BTRY 6020 Homework VI

NAME: ANDRES CASTANO

NETID: ac986

DUE DATE: 8:40 am Monday April 24

Question 1.

An experiment was run to study how long mung bean seeds should be soaked prior to planting in order to promote early growth of bean sprouts. The experiment was run using a completely randomized design. Soaking levels used in this experiment were as follows: A= low, B= medium, C = high, and D = very high. For each treatment level, 17 beans were used and the mean shoot length (Y in mm) was measured 48 hours following soaking. Data appears in the file Hwk7Quest4Sp04.txt.

A) Perform analysis of variance to test the hypothesis that the four treatments' means are equal. State carefully your conclusions.

```
library(readxl)
data_bean = read_excel("Hwk6Q1DatSp17.xlsx")
head(data_bean)
```

```
##   Obs Treatment Length
## 1   1         A      5
## 2   2         A      8
## 3   3         A      5
## 4   4         A     11
## 5   5         A      3
## 6   6         A      4
```

```
data_bean$Treatment = factor(data_bean$Treatment)
model1.lm = lm(Length~Treatment, data=data_bean)
anova(model1.lm)
```

```
## Analysis of Variance Table
##
## Response: Length
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Treatment  3 2501.29  833.76  75.924 < 2.2e-16 ***
## Residuals 64  702.82   10.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results shows that we reject the null hypothesis that the four treatments means are equal (p-value $<$ 0.05), this means that there is at least one mean that is different from the others. To determine which means are different from the others, we need to perform a multiple comparison procedure.

B) Give a statistical model appropriate for describing the response variable in this study and explain each term in the model.

The inferential model for this study (cell means model) for this study is a oneway ANOVA:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Where y_{ij} is the jth observation in cell i (e.g., the cell receiving treatment i), with i=1,2,. . . ,4, and j=1,2,. . . ,17. In our case y_{ij} is the shoot length for the observation jth in the treatment i. μ_i is the mean shoot length in treatment i. ϵ_i is an error term. We also assume that $y_{ij} \sim indN(\mu_i, \sigma^2)$ and that $\epsilon_{ij} \sim i.i.dN(0, \sigma^2_\epsilon)$

C) Assess the validity of assumptions underlying analysis of variance in this study.

The assumptions of ANOVA are:

1) The observations y_{ij}, $i = 1, ..., k$ and $j = 1, ...n_k$ are independent between and within groups (treatments)

2) $y_{ij} \sim N(\mu_i, \sigma^2)$, this implies:

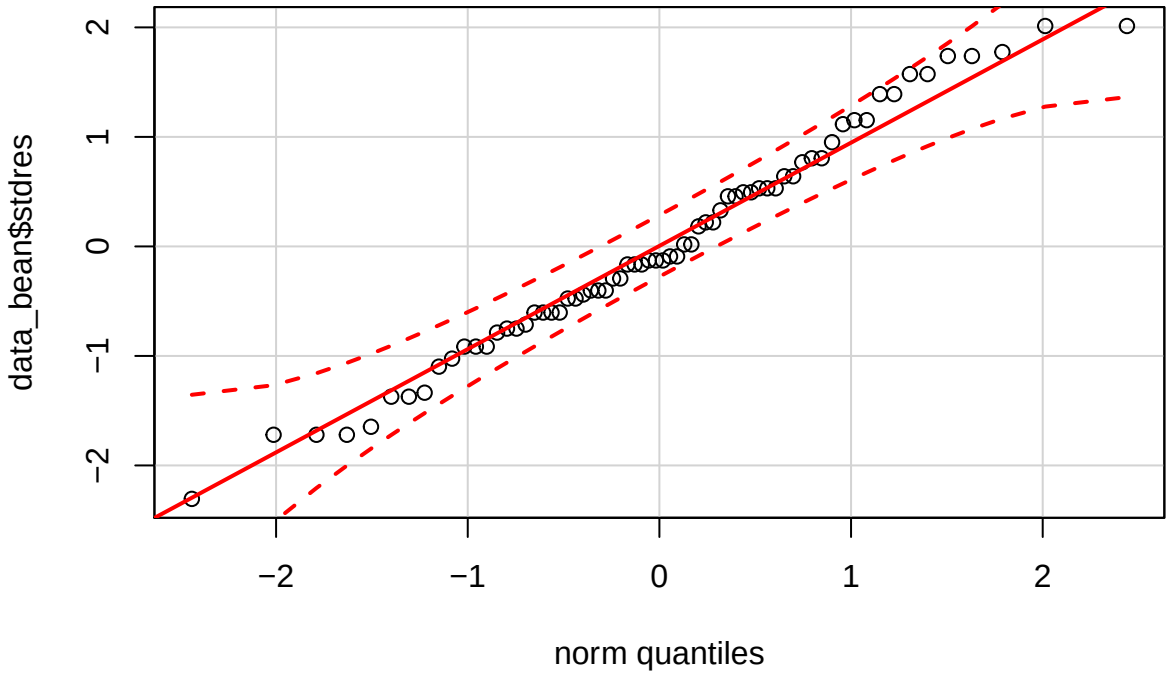
2.1) All n_i observations associated with treatment i are draw from a normal distribution with mean μ_i, or we cal also check whether the residuals of the regression are distributed normal.

2.2) The variance is the same for all groups (treatments) or equivalently constancy of error variance

In our case, the assumption of independence holds due that the experiment was implemented using a completely randomized design.

To check normality we can make a qqPlot for the shoot length by treatment:

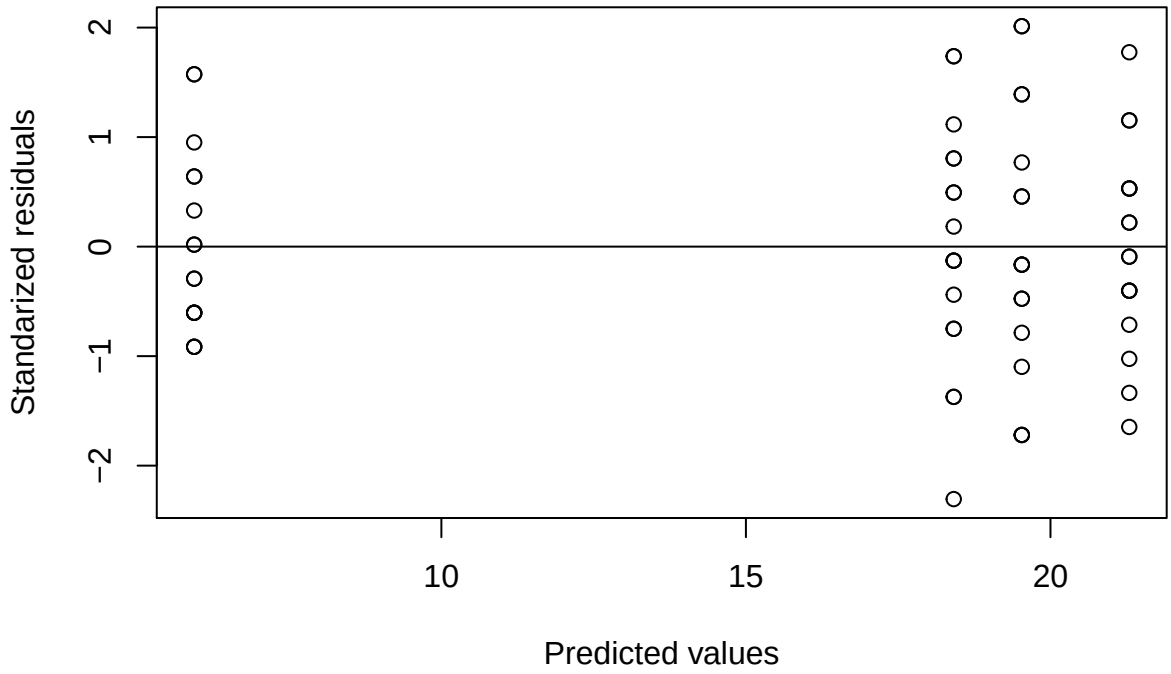
```
library(car)
data_bean$stdres=rstandard(model1.lm)
qqPlot(data_bean$stdres)
```



We can see that the assumption of normality holds.

To check the constancy of the error variance assumption, we can make a plot between the standardized residuals vs the predicted values:

```r
plot(model1.lm$fitted.values, data_bean$stdres, ylab="Standarized residuals", xlab="Predicted values", a
```



As we can see the scatter for each treatment seems to be similar, which means that the constancy of the error variance assumption is feasible. On the other hand, there is not seem to be outliers driving the conclusions.

D) i) Compare all pairs of means, using Bonferoni's method and make an interpretation of your results. Use $\alpha_{overall} = .05$.

```r
with(data_bean, pairwise.t.test(x=Length, g=Treatment, p.adjust.method = "bonferroni") )
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Length and Treatment
##
##   A       B     C
## B 1.4e-15 -     -
## C < 2e-16 1.000 -
## D < 2e-16 0.082 0.753
##
## P value adjustment method: bonferroni
```

The results shows:

- The mean lenght shoot for Treatment A (low) is different (lower) from the other three treatments (B=medium, C= high, and D=very high).
- There are not differences in the other pairwise comparisons.

ii) What are the advantages and disadvantages of using this method of pairwise comparisons?

The advantages of this methods are:

1) Is a good option when we want to perform a small number of pre-planned, non-orthogonal contrasts
2) Is easy to use
3) Widely implemented in the majority of softwares

The disadvantages:

1) Is too conservative and therefore with low power when we want to implement a large number of constrast.

E) i) Compare all pairs of means, using Tukey's method and make an interpretation of your results. Use $\alpha_{overall} = .05$.

