BTRY 6020 Homework VIII

NAME: student name

NETID: student NetID

DUE DATE: May 10, by 8:40 am

Question 1.

A study was conducted to investigate the effect of fertilization on the yield of a commercial variety of tomato. In this study, a completely randomized design was used as the experimental plot was homogeneous both in humidity and nutrient content. The treatments of interest in this study consisted of a control (E=no fertilizer used), very slow release (A), a slow release fertilizer (B), a moderate release fertilizer (C) and a fast release fertilizer (D). These treatment levels were each assigned at random to 20 plots with 15 plants per plot. Prior to planting, the researchers noticed big differences in plant heights. There was no record of when the seeds were planted, so nobody knew if the plants were the same age or not. They therefore decided to record information on mean plant height on each plot. The response variable of interest was the total weight of tomatoes harvested per plot. Data for this study can be downloaded from the course web site. The file name is Hwk8Q1DatSp17.

```r
#load data
library(readxl)
plant.data <- read_excel("Hwk8Q1DatSp17.xlsx")

plant.data$Treatment = as.factor(plant.data$Treatment)
#change the control to the first level in Treatment:
plant.data$Treatment = relevel(plant.data$Treatment, "E")
```

Answer the following questions.

A) Give a model for the analysis of covariance, explain each term in the model and formulate appropriate assumptions.

$y_{ij} = \mu_{.} + \alpha_i + \beta_j (x_{ij} - \bar{x}_{..}) + \epsilon_{ij}$

where y_{ij} is the weight of tomoatoes from the jth plot receiving treatment i

α_i is the effect of fertilizer type

β_j is the slope for the covariate mean height

x_{ij} is the covariate value of mean height of plants on the jth plot receiving treatment i

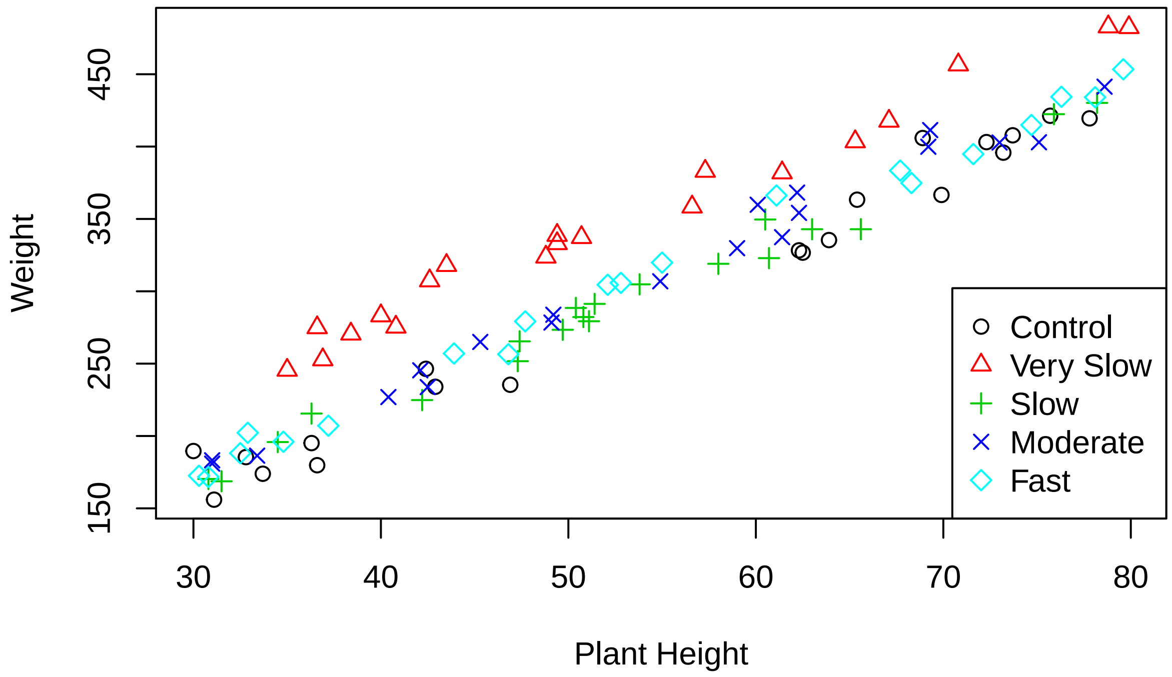
the model is subject to the following considerations/assumptions:

$\epsilon_{ij} \sim iidN(0, \sigma^2)$

So, $y_{ij} \sim indN(\mu_{ij}, \sigma^2)$

B) Plot the response variable weight against the covariate height using levels of the treatments as different plotting symbols. What relationship is there between these two variables?

```
plot(plant.data$Height, plant.data$Weight, pch = as.numeric(plant.data$Treatment),
     col = as.numeric(plant.data$Treatment), xlab = "Plant Height", ylab = "Weight")
legend("bottomright", pch = c(1, 2, 3, 4, 5), col = c(1, 2, 3, 4, 5),
       legend = c("Control", "Very Slow", "Slow", "Moderate", "Fast"))
```



As plant height increases, the weight of harvested tomatoes also increases. Further, it looks as though this relationship holds true regardless of fertilizer type use. So there does not appear to be an interaction between fertilizer type and plant height.

C) Based on ordinary one-way ANOVA perform a significance test for the equality of the five treatment means. State hypotheses, test statistic, p-value, and your conclusions.

```
plant.aov = aov(Weight ~ Treatment, data = plant.data)
anova(plant.aov)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  4  39818  9954.6  1.3708 0.2499
## Residuals 95 689866  7261.8
```

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

$H_a :$ not H_0

test statistic $= 1.3708$

p-value $= p(F_{4,95} > 1.3708) = 0.2499 > 0.05$.

Fail to reject the null hypothesis at a 5% significance level. Conclude that there is no effect of fertilizer on weight fo tomato harvest. So all five treatment means are equal.

D) Determine if the interaction between treatment and initial plant height is significant. State hypotheses, test statistic, p-value, and your conclusions.

We can try fitting an interaction model:

```
plant.aov2 = lm(Weight ~ Treatment * Height, data = plant.data)
anova(plant.aov2)
```

```
## Analysis of Variance Table
##
## Response: Weight
##                  Df Sum Sq Mean Sq   F value Pr(>F)
## Treatment         4  39818    9955   85.5258 <2e-16 ***
## Height            1 678975  678975 5833.4655 <2e-16 ***
## Treatment:Height  4    416     104    0.8928 0.4717
## Residuals        90  10475     116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : (\alpha\beta)_{ij} = 0$ for all i, j

$H_a :$ not H_0

test statistic = F = 0.8928 (Note: F < 1 ALWAYS means a lack of significance)

p-value = p(F_{4, 90} > .8928) = 0.4717 >> 0.05

Fail to reject the null hypothesis at a 5% significance level. There is no evidence of an interaction.

E) Regardless of your answer to Part D above, perform analysis of covariance and summarize your results.

```
plant.aov3 = lm(Weight ~ Treatment + Height, data = plant.data)
anova(plant.aov3)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Treatment  4  39818    9955   85.918 < 2.2e-16 ***
## Height     1 678975  678975 5860.194 < 2.2e-16 ***
## Residuals 94  10891     116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

$H_a :$ not H_0

test statistic = 85.918

p-value $= p(F_{4,94} > 85.918) < 2.2 * 10^{-16}$

Reject the null hypothesis and conclude that the treatments (different fertilizers) have an effect on weight of harvested tomatoes.

