SDS230_FinalProject

Introduction

Our project is an analysis of the factors that may be correlated with the grades earned by students in two secondary schools in Portugal. We explore factors including alcohol consumption on weekends and weekdays, sex, studytime and freetime, etc. Through this analysis we hope to elucidate potential determinants of grades in this population. We use descriptive plots to provide some initial insight into the data, provide information about significant correlations between variables, and perform grade-predicting multiple regression. We perform analysis of variance to identify differences in grades between groups that consume alcohol on weekends.

Data

The data used in this analysis were collected in a survey of students in two secondary schools. It includes information about gender, social habits, alcohol consumption, grades, etc.

The variables used in this analysis are: 1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira); categorical 2. sex - student's sex (binary: 'F' - female or 'M' - male); categorical 3. age - student's age (numeric: from 15 to 22); continuous* 4. studytime - weekly study time (numeric: 1 - 10 hours); continuous* 5. higher - wants to take higher education (binary: yes or no); categorical 6. freetime - free time after school (numeric: from 1 - very low to 5 - very high); continuous* 7. goout - going out with friends (numeric: from 1 - very low to 5 - very high); continuous* 8. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high); continuous* 9. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high); continuous* 10. health - current health status (numeric: from 1 - very bad to 5 - very good); continuous* 11. absences - number of school absences (numeric: from 0 to 93); continuous* 12. grade - final grade (numeric: from 0 to 20); continuous* *(discrete integers, makes sense to be treated as continuous variable, because of the nature of the variable)

Data Cleaning Process

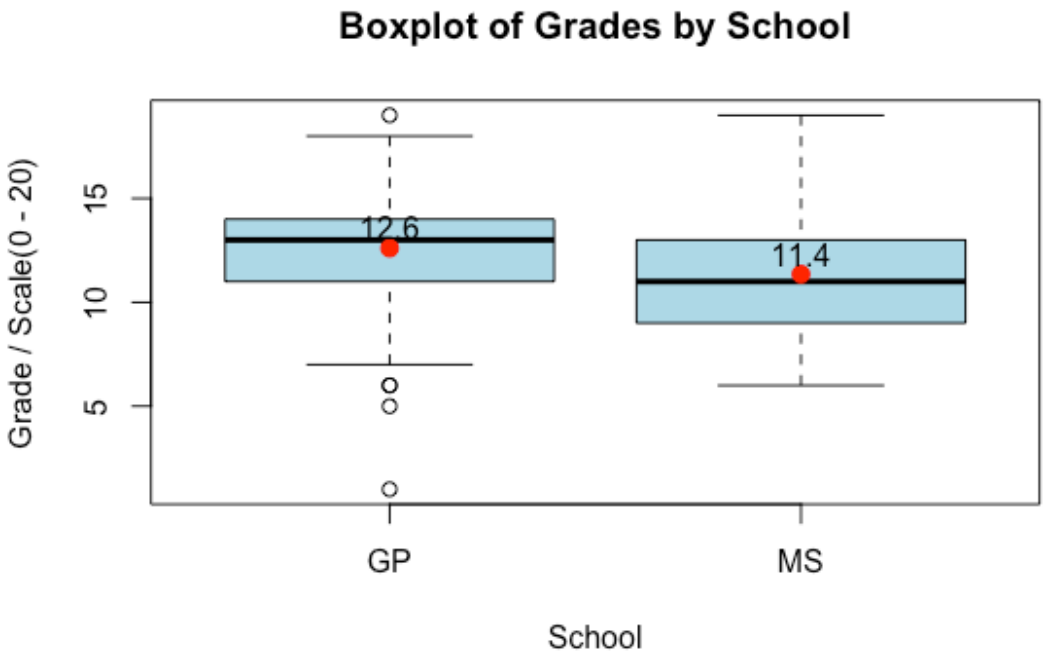
Our data did not require much cleaning. We renamed the column G3 to "grade" to provide a more intuitive and clear description of what the variable was and we removed observations where grade was 0 because this unreasonable grade indicated that the data for these observations was incomplete. Later, for some analyses we recode Dalc and Walc to be binary because many students selected the lowest option on the scale.

```
all_data <- read.csv("student-por.csv", header=TRUE)
selected_data <- all_data[, c("school", "sex", "age", "studytime", "higher",
"freetime", "goout", "Dalc", "Walc", "health", "absences", "G3")]
# rename G3 to grade
selected_data <- rename(selected_data, c("G3"="grade"))
```

