BTRY 6020 Homework VI

NAME: student name

NETID: student NetID

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.

We wish to test the hypotheses:

H_0 : there is no different in average shoot length across soaking levels

vs

H_a : the average shoot length is different for at least one soaking level.

We run this test by loading the data and running a full anova test.

```
library(readxl)
q1dat = read_excel("Hwk6Q1DatSp17.xlsx")
q1dat$Treatment = as.factor(q1dat$Treatment)

lm1 = lm(Length~Treatment, data = q1dat)
anova(lm1)
```

```
## 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
```

Since the p-value is < 0.001 we can conclude that at any reasonable significance level we have evidence that the average shoot length is different for at least one soaking level.

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

The oneway ANOVA model is

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

where

Y_{ij} - j^{th} shoot length for soaking level $i = A, B, C, D$.

μ_i - average shoot length for soaking level $i = A, B, C, D$.

ϵ_{ij} - j^{th} error for soaking level $i = A, B, C, D$. Assumed to be $N(0, \sigma^2)$ and independent across observations.

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

Independence should hold as we use a completely randomized design. **(Note: Independence could fail to hold if after assigning the treatment, all the experiments were undertaken in order of treatment. If all treatment A beans were grown on a specific day, then environmental factors may be accounting for the observed difference. Without further information on the conditions of the experiment, however, we are not able to comment on this.)**

