MATH 227 Final Project

Devin Halliburton, Myra Anigbo, Krystell Ewing, & Yacine Bouabida

December 3, 2021

Introduction:

Our project is about observing first trimester grades from secondary students in Portugal. We also take into consideration the amount of alcohol consumption the students indulge on a daily basis and weekly basis. Our response variable is the first trimester grades these Portugal students receive, and the explanatory variables are the daily amount of alcohol consumption and their second trimester grades.

```
data = read.csv2("student-mat.csv", header = TRUE, sep=",")
str(data)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
```

```
##  $ G3         : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
data$Dalc = factor(data$Dalc, levels = c("1", "2", "3", "4", "5"), labels = c("Very Low", "Low", "Modera

data$Walc = factor(data$Walc, levels = c("1", "2", "3", "4", "5"), labels = c("Very Low", "Low", "Modera
```

Research Question #1‘

We will examine how the amount of alcohol consumption will effect the first trimester grade for students.

Quantitative Responsive Variable:

```
summary(data$Dalc)
```

```
##  Very Low       Low  Moderate      High Very High
##       276        75        26         9         9
```

```
summary(data$G1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   11.00   10.91   13.00   19.00
```

```
sd(data$G1)
```

```
## [1] 3.319195
```

```
hist(data$G1, main = "First Trimester Grade Based of Alcohol Consumption", ylab = "Percentage of Studen
```

Based on the histogram most students in their first trimester were failing or sufficient.

Check for Outliers:

```
low.out = quantile(data$G1, .25) - 1.5*IQR(data$G1)
high.out = quantile(data$G1, .75) + 1.5*IQR(data$G1)
```

```
sum(data$G1 < low.out)
```

```
## [1] 0
```

```
sum(data$G1 > high.out)
```

```
## [1] 0
```

The shape of the distribution is appears to be *approximatelydistributed* based off the histogram and because of the central limit theorem, the size is greater than 30 $395 > 30$. The center of the distribution is 10.91 which means that the average student that is consuming alcohol is barely "sufficient" (10-11 is the sufficient grade range). The spread of the distribution is symmetrical around the mean of the sample. We have zero outliers in our dataset whether it is a high outlier or low outlier.

Categorical Explanatory Variable:

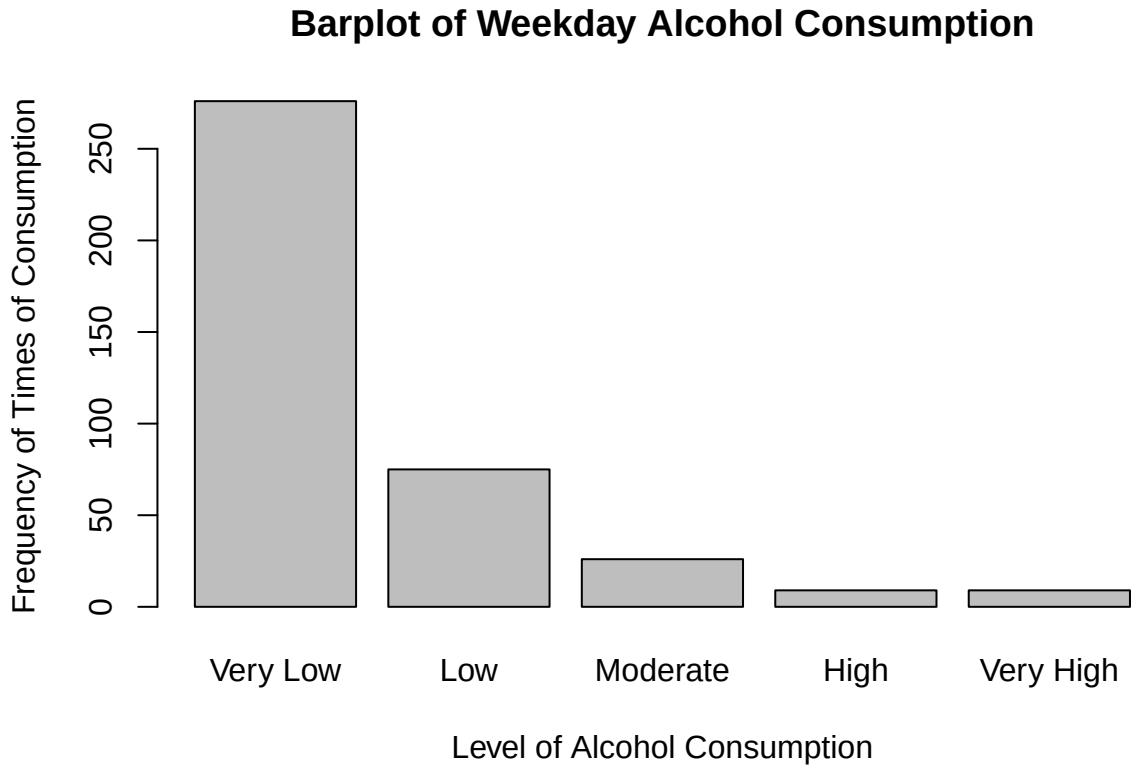
```
table(data$Dalc)
```

```
##
##  Very Low       Low  Moderate      High Very High
##       276        75        26         9         9
```

```
sum(table(data$Dalc))
```

```
## [1] 395
```

```
barplot(table(data$Dalc), main = "Barplot of Weekday Alcohol Consumption", xlab = "Level of Alcohol Cons
```



$VeryLow$ has the highest frequency which means there are NOT a lot of student consuming alcohol during the weekdays. Low is the second highest frequency which explains that there are descent amount of students drinking a little bit during the weekdays. For $Moderate$, $High$, and $VeryHigh$ are low in frequency which means that there are not a lot of students drinking a good or high amount of alcohol during the week.