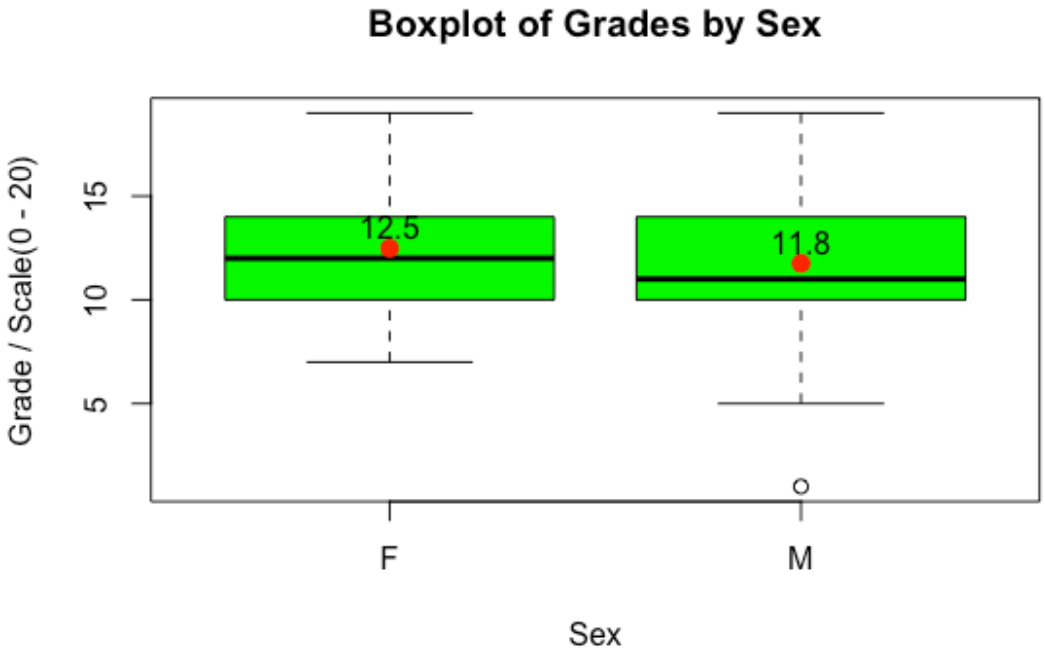
```r
# remove observations where grade is 0
selected_data <- selected_data[selected_data$grade != 0,]
nrow(selected_data)

## [1] 634

names(selected_data)

##  [1] "school"    "sex"       "age"       "studytime" "higher"
"freetime"
##  [7] "goout"     "Dalc"      "Walc"      "health"    "absences"  "grade"

str(selected_data)

## 'data.frame':    634 obs. of  12 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 ...
##  $ studytime: int  2 2 2 3 2 2 2 2 2 2 ...
##  $ higher   : chr  "yes" "yes" "yes" "yes" ...
##  $ 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  4 2 6 0 0 6 0 2 0 0 ...
##  $ grade    : int  11 11 12 14 13 13 13 13 17 13 ...

attach(selected_data)
```

Descriptive Plots

```r
# Graphics - Boxplot (grade by school and by sex with superimposed mean)
boxplot(grade ~ school, main = "Boxplot of Grades by School", ylab = "Grade /
Scale(0 - 20)", col = "lightblue", xlab = "School")
means <- tapply(grade, school, mean)
points(means, col = "red", pch = 19, cex = 1.2)
text(x = c(1:6), y = means + 1, labels = round(means,1))
```