```
levels(data_bean$Treatment) = c("Low", "Medium", "High", "Very High")
model2.lm = aov(Length~Treatment, data=data_bean)
TukeyHSD(model2.lm, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Length ~ Treatment, data = data_bean)
##
## $Treatment
##                        diff        lwr       upr     p adj
## Medium-Low       12.470588  9.4723100 15.468866 0.0000000
## High-Low         13.588235 10.5899571 16.586513 0.0000000
## Very High-Low    15.352941 12.3546630 18.351219 0.0000000
## High-Medium       1.117647 -1.8806311  4.115925 0.7594325
## Very High-Medium  2.882353 -0.1159253  5.880631 0.0638738
## Very High-High    1.764706 -1.2335723  4.762984 0.4128068
```

```
#plot(TukeyHSD(model2.lm))
```

- The mean lenght shoot for Treatment A (low) is different (lower) compared to the other three treatments (B=medium, C= high, and D=very high).
- There are not differences in the other pairwise comparisons.

ii) What are the advantages and disadvantages of using this method of pairwise comparisons?

Advantages:

1) Is the most powerful method for comparing all (or almost all) pairs of means while we absolutely controlling the overall significance level ($\alpha_{overall}$)
2) When doing all pairwise comparisons, this method is considered the best available when confidence intervals are needed or sample sizes are not equal

Disadvantages:

1) If is used properly none. However, when we have a case in which all constrasts or many of them might be of interest, Scheffe's method tends to give narrower confidence intervals than Tukey and therefore is a prefered method.

F) i) Compare all pairs of means, using Fisher's Protected LSD and make an interpretation of your results. Use $\alpha_{overall} = .05$.

```
with(data_bean, pairwise.t.test(x=Length, g=Treatment, p.adjust.method = "none") )
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Length and Treatment
##
##           Low     Medium High
## Medium    2.4e-16 -      -
## High      < 2e-16 0.329  -
```

```
## Very High < 2e-16 0.014  0.125
##
## P value adjustment method: none
```

- The mean lenght shoot for Treatment A (low) is different (lower) compared to the other three treatments (B=medium, C= high, and D=very high).
- The mean lenght shoot for treatment B (medium) is different (lower) compared to treatment D (very high). With the Tukey method we were unable to get this test significant.

ii) What are the advantages and disadvantages of using this method of pairwise comparisons?

Advantages:

1) More powerful than Scheffe´s or Tukey´s procedures

Disadvantages:

1) We have less control over the $\alpha_{overall}$, since it is only controlled at approximately the α of the overall F test.
2) Because of 1, if we fail to reject the F test for all the means, we need to use the protection and use the Tukey test.

G) Consider the first two levels (low, medium) as "short" soaking periods and the two higher levels (High, very high) as "long" soaking periods. You want to determine the difference in mean sprout length between the short and long soaking periods.

```
# Short
data_bean_short= subset(data_bean, Treatment=="Low" | Treatment=="Medium", select = Length)
data_bean_short$Length=as.numeric(data_bean_short$Length)
mean(data_bean_short$Length)
```

```
## [1] 12.17647
```

```
# Long
data_bean_long= subset(data_bean, Treatment=="High" | Treatment=="Very High", select = Length)
data_bean_long$Length=as.numeric(data_bean_long$Length)
mean(data_bean_long$Length)
```

```
## [1] 20.41176
```

```
i) Give a 90% confidence interval for the difference in mean sprout length between short and long soaki
```

```
test=t.test(data_bean_long$Length, data_bean_short$Length, conf.level=0.90)
test
```

```
##
##  Welch Two Sample t-test
##
## data:  data_bean_long$Length and data_bean_short$Length
## t = 6.0908, df = 48.696, p-value = 1.733e-07
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   5.968177 10.502411
## sample estimates:
## mean of x mean of y
##  20.41176  12.17647
```

```
# 90% confidence interval
c(test$conf.int[1],test$conf.int[2])
```

```
## [1]  5.968177 10.502411
```

ii) Test to see if the long soaking periods produce higher mean sprout length than the short periods, us

Here we are interested in determine:

$$H_0 : \mu_{long} - \mu_{short} = 0$$
$$H_A : \mu_{long} - \mu_{short} > 0$$

The test statistic is defined as:

$$Test = \frac{(\bar{X}_{long} - \bar{X}_{short}) - (\mu_{long} - \mu_{short})}{\sqrt{\frac{S^2_{long}}{n_{long}} + \frac{S^2_{short}}{n_{short}}}}$$

Where: $\bar{X}_{long}$ is the sample mean in sprout lenght for long soaking soaking periods; $\bar{X}_{short}$ is the sample mean lenght for short soaking periods; S^2_{long} and S^2_{short} are the sample variances in sprout lenght for the long and short soaking periods, respectively. n_{long} and n_{short} are the sample size for each group (34 for both groups). Under the null $\mu_{long} - \mu_{short} = 0$

Normally this test is distributed t student with degrees of freedom equal to $min(n_{long} - 1, n_{short} - 1)$ (according to my previous notes from 6010). However, in R the degrees of freedom for the test are calculated in another way (I do not know why). The test in R is implemented as follows:

```
test=t.test(data_bean_long$Length, data_bean_short$Length, conf.level=0.95, alternative = "greater")
test
```

```
##
##  Welch Two Sample t-test
##
## data:  data_bean_long$Length and data_bean_short$Length
## t = 6.0908, df = 48.696, p-value = 8.663e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.968177      Inf
## sample estimates:
## mean of x mean of y
##  20.41176  12.17647
```

The p-value of the test is P(t_{df}>6.0998) = 0.00000008663. Then, we reject the null hypothesis and conclude that long soaking periods produce higher mean sprout length than the short periods.

Question 1 continued on the following page.

H) Using the values corresponding to the levels of the treatments: A = 12 hours, B= 18 hours, C = 24 hours, and D = 30 hours,

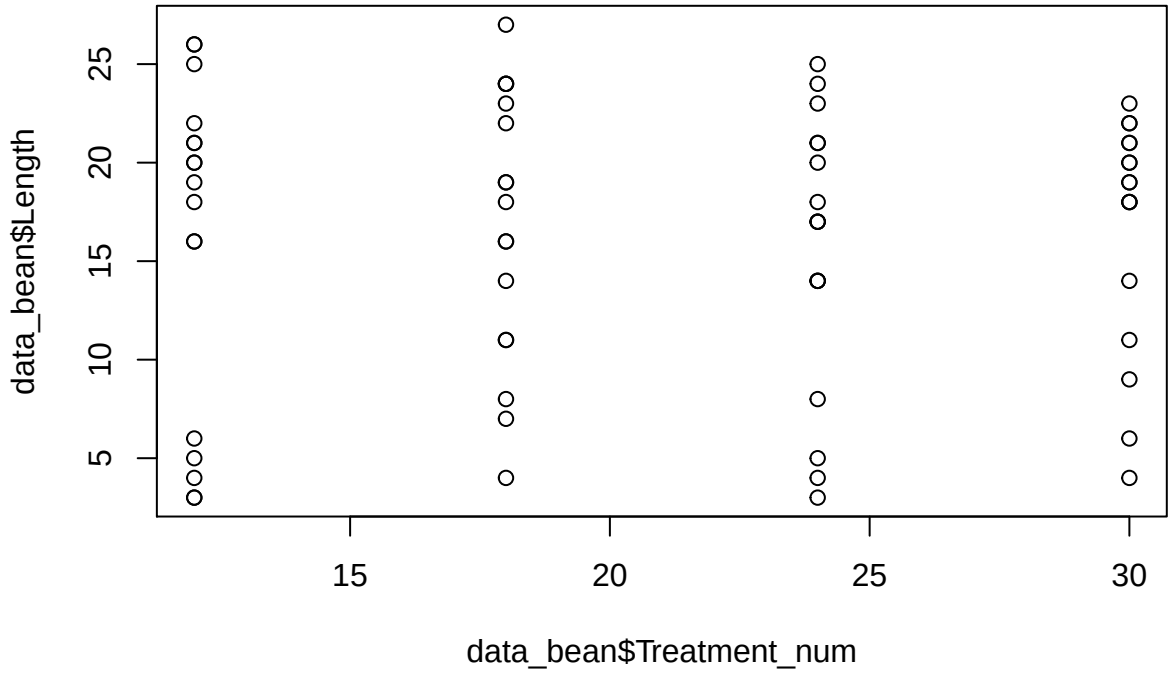
i) Fit a polynomial regression in hours to this data; report what you get and how you got there (show all steps and tests).

Step 1: Transform the treatment into a nuerical variable

```
data_bean$Treatment_num=data_bean$Treatment
data_bean$Treatment_num=c(12,18,24,30)
data_bean$Treatment_num=as.numeric(data_bean$Treatment_num)
is.numeric(data_bean$Treatment_num)
```

```
## [1] TRUE
```

```
plot(data_bean$Treatment_num, data_bean$Length)
```



Step 2: Run a polynomial regression, let´s start with a cubic specification

```
model_poly.lm = lm(Length~Treatment_num + I(Treatment_num^2) + I(Treatment_num^3), data=data_bean)
summary(model_poly.lm)
```

