Homework 9: ANOVA

NAME: Andres Castano

NETID: ac986

DUE DATE: November 22, 2016 by 1:00pm

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
`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

Initually, let's load the data into R:

```
Sleep_data <- read.csv("SleepStudy.csv")
dim(Sleep_data)
```

```
## [1] 253  27
```

```
Sleep_data$ClassYear=factor(Sleep_data$ClassYear)
```

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:

a) State the null and alternative hypotheses.

There are 4 categories for class year, then the null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A : At least one of the means is not the same as the others

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.

Total degrees of freedon are:

```
ntotal= 253
total_df = ntotal - 1
total_df
```

```
## [1] 252
```

Whereas the degrees of freedon associated with the Mean Square Error (mse), e.g. the measure of the variability within groups are:

```
k = 4
mse_df = ntotal - k
mse_df
```

```
## [1] 249
```

Finally the degrees of freedon associated with Mean Square Between Groups (msb), e.g. the measure of variability between groups are:

```
k = 4
msb_df = k-1
msb_df
```

```
## [1] 3
```

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

Strategy 1:

```
gpa.lm = lm(Sleep_data$GPA ~ Sleep_data$ClassYear, data = Sleep_data)
anova(gpa.lm)
```

```
## Analysis of Variance Table
##
## Response: Sleep_data$GPA
##                        Df Sum Sq Mean Sq F value    Pr(>F)
## Sleep_data$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
```

Strategy 2:

```
fit = aov(Sleep_data$GPA ~ Sleep_data$ClassYear)
summary(fit)
```

```
##                        Df Sum Sq Mean Sq F value   Pr(>F)
## Sleep_data$ClassYear    3   5.13  1.7109   11.82 2.91e-07 ***
## Residuals             249  36.06  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.

The sse is the sum of squares within groups and is calculated as:

$sse = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$

According to the ANOVA output the sse is 36.06. When we divided this result by $n_{total} - k$, we get a Pooled Error Variance, also called Mean Square Error (mse):

$mse = sse/n_{total} - k$, where n_{total} is the total of observations and k is the number of groups, in our case k is the class year of the students.

```
sse = 36.06
mse = sse / mse_df
mse
```

```
## [1] 0.1448193
```

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

The ssb is the sum of squares between groups and is calculated as:

$ssb = \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$

According to the ANOVA output this values is 5.13. When we want to know how much on average a group vary from the grand mean, we compute the Mean Square Between Groups (msb) as:

$msb = ssb/k - 1$

```
ssb = 5.13
msb = ssb / msb_df
msb
```

```
## [1] 1.71
```

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

Under H_0 a realization of the F statistic is $F = msb/mse$, which has a $F_{df1,df2}$ distribution.

```
F =  msb/mse
F
```

```
## [1] 11.80782
```

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?

