STAT850 HW5

Stewart Kerr

February 25, 2019

Problem 1

See attached sheet.

Problem 2

a. My interpretation of the instructions for this problem are that we should have one row per branch. I will first reshape the data as such.

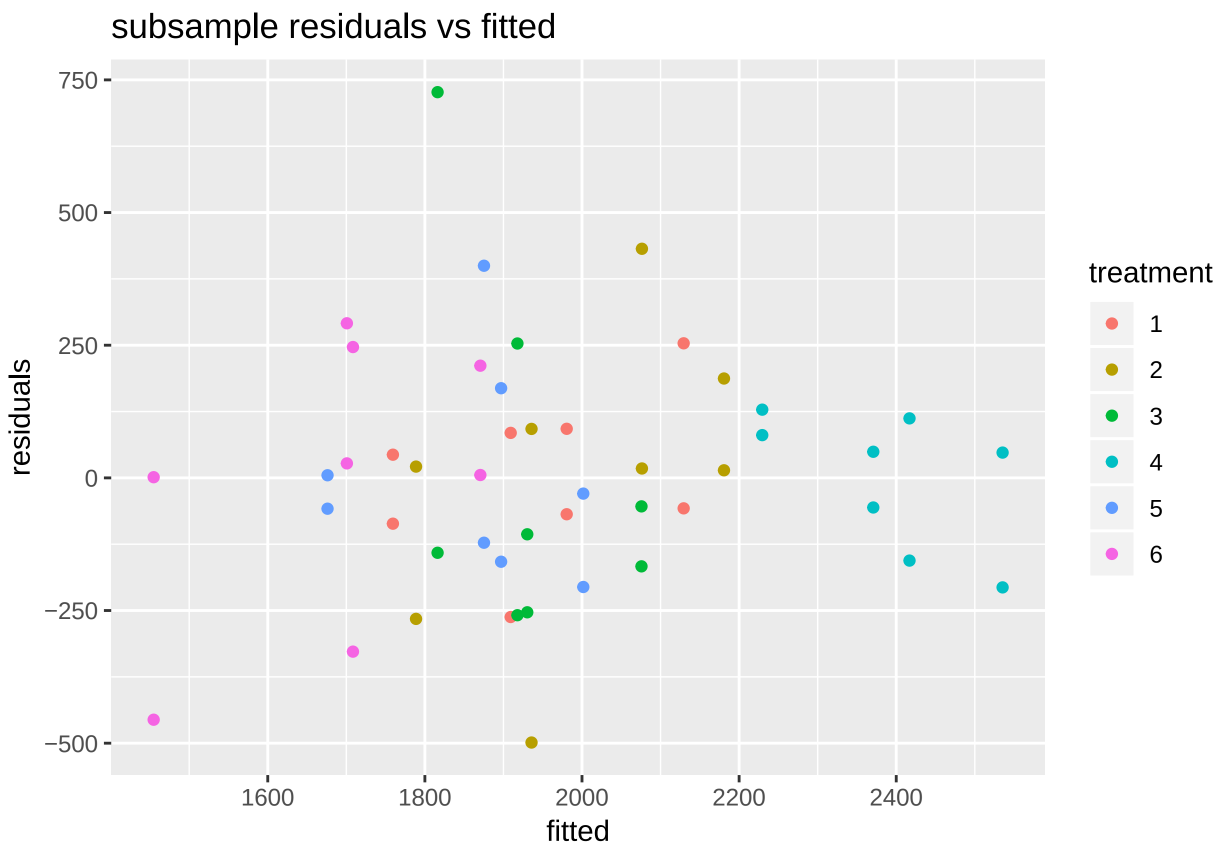
```
#Problem 2a
#Load the data in
apple <- read.csv("~/2019spring/STAT850/hw4/apple.csv")
apple_clean = apple
apple$treatment = factor(apple$treatment)
apple$block = factor(apple$block)
apple = mutate(apple, tree = row_number())

#Reshape data so that we have 1 row per branch
apple1 = dplyr::select(apple, -weight_b2) %>% rename(weight = weight_b1)
apple2 = dplyr::select(apple, -weight_b1) %>% rename(weight = weight_b2)
apple = rbind(apple1,apple2)
```