Bivariate EDA:

```
summary(data$G1[data$Dalc == "Very Low"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   11.00   11.16   14.00   19.00
```

```
summary(data$G1[data$Dalc == "Low"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00    8.00   10.00   10.31   12.00   18.00
```

```
summary(data$G1[data$Dalc == "Moderate"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00    8.25   10.00   10.58   11.75   17.00
```

```
summary(data$G1[data$Dalc == "High"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.000   9.000  10.000   9.778  10.000  14.000
```

```
summary(data$G1[data$Dalc == "Very High"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   10.00   11.00   10.44   12.00   14.00
```

```
sd(data$G1[data$Dalc == "Very Low"])
```

```
## [1] 3.454799
```

```
sd(data$G1[data$Dalc == "Low"])
```

```
## [1] 3.119136
```

```
sd(data$G1[data$Dalc == "Moderate"])
```

```
## [1] 2.670926
```

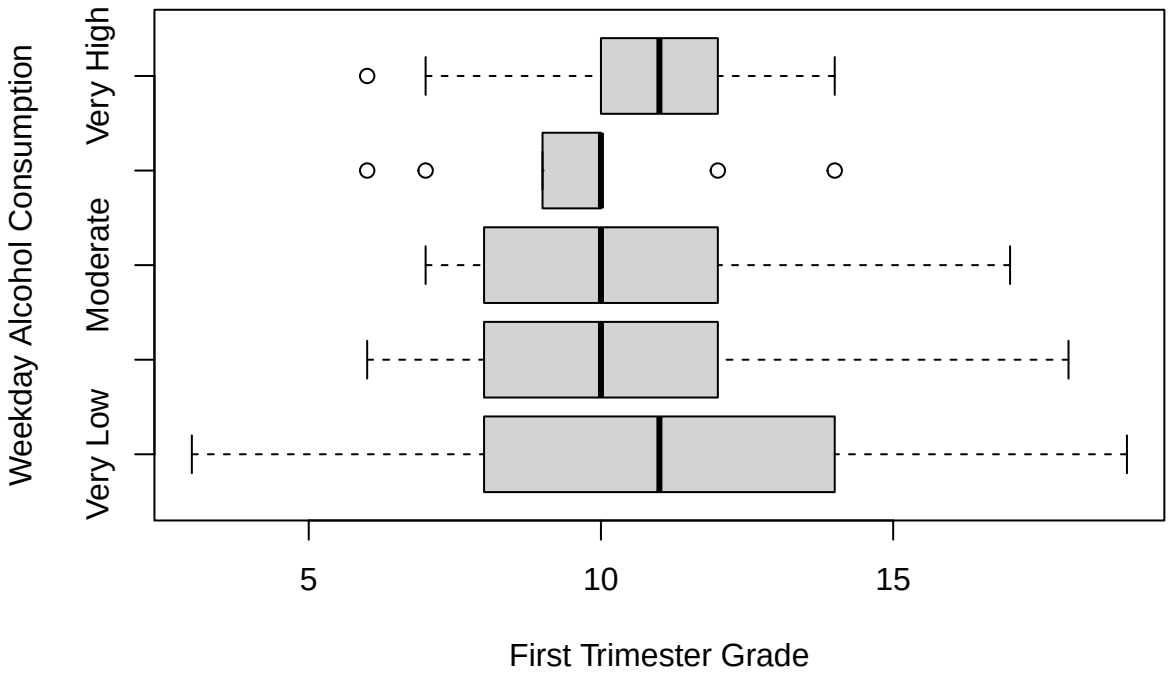
```
sd(data$G1[data$Dalc == "High"])
```

```
## [1] 2.386304
```

```
sd(data$G1[data$Dalc == "Very High"])
```

```
## [1] 2.603417
```