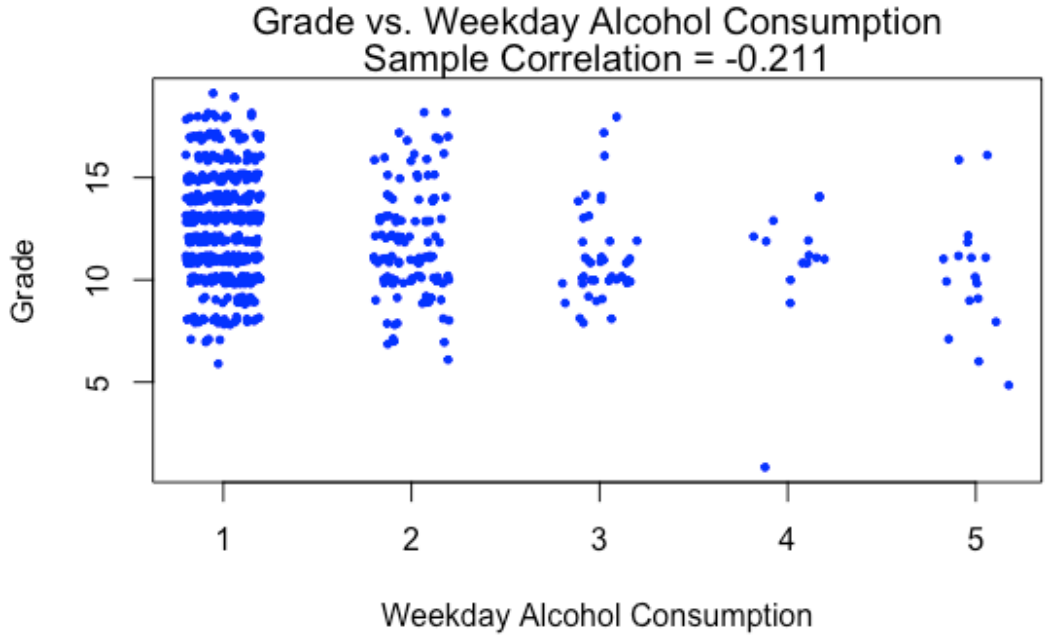



```
boxplot(grade ~ sex, main = "Boxplot of Grades by Sex", ylab = "Grade /
Scale(0 - 20)", col = "green", xlab = "Sex")
means <- tapply(grade, sex, mean)
points(means, col = "red", pch = 19, cex = 1.2)
text(x = c(1:6), y = means + 1, labels = round(means,1))
```

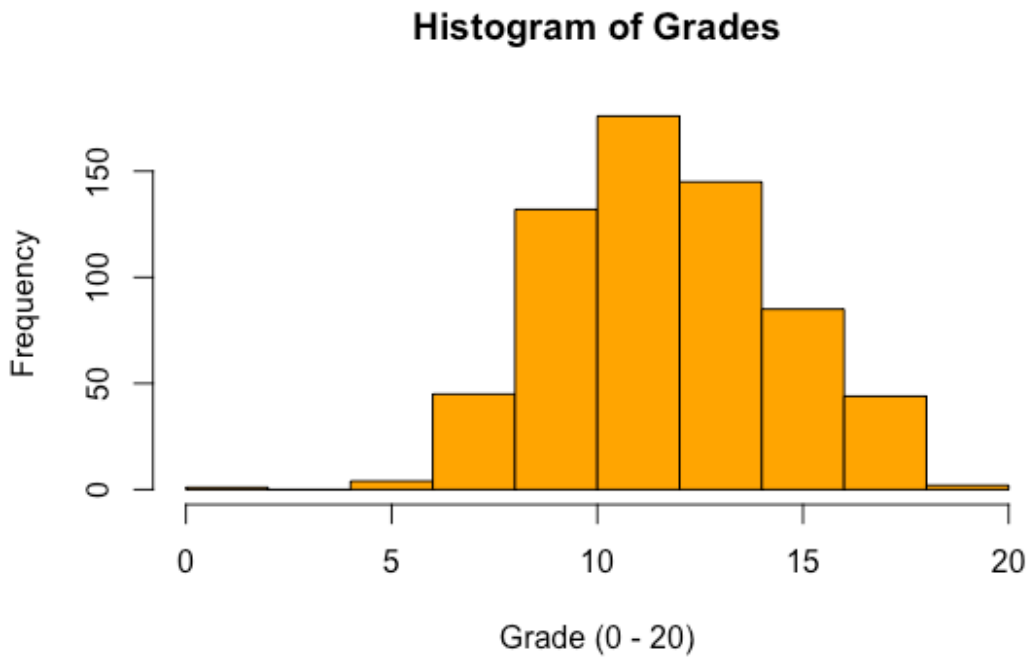


```
plot(jitter(Walc, factor = 1), jitter(grade), pch = 19, col = "red", xlab =
"Weekend Alcohol Consumption",
     ylab = "Grade", cex = 0.5)
mtext("Grade vs. Weekend Alcohol Consumption", cex = 1.2, line = 1)
mtext(paste("Sample Correlation =", round(cor(grade, Walc), 3)), cex = 1.2,
line = 0)
```