```
pvalue_F = 1-pf(F,msb_df,mse_df)
pvalue_F
```

```
## [1] 2.943214e-07
```

Given that $0.000000291 < 0.05$, we should reject H_0. We are 95% confident that mean GPA is not the same for all four class year groups.

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

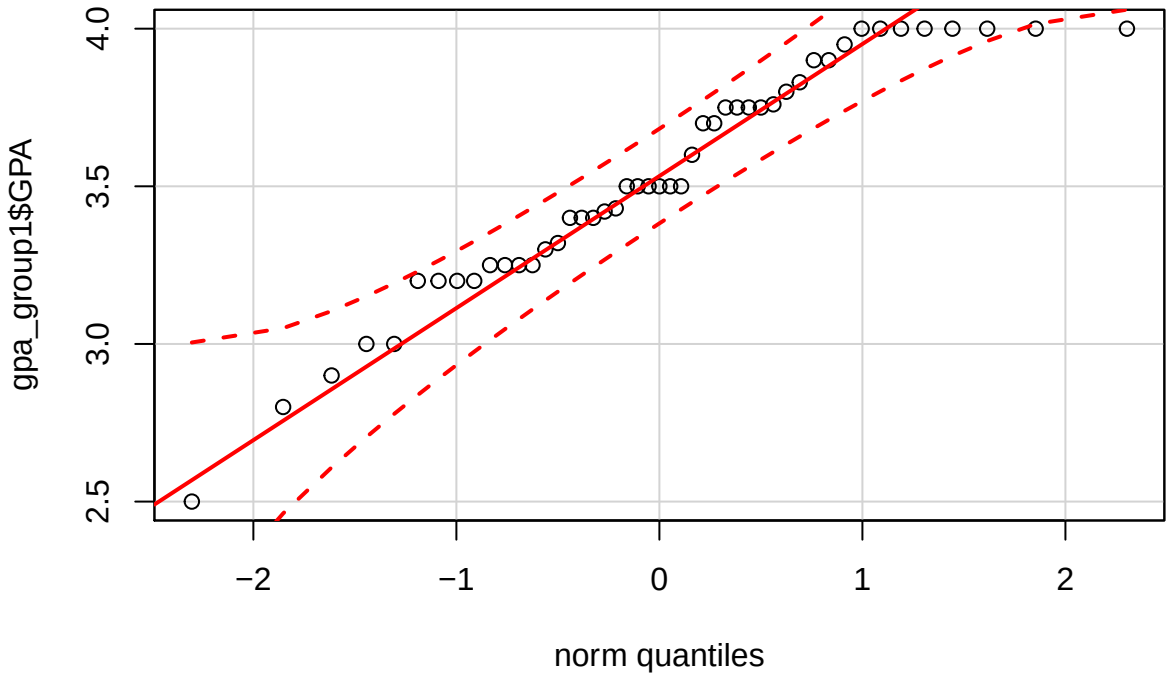
i) Does the assumption of independence hold? (Suppose you could talk to the researchers who collected these data. What would be relevant to ask them?)

it will be relevant to ask the researchers how the k different populations (class year groups) were sampled? and If the researchers point out that the data come from an independent SRS (Simple Random Sampling) of the k different class years, we can defend that the GPA of the students is independent within and between groups.