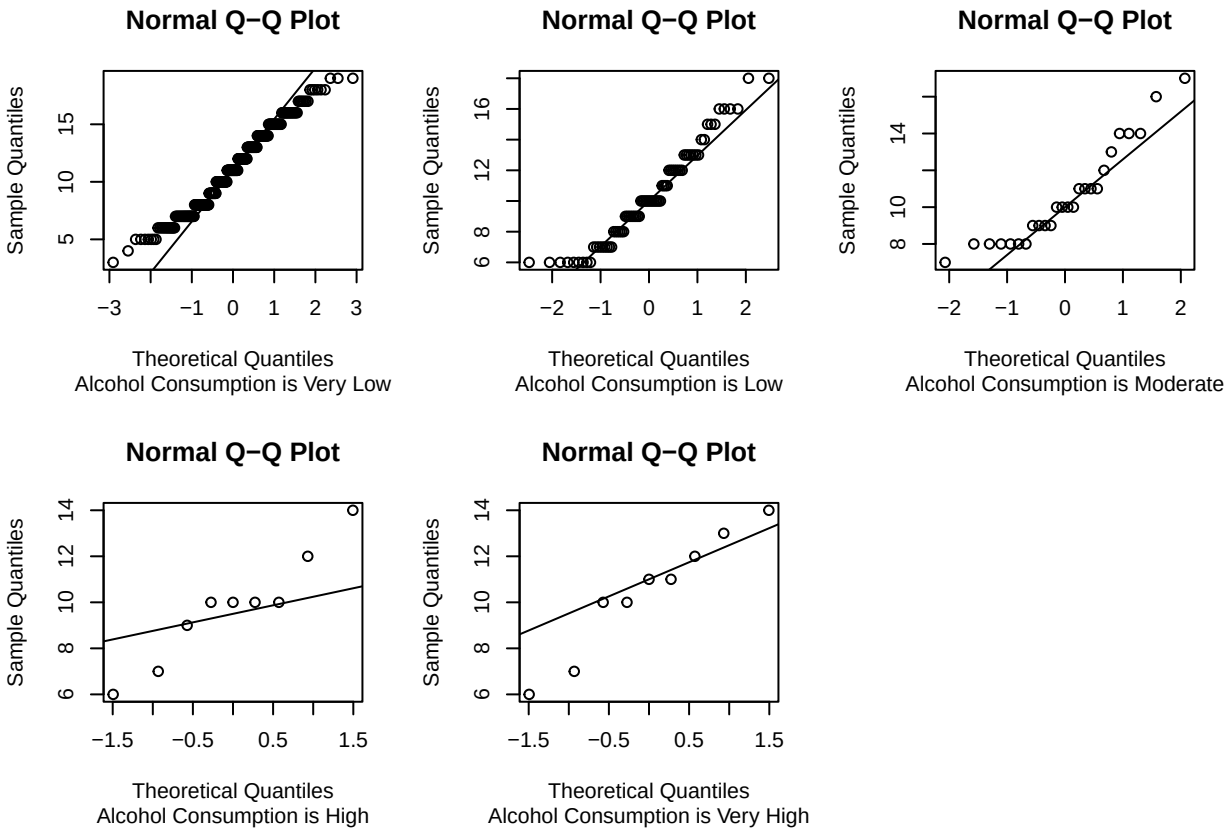
```
boxplot(data$G1 ~ data$Dalc, horizontal = TRUE, xlab = "First Trimester Grade", ylab = "Weekday Alcohol
```



It appears that there is NOT much of a difference between the first trimester grade based on students alcohol consumption. Actually the $VeryHigh$ and $VeryLow$ have about the same mean for students that hace those alcohol consumptions.

In order to carry out the ANOVA, we will need to check some assumptions. One being that the largest standard deviation is no more than twice the smallest. We can see from the sd's above that this condition is met. We also need to check to make sure that each sample (separated by category) is approximately Normal. This is achieved with qqplots. If either of these conditions are note met, let me know so we can discussion options moving forward.

```
par(mfrow = c(2, 3))

qqnorm(data$G1[data$Dalc == "Very Low"], sub = "Alcohol Consumption is Very Low")
qqline(data$G1[data$Dalc == "Very Low"])

qqnorm(data$G1[data$Dalc == "Low"], sub = "Alcohol Consumption is Low")
qqline(data$G1[data$Dalc == "Low"])

qqnorm(data$G1[data$Dalc == "Moderate"], sub = "Alcohol Consumption is Moderate")
qqline(data$G1[data$Dalc == "Moderate"])

qqnorm(data$G1[data$Dalc == "High"], sub = "Alcohol Consumption is High")
qqline(data$G1[data$Dalc == "High"])

qqnorm(data$G1[data$Dalc == "Very High"], sub = "Alcohol Consumption is Very High")
qqline(data$G1[data$Dalc == "Very High"])
```



Research Question #2

We will examine whether the second trimester grades can be explained by the alcohol consumption on the weekdays and weekends.

Trivate EDA:

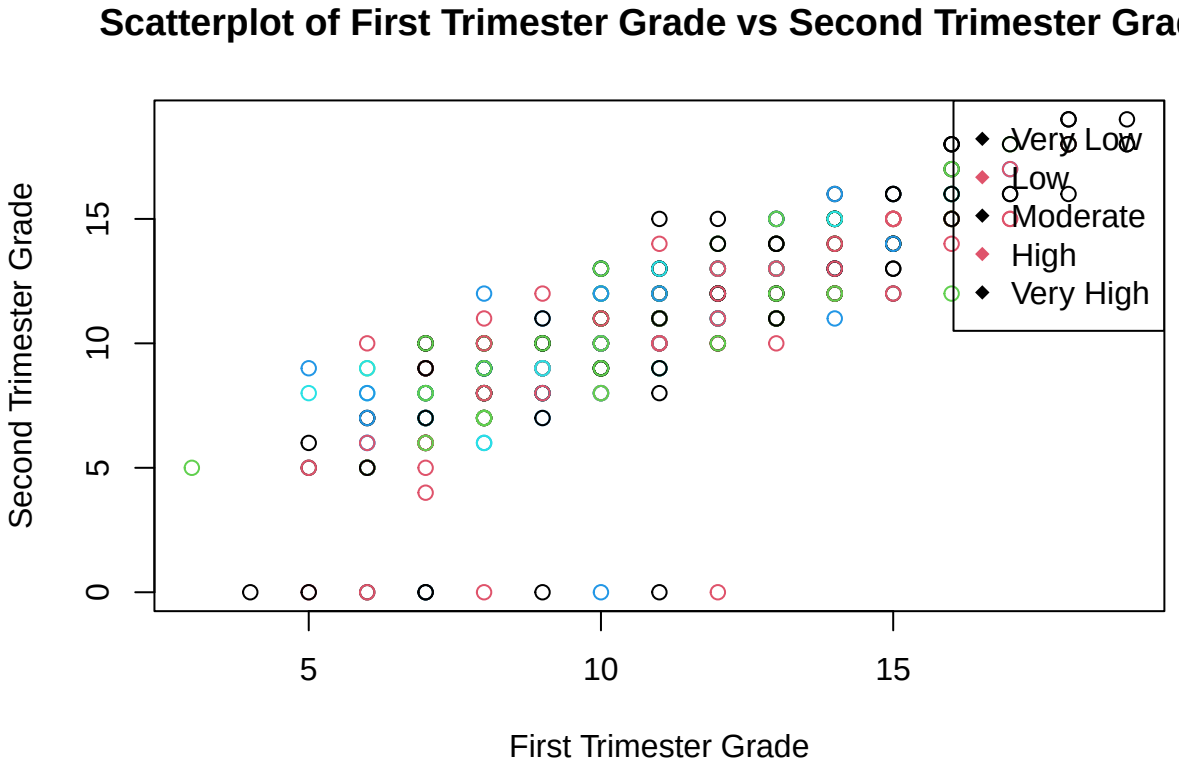
```
cor(data$G1, data$G2)
```

```
## [1] 0.8521181
```

```
summary(data$G2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.71   13.00   19.00
```

```
plot(data$G1, data$G2, col = data$Walc, xlab = "First Trimester Grade", ylab = "Second Trimester Grade"
legend("topright", c("Very Low", "Low", "Moderate", "High", "Very High"), col = c(1, 2), pch = 18)
```