Now, the experimental design is a CRBD with subsampling. First, I will conduct an overall F-test to determine if the treatments are different. I will also perform residual analysis to examine the validity of my method.

```
#ANOVA LM with block
apple_lmer <- lmer(weight ~ treatment + block + (1|tree), data = apple)
anova(apple_lmer)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##            Sum Sq Mean Sq NumDF DenDF F value   Pr(>F)
## treatment 1788674  357735     5    15  5.4385 0.004738 **
## block      604030  201343     3    15  3.0609 0.060540 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

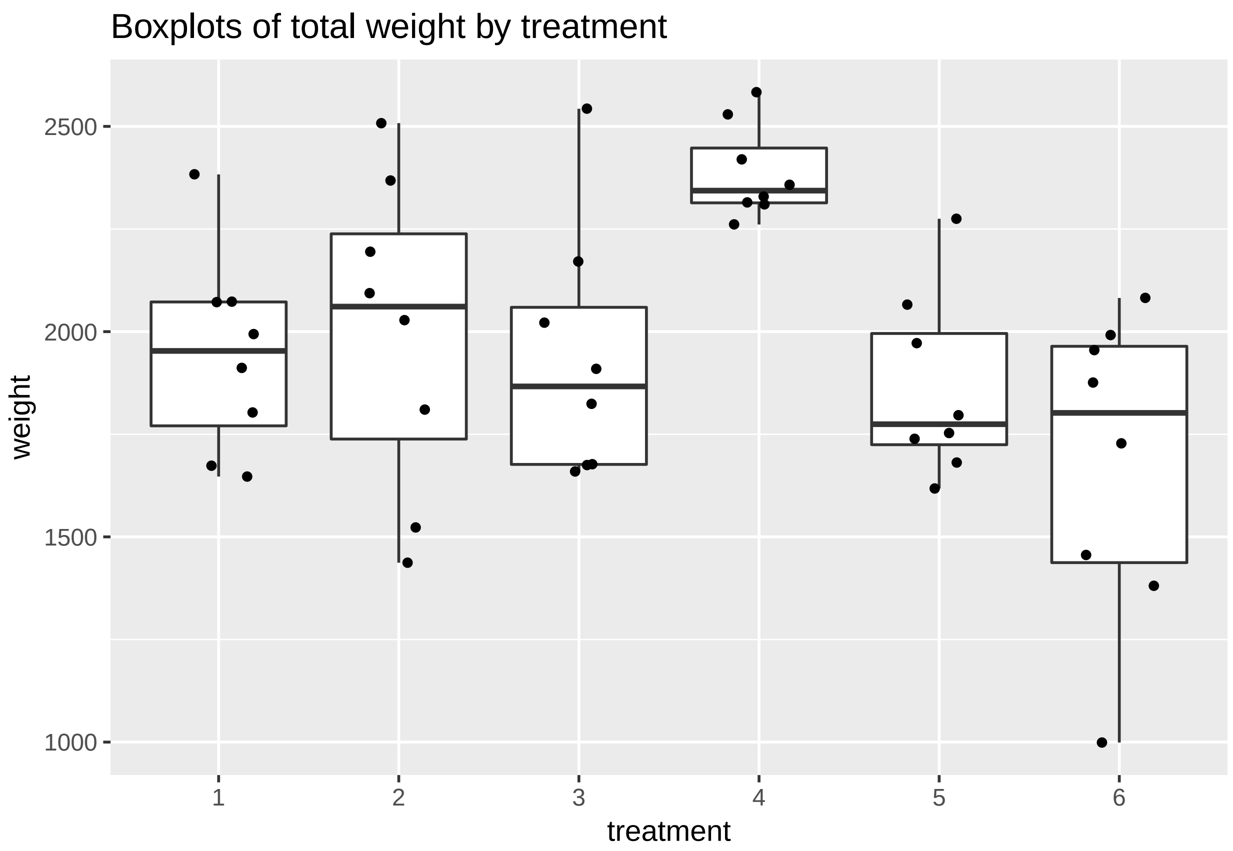