We now look for normality and for constant variance

```
library(car)
qqPlot(lm1)
```
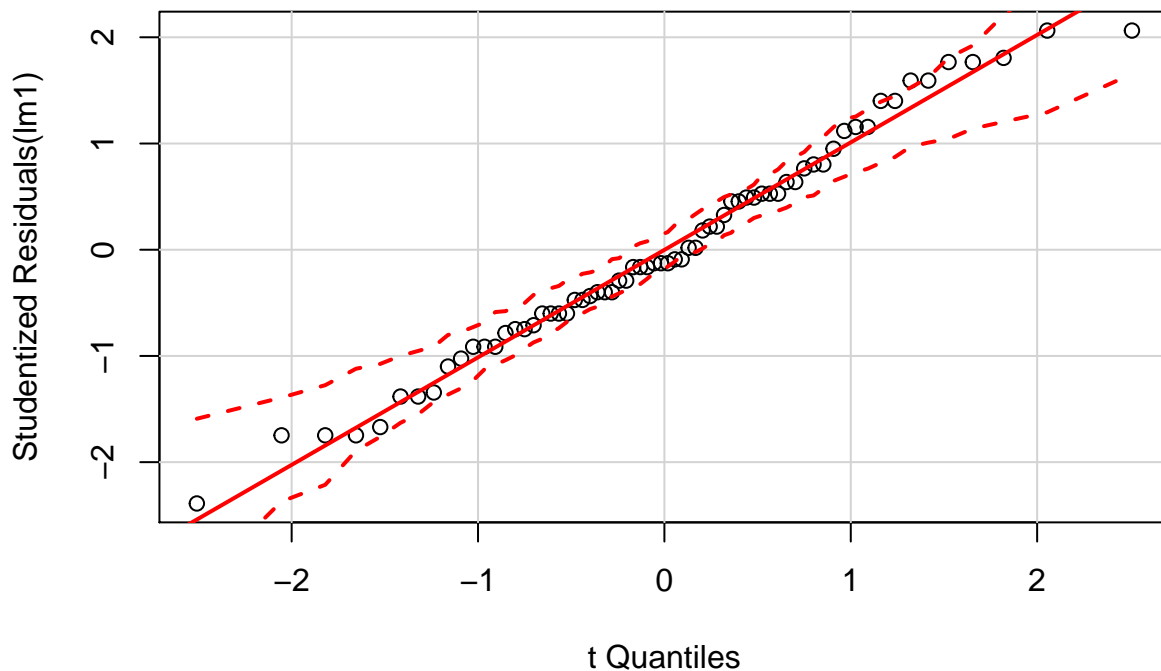


```
plot(as.numeric(q1dat$Treatment),rstandard(lm1))
```
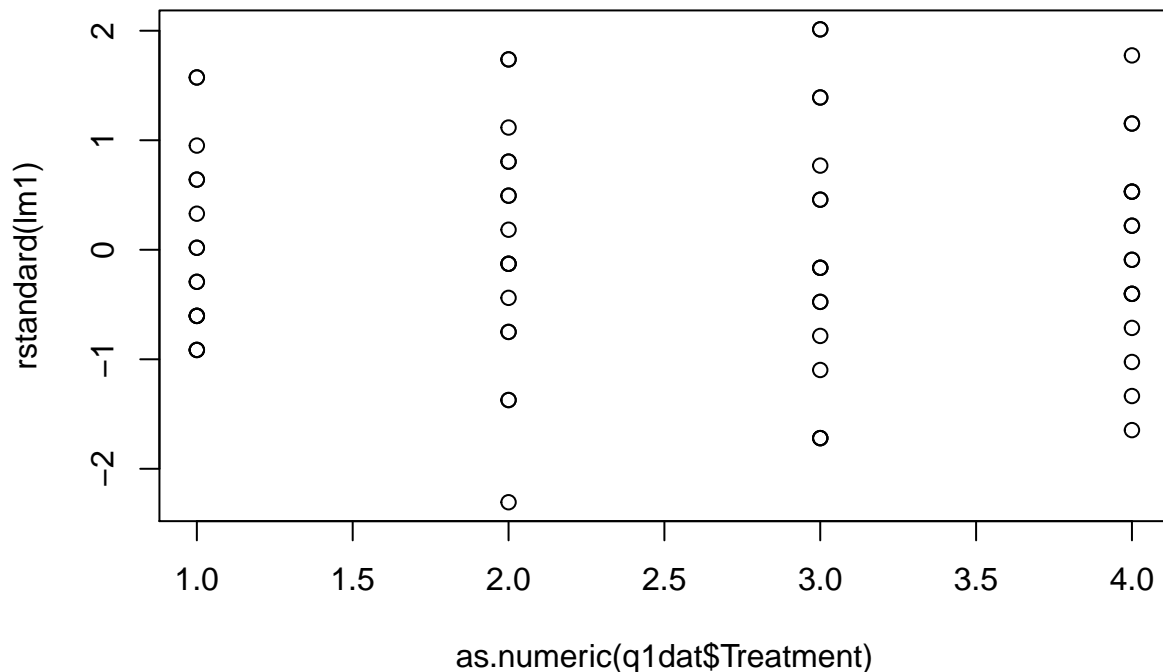
We see from the qqPlot that normality appears to hold, and from the residual plot that constant variance appears to be a reasonable assumption.

Lastly, no outliers are appearent in either diagnostic plot, therefore there outliers should not be driving the conclusions of these analyses.

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

One way that this can be done is with the `pairwise.t.test` function in R.

```
pairwise.t.test(q1dat$Length, q1dat$Treatment, p.adj = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  q1dat$Length and q1dat$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
```

Note that here we used a bonferonni correction on 6 tests.

At the 0.05 level, we conclude that treatment A has a significantly different average shoot length from treatment levels B, C, and D. We do not have significant evidence that there is any difference in average

3

shoot length between the treatments B, C, and D.

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

The advantages are: easy to use, widely applicable

The disadvantages are: Too conservative (low power) with a large number of contrasts.