```
##
## Call:
## lm(formula = Length ~ Treatment_num + I(Treatment_num^2) + I(Treatment_num^3),
##     data = data_bean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.941  -5.794   1.735   5.147  10.118
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -11.470588  46.566453  -0.246    0.806
## Treatment_num        4.486928   7.442354   0.603    0.549
## I(Treatment_num^2)  -0.227124   0.372648  -0.609    0.544
```

```
## I(Treatment_num^3)   0.003631   0.005903   0.615    0.541
##
## Residual standard error: 7.053 on 64 degrees of freedom
## Multiple R-squared:  0.006315,   Adjusted R-squared:  -0.04026
## F-statistic: 0.1356 on 3 and 64 DF,  p-value: 0.9385
```

```
#anova(model1.lm)
```

Since the cubic is not significant, we can drop it individually and re-run only with the quadratic specification:

```
model_poly2.lm = lm(Length~Treatment_num + I(Treatment_num^2)  , data=data_bean)
summary(model_poly2.lm)
```

```
##
## Call:
## lm(formula = Length ~ Treatment_num + I(Treatment_num^2), data = data_bean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.294  -5.176   1.765   4.824  10.823
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        16.529412   9.772487   1.691   0.0955 .
## Treatment_num      -0.049020   1.001177  -0.049   0.9611
## I(Treatment_num^2)  0.001634   0.023645   0.069   0.9451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.019 on 65 degrees of freedom
## Multiple R-squared:  0.0004406,  Adjusted R-squared:  -0.03032
## F-statistic: 0.01433 on 2 and 65 DF,  p-value: 0.9858
```

```
#anova(model1.lm)
```

The quadratic is also no significant, let´s drop a rerun the model only with linear relationship:

```
model_poly3.lm = lm(Length~Treatment_num , data=data_bean)
summary(model_poly3.lm)
```

```
##
## Call:
## lm(formula = Length ~ Treatment_num, data = data_bean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.353  -5.235   1.706   4.882  10.765
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   15.88235    2.77626   5.721  2.8e-07 ***
## Treatment_num  0.01961    0.12593   0.156    0.877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.966 on 66 degrees of freedom
## Multiple R-squared:  0.0003672,  Adjusted R-squared:  -0.01478
```

```
## F-statistic: 0.02424 on 1 and 66 DF,  p-value: 0.8767
```

```
#anova(model1.lm)
```

This should be final model. As we can see, using treatment as numerical variable does not show any effect on the length. In particular, the adjustment is very poor (R^2=0.0003672).

ii) Compare the MSE you got from using the treatments as categorical predictors and the polynoimial pred

In this question it is almost assumed that some cuadratic or cubic terms remain in the model after fitti

```
# using treatment as numerical
#summary(model_poly3.lm)
anova(model_poly3.lm)
```

```
## Analysis of Variance Table
##
## Response: Length
##               Df Sum Sq Mean Sq F value Pr(>F)
## Treatment_num  1    1.2   1.176  0.0242 0.8767
## Residuals     66 3202.9  48.529
```

```
# using treatment as categorical
#summary(model1.lm)
anova(model1.lm)
```

```
## Analysis of Variance Table
##
## Response: Length
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Treatment  3 2501.29  833.76  75.924 < 2.2e-16 ***
## Residuals 64  702.82   10.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we use the treatment as a quantitative predictor we are only able to explain $(1.2/(1.2+3202.9) = 0.00037 \approx 0.037\%)$ of the variability observed in y. On the other hand, when we use treatment as a categorical predictor, we are able to explain $(2501.29/(2501.29+702.82) = 0.7806 \approx 78\%)$ of the total variability observed in Y. Then, by using treatment as a quantitative variable we have lost almost 78% of the explanation power. This is also reflected in a mean square error (MSE or mean square residuals): when we use the treatments as a quantitative variable the MSE is quite greater in comparison to the MSB (mean square treatments), then we are not able to reject the null hypothesis in the F test. On the other hand, when we use treatments as a categorical variable, the MSB is quite greater in comparison with the MSE, which means that F test is able to identify the presence of differences.

iii) What mean sprout length could you expect for 15 hours of soaking (use a 95% interval). Could you ha

```
newdata=data.frame(Treatment_num=15)
predict(model_poly3.lm, newdata, type="response", interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 16.17647 13.91356 18.43938
```

```
#the 95% con
```

We would expect a mean sprout length of 16.17647 mm. The 95% confidence interval is (13.91, 18.4).

Finally, we will also be able to obtain a confidence interval for the mean sprout length in the case of assume the treatments as a categorical variable, the difference is that we will obtain the expect mean sprout for the desired level of treatment, in tha case case the treatment only have 3 possible values (because with compared

the mean of the treatments with a reference). for instance, we can obtain what we would expect to be the mean sprout length for the treatment low. Obviusly the result will be very different depending on how we define the treatment to be (categorical vs quantitative).