```
#Plot residuals
ggplot(apple, aes(y=weight-fitted(apple_lmer), x=fitted(apple_lmer), color = treatment)) + geom_point()
```



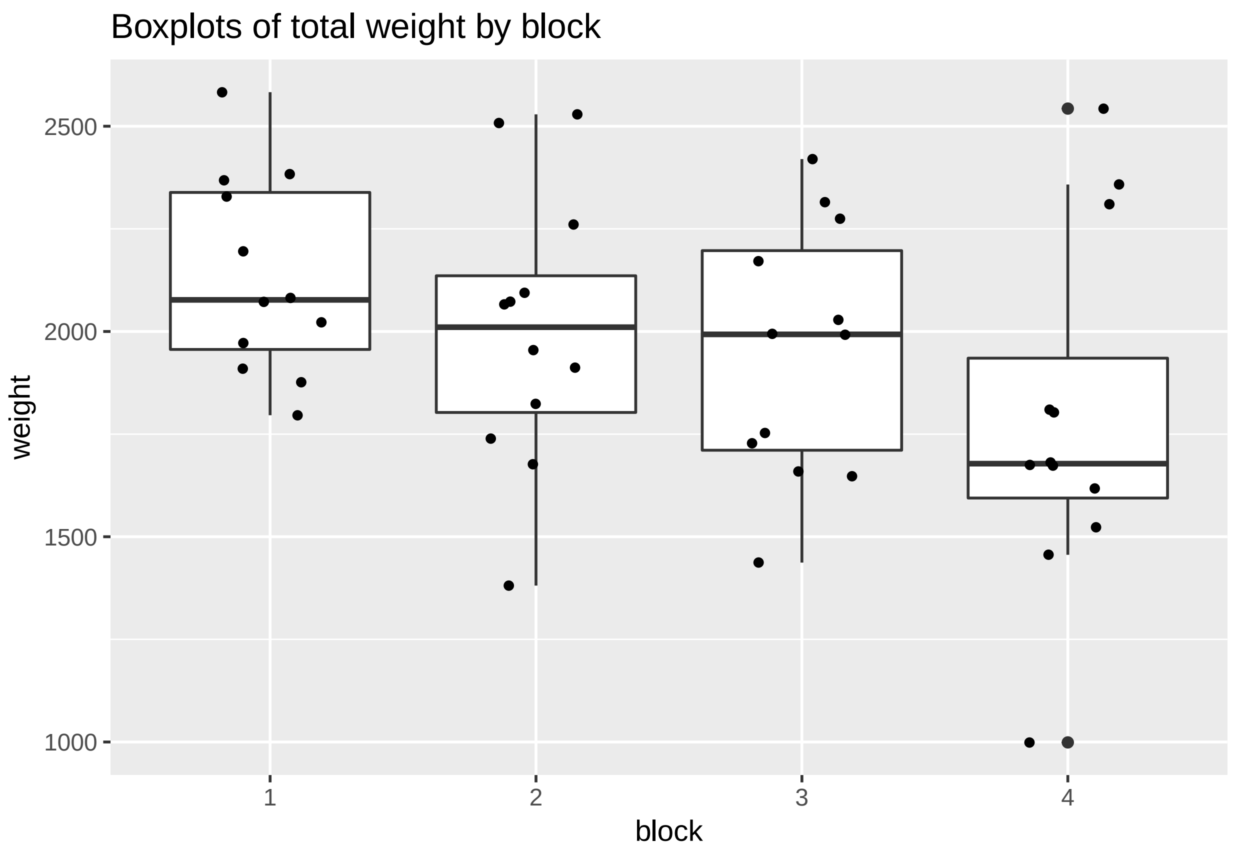
From the ANOVA table, we are testing the hypothesis that H_0 : The mean of each treatment is the same vs. they are different. From the ANOVA function, we get an F-statistic of 5.44 and a p-value of 0.0047. Thus, there is strong evidence to reject the null hypothesis. If we look at the residual plot for samples, we see fairly even spread and no clear patterns. There may be a few outliers, but overall the residuals look good.

b.