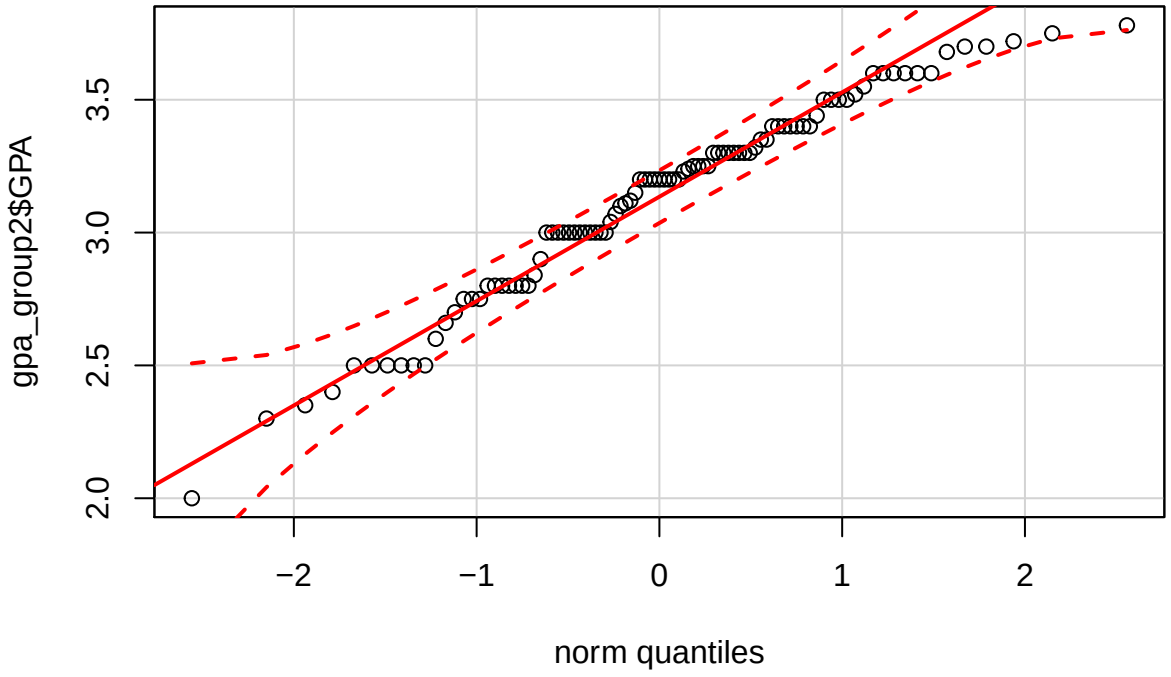
`ii) Check using a graphical method if the assumption of normality holds. Does it hold?`

Using a qqplot of the GPS by class we can discuss if the assumption of normality holds:

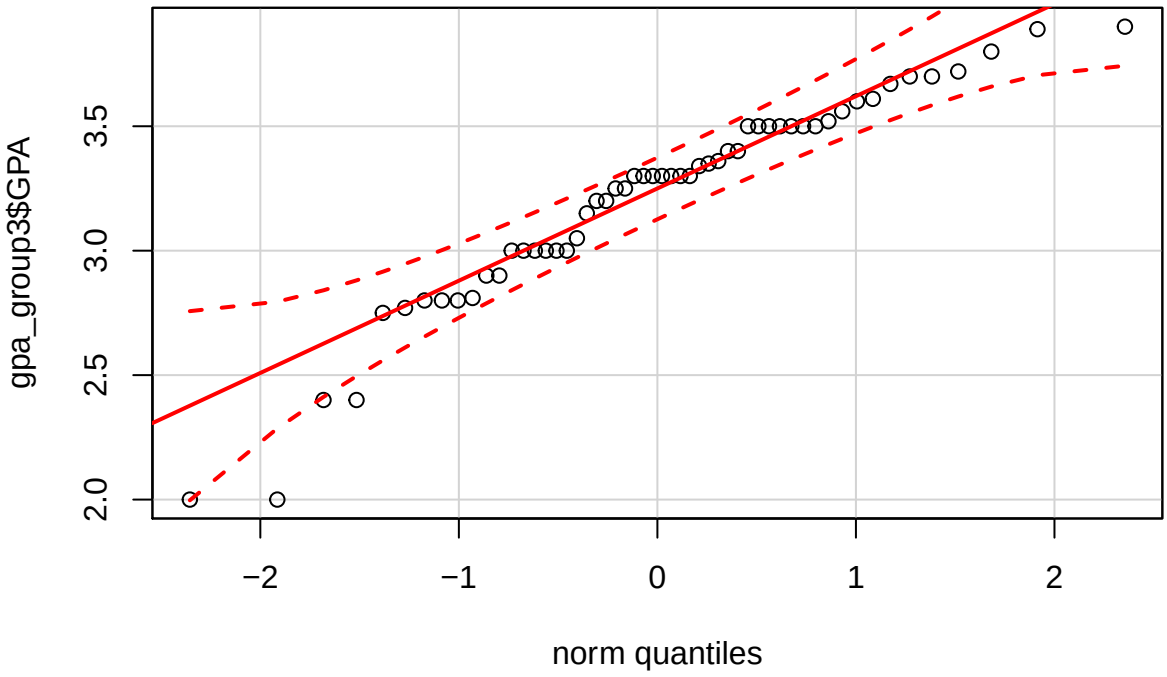
```
gpa_group1 <- subset(Sleep_data, ClassYear==1, select = GPA)
gpa_group2 <- subset(Sleep_data, ClassYear==2, select = GPA)
gpa_group3 <- subset(Sleep_data, ClassYear==3, select = GPA)
gpa_group4 <- subset(Sleep_data, ClassYear==4, select = GPA)
library(car)
qq_group1=qqPlot(gpa_group1$GPA)
```



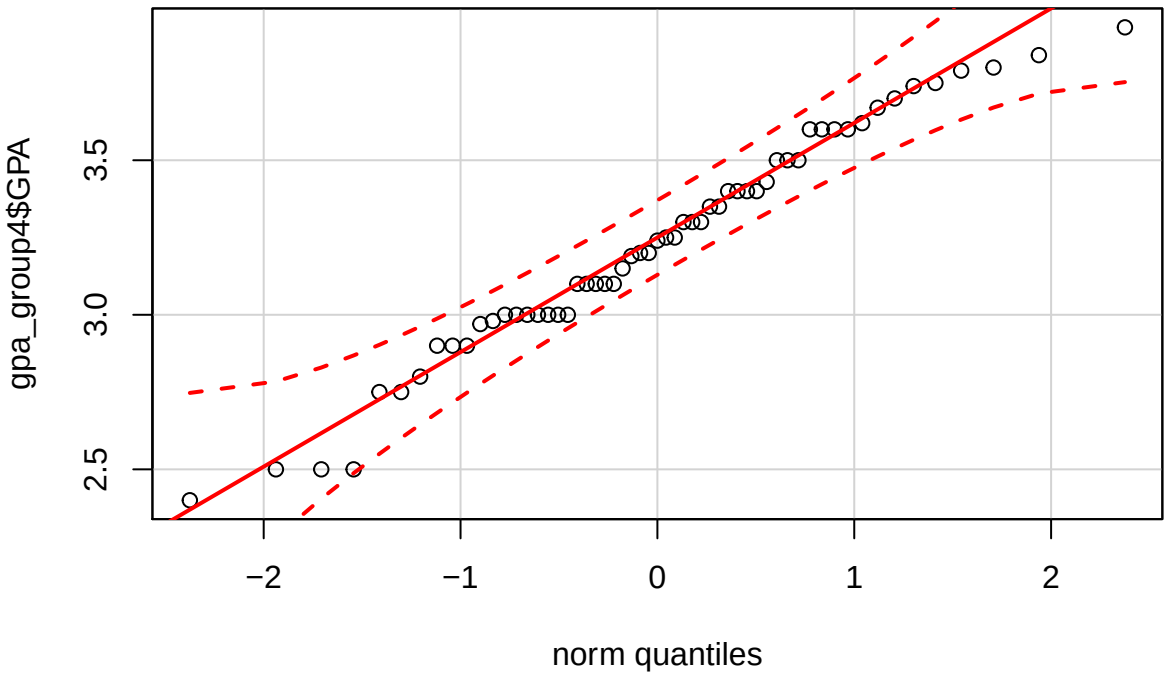