The explanatory variables are related to the response because it showcases how students grades have moved since the their first trimester grade to their second trimester grade, based off the amount of alcohol consumption on the weekends. Majority of students remained in the same grade level when it came to transition from the first trimester to the second trimester, based of their alcohol consumption.

ANOVA:

Parameters:

μ = The average first trimester grade based off alcohol consumption.

Hypotheses:

The null hypothesis H_o is that the students that have higher alcohol consumption have a first trimester grade mean that is equal to the students who have lower or no alcohol consumption $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. The alternative hypothesis H_a is that there is a difference between one of the means of first trimester grade from students that have high alcohol consumption, and students who have low or no alcohol consumption i ϵ {1, 2, 3, 4, 5}.

ANOVA Requirements:

1. There are i independent normal populations

QQPlots performed in Progress Report #1.

3. Checking to make sure each independent population has the same standard deviation σ.

```
sd(data$G1[data$Dalc == "Very Low"]) / sd(data$G1[data$Dalc == "High"])
```

```
## [1] 1.447762
```

Rule of thumb is that if the largest S divided by the smallest S is less than 2, than each independent population has similar S's. Above the highest S came from the $VeryLow$ alcohol consumption population,

while the lowest came from the $High$. When you divide those two population's Ss you get 1.447762, which is less than 2.

ANOVA:

```
aov(data$G1 ~ data$Dalc)
```

```
## Call:
##    aov(formula = data$G1 ~ data$Dalc)
##
## Terms:
##                 data$Dalc Residuals
## Sum of Squares     60.348  4280.371
## Deg. of Freedom         4       390
##
## Residual standard error: 3.312901
## Estimated effects may be unbalanced
```

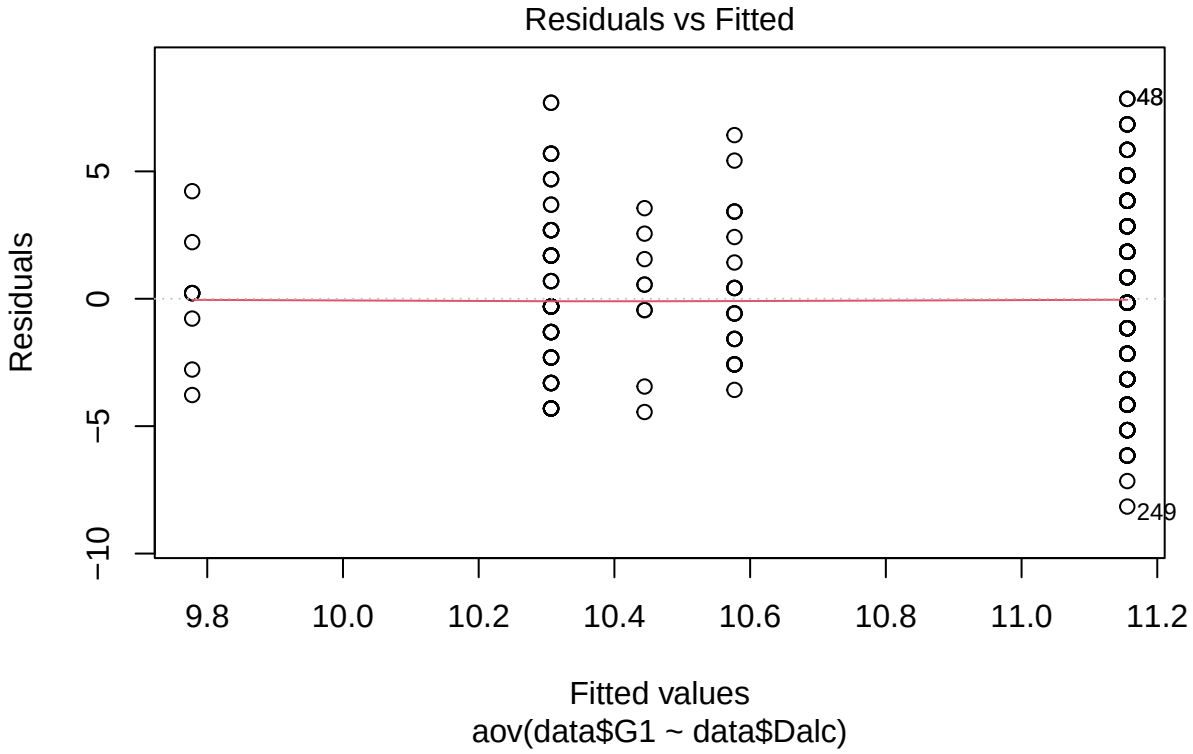
```
summary(aov(data$G1 ~ data$Dalc))
```