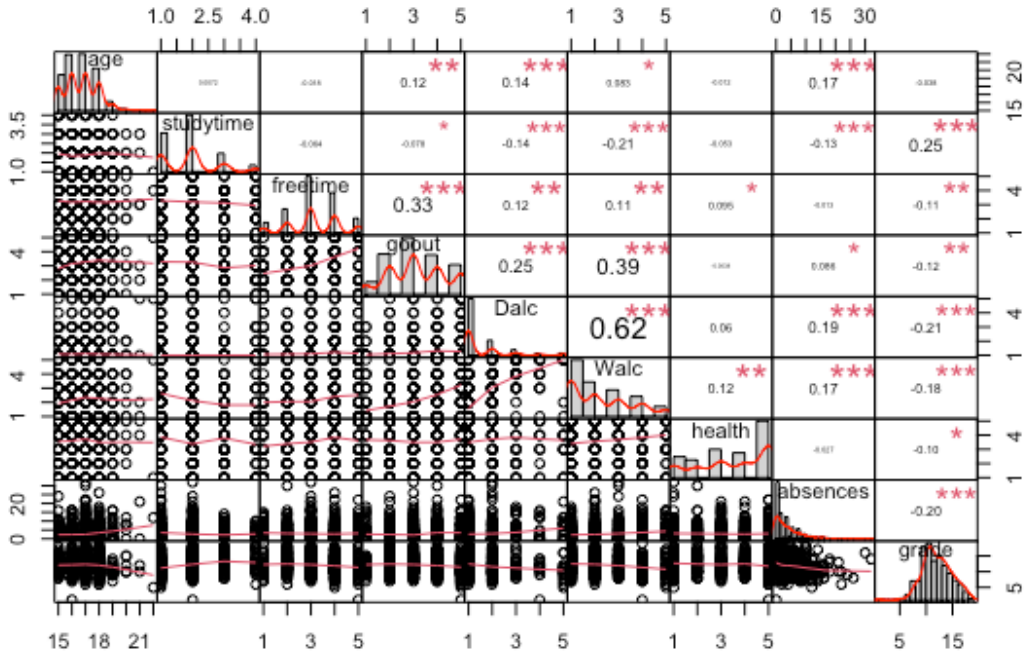

```
plot(jitter(Dalc, factor = 1), jitter(grade), pch = 19, col = "blue", xlab =
"Weekday Alcohol Consumption",
     ylab = "Grade", cex = 0.5)
mtext("Grade vs. Weekday Alcohol Consumption", cex = 1.2, line = 1)
mtext(paste("Sample Correlation =", round(cor(grade, Dalc), 3)), cex = 1.2,
line = 0)
```