$H_0 : \beta_j = 0$

$H_a :$ not H_0

test statistic $= 5860.194$

p-value $= p(F_{1,94} = 5860.194) < 2.2 * 10^{-16}$

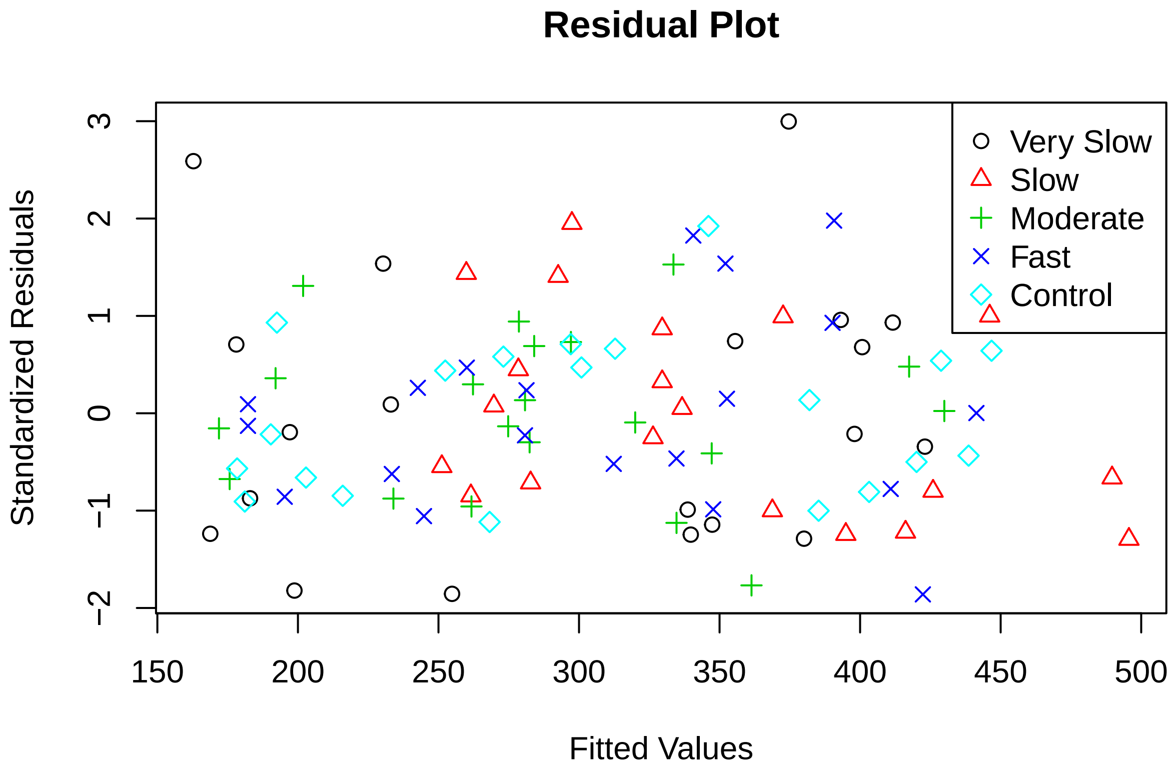
Reject the null hypothesis and conclude that height has an effect on weight of harvested tomatoes

F) Based on the model in Part A, make use of standardized residual plots to assess validity of the assumptions of independence, equal variance, and normality.

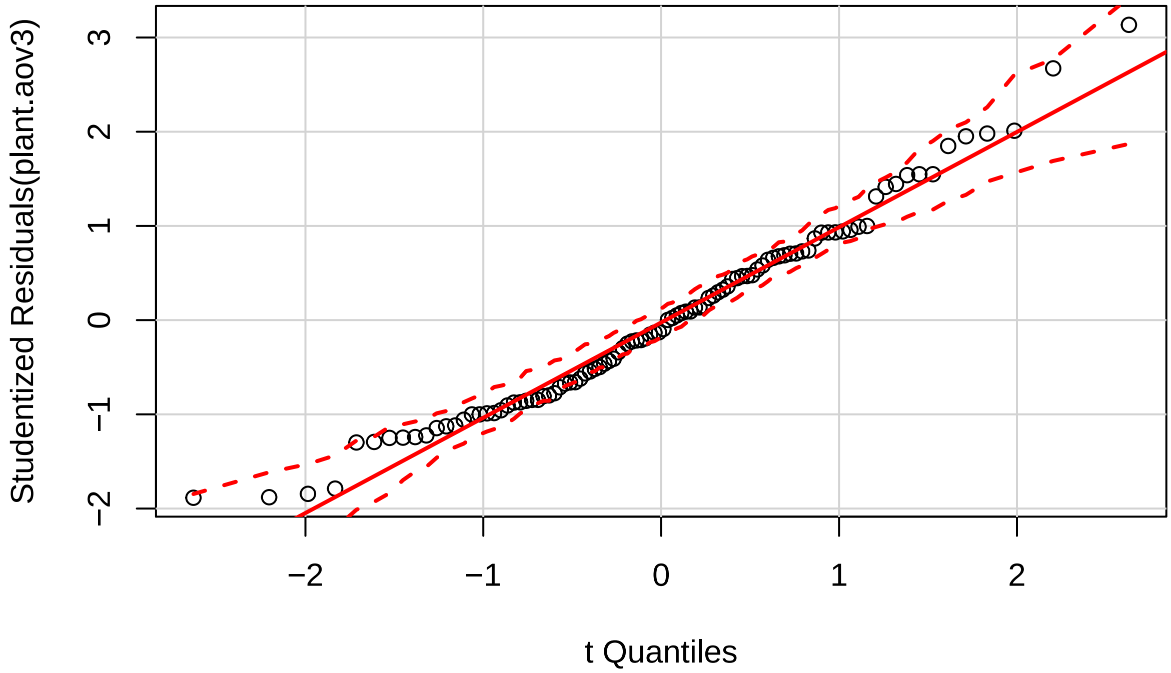
```r
plant.resid = rstandard(plant.aov3)

plot(plant.aov3$fitted.values, plant.resid, pch = as.numeric(plant.data$Treatment),
     col = as.numeric(plant.data$Treatment), xlab = "Fitted Values",
     ylab = "Standardized Residuals", main = "Residual Plot")
legend("topright", pch = c(1, 2, 3, 4, 5), col = c(1, 2, 3, 4, 5),
       legend = c("Very Slow", "Slow", "Moderate", "Fast", "Control"))

library(car)
```



```r
qqPlot(plant.aov3)
```