Question 2.

For newly planted strawberries, the development of flower clusters decreases the plant vigor. It is common practice to remove the flower stalks by hand, but this is a laborious and time-consuming procedure. To investigate the effect of flower clusters on the plant vigor, an experiment consisting of four treatments was conducted. This experiment was completely randomized and consisted of the following treatments: A = Control (no flower removal), B = Hand removal, C = Regulator G1, and D = Regulator G2 (note that G1 and G2 are hormone-based regulators). A plot of 10 plants was treated and the average number of runners per mother plant, a measure of vigor, was recorded on each plot.

The layout of the experiment and the measures of vigor are provided below for each plot.

C. 3.6 (plot 1) A. 1.4 (plot 6) A. 0.8 (plot 11) B. 5.2 (plot 16)

C. 2.4 (plot 2) D. 7.3 (plot 7) B. 6.8 (plot 12) C. 1.8 (plot 17)

A. 0.6 (plot 3) C. 4.6 (plot 8) B. 3.0 (plot 13) D.6.2 (plot 18)

D. 3.8 (plot 4) D. 4.1 (plot 9) A. 1.2 (plot 14) B. 5.0 (plot 19)

B. 6.0 (plot 5) B. 4.0 (plot 10) A. 0.5 (plot 15) A. 1.5 (plot 20)

Note: This data set is not provided, so you need to create it.

```
library(readxl)
data_vigor = read_excel("Hwk6Q2.xlsx")
head(data_vigor)
```

```
##   Obs   Treatment Vigor
## 1   1 RegulatorG1   3.6
## 2   2 RegulatorG1   2.4
## 3   3     Control   0.6
## 4   4 RegulatorG2   3.8
## 5   5    Hremoval   6.0
## 6   6     Control   1.4
```

```
data_vigor$Treatment = factor(data_vigor$Treatment)
```

A) Construct a set of 3 contrasts that are suggested by the treatment structure in this experiment to be orthogonal.

The treatments are defined as follows:

1= Control 2= Hand removal 3= Regulator 1 4= Regulator 2

The three contrast are:

1) control (1) vs others (2,3,4):

$$L_1 = 1\mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3 - \frac{1}{3}\mu_4$$

2) Hand removal (2) vs regulators (3,4):

$$L_2 = 0\mu_1 + 1\mu_2 - \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$$

3) Regulator 1 (3) vs regulator 2 (4):

$$L_3 = 0\mu_1 + 0\mu_2 + \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$$

As we can see this definition is consistent with the linear combination of means in the form $L = a_1\mu_1 + a_2\mu_2, ..., +a_t\mu_t = \sum_{i=1}^{t} a_i\mu_i$, and since $\sum_{i=1}^{t} a_i = 0$ we have a linear contrast. On the other hand, all the linear constrast above are orthogonal since for all of them $a_1 * b_1 + a_2 * b_2 + ... + a_t * b_t = 0$. For instance:

- Linear contrast 1 is orthogonal to linear constrast 2 since:

$$(1 * 0) + (-\frac{1}{3} * 1) + (-\frac{1}{3} * -\frac{1}{2}) + (-\frac{1}{3} * -\frac{1}{2}) = 0$$

- Linear contrast 1 is orthogonal to linear contrast 3 since:

$$(1 * 0) + (-\frac{1}{3} * 0) + (-\frac{1}{3} * \frac{1}{2}) + (-\frac{1}{3} * -\frac{1}{2}) = 0$$

- Linear contrast 2 is ortoghonal to linear contrast 3 since:

$$(0 * 0) + (1 * 0) + (-\frac{1}{2} * \frac{1}{2}) + (-\frac{1}{2} * -\frac{1}{2}) = 0$$

B) Verbally define each of the three contrasts above.

1) Is the mean strawberries vigor different when not flower removal is implemented, compared to the other. treatments (removal by hand or with regulators).

2) Is the mean strawberries vigor different when the flower removal is by hand compared to the regulators.

3) Is the mean strawberries vigor different between the two type of regulators.