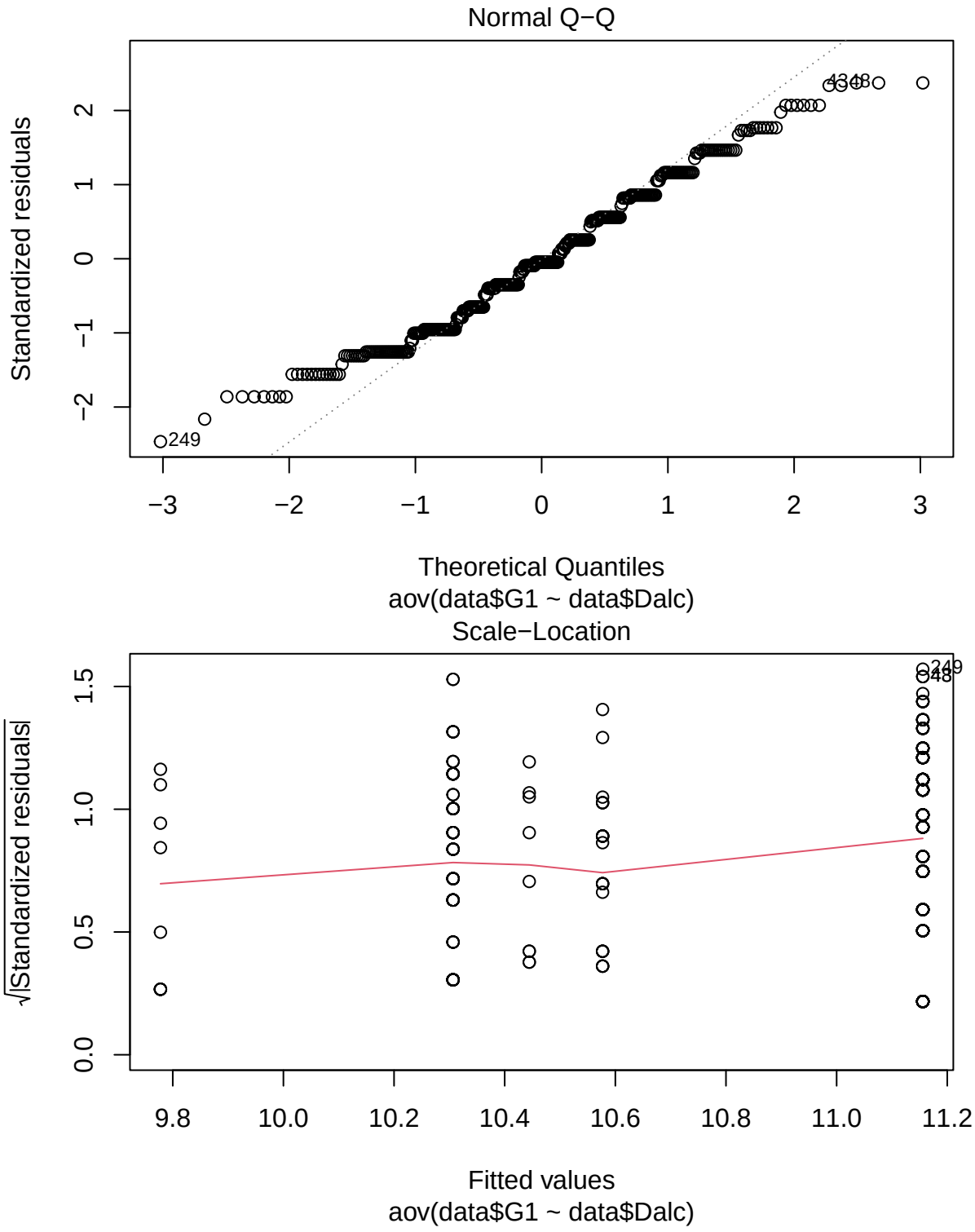
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## data$Dalc     4     60   15.09   1.375  0.242
## Residuals   390   4280   10.97
```

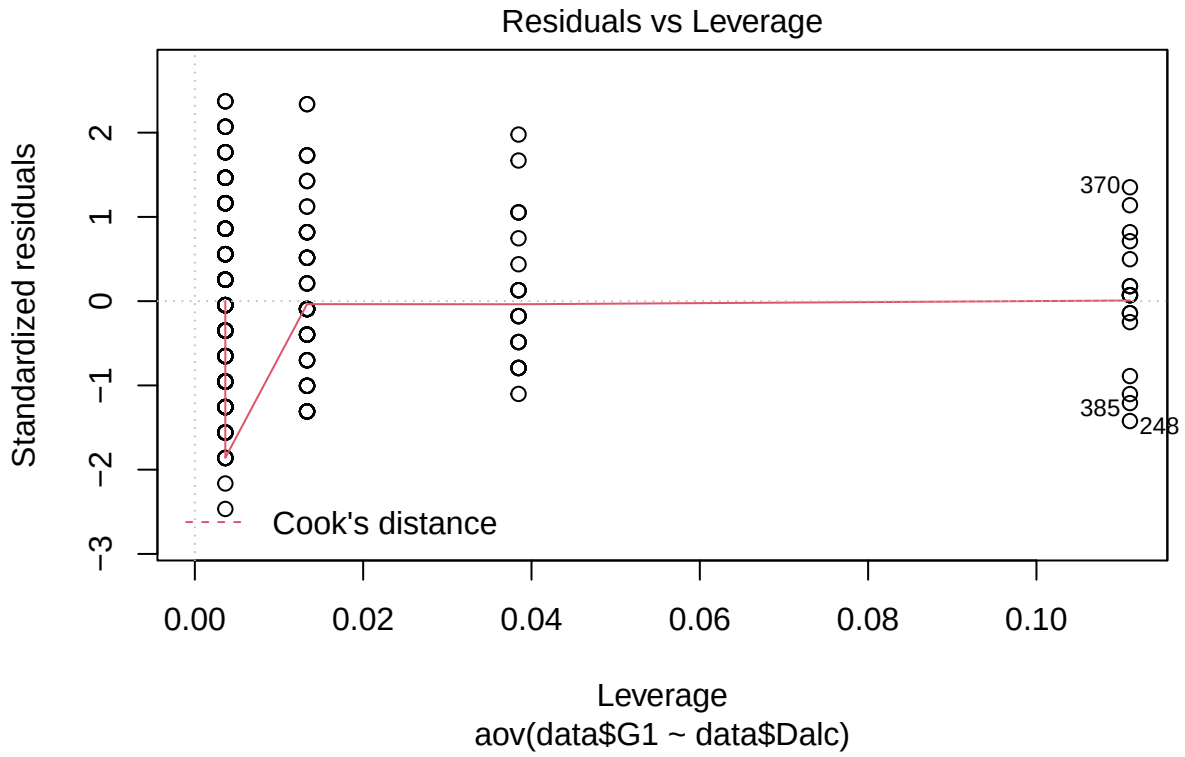
The p-value = 0.242 which is greater than alpha = 0.05 $0.242 > 0.05$, which means that we are going to fail to reject our null hypothesis. There is sufficient evidence that the there is no difference between the true average of first trimester grades based off daily alcohol consumption from students.

Residuals:

```
plot(aov(data$G1 ~ data$Dalc))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(data$G1 ~ data$Dalc)
Scale−Location
√|Standardized residuals|
Fitted values
aov(data$G1 ~ data$Dalc)



Since the residuals are randomly scattered around zero and are approximately normal, the after conditions are met.

Regression:

Define your parameters:

B_1 = True slope of second trimester grades based off alcohol consumption.

B_2 = True slope of second trimester grades based off first trimester grades.

State your Hypothesis:

The null hypothesis H_o is that the slope of second trimester grades is equal to zero based off alcohol consumption.: $B_1 = 0$.

The alternative hypothesis H_a is that the slope of second trimester grades is not equal to zero based off alcohol consumption: B_1 != 0.

The null hypothesis H_o is that the slope of second trimester grades is equal to zero based off first trimester grades: $B_2 = 0$.

The alternative hypothesis H_a is that the slope of second trimester grades is not equal to zero based off first trimester grades: B_2 != 0.

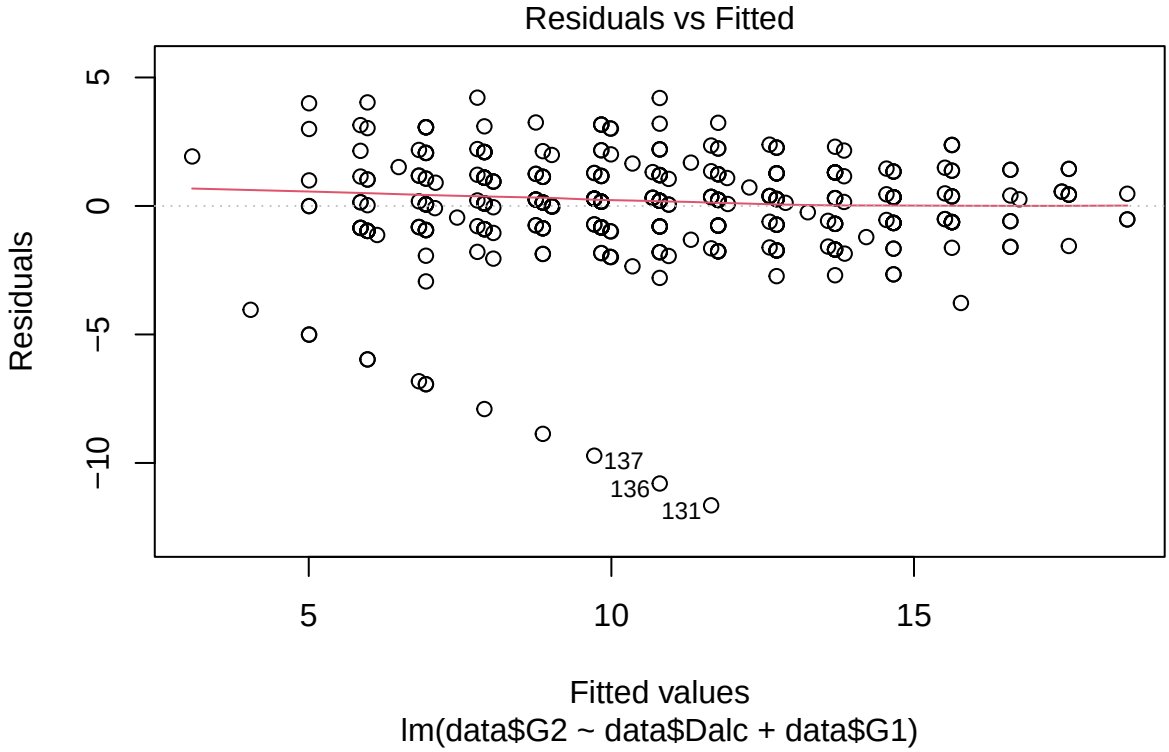
Checking the Residuals:

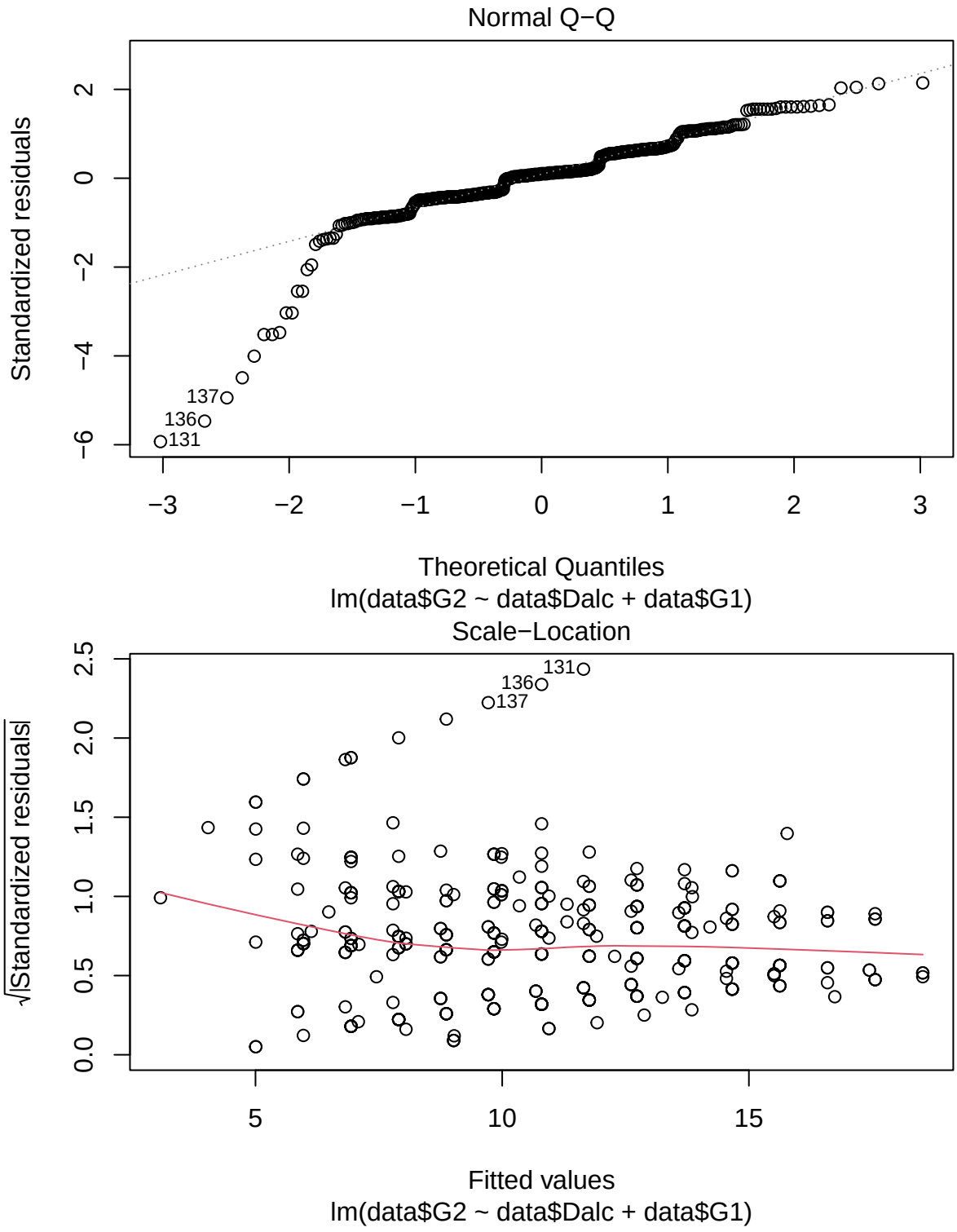
```
summary(lm(data$G2 ~ data$Dalc + data$G1))
```

