Homework 9: ANOVA

NAME: Michael Darfler

NETID: mbd25

DUE DATE: November 27, 2019 by 11:59pm

Instructions

For this homework:

1. All calculations must be done within your document in code chunks. Provide all intermediate steps.
2. DO NOT JUST INCLUDE A CALCULATION: Incude any formulas you are using for a calculation. You can put these immediately before the code chunk where you actually do the calculation.

Sleep Study Data

This dataset includes observations of the following variables from a random sample of 253 college students.

Variable	Description
`Gender`	1 = Male 0 = Female
`ClassYear`	Year in school, 1=first year,…, 4=Senior
`LarkOwl`	Early riser or night owl? Responses: `Lark` , `Owl` , or `Neither`
`NumEarlyClass`	Number of early classes each week (before 9am)
`EarlyClass`	Indicator for at least 1 early class
`GPA`	Grade Point Average
`ClassesMissed`	Number of classes missed in a semester
`CognitionZscore`	Z-score on a test of cognitive skills
`PoorSleepQuality`	Higher values indicate poorer sleep
`DepressionScore`	Measure of degree of depression
`AnxietyScore`	Measure of amount of anxiety

Variable	Description
`StressScore`	Measure of amount of stress
`DepressionStatus`	`normal` , `moderate` , or `severe`
`AnxietyStatus`	`normal` , `moderate` , or `severe`
`Stress`	`normal` or `high`
`DASScore`	Combined score for depression, anxiety and stress
`Happiness`	Measure of degree of happiness
`AlcoholUse`	`Abstain` , `Light` , `Moderate` , or `Heavy`
`Drinks`	Number of alcoholic drinks per week
`WeekdaySleep`	Average hours of sleep on the weekdays
`WeekendSleep`	Average hours of sleep on the weekend days
`AverageSleep`	Average hours of sleep for all days
`AllNighter`	Had an all-nighter this semester? 1 = yes 0 = no

Problem 1

Does mean GPA for college students change by class year? At a significance level of 0.05, test this hypothesis. The first step is reading in the `SleepStudy` data from the folder for Homework 9. Then, do the following:

```
student_data <- read.csv("https://raw.githubusercontent.com/mdarfler/BTRY_6010/master/Homework/HW%209/SleepStudy(5).csv")
student_data$ClassYear <- as.factor(student_data$ClassYear)
```

a. State the null and alternative hypotheses.

H0: All means are the same
HA: At least one mean is different

b. Compute the total degrees of freedom, the degrees of freedom associated with the numerator of the F statistic, and the error degrees of freedom.

DoF associated with the numerator is the number of factors - 1
DoF associated with the error is the number of observations - number of categories
The total DoF is the total number of observations - 1

```
dof_n <- length(unique(student_data$ClassYear)) - 1
dof_e <- nrow(student_data) - 3


paste0("The number of degrees of freedom associated with the numberator of the F
statistic is; ", dof_n)
```

```
## [1] "The number of degrees of freedom associated with the numberator of the F
statistic is; 3"
```

```
paste0("The number of error degrees of freedom is: ", dof_e)
```

```
## [1] "The number of error degrees of freedom is: 250"
```

```
paste0("The total number of degrees of freedom is: ", nrow(student_data) - 1)
```

```
## [1] "The total number of degrees of freedom is: 252"
```

c. Perform the test in R and include the ANOVA table here.

```
student_data.lm <- lm(GPA ~ ClassYear, data = student_data)
aov <- anova(student_data.lm)
print(aov)
```

```
## Analysis of Variance Table
##
## Response: GPA
##             Df Sum Sq Mean Sq F value    Pr(>F)
## ClassYear    3  5.133  1.7109  11.816 2.914e-07 ***
## Residuals 249 36.056  0.1448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d. What is the sse and mse? Include how you would calculate the mse from the sse.

SSE is the sum of squares of error and is calculated as
$SSE = \sum_{j=i}^{k} \sum_{i=1}^{k} (Y_{ij} - \bar{Y}_{i.})^2$
MSE is the mean square of errors and is calculated as $MSE = \frac{SSE}{df2}$