```
qq_group2=qqPlot(gpa_group2$GPA)
```




```
qq_group3=qqPlot(gpa_group3$GPA)
```

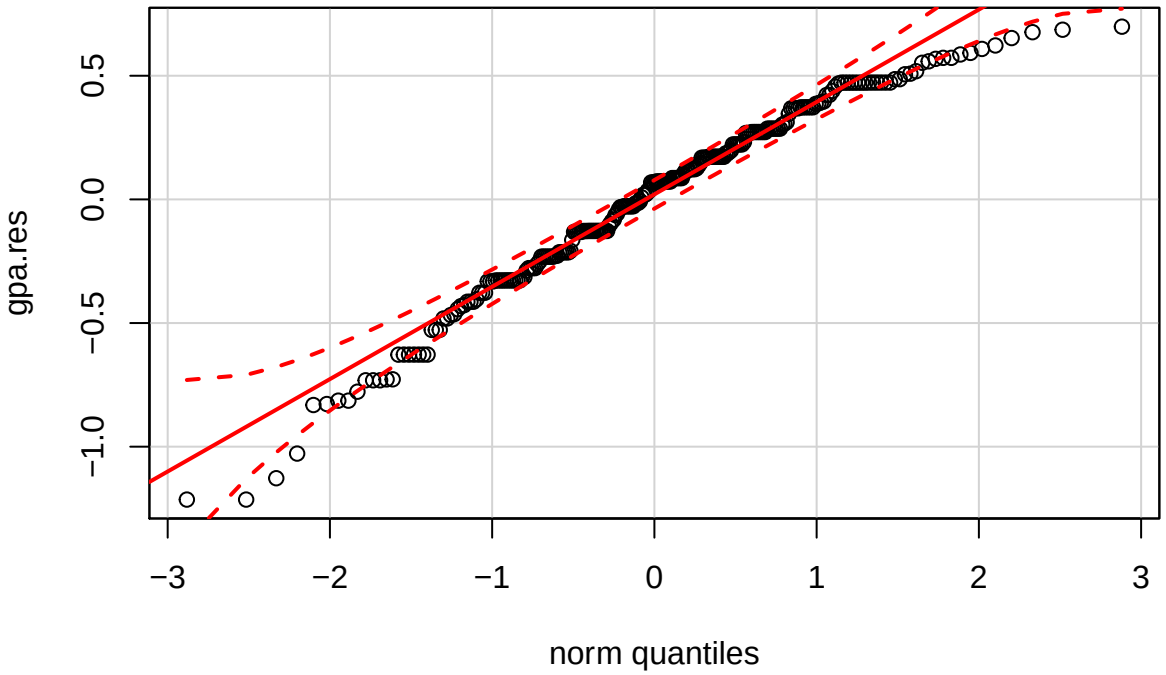