The assumption of independence holds, since complete randomization techniques were used.

From the above residual plot, we verify the validity of equal variance assumption.

From the above qqplot, we verify the validity of normality assumption.

G) What multiple comparison method that we've used would be more appropriate to compare each treatment to the control treatment? Use such a method at $\alpha_{overall} = 0.05$ and state carefully your conclusions.

We can use Dunnett's to figure out which treatments differ from the control.

```
#DUNNETT
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 3.3.3

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```r
summary(glht(plant.aov3, linfct = mcp(Treatment = 'Dunnett')))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = Weight ~ Treatment + Height, data = plant.data)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## A - E == 0   61.153      3.408  17.942  < 1e-04 ***
## B - E == 0    4.730      3.410   1.387 0.443255
## C - E == 0   13.969      3.404   4.104 0.000314 ***
## D - E == 0   13.915      3.405   4.087 0.000345 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Since we want to see which treatments differ from the control, we will look at those that differ from treatment "E" (the control).

We see that treatment A (very slow release) is significantly different from the control ($p < 0.001$).

Treatment B (slow release) is not significantly different from the control ($p = 0.443$)

Treatment C (moderate release) is significantly different from the control ($p < 0.001$)

Treatment D (fast release) is significantly different from the control ($p < 0.001$)

Note: We could have used Tukey's here, and would probably have drawn the same conclusions (with such low p-values in all significant differences). However, since Dunnett's only controls $\alpha_{overall}$ for comparisons with the control, while Tukey's controls for all possible pairwise comparisons, using Tukey's would cause a loss of power. (Tukey's controls for MANY more pairwise comparisons than Dunnett's-and therefore has less power.)

Question 2.

A field trial is run to test the productivity of three different varieties of strawberries in an experimental field station in New York State. Four equally sized fields are available for use, and one-third of each field is planted in one variety of strawberry, the density of plants kept the same throughout the trial. Each 1/3 field is then randomly assigned one of the three varieties in such a way that each field has all three varieties planted within it. The Yield of strawberries (kg) over a two week period is then recorded for each 1/3 field.

A) GiVe a model statement for this experiment. Define each term and any constraints or conditions that are attached to it.

```r
#load the data
library(readxl)
berry.data <- read_excel("Hwk8Q2DatSp17.xlsx")

berry.data$Var = as.factor(berry.data$Var)
berry.data$Blk = as.factor(berry.data$Blk)
```

$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

where y_{ij} = the yield in kg from block j and treatment i

μ is the overall mean

α_i is the fixed effect of the ith treatment (variety of strawberry)

β_j is the random effect of jth block (blocks are the fields)

ϵ_{ij} is the error term

the model is subject to the following:

$\beta_j \sim iidN(0, \sigma^2_{Blocks})$ and $\epsilon_{ij} \sim iidN(0, \sigma^2)$.

B) Test to see if there are any differences between the yields of the various strawberries. State hypotheses, teststatistic, p-value, and conclusions.

Including the random block in the analysis, we find:

```r
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.3.3
```

```
## Loading required package: Matrix
```

```r
Full = lmer(Yield ~ Var + (1|Blk), data = berry.data, REML = F)
Red = lmer(Yield ~ (1|Blk), data = berry.data, REML = F)
anova(Red, Full)
```

```
## Data: berry.data
## Models:
## Red: Yield ~ (1 | Blk)
## Full: Yield ~ Var + (1 | Blk)
##      Df    AIC    BIC   logLik deviance Chisq Chi Df Pr(>Chisq)
## Red   3 53.896 55.351 -23.9479   47.896
## Full  5 21.476 23.901  -5.7381   11.476 36.42      2  1.235e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$

$H_a :$ not H_0

test statistic = $\chi^2 = 36.42$

Reject the null hypothesis with p = $1.235 * 10^{-8}$ and conclude that strawberry variety does affect yield.

C) Use Tukey's HSD to find the varieties of strawberries which produce different yields. Whiuch would you recommend to be used in New York State (based solely on this yield criteria).

```
library(multcomp)
library(MASS)
diffs = (glht(Full, linfct = mcp(Var = 'Tukey')))
summary(diffs)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lmer(formula = Yield ~ Var + (1 | Blk), data = berry.data, REML = F)
##
## Linear Hypotheses:
##            Estimate Std. Error z value Pr(>|z|)
## B - A == 0  -4.1250     0.2129 -19.378   <1e-08 ***
## C - A == 0  -1.3000     0.2129  -6.107   <1e-08 ***
## C - B == 0   2.8250     0.2129  13.271   <1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We can conclude that B has significantly smaller yields than A ($p < 1 * 10^{-9}$)

C has significantly smaller yields than A (p = 1.86*10^{-9})

C has significantly larger yields than B (p = 1*10^{-9})

Since larger yields would be of more interest, I would recommend variety A be used in New York state.

D) What are the the values of the two compnents of variance in this study? What proportion of the variance of yield is attributable to the differences in fields?

```
summary(Full)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Yield ~ Var + (1 | Blk)
##    Data: berry.data
##
##      AIC      BIC   logLik deviance df.resid
##     21.5     23.9     -5.7     11.5        7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.70450 -0.33109  0.08301  0.46652  1.20209
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Blk      (Intercept) 0.11333  0.3367
##  Residual             0.09063  0.3010
## Number of obs: 12, groups:  Blk, 4
##
```

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  10.3000     0.2258   45.61
## VarB         -4.1250     0.2129  -19.38
## VarC         -1.3000     0.2129   -6.11
##
## Correlation of Fixed Effects:
##      (Intr) VarB
## VarB -0.471
## VarC -0.471  0.500
```

the components of variance are:

variance of the filed: 0.11333

variance of the residual: 0.09063

the proportion of the variance of yield attributable to differences in fields is: $\frac{\sigma^2_{blocks}}{\sigma^2_{blocks}+\sigma^2_{error}} = 0.11333/(0.11333+0.09063) = 0.5556482$

```
0.11333 / (0.11333 + 0.09063)
```

```
## [1] 0.5556482
```

E) Use library(lme4) and library(LmerTest) to test to see if the random effects are statistically significant in this model. (Note: There is some debate amongst statisticians that this procedure is appropriate when using REML. Some say it is OK to use REML with this test IF THE FIXED EFFECTS ARE THE SAME IN BOTH THE FULL AND REDUCED MODELS, which how it is stated in our text.)

```
library(nlme)
```

```
## Warning: package 'nlme' was built under R version 3.3.3
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
Full2 = lme(Yield ~ Var, random = ~1|Blk, data = berry.data, method = "ML")
Red2 = gls(Yield ~ Var, data = berry.data, method = "ML")
anova(Red2, Full2)
```

```
##       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## Red2      1  4 22.97645 24.91608 -7.488225
## Full2     2  5 21.47625 23.90079 -5.738126 1 vs 2 3.500197  0.0614
```

$H_0 : \sigma^2_{blocks} = 0$

$H_a : \sigma^2_{blocks} > 0$

Fail to reject the null hypothesis at $\alpha = .05$ (p = 0.0614). Cannot conclude that the blocks (fields) affect strawberry yield at a 5% significance level.

F) State which of our observations are correlated in this situation.

Since the random effect of blocks was found *not* to be significant in part E, then we don't have any evidence that the observations (i.e. observations from same field) are correlated.