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

To do this, first we fit with the `aov` function in `R` and then run the `TukeyHSD` function.

```
aov1 = aov(Length~Treatment, data = q1dat)
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Length ~ Treatment, data = q1dat)
##
## $Treatment
##          diff         lwr        upr     p adj
## B-A 12.470588   9.4723100  15.468866 0.0000000
## C-A 13.588235  10.5899571  16.586513 0.0000000
## D-A 15.352941  12.3546630  18.351219 0.0000000
## C-B  1.117647  -1.8806311   4.115925 0.7594325
## D-B  2.882353  -0.1159253   5.880631 0.0638738
## D-C  1.764706  -1.2335723   4.762984 0.4128068
```

At the 0.05 level, we conclude that treatment A has a significantly different average shoot length from treatment levels B, C, and D. We do not have significant evidence that there is any difference in average shoot length between the treatments B, C, and D.

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

The advantages are: most powerful procedure for comparing all pairs of means while absolutely controlling overall significance.

The disadvantages are: is not widely applicable.

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

Not that since we rejected the overall F-test in part A we can continue with this method.

```
with(q1dat, pairwise.t.test(x = Length, g = Treatment, p.adjust = 'none'))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Length and Treatment
##
##   A       B     C
## B 2.4e-16 -     -
## C < 2e-16 0.329 -
## D < 2e-16 0.014 0.125
##
## P value adjustment method: none
```

At the 0.05 level, we conclude that treatment A has a significantly different average shoot length from treatment levels B, C, and D, and treatment level B is significantly different than treatment level D. We do not have significant evidence that there is any difference in average shoot length between the treatments B and C, nor between C and D.

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

The advantages are: easy to use, and more powerful than Scheffe's or Tukey's method.

The disadvantages are: Less control over significance level; only controlled approximately at α of overall F-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.

   i) Give a 90% confidence interval for the difference in mean sprout length between short and long soaking periods.

First create a variable in our data set corresponding to Soaking Periods. Set it to `'short'` for all observations, and then edit observations where `Treatment = C or D` to have `'long'` soaking periods.

```
q1dat$Soak = 'short'
q1dat$Soak[q1dat$Treatment == 'C'] = 'long'
q1dat$Soak[q1dat$Treatment == 'D'] = 'long'
q1dat$Soak = as.factor(q1dat$Soak)
```

We can obtain a 90% confidence interval for the difference by looking at a confidence interval for the slope term with the `lm` function.

```
lm2 = lm(Length~Soak, data = q1dat)
confint(lm2, level = 0.9)
```

```
##                    5 %       95 %
## (Intercept)   18.81678 22.006750
## Soakshort    -10.49094 -5.979645
```

Therefore we are 90% confident that the stock will be on average between 5.98mm and 10.49mm longer if it has a long soak instead of a short soak.

Note that this could have been done with a more traditional contrast approach; the approach above will boil down to the same thing.

   ii) Test to see if the long soaking periods produce higher mean sprout length than the short periods, using $\alpha = .05$. State hypotheses, test statistic, p-value, and conclusions.

Testing this is equivalent to testing: H_0 : estimate for `Soakshort` $= 0$ vs. H_a: estimate for `Soakshort` < 0.

```
summary(lm2)
```

```
##
## Call:
## lm(formula = Length ~ Soak, data = q1dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1765 -4.2353 -0.4118  4.0147 11.8235
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.4118     0.9561  21.350  < 2e-16 ***
```

```
## Soakshort     -8.2353      1.3521  -6.091 6.45e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.575 on 66 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3501
## F-statistic:  37.1 on 1 and 66 DF,  p-value: 6.449e-08
```

From the above table, the t-statistic for this test is -6.091. Therefore the pvalue for this one-sided test can be calculated as follows

```
pt(-6.091, 66)
```

```
## [1] 3.221741e-08
```

Therefore we have significant evidence at the 0.05 significace level that long soaking periods produce higher mean sprout lenght than short periods.

  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).