```r
#Problem 2b
#Plot the data
#Create a grouped boxplot
ggplot(apple, aes(x=treatment, y=weight, group=treatment)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  ggtitle('Boxplots of total weight by treatment')
```



```r
ggplot(apple, aes(x=block, y=weight, group=block)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  ggtitle('Boxplots of total weight by block')
```



Again, I would definitely recommend treatment 4 in order to maximize the apple yield. From the boxplots, it's clear that treatment 4 outperforms the other treatments. However, with the increase in variation from subsampling, a few of the other treatments come close to the performance of treatment 4. Treatment 2 in particular is pretty close.

It's also worth noting that with subsampling, block 4 no longer appears to have significantly performed much worse than the other blocks. I would still recommend not planting in the low ground if possible.

c. At first glance, it appears that there is a strong, positive interaction between planting flowers and mowing. This is because each of treatments 1-3 have about the same average branch weight (indicating little to no main effects of mowing and flowers by themselves) yet treatment 4 has a much larger average branch weight. In order to confirm this hypothesis, I will calculate Fisher's LSD and see if treatment 4 differs significantly from treatments 1,2, and 3.

```
#Problem 2c
apple_aov = aov(weight ~ treatment + block + (1|tree), data = apple)
model.tables(apple_aov, type = "means", se = TRUE)
```

```
## Tables of means
## Grand mean
##
## 1968.208
##
##  treatment
## treatment
##      1      2      3      4      5      6
## 1944.6 1995.4 1935.0 2388.1 1862.5 1683.6
##
```

```
##  block
## block
##      1      2      3      4
## 2132.3 2001.6 1951.6 1787.4
##
## Standard errors for differences of means
##         treatment block
##             133.4 108.9
## replic.         8    12
```

```
LSD = qt(0.975, df = 18)*327.2*sqrt(2/4)
paste("The LSD for the apple data is: ", round(LSD,1))
```

```
## [1] "The LSD for the apple data is:  486.1"
```

Based on this calculated LSD, it appears that treatment 4 no longer differs significantly from the treatments 2 - indicating that there may not be a significant interaction between planting flowers and mowing. I believe this discrepancy to be due to the increased variance that comes from subsampling. However, I think this analysis is not as strong as the analysis performed in hw4 before there is a lot of correlation among branches that violates our independence assumption for the subsampling term.

Problem 3

a.

```
#Problem 3a
#Read data in and look for "typos"
prairie <- read.csv("~/2019spring/STAT850/hw5/prairiespecies.csv")

#Make sure number of treatments = 20
kable(count(group_by(prairie,trt)))
```

trt	n
T1	20
T2	20
T3	20
T4	20

```
#Make sure number of each species = 20
kable(count(group_by(prairie,species)))
```

species	n
A	16
B	20
C	24
D	20

```
#Looks like some of our As are labeled as Cs - fixing below
prairie = bind_rows(filter(prairie, trt != "T1"), mutate(filter(prairie, trt == "T1"), species = "A"))
prairie$species = factor(prairie$species)