C) Using the contrasts in a, assess the statistical significance of each contrast based on p-values from an appropriate test.

```
# defining contrasts
Con1 = c(1, -1/3, -1/3, -1/3)
Con2 = c(0, 1, -1/2, -1/2)
Con3 = c(0, 0, 1/2, -1/2)
Cons = cbind(Con1, Con2, Con3)
rownames(Cons) = c ("Control", "Hremoval", "RegulatorG1", "RegulatorG2")
Cons
```

```
##                   Con1 Con2 Con3
## Control      1.0000000  0.0  0.0
## Hremoval    -0.3333333  1.0  0.0
## RegulatorG1 -0.3333333 -0.5  0.5
## RegulatorG2 -0.3333333 -0.5 -0.5
```

```
# Check that constrast are orthogonal
t(Cons)%*%Cons
```

```
##          Con1 Con2 Con3
## Con1 1.333333  0.0  0.0
## Con2 0.000000  1.5  0.0
## Con3 0.000000  0.0  0.5
```

```
# Define contrast before Anova
contrasts(data_vigor$Treatment) = Cons
# Run Anova
aov.cons = aov(Vigor~Treatment, data_vigor)
#summary(aov.cons)
summary(aov.cons, split=list(Treatment=list("Control - Others"=1, "Hand - Regulators"=2, "Regulator1-Reg
```

```
##                                    Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment                           3  65.33   21.78  14.921 6.75e-05 ***
##   Treatment: Control - Others       1  53.14   53.14  36.415 1.74e-05 ***
##   Treatment: Hand - Regulators      1   2.06    2.06   1.411    0.252
##   Treatment: Regulator1-Regulator2  1  10.13   10.13   6.938    0.018 *
## Residuals                          16  23.35    1.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results shows:

1) The first linear contrast is statistically significant (p-value < 0.05): there is a difference in the strawberries mean vigor when not flower removal is implemented (control), compared to the other treatments (removal by hand or with regulators).

2) The second linear contrast is not significant: The mean strawberries vigor is not different when the flower removal is by hand compared to the regulators.

3) The third linear contrast is significant: The mean strawberries vigor is different between the two type of regulators.

D) Demonstrate that the three contrast sums of squares do not add up to the treatment sum of squares (there is more than one way to do this). Are you surprised by your results? Why or why not? Are these contrasts orthogonal? Why or why not?

```
means = aggregate(data_vigor[, 3], list(data_vigor$Treatment), mean)
means
```

```
##       Group.1    x
## 1     Control 1.00
## 2    Hremoval 5.00
## 3 RegulatorG1 3.10
## 4 RegulatorG2 5.35
```

```
u_control = means$x[1]
u_Hremoval = means$x[2]
u_regulator1 = means$x[3]
u_regulator2 = means$x[4]
n_control = 6
n_Hremoval = 6
n_regulator1 = 4
n_regulator2 = 4
# sum squares linear contrast 1 (SSL1)  = (L-hat)^2 / summ ((ai^2)/ni)
Lhat1 = (1*u_control) - (1/3)*u_Hremoval - (1/3)*u_regulator1 - (1/3)*u_regulator2
sum_ai1 = ((1^2) / n_control) +  ((1/3)^2 / n_Hremoval) + ((1/3)^2 / n_regulator1) + ((1/3)^2 / n_regula
SSLhat1 = ((Lhat1)^2) / sum_ai1
SSLhat1
```

```
## [1] 50.40115
```

```r
# summ squares linear contrast 2 (SSL2)
Lhat2 = (0*u_control) + (1)*u_Hremoval - (1/2)*u_regulator1 - (1/2)*u_regulator2
sum_ai2 = ((0^2) / n_control) +  ((1)^2 / n_Hremoval) + ((1/2)^2 / n_regulator1) + ((1/2)^2 / n_regulat
SSLhat2 = ((Lhat2)^2) / sum_ai2
SSLhat2
```

```
## [1] 2.059286
```

```r
# summ squares linear contrast 3 (SSL3)
Lhat3 = (0*u_control) + (0)*u_Hremoval + (1/2)*u_regulator1 - (1/2)*u_regulator2
sum_ai3 = ((0^2) / n_control) +  ((0)^2 / n_Hremoval) + ((1/2)^2 / n_regulator1) + ((1/2)^2 / n_regulat
SSLhat3 = ((Lhat3)^2) / sum_ai3
SSLhat3
```

```
## [1] 10.125
```

```r
# linear contrast summ of squares
SStreatment_contrasts = SSLhat1 + SSLhat2 + SSLhat3
SStreatment_contrasts
```

```
## [1] 62.58544
```

As we can see the three contrast sums of squares (62.58544) do not add up to the treatment sum of squares (65.33). Initially, this result is surprising because by construction it seems that linear contrast that we have used before were orthogonal, however it seems that the unbalance design (simple size unequal) is affecting the orthogonality of the linear contrast to the point that they seem to be orthogonal, but they really are not.

E) Remove the observations for plots 5, 10, 15, and 20. Re-compute the treatment and contrast sums of squares. Demonstrate that the three contrast sums of squares add up to the treatment sum of squares. Are you surprised by your results? Why or why not? Construct an ANOVA table which shows that with this balanced design, the sums of squares for treatments partitions into the sums of squares for the three contrasts. Are these contrasts now orthogonal? Why or why not?