First create this variable in our dataset

```
q1dat$Hours = 12
q1dat$Hours[q1dat$Treatment == 'B'] = 18
q1dat$Hours[q1dat$Treatment == 'C'] = 24
q1dat$Hours[q1dat$Treatment == 'D'] = 30
```

Now plot the data to determine an appropriate polynomial degree for our regression.

```
plot(q1dat$Hours, q1dat$Length)
```
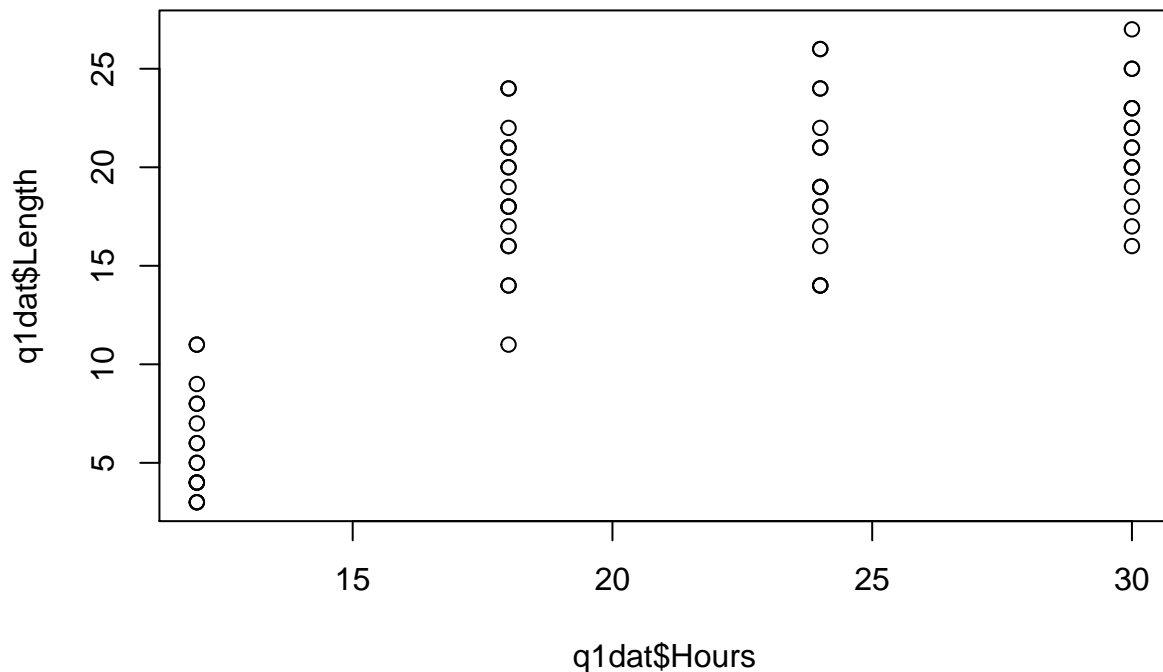
A degree 2 polynomial should potentially be appropriate, but first look at a degree 3 polynomial.

```
lmP3 =lm(Length~Hours + I(Hours^2) + I(Hours^3), data = q1dat)
summary(lmP3)
```

```
##
## Call:
## lm(formula = Length ~ Hours + I(Hours^2) + I(Hours^3), data = q1dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4118 -2.0294 -0.4118  2.0588  6.4706
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.011e+02  2.188e+01  -4.619 1.91e-05 ***
## Hours        1.548e+01  3.497e+00   4.426 3.82e-05 ***
## I(Hours^2)  -6.577e-01  1.751e-01  -3.756 0.000375 ***
## I(Hours^3)   9.259e-03  2.773e-03   3.339 0.001407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.314 on 64 degrees of freedom
## Multiple R-squared:  0.7806, Adjusted R-squared:  0.7704
## F-statistic: 75.92 on 3 and 64 DF,  p-value: < 2.2e-16
```