#Make sure we have 4 data points for each plot
```

```r
kable(count(group_by(prairie, plot)))
```

plot	n
P1	4
P10	4
P11	4
P12	4
P13	4
P14	4
P15	4
P16	4
P17	4
P18	4
P19	4
P2	4
P20	4
P3	4
P4	4
P5	4
P6	4
P7	4
P8	4
P9	4

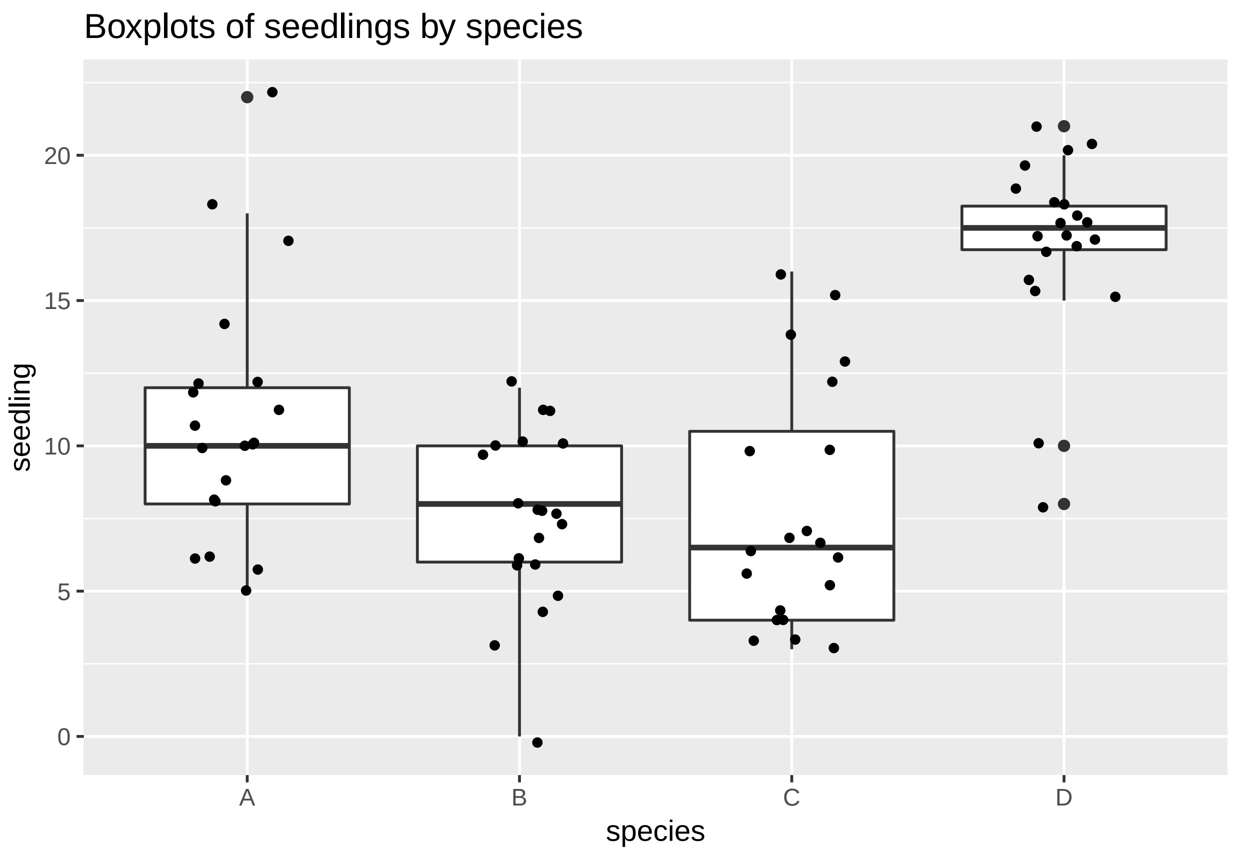
```r
#Make sure we have 20 points in each square
kable(count(group_by(prairie,square)))
```

square	n
1	20
2	20
3	20
4	20

Above are a few checks to make sure that the data doesn't have any clear errors. An error was