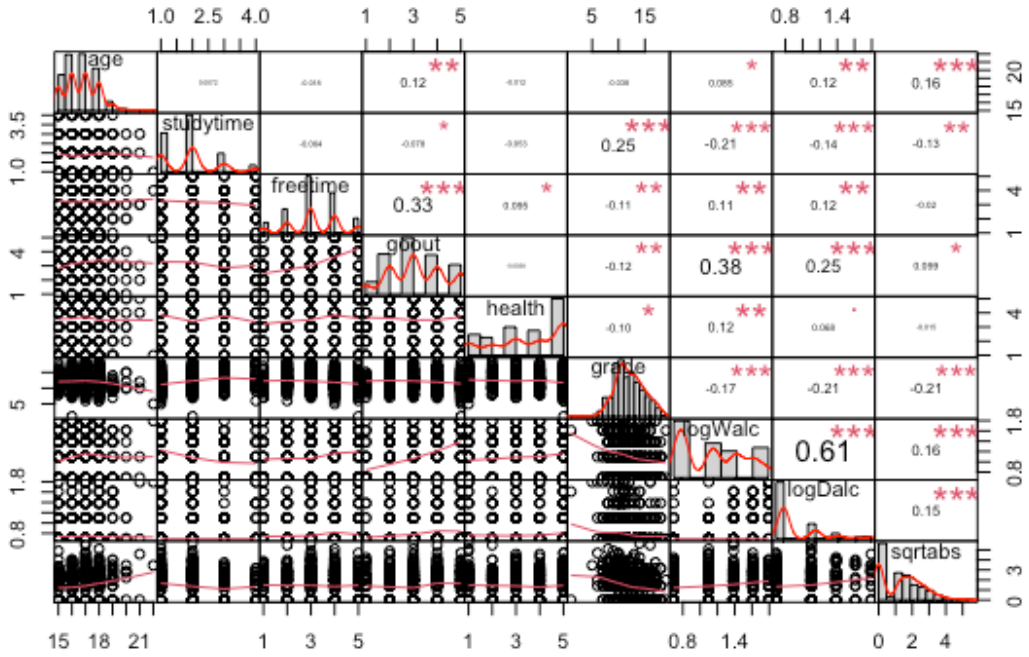

```
hist(grade, main = "Histogram of Grades", xlab = "Grade (0 - 20)", xlim =
c(0, 20), col = "orange")
```




```
chart.Correlation(selected_data[, c(3,4,6:12)], histogram = TRUE, pch = 19)
```




```
##might be useful to transform Walc, Dalc, and absences. for now, I will try
a sqrt transformation of abs, and log transformation of Walc and Dalc.

transdata <- selected_data[, c(3,4,6:12)]
transdata$logWalc <- log(transdata$Walc + 1)
transdata$logDalc <- log(transdata$Dalc + 1)
transdata$sqrtabs <- sqrt(transdata$absences)
chart.Correlation(transdata[, c(1:4,7,9:12)], histogram = TRUE, pch = 19)
```



Summary Information

Our preliminary descriptive plots show that grades, freetime, and time going out appear to be approximately normally distributed. The box plots indicate there may be significant difference in the mean of grades between the two schools and between females and males. Our initial matrix plot indicates that transformations of the variables absences, Dalc, and Walc may be helpful. These three variables all appear strongly right skewed. This was expected for Walc and Dalc, which measure alcohol consumption, because many students reported that they do not consume any alcohol. As expected, there is a strong correlation between Walc and Dalc on both the raw and log scale and this indicates that collinearity between these variables may need to be considered in our models. Interestingly, there does not appear to be a significant correlation between age and grades or studytime or freetime.

Analysis

Basic Tests - T-test

```
t.test(Walc ~ sex)

##
##  Welch Two Sample t-test
##
## data:  Walc by sex
## t = -8.0002, df = 446.11, p-value = 1.081e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0339620 -0.6261485
## sample estimates:
```

```
## mean in group F mean in group M
##        1.933511        2.763566

t.test(grade ~ sex)

##
##  Welch Two Sample t-test
##
## data:  grade by sex
## t = 3.3383, df = 549.72, p-value = 0.0009001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2970362 1.1463499
## sample estimates:
## mean in group F mean in group M
##        12.48138        11.75969

t.test(grade ~ school)
t.test(Walc ~ school)
t.test(studytime ~ sex)
t.test(Dalc ~ sex)
t.test(log10(Walc) ~ sex)
t.test(log10(Dalc) ~ sex)
t.test(log10(Walc) ~ school)
t.test(Dalc ~ school)
t.test(log10(Dalc) ~ school)
```

Discussion of t-test results The results of a two-sample t-test showed that there was enough evidence to reject the null hypothesis that the difference in means of weekend and weekday alcohol consumption for Females and Males was equal to zero. This was also true on the log scale. Additionally, a t-test showed significant difference in study time and grades for females and males, and showed significant differences in grades between schools. There was not enough evidence to reject the null hypothesis that the difference in mean weekday alcohol consumption between schools was equal to zero. This was true on the log scale and for weekend alcohol consumption as well.