```
qq_group4=qqPlot(gpa_group4$GPA)
```



As we can observe, there is not notable outliers across the groups, with exception of one particular point for group 3 and another for group 2, then the normality assumption is reasonable. However, maybe a more compelling way to check if the normality assumption holds is making the qqplot over the residuals estimated for the ANOVA model:

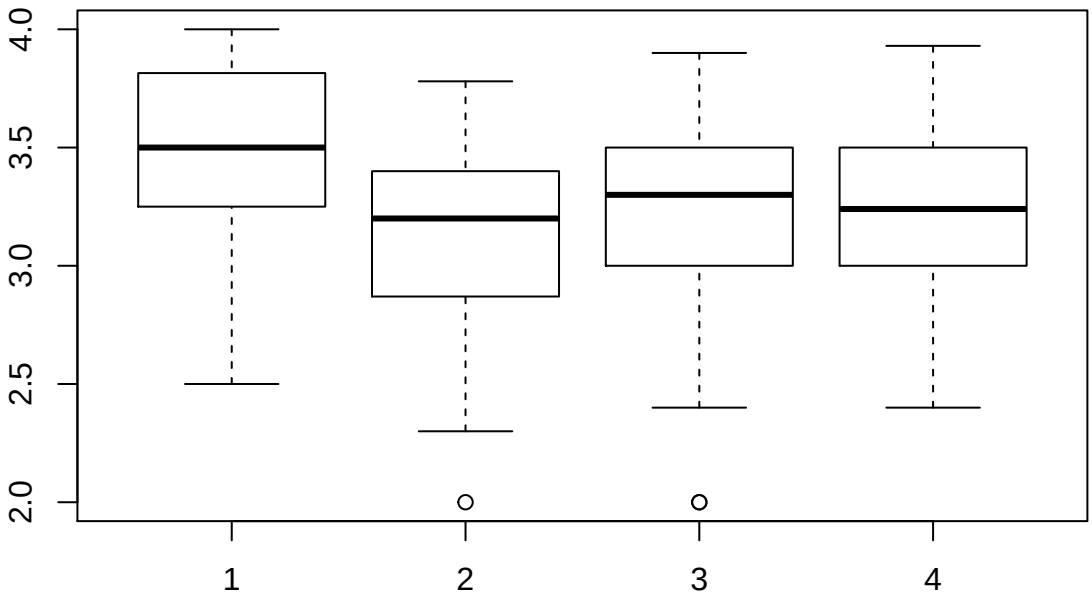
```
gpa.res=gpa.lm$residuals  #Stores Estimated Residuals
qqPlot(gpa.res)
```



With this alternative method there are some outliers as well, but since we have sample sizes across groups large enough it seems reasonable to defend the normality assumption.

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

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



Using the above boxplots it seems reasonable to assume equal variance. But it will be useful to make the ANOVA analysis with and without the outliers for group 2 and 3, just to be sure.

```
leveneTest(gpa.lm)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
```

```
##        Df F value Pr(>F)
## group   3  0.0799 0.9709
##       249
```

With the formal formal test (Levene's test) at a significance level of 0.05, we fail to reject the null hypothesis that the 4 populations have the same variance.

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

Yes, according the results and analysis above, all the assumptions are reasonable.

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.