found where some of the data for species A was transcribed as data for species C. This typo was fixed. I could have performed a few other checks (such as making sure each plot had each square) - but I didn't want the output to be too long.

b.

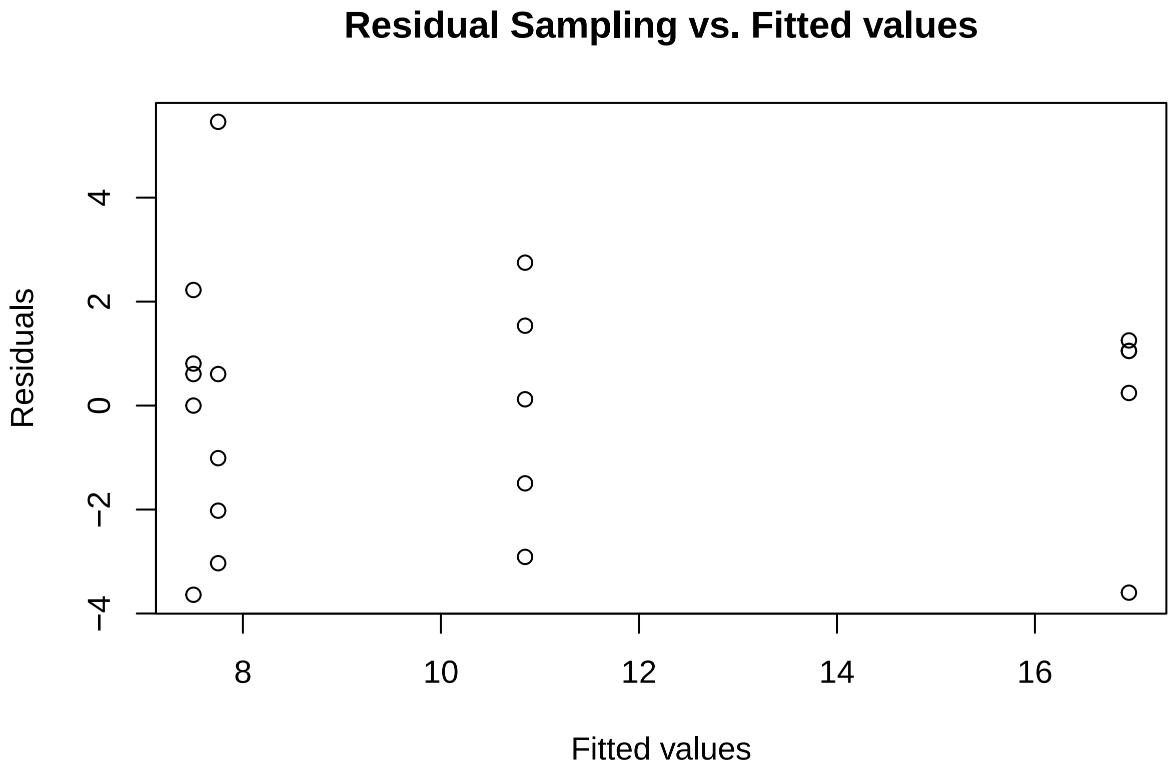
```r
#Problem 3b
#Graph data
ggplot(prairie, aes(x=species, y=seedling, group=species)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  ggtitle('Boxplots of seedlings by species')
```