```
fr = subset(student_data, ClassYear == 1, select = `GPA`)
so = subset(student_data, ClassYear == 2, select = `GPA`)
jr =  subset(student_data, ClassYear == 3, select = `GPA`)
sr = subset(student_data, ClassYear == 4, select = `GPA`)

SSE = sum(c(sum((fr$GPA - mean(fr$GPA))^2),sum((so$GPA - mean(so$GPA))^2),sum((j
r$GPA - mean(jr$GPA))^2),sum((sr$GPA - mean(sr$GPA))^2)))
print(SSE)
```

```
## [1] 36.05557
```

```
df2 = nrow(student_data) - 4

SSE/df2
```

```
## [1] 0.1448015
```

e. What is the ssb and msb? Include how you would calculate the msb from the ssb.

SSB is the sum of squares between and is calculated as $\sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$
MSB is the mean square between and is calculated as $\frac{SSB}{df1}$

```
df1 = aov$Df[1]
paste0("SSB = ", aov$`Sum Sq`[1])
```

```
## [1] "SSB = 5.1327856922455"
```

```
paste0("MSB = ", aov$`Mean Sq`[1])
```

```
## [1] "MSB = 1.71092856408183"
```

f. What is the realization of the F statistic? Include a formula.

F = MSB/MSE

```
f <- aov$`F value`[1]
paste0("The realization of the F statistic is ", f)
```

```
## [1] "The realization of the F statistic is 11.8156832166259"
```

g. What is the p-value of this test? Calculate the p-value using the `pf()` function in a code chunk here. Verify that this p-value matches the p-value in the ANOVA table. Should the null hypothesis be rejected based on this p-value? Interpret your conclusion in the context of this problem.

```
1 - pf(f,df1,df2)
```

```
## [1] 2.913781e-07
```

```
aov$`Pr(>F)`[1]
```

```
## [1] 2.913781e-07
```

h. Follow these steps to check the assumptions required to perform an ANOVA test.

```
i) Does the assumption of independence hold? (Suppose you could talk to the resea
rchers who collected these data.  What would be relevant to ask them?)
```

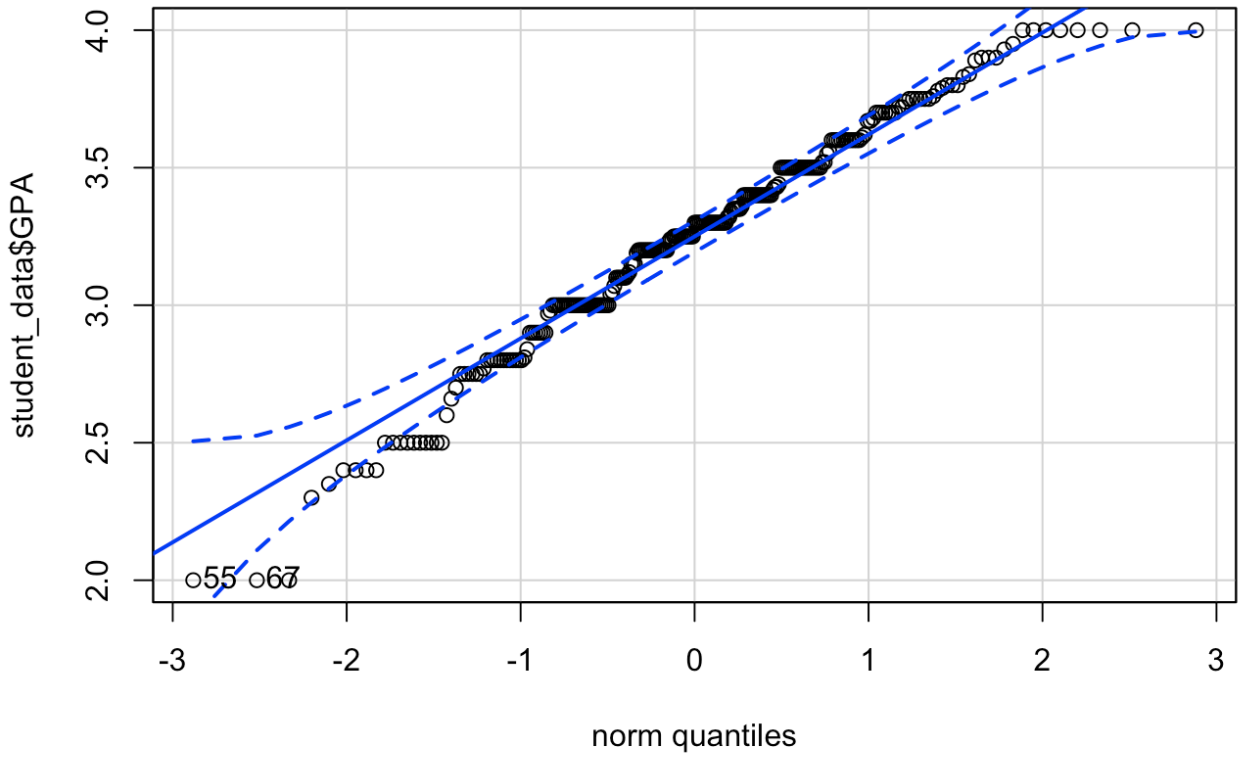
I wouls ask if the student GPA was based primarily on individual work an that group projects were a minimal factor in calculating the GPA.