```r
# new dataset without the points (I have done this in excel)
library(readxl)
data_vigor_2 = read_excel("Hwk6Q2-2.xlsx")
head(data_vigor_2)
```

```
##   Obs   Treatment Vigor
## 1   1 RegulatorG1   3.6
## 2   2 RegulatorG1   2.4
## 3   3     Control   0.6
## 4   4 RegulatorG2   3.8
## 5   6     Control   1.4
## 6   7 RegulatorG2   7.3
```

```r
data_vigor_2$Treatment = factor(data_vigor_2$Treatment)
```

The new calculations for the contrast summ of squares are as follows:

```r
means_1 = aggregate(data_vigor_2[, 3], list(data_vigor_2$Treatment), mean)
means_1
```

```
##       Group.1    x
## 1     Control 1.00
## 2    Hremoval 5.00
## 3 RegulatorG1 3.10
## 4 RegulatorG2 5.35
```

```
u_control = means_1$x[1]
u_Hremoval = means_1$x[2]
u_regulator1 = means_1$x[3]
u_regulator2 = means_1$x[4]
n_control = 4
n_Hremoval = 4
n_regulator1 = 4
n_regulator2 = 4
# sum squares linear contrast 1 (SSL1)  = (L-hat)^2 / summ ((ai^2)/ni)
Lhat1 = (1*u_control) - (1/3)*u_Hremoval - (1/3)*u_regulator1 - (1/3)*u_regulator2
sum_ai1 = ((1^2) / n_control) +  ((1/3)^2 / n_Hremoval) + ((1/3)^2 / n_regulator1) + ((1/3)^2 / n_regula
SSLhat1 = ((Lhat1)^2) / sum_ai1
SSLhat1
```

```
## [1] 36.40083
```

```
# summ squares linear contrast 2 (SSL2)
Lhat2 = (0*u_control) + (1)*u_Hremoval - (1/2)*u_regulator1 - (1/2)*u_regulator2
sum_ai2 = ((0^2) / n_control) +  ((1)^2 / n_Hremoval) + ((1/2)^2 / n_regulator1) + ((1/2)^2 / n_regulat
SSLhat2 = ((Lhat2)^2) / sum_ai2
SSLhat2
```

```
## [1] 1.601667
```

```
# summ squares linear contrast 3 (SSL3)
Lhat3 = (0*u_control) + (0)*u_Hremoval + (1/2)*u_regulator1 - (1/2)*u_regulator2
sum_ai3 = ((0^2) / n_control) +  ((0)^2 / n_Hremoval) + ((1/2)^2 / n_regulator1) + ((1/2)^2 / n_regulat
SSLhat3 = ((Lhat3)^2) / sum_ai3
SSLhat3
```

```
## [1] 10.125
```

```
# linear contrast summ of squares
SStreatment_contrasts = SSLhat1 + SSLhat2 + SSLhat3
SStreatment_contrasts
```

```
## [1] 48.1275
```

Now, the treatment summ of squares and the calculations for the contrasts are:

```
# defining contrasts
Con1 = c(1, -1/3, -1/3, -1/3)
Con2 = c(0, 1, -1/2, -1/2)
Con3 = c(0, 0, 1/2, -1/2)
Cons = cbind(Con1, Con2, Con3)
rownames(Cons) = c ("Control", "Hremoval", "RegulatorG1", "RegulatorG2")
Cons
```

```
##                   Con1 Con2 Con3
## Control      1.0000000  0.0  0.0
## Hremoval    -0.3333333  1.0  0.0
## RegulatorG1 -0.3333333 -0.5  0.5
## RegulatorG2 -0.3333333 -0.5 -0.5
```

```
# Check that constrast are orthogonal
t(Cons)%*%Cons
```

```
##          Con1 Con2 Con3
## Con1 1.333333  0.0  0.0
```

```
## Con2 0.000000  1.5  0.0
## Con3 0.000000  0.0  0.5
```

```r
# Define contrast before Anova
contrasts(data_vigor_2$Treatment) = Cons
# Run Anova
aov.cons = aov(Vigor~Treatment, data_vigor_2)
#summary(aov.cons)
summary(aov.cons, split=list(Treatment=list("Control - Others"=1, "Hand - Regulators"=2, "Regulator1-Reg
```

```
##                                   Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment                          3  48.13   16.04   9.233 0.001925 **
##   Treatment: Control - Others      1  36.40   36.40  20.950 0.000636 ***
##   Treatment: Hand - Regulators     1   1.60    1.60   0.922 0.355944
##   Treatment: Regulator1-Regulator2 1  10.13   10.13   5.827 0.032674 *
## Residuals                         12  20.85    1.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see the Treatment summ of squares (48.13) add up to the constrast summ squares ($48.1275 \approx 48.13$). Now, I am less surprised by the results since we have a balance design: The linear constrasts are now really orthogonal.