```
#Analyze difference among species with random effect of plot
prairie_lmer <- lmer(seedling ~ species + (1 | species:plot), data = prairie)
anova(prairie_lmer)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##         Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
## species 221.98  73.993     3    16  10.005 0.0005929 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Look at residuals, LSD
prairie_re = ranef(prairie_lmer)
tmp = data.frame(resid = prairie_re[[1]][,1], species = substr(rownames(prairie_re[[1]]),1,1))
tmp$fitted = predict(prairie_lmer, tmp, re.form = NA)
plot(tmp$resid ~ tmp$fitted, main = "Residual Sampling vs. Fitted values", xlab = "Fitted values", ylab
```



```
difflsmeans(prairie_lmer, test.effs = "Group")
```

```
## Least Squares Means table:
##
##                      Estimate Std. Error df t value     lower     upper
## speciesA - speciesB   3.35000    1.96612 16  1.7039  -0.81799   7.51799
## speciesA - speciesC   3.10000    1.96612 16  1.5767  -1.06799   7.26799
## speciesA - speciesD  -6.10000    1.96612 16 -3.1026 -10.26799  -1.93201
## speciesB - speciesC  -0.25000    1.96612 16 -0.1272  -4.41799   3.91799
## speciesB - speciesD  -9.45000    1.96612 16 -4.8064 -13.61799  -5.28201
## speciesC - speciesD  -9.20000    1.96612 16 -4.6793 -13.36799  -5.03201
##                      Pr(>|t|)
## speciesA - speciesB 0.1077465
## speciesA - speciesC 0.1344264
## speciesA - speciesD 0.0068431 **
## speciesB - speciesC 0.9004028
## speciesB - speciesD 0.0001938 ***
## speciesC - speciesD 0.0002513 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##   Confidence level: 95%
##   Degrees of freedom method: Satterthwaite
```

First, I will test the hypothesis $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$. I will use a linear model with treatment as a fixed effect and plot as a random effect. From the ANOVA output, we have a test statistic of F = 10.0 and a p-value of 0.0006. There is very strong evidence that the average number of seedlings that sprouted differs based on species. From the boxplot, it appears that species D may be the best performing species.The difflsmeans function gives us confidence intervals for the difference in each group mean. From the confidence intervals, we confirm that species D significantly differs from the others, while the others do not differ significantly among themselves. From the residual plot, it appears that are residuals are fairly normally distributed around 0 and homoscedastic for each species. I have a little concern that we have residuals that are around 4 - these might be outliers. However, I will accept these residuals as valid for my ANOVA analysis. Next, I will group the "mature" species and "transitional" species and repeat my analysis.

```
#Problem 3b
#Group the mature and transitional species
prairie <- mutate(prairie, species_type = case_when(
  species == "A" | species == "D" ~ "mature",
  species == "C" | species == "B" ~ "transitional"))
prairie$species_type = factor(prairie$species_type)

#Perform a t-test on the grouped means
t.test(seedling ~ species_type, data = prairie, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  seedling by species_type
## t = 6.5427, df = 72.37, p-value = 3.74e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.676979      Inf
## sample estimates:
##       mean in group mature mean in group transitional
##                     13.900                      7.625
```

After grouping the species into the mature and transitional species_type groups, a t-test was conducted to examine the null hypothesis $H_0 : \mu_{mature} = \mu_{transitional}$ vs $H_A : \mu_{mature} > \mu_{transitional}$. The calculated t-test statistic is 6.54 with 72 degrees of freedom. Thus, a p-value of 3.74e-9 was calculated for this test. There is very strong evidence to suggest that the mature species do have a higher average number of seedlings than the transitional group.

c. The model for the response of treatment A is $Y = \mu + \alpha_1 + \bar{\epsilon_{1..}} + \bar{\delta_{1...}}$. Here, both μ and α_i are fixed effects, thus the variance is only due to $\bar{\epsilon_{1..}} + \bar{\delta_{1...}}$. The variance of $\bar{\epsilon_{1..}}$ is σ_ϵ^2/n while the variance of $\bar{\delta_{1..}}$ is σ_δ^2/ns. Thus the variance of $\hat{\mu_A}$ is $(\sigma_\epsilon^2/n) + (\sigma_\delta^2/ns) = (s\sigma_\epsilon^2 + \sigma_\delta^2/ns)$ which is estimated with (MS Plot Error)/ns. We can get MS plot error from the ANOVA table by looking at the row for species:plot.

```
#Problem 3c
prairie_lm = lm(seedling ~ species + species:plot, data = prairie)
anova(prairie_lm)
```

```
## Analysis of Variance Table
##
## Response: seedling
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## species       3 1160.24  386.75 52.2924 < 2.2e-16 ***
## species:plot 16  618.50   38.66  5.2268 1.175e-06 ***
## Residuals    60  443.75    7.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, MS Plot Error = 38.66. Then, we estimated the the variance of $\hat{\mu_A}$ to be $38.66/(8 * 2) = 2.42$.

Problem 4

See attached sheet.