```
ii) Check using a graphical method if the assumption of normality holds. Does it
hold?
```

```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(student_data$GPA)
```




```
## [1] 55 67
```

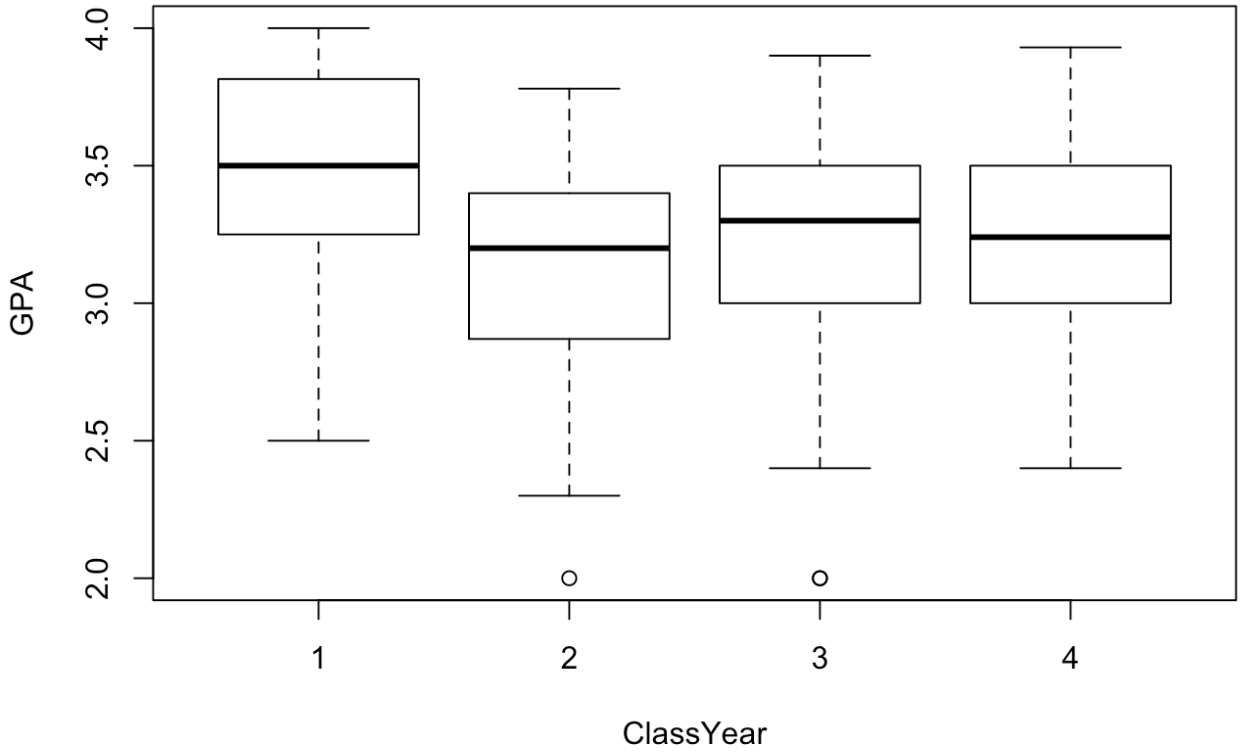
The qqplot shows that the data appear to be normally distributed

```
iii) Check using a graphical method and a formal test if the assumption of equal
variance holds.  Does it hold?
```

```
leveneTest(student_data.lm)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.0799 0.9709
##       249
```

```
boxplot(GPA~ClassYear, data = student_data)
```



Leven's test for homogeneity of Variance returned a p value of .97 so that we can accept that the variance is the same for all populations.