Basic Tests - Correlation

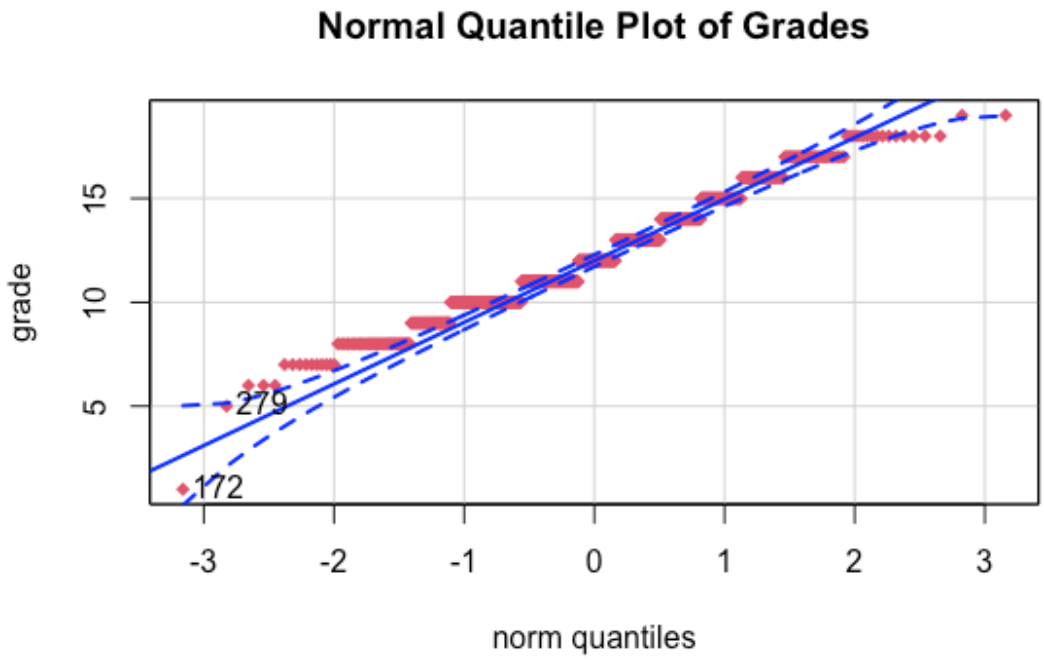
```
cor.test(Walc, grade)

##
##  Pearson's product-moment correlation
##
## data:  Walc and grade
## t = -4.7322, df = 632, p-value = 2.743e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2591219 -0.1086860
## sample estimates:
##        cor
## -0.1849874
```

```
cor.test(Dalc, grade)
cor.test(Dalc, Walc)
cor.test(age, grade)
cor.test(freetime, goout)
cor.test(goout, Walc)
```

Discussion of correlation results (See Matrix Plot for more correlations.) We found statistically significant positive correlations between Dalc and Walc (p-value < 2.2e-16), freetime and goout (p-value < 2.2e-16), and goout and Walc (p-value < 2.2e-16). For these, we reject the null hypothesis that the true correlation is equal to 0. The correlation between age and grade was not statistically significant (p-value = 0.3393, above alpha 0.05), and we do not reject the null hypothesis that the correlation is equal to zero. We found statistically significant negative correlations between Dalc and grade (p-value = 8.19e-08) and Walc and grade (p-value = 2.743e-06). For these, we reject the null hypothesis that the truee correlation is equal to 0.