Code

```
## ----include = FALSE-----------------------------------------------------
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(knitr)
library(car)
library(ggplot2)
library(MASS)
library(lme4)
library(lmerTest)
library(tidyr)

set.seed(1104)             # make random results reproducible

this_file <- "kerr_stat850_hw05.Rmd"  # used to automatically generate code appendix

## ------------------------------------------------------------------------
#Problem 2a
#Load the data in
apple <- read.csv("~/2019spring/STAT850/hw4/apple.csv")
apple_clean = apple
apple$treatment = factor(apple$treatment)
apple$block = factor(apple$block)
apple = mutate(apple, tree = row_number())

#Reshape data so that we have 1 row per branch
apple1 = dplyr::select(apple, -weight_b2) %>% rename(weight = weight_b1)
apple2 = dplyr::select(apple, -weight_b1) %>% rename(weight = weight_b2)
apple = rbind(apple1,apple2)

## ------------------------------------------------------------------------
#ANOVA LM with block
apple_lmer <- lmer(weight ~ treatment + block + (1|tree), data = apple)
anova(apple_lmer)

#Plot residuals
ggplot(apple, aes(y=weight-fitted(apple_lmer), x=fitted(apple_lmer), color = treatment)) + geom_point()


## ------------------------------------------------------------------------
#Problem 2b
#Plot the data
#Create a grouped boxplot
ggplot(apple, aes(x=treatment, y=weight, group=treatment)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  ggtitle('Boxplots of total weight by treatment')
ggplot(apple, aes(x=block, y=weight, group=block)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  ggtitle('Boxplots of total weight by block')

## ----------------------------------------------------------------------
#Problem 2c
apple_aov = aov(weight ~ treatment + block + (1|tree), data = apple)
model.tables(apple_aov, type = "means", se = TRUE)

LSD = qt(0.975, df = 18)*327.2*sqrt(2/4)
paste("The LSD for the apple data is: ", round(LSD,1))

## ---- warning = FALSE--------------------------------------------------
#Problem 3a
#Read data in and look for "typos"
prairie <- read.csv("~/2019spring/STAT850/hw5/prairiespecies.csv")

#Make sure number of treatments = 20
kable(count(group_by(prairie,trt)))

#Make sure number of each species = 20
kable(count(group_by(prairie,species)))
#Looks like some of our As are labeled as Cs - fixing below
prairie = bind_rows(filter(prairie, trt != "T1"), mutate(filter(prairie, trt == "T1"), species = "A"))
prairie$species = factor(prairie$species)

#Make sure we have 4 data points for each plot
kable(count(group_by(prairie, plot)))

#Make sure we have 20 points in each square
kable(count(group_by(prairie,square)))

## ----------------------------------------------------------------------
#Problem 3b
#Graph data
ggplot(prairie, aes(x=species, y=seedling, group=species)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  ggtitle('Boxplots of seedlings by species')

#Analyze difference among species with random effect of plot
prairie_lmer <- lmer(seedling ~ species + (1 | species:plot), data = prairie)
anova(prairie_lmer)

#Look at residuals, LSD
prairie_re = ranef(prairie_lmer)
tmp = data.frame(resid = prairie_re[[1]][,1], species = substr(rownames(prairie_re[[1]]),1,1))
tmp$fitted = predict(prairie_lmer, tmp, re.form = NA)
plot(tmp$resid ~ tmp$fitted, main = "Residual Sampling vs. Fitted values", xlab = "Fitted values", ylab
difflsmeans(prairie_lmer, test.effs = "Group")

## ----------------------------------------------------------------------
#Problem 3b
```

```
#Group the mature and transitional species
prairie <- mutate(prairie, species_type = case_when(
  species == "A" | species == "D" ~ "mature",
  species == "C" | species == "B" ~ "transitional"))
prairie$species_type = factor(prairie$species_type)

#Perform a t-test on the grouped means
t.test(seedling ~ species_type, data = prairie, alternative = "greater")

## -----------------------------------------------------------------------
#Problem 3c
prairie_lm = lm(seedling ~ species + species:plot, data = prairie)
anova(prairie_lm)

## ----code = readLines(purl(this_file, documentation = 1)), echo = T, eval = F----
## # this R markdown chunk generates a code appendix
```