7

```
lmP2 = lm(Length~Hours + I(Hours^2), data = q1dat)
summary(lmP2)
```

```
##
## Call:
## lm(formula = Length ~ Hours + I(Hours^2), data = q1dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3294 -2.5412 -0.3294  2.4059  7.3882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29.65882    4.96058  -5.979 1.05e-07 ***
## Hours         3.90882    0.50820   7.691 1.03e-10 ***
## I(Hours^2)   -0.07435    0.01200  -6.194 4.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.563 on 65 degrees of freedom
## Multiple R-squared:  0.7424, Adjusted R-squared:  0.7345
## F-statistic: 93.69 on 2 and 65 DF,  p-value: < 2.2e-16
```

All terms are significant in this model, however including the 3rd degree term only increases the R^2 by 3.82%, so we use the 2nd order polynomial. **(Note: that since there are only 4 distinct predictor values in the Hours variable, we cannot fit a higher than order 3 polynomial.**

Also note: depending on the data situation, this 3.8% increase in explained variance may be considered significant.

Our final model is estimated as:

Length = -29.66 + 3.91Hours -0.074Hours^2

    ii) Compare the MSE you got from using the treatments as categorical predictors and the polynoimial predictors; assess how much explained variation you lost by forcing the means to follow the regression "line" compared to letting them "float". The MSE increases from 10.89 to 12.69 when forcing the means to follow the regression "line". This corresponds to a 3.82% increase in explained variance.

(Note: if a 3rd order polynomial is used, then these MSE values will be the same)

    iii) What mean sprout length could you expect for 15 hours of soaking (use a 95% interval). Could you have gotten this by using the treatments as categorical predictors?

A 95% confidence interval at Hours = 15 is obtained as

```
newdata = data.frame(Hours = 15)
predict(lmP2, newdata, interval = 'confidence')
```

```
##        fit     lwr      upr
## 1 12.24559 11.0679 13.42328
```

Therefore we are 95% confdient that the mean sprout length for a bean that was soaking for 15 hours will be between 11.07mm and 13.42mm.

This could not be done when treating the treatments as categorical predictors.

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.

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

First we create the dataset

```
vigor = c(3.6, 2.4, 0.6, 3.8, 6.0, 1.4, 7.3, 4.6, 4.1,
          4, 0.8, 6.8, 3, 1.2, 0.5, 5.2, 1.8, 6.2, 5, 1.5)
treat = c("C", "C", "A", "D", "B", "A", "D", "C", "D", "B",
          "A", "B", "B", "A", "A", "B", "C", "D", "B", "A")
q2dat = data.frame(vigor, as.factor(treat))
names(q2dat) = c("vigor", "treat")
```

An example of 3 contrasts are

```
Con1 = c(1, -1/3, -1/3, -1/3)
Con2 = c(0, 0, 1/2, -1/2)
Con3 = c(0, 1, -1/2, -1/2)
Cons = cbind(Con1, Con2, Con3)
rownames(Cons) = c("Control", "HandRemove", "G1", "G2")
```

B) Verbally define each of the three contrasts above.

`Con1` looks to see if the Control average is different than the treatment average.

`Con2` looks to see if the average between regulator G1 is different than regulator G2.

`Con3` looks to see if the Hand Removal average is different than the Regulator (both G1 and G2) average.

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

First define contrasts, then run anova.

```
contrasts(q2dat$treat) = Cons
aov.cons = aov(vigor~treat, data = q2dat)
summary(aov.cons, split = list(treat =list('Control-Rest'= 1, 'G1-G2' = 2, 'Hand-Reg' = 3)))
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## treat             3  65.33   21.78  14.921 6.75e-05 ***
```

```
##    treat: Control-Rest  1  53.14    53.14  36.415 1.74e-05 ***
##    treat: G1-G2          1  10.13    10.13   6.938    0.018 *
##    treat: Hand-Reg       1   2.06     2.06   1.411    0.252
## Residuals              16  23.35     1.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since we are running three tests, we do a bonferonni correction ($\alpha_{individual} = .05/3 = .01667$. With the bonferonni correction on three tests applied, we only have significance for the first constrast.

**In the analysis done above, Type I SSs were used (sequential SSs), and so these will add up to SSTr and always look orthogonal. They are not since sample sizes are different; let's rerun this ANOVA with Type III SSs (Extra SSs):

