hat{\theta_1}) = \frac{\hat{\theta_1}^2}{\sum_{i=1}^{3} \frac{\frac{c^2}{i}}{n_i}} = \frac{235.1667^2}{\frac{-1^2}{6} + \frac{0^2}{6} \frac{1^2}{6}} = \frac{55303.38}{\frac{1}{6} + \frac{1}{6}} = \frac{55303.38}{\frac{2}{6}} = 165910.1$. Now that we have $SS(\hat{\theta_1}) = 165910.1$ we will plug it into our F-value formula to do our test. Here we see that $F = \frac{SS(\hat{\theta})/2}{2903.778} = \frac{165910.1/2}{2903.778} \approx 28.568$. This value we obtain is much greater than our F-critical value of 3.4028261 so θ_1 differs significantly from 0.
- θ_2: The only part we are missing from our F-test is the $SS(\hat{\theta_2})$. We can get that by saying that $SS(\hat{\theta_2}) = \frac{\hat{\theta_2}^2}{\sum_{i=1}^{3} \frac{\frac{c^2}{i}}{n_i}} = \frac{-132.6667^2}{\frac{-1^2}{6} + \frac{0^2}{6} \frac{1^2}{6}} = \frac{17600.45}{\frac{1}{6} + \frac{1}{6}} = \frac{17600.45}{\frac{2}{6}} = 52801.35$. Now that we have $SS(\hat{\theta_2}) = 52801.35$ we will

plug it into our F-value formula to do our test. Here we see that $F = \frac{SS(\hat{\theta_2})/2}{2903.778} = \frac{52801.35/2}{2903.778} \approx 9.092$. This value we obtain is much greater than our F-critical value of 3.4028261 so θ_2 differs significantly from 0.

- θ_3: The only part we are missing from our F-test is the $SS(\hat{\theta_3})$. We can get that by saying that $SS(\hat{\theta_3}) = \frac{\hat{\theta_3}^2}{\sum_{i=1}^{3} \frac{c_i^2}{n_i}} = \frac{-422.5^2}{\frac{1^2}{6} + \frac{2^2}{6} \frac{1^2}{6}} = \frac{178506.2}{\frac{1}{6}+\frac{4}{6}+\frac{1}{6}} = \frac{178506.2}{\frac{6}{6}} = 178506.2$. Now that we have $SS(\hat{\theta_3}) = 178506.2$ we will plug it into our F-value formula to do our test. Here we see that $F = \frac{SS(\hat{\theta_3})/2}{2903.778} = \frac{178506.2/2}{2903.778} \approx 30.737$. This value we obtain is much greater than our F-critical value of 3.4028261 so θ_3 differs significantly from 0.
- θ_4: The only part we are missing from our F-test is the $SS(\hat{\theta_4})$. We can get that by saying that $SS(\hat{\theta_4}) = \frac{\hat{\theta_4}^2}{\sum_{i=1}^{3} \frac{c_i^2}{n_i}} = \frac{20^2}{\frac{1^2}{6} + \frac{2^2}{6} \frac{1^2}{6}} = \frac{400}{\frac{1}{6}+\frac{4}{6}+\frac{1}{6}} = \frac{400}{\frac{6}{6}} = 400$. Now that we have $SS(\hat{\theta_2}) = 400$ we will plug it into our F-value formula to do our test. Here we see that $F = \frac{SS(\hat{\theta_4})/2}{2903.778} = \frac{400/2}{2903.778} \approx 0.06887579$. This value we obtain is much less than our F-critical value of 3.4028261 so we **cannot** say with statistical evidence that θ_4 differs significantly from 0.
- θ_5: The only part we are missing from our F-test is the $SS(\hat{\theta_5})$. We can get that by saying that $SS(\hat{\theta_5}) = \frac{\hat{\theta_5}^2}{\sum_{i=1}^{6} \frac{c_i^2}{n_i}} = \frac{367.8333^2}{\frac{-1^2}{6} + \frac{0^2}{6} \frac{1^2}{6} + \frac{1^2}{6} + \frac{0^2}{6} \frac{-1^2}{6}} = \frac{135301.3}{\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}} = \frac{135301.3}{\frac{4}{6}} = 202951.9$. Now that we have $SS(\hat{\theta_5}) = 202951.9$ we will plug it into our F-value formula to do our test. Here we see that $F = \frac{SS(\hat{\theta_5})/2}{2903.778} = \frac{202951.9/2}{2903.778} \approx 34.946$. This value we obtain is much greater than our F-critical value of 3.4028261 so θ_5 differs significantly from 0.
- θ_6: The only part we are missing from our F-test is the $SS(\hat{\theta_6})$. We can get that by saying that $SS(\hat{\theta_6}) = \frac{\hat{\theta_6}^2}{\sum_{i=1}^{6} \frac{c_i^2}{n_i}} = \frac{-442.5^2}{\frac{1^2}{6} + \frac{-2^2}{6} \frac{1^2}{6} + \frac{-1^2}{6} + \frac{2^2}{6} \frac{-1^2}{6}} = \frac{195806.2}{\frac{1}{6}+\frac{4}{6}+\frac{1}{6}+\frac{4}{6}+\frac{1}{6}+\frac{1}{6}} = \frac{195806.2}{\frac{12}{6}} = 97903.1$. Now that we have $SS(\hat{\theta_6}) = 97903.1$ we will plug it into our F-value formula to do our test. Here we see that $F = \frac{SS(\hat{\theta_6})/2}{2903.778} = \frac{97903.1/2}{2903.778} \approx 16.858$. This value we obtain is much greater than our F-critical value of 3.4028261 so θ_6 differs significantly from 0.

In conclusion, $\theta_1, \theta_2, \theta_3, \theta_5, \theta_6$ all differ while significantly from 0 while θ_4 does not. Since most of the contrasts show statistically significant evidence differing from 0 , we can say that interaction between temperature and salinity is present. We can also see in our ANOVA table in **Part A** that the interaction term between temperature and salinity was significant so this confirms our belief that interaction between these two variables remains present.