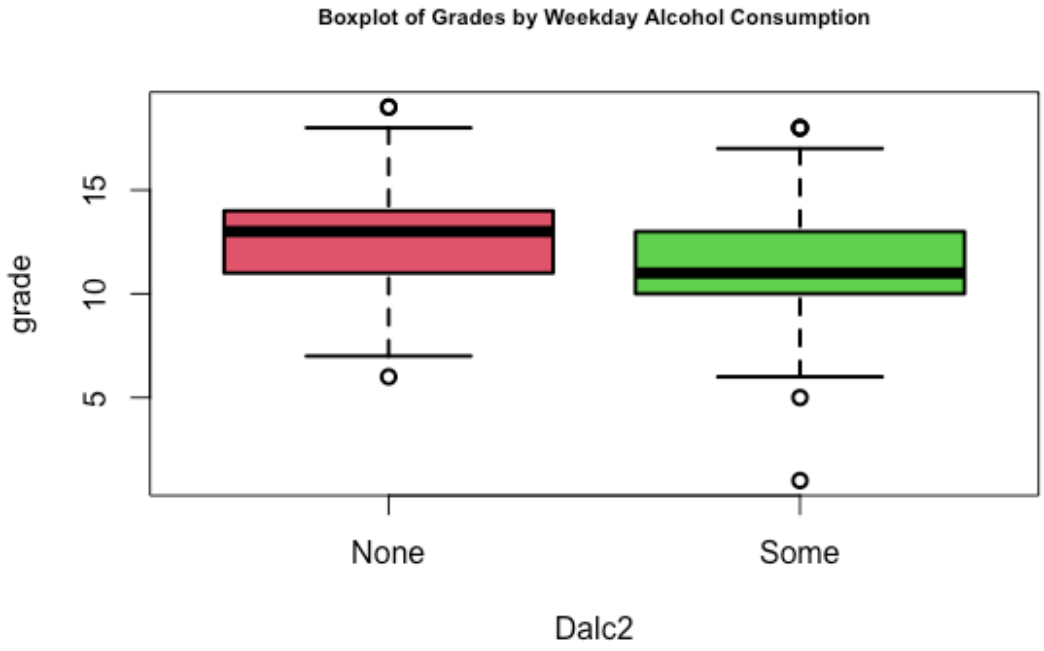
Basic Tests - Bootstrap

```
#Recoding Dalc and Walc to a binary, for later use
Dalc2 <- ifelse(Dalc == 1, "None", "Some")
Walc2 <- ifelse(Walc == 1, "None", "Some")

t.test(grade ~ Dalc2)

##
##  Welch Two Sample t-test
##
## data:  grade by Dalc2
## t = 4.8455, df = 361.66, p-value = 1.877e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6582174 1.5574469
## sample estimates:
## mean in group None mean in group Some
##           12.52144           11.41361

qqPlot(grade, col = 2, pch = 18, main = "Normal Quantile Plot of Grades")
```




```
## [1] 172 279
```

```
boxplot(grade ~ Dalc2, main = "Boxplot of Grades by Weekday Alcohol
Consumption", cex.main = 0.7, col = c(2:4), lwd = 2)
```



```r
sum(Dalc2 == "None")
```

```
## [1] 443
```

```r
sum(Dalc2 == "Some")
```

```
## [1] 191
```

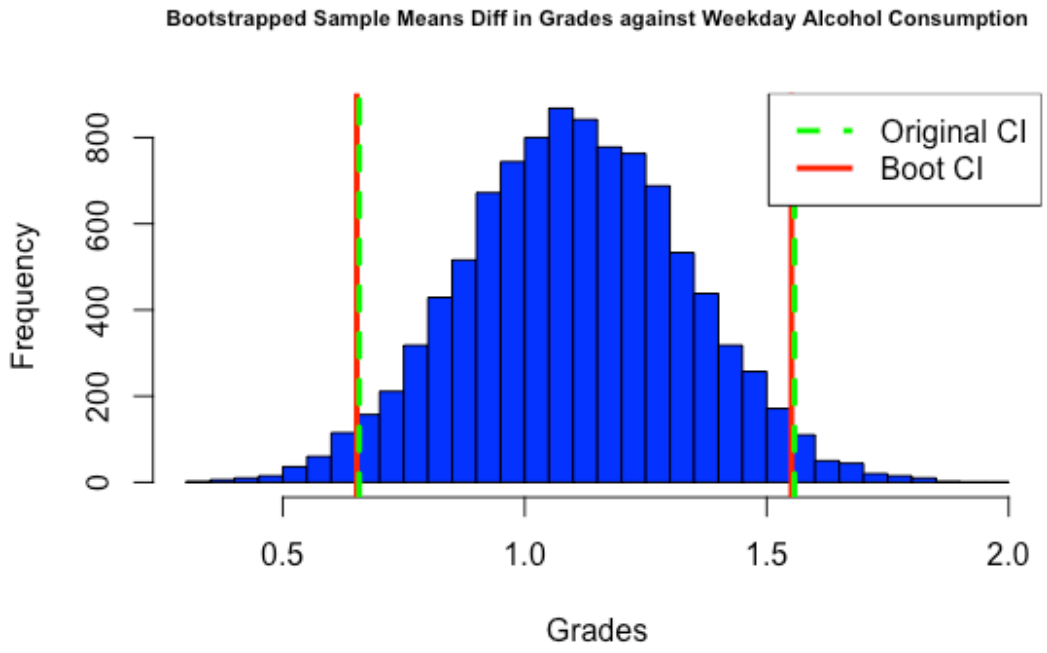
```r
N <- 10000
diffGrade <- rep(NA, N)
for (i in 1:N) {
  sN <- sample(grade[Dalc2 =="None"], sum(Dalc2 == "None"), replace = TRUE)
  sS <- sample(grade[Dalc2 =="Some"], sum(Dalc2 == "Some"), replace = TRUE)
  diffGrade[i] <- mean(sN) - mean(sS)
}

ci <- quantile(diffGrade, c(0.025, 0.975))
round(ci,1)
```

```
##  2.5% 97.5%
##   0.7   1.6
```