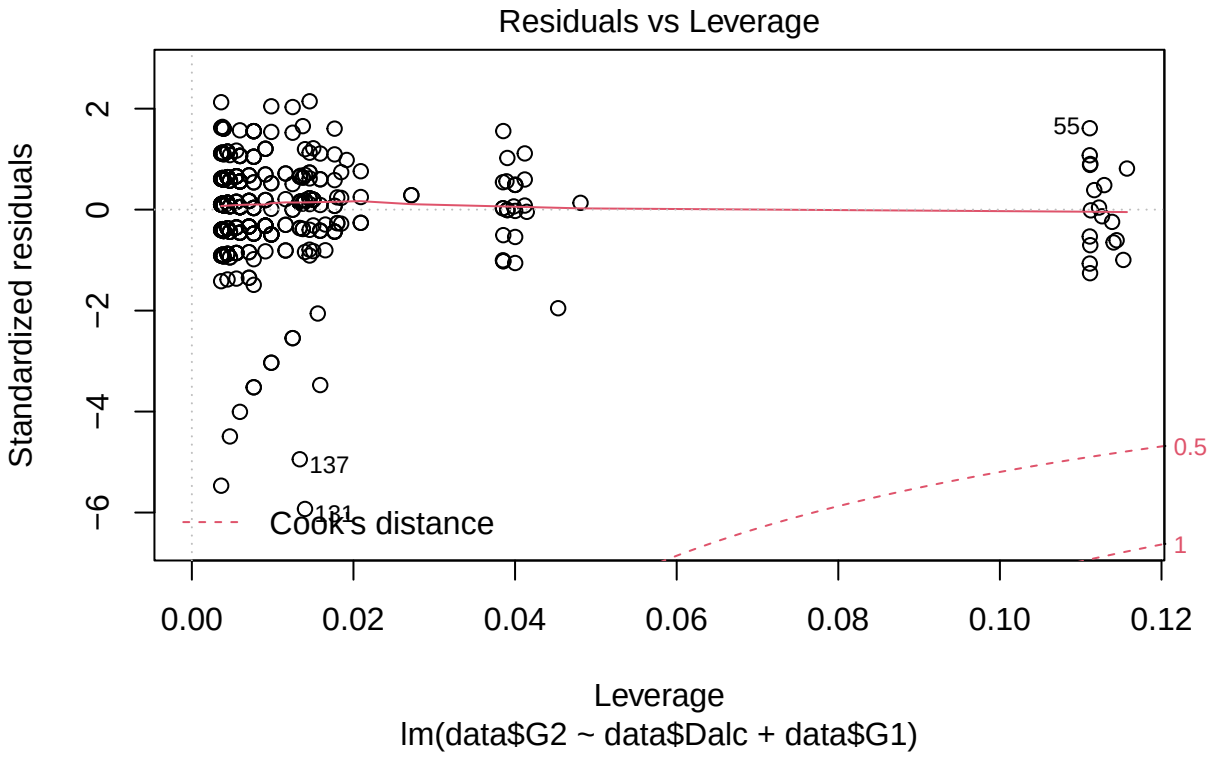
```
##
## Call:
## lm(formula = data$G2 ~ data$Dalc + data$G1)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6485  -0.8268   0.2012   1.1732   4.2140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.31828    0.37866   0.841    0.401
## data$Dalc.L  0.41298    0.47823   0.864    0.388
## data$Dalc.Q  0.18476    0.45429   0.407    0.684
## data$Dalc.C -0.01104    0.48966  -0.023    0.982
## data$Dalc^4  0.14722    0.44213   0.333    0.739
## data$G1      0.96563    0.03024  31.932   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.978 on 389 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.7233
## F-statistic:   207 on 5 and 389 DF,  p-value: < 2.2e-16
```

```
plot(lm(data$G2 ~ data$Dalc + data$G1))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(data$G2 ~ data$Dalc + data$G1)
Scale−Location
√|Standardized residuals|
Fitted values
lm(data$G2 ~ data$Dalc + data$G1)



Slope Equation:

$\hat{y} = 0.31828 + 0.41298x_1 + 0.18476x_2 + \text{-}0.01104x_3 + 0.14722x_4 + 0.96563x_5$

Variables:

x_1 = "Low" Daily Alcohol Consumption x_2 = "Moderate" Daily Alcohol Consumption x_3 = "High" Daily Alcohol Consumption x_4 = "Very High" Daily Alcohol Consumption x_5 = First Trimester Grades

Coefficients/Slopes:

$\hat{B}_i$ = The true slope for predicting the second trimester grades based off low, moderate, high, very high and first trimester grades.

Interpretations:

For every 1 unit increase in first trimester grades, second trimester increases by $\hat{B}_5 = 0.96563$ slope on average.

For any student with "Low" daily alcohol consumption, their second trimester grade increases by $\hat{B}_1 = 0.41298$ slope on average, compared to students with "Very Low" alcohol consumption.

For any student with "Moderate" daily alcohol consumption, their second trimester grade increases by $\hat{B}_2 = 0.18476$ slope on average, compared to students with "Very Low" alcohol consumption.

For any student with "High" daily alcohol consumption, their second trimester grade decreases by $\hat{B}_3 = -0.01104$ slope on average, compared to students with "Very Low" alcohol consumption.

For any student with "High" daily alcohol consumption, their second trimester grade increases by $\hat{B}_4 = 0.14722$ slope on average, compared to students with "Very Low" alcohol consumption.

For any student with "High" daily alcohol consumption, their second trimester grade increases by $\hat{B}_5 = 0.96563$ slope on average, compared to students with "Very Low" alcohol consumption.