```
TukeyHSD(fit, 'Sleep_data$ClassYear', conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = Sleep_data$GPA ~ Sleep_data$ClassYear)
##
## $`Sleep_data$ClassYear`
##            diff         lwr         upr     p adj
## 2-1 -0.40029339 -0.61372199 -0.18686480 0.0000001
## 3-1 -0.31398345 -0.55272832 -0.07523859 0.0002808
## 4-1 -0.29629339 -0.53209647 -0.06049031 0.0005822
## 3-2  0.08630994 -0.11765410  0.29027398 0.5441416
## 4-2  0.10400000 -0.09651263  0.30451263 0.3629506
## 4-3  0.01769006 -0.20958216  0.24496227 0.9948334
```

According to the simultaneosly pairwise comparisons using the Tukey HSD test, we can observe statistical significant differences between groups 1 (Freshmen) and 2 (Sophomores), 1 (freshmen) and 3 (juniors), and, 1 (freshmen) and 4 (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.

```
fit = aov(Sleep_data$GPA ~ Sleep_data$AlcoholUse)
summary(fit)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## Sleep_data$AlcoholUse    3   0.60  0.2004    1.23  0.299
## Residuals              249  40.59  0.1630
```

b) What is the realization of the test statistic?

Under H_0 a realization of the F statistic is $F = msb/mse$, which has a $F_{df1,df2}$ distribution.

```
msb=0.2004
mse=0.1630
F =  msb/mse
F
```

```
## [1] 1.229448
```

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.

The p-value of the statistic is:

$$P(F_{K-1,n_{total}-k} > msb/mse)$$

```
msb_df = 3
mse_df = 249
pvalue = 1-pf(F,msb_df,mse_df)
pvalue
```

```
## [1] 0.2995416
```

Since $0.29 > 0.05$, we should not reject H_0. We are 95% confident that mean GPA is the same for all four types of alcohol consumers.

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

There is not differences between the GPA's of the students according to their use of alcohol, or in other words, the alcohol consumption is not related with academic performance (measured with the GPA).

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 is the degrees of freedon of the msb, e.g. the number of groups minus 1 (k-1) . A is the degrees of freedon of the mse, e.g. the number of observation minus the number of groups (n-k). Finally, H is n-1.

```
k=5
n=40
G=k-1
G
```

```
## [1] 4
```

```
A=n-k
A
```

```
## [1] 35
```

```
H=n-1
H
```

```
## [1] 39
```

b) Determine B.

B sum of squares between groups (ssb). ssb=$\sum_{i=1}^{k} n_i \ (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$, then using the identity sst = ssb + sse, we can get ssb as ssb= sst-sse.

```
sst = 3037
sse = 1837
B = sst-sse
B
```

```
## [1] 1200
```

c) Determine C.

C is how much on average a group vary from the grand mean, also called Mean Square Between Groups (msb) and it's calculated as $msb = ssb/k - 1$

```
C = B/G
C
```

```
## [1] 300
```

d) Determine D.

D is the Pooled Error Variance, also called Mean Square Error (mse):

$mse = sse/n_{total} - k$, where n_{total} is the total of observations and k is the number of groups.

```
D = sse/A
D
```

```
## [1] 52.48571
```

e) Determine E.

Under H_0, E is a realization of the F statistic and it's calculated as $F = msb/mse$, which has a $F_{df1,df2}$.

```
E = C/D
E
```

```
## [1] 5.715841
```

f) Determine F.

F is the p-value associated to the F statistic, formally:

$$P(F_{K-1,n_{total}-k} > msb/mse)$$

```
F = 1-pf(E,G,A)
F
```

```
## [1] 0.001195534
```

Then, our complete table should be:

```
tbl <- matrix(c("4", "35", "39",
                "1200", "1837", "3037",
                "300", "52.49", "",
                "5.72", "", "",
                "0.001195", "", ""), 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	4	1200	300	5.72	0.001195
Error	35	1837	52.49		
Total	39	3037			