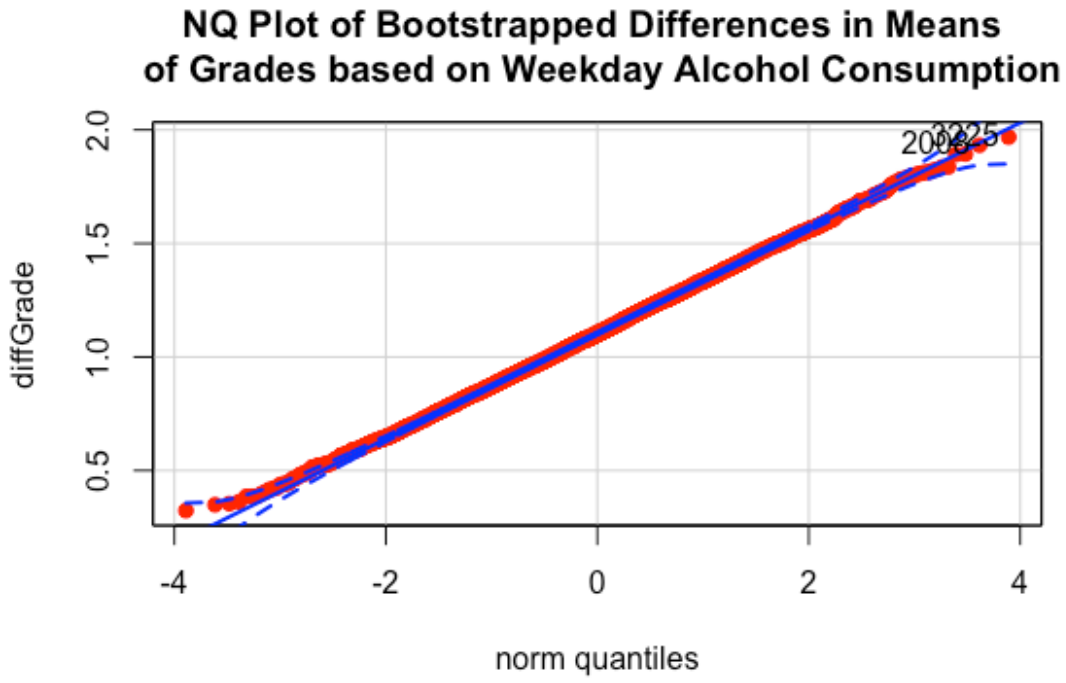
```r
Gradet_test <- t.test(grade ~ Dalc2)$conf.int
round(Gradet_test,2)
```

```
## [1] 0.66 1.56
## attr(,"conf.level")
## [1] 0.95
```

```r
hist(diffGrade, col = "blue", main = "Bootstrapped Sample Means Diff in
Grades against Weekday Alcohol Consumption", cex.main = 0.7, xlab = "Grades",
breaks = 50)

abline(v = ci, lwd = 3, col = "red")
abline(v = Gradet_test, lwd = 3, col = "green", lty = 2)
legend("topright", c("Original CI","Boot CI"), lwd = 3, col =
c("green","red"), lty = c(2,1))
```




```
qqPlot(diffGrade, pch = 19