```
library(car)
Con1 = c(1, -1/3, -1/3, -1/3)
Con2 = c(0, 0, 1/2, -1/2)
Con3 = c(0, 1, -1/2, -1/2)
Cons = cbind(Con1, Con2, Con3)
rownames(Cons) = c("Control", "HandRemove", "G1", "G2")
Anova(lm(vigor~C(treat, Con1, 1)+C(treat, Con2, 1) + C(treat, Con3, 1), data = q2dat), type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: vigor
##                   Sum Sq Df  F value     Pr(>F)
## (Intercept)      250.563  1 171.6920 5.697e-10 ***
## C(treat, Con1, 1)  50.401  1  34.5361 2.340e-05 ***
## C(treat, Con2, 1)  10.125  1   6.9379   0.01805 *
## C(treat, Con3, 1)   2.059  1   1.4111   0.25221
## Residuals          23.350 16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we see the SSs for the three contrasts add up to 62.585 *ne* 65.33; that is , the SSs for these three contrasts do not add up to the SSTr from the previous ANOVA, indicating the lack of orthogonality.

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?

From the anova summary table we have that the contrast cum of squares add up to $53.14 + 10.13 + 2.06 = 65.33$. Therefore the three contrast sum of squares add up to the treatment sum of squares.

This is expected as our contrasts were orthogonal, as can be seen by noting that the off-diagonal elements in the following code is equal to 0.

```
t(Cons)%*%Cons
```

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

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?

We obtain the newdata, and reobtain the anova summary table

```
#delete observations 5, 10, 15, 20
q2datNew = q2dat[-c(5, 10, 15, 20),]

#rerun anova
contrasts(q2datNew$treat) = Cons
aov.consNew = aov(vigor~treat, data = q2datNew)
summary(aov.consNew, split = list(treat =list('Control-Rest'= 1, 'G1-G2' = 2, 'Hand-Reg' = 3)))
```

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## treat                     3  48.13   16.04   9.233 0.001925 **
##   treat: Control-Rest      1  36.40   36.40  20.950 0.000636 ***
##   treat: G1-G2             1  10.12   10.12   5.827 0.032674 *
##   treat: Hand-Reg          1   1.60    1.60   0.922 0.355944
## Residuals                12  20.85    1.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The three constrast sum of squares add up to $36.40 + 10.12 + 1.60 = 48.13 = SSTr$ (from previous ANOVA table), as expected, since we are using Type I SSs. To determine true orthogonality, we must use Type III SSs, and still have this condition met.

```
library(car)
q2datNew = q2dat[-c(5, 10, 15, 20),]
Con1 = c(1, -1/3, -1/3, -1/3)
Con2 = c(0, 0, 1/2, -1/2)
Con3 = c(0, 1, -1/2, -1/2)
Cons = cbind(Con1, Con2, Con3)
rownames(Cons) = c("Control", "HandRemove", "G1", "G2")
Anova(lm(vigor~C(treat, Con1, 1)+C(treat, Con2, 1) + C(treat, Con3, 1), data = q2datNew), type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: vigor
##                    Sum Sq Df  F value    Pr(>F)
## (Intercept)       208.803  1 120.1741 1.316e-07 ***
## C(treat, Con1, 1)  36.401  1  20.9501 0.0006356 ***
## C(treat, Con2, 1)  10.125  1   5.8273 0.0326741 *
## C(treat, Con3, 1)   1.602  1   0.9218 0.3559437
## Residuals          20.850 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now see that SSs add up to $36.401 + 10.125 + 1.602 = 48.128 = 48.13 = SStr$ from previous ANOVA table, indicating orthogonality of these contrasts. "'