```
iv) Are the assumptions required to perform an ANOVA hypothesis test met?
```

Yes. All assumptions are met.

Problem 2

Which of the class years have significantly different mean GPAs? Conduct all pairwise tests using the Tukey HSD method. Assume the significance level of these tests is 0.01. State your conclusion.

```
library(multcomp)
```

```
## 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
```

```
student_data.tukey = glht(student_data.lm, linfct = mcp(ClassYear = "Tukey"))
summary(student_data.tukey)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = GPA ~ ClassYear, data = student_data)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0 -0.40029    0.06786  -5.899   <0.001 ***
## 3 - 1 == 0 -0.31398    0.07591  -4.136   <0.001 ***
## 4 - 1 == 0 -0.29629    0.07498  -3.952   <0.001 ***
## 3 - 2 == 0  0.08631    0.06485   1.331    0.542
## 4 - 2 == 0  0.10400    0.06375   1.631    0.361
## 4 - 3 == 0  0.01769    0.07226   0.245    0.995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

The following class years have statistically significant difference at the level 0.01 Freshman and Sophomore
Freshman and Juniors
Freshman and Seniors

Problem 3

Does mean GPA for college students change by `AlcoholUse` ? At a significance level of 0.05, test this

hypothesis. For this research question, do the following:

a. Perform an ANOVA test in R and include the ANOVA table here.

```
alcohol.lm = lm(GPA~AlcoholUse, data = student_data)
aov_alcohol = anova(alcohol.lm)
print(aov_alcohol)
```

```
## Analysis of Variance Table
##
## Response: GPA
##             Df Sum Sq Mean Sq F value Pr(>F)
## AlcoholUse   3  0.601 0.20042  1.2296 0.2995
## Residuals  249 40.587 0.16300
```

b. What is the realization of the test statistic?

```
paste0("The realization of the test statistic is: ", aov_alcohol$`F value`[1])
```

```
## [1] "The realization of the test statistic is: 1.22957303879623"
```

c. What is the p-value of this test? Include a formula for the p-value with your answer. Should the null hypothesis be rejected based on the p-value? Use a significance level of 0.05.

```
1 - pf(aov_alcohol$`F value`[1],aov_alcohol$Df[1],aov_alcohol$Df[2])
```

```
## [1] 0.2994964
```

```
aov_alcohol$`Pr(>F)`[1]
```

```
## [1] 0.2994964
```

d. Assuming the assumptions for ANOVA are met, state the conclusion of this test in the context of the research question.

> At the significance level of 0.05 we would reject the accept the null hypothesis that all means are the same.

Problem 4

Consider the partially filled in ANOVA table below. In this study, the means of 5 different treatments were compared. There was 8 observations for each of the 5 populations.

```
tbl <- matrix(c("G", "A", "H",
                "B", "1837", "3037",
                "C", "D", "",
                "E", "", "",
                "F", "", ""), nrow = 3)
rownames(tbl) <- c("Treatment", "Error", "Total")
colnames(tbl) <- c("Df", "Sum Sq", "Mean Sq", "F value", "Pr(>F)")
knitr::kable(tbl)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	G	B	C	E	F
Error	A	1837	D		
Total	H	3037			

a. Determine G, A, and H.

G = 4, A = 35, H = 39

b. Determine B.

ssb = tss - sse = 3037 - 1837 = 1200

c. Determine C.

MSB = SSB/df1 = 1200/4 = 300

d. Determine D.

MSE = SSE/df2 = 1837/35 = 52.4871

e. Determine E.

F* = MSB/MSE = 300/52.4871 5.506

f. Determine F.

```
1 - pf(5.506, 4, 35)
```

```
## [1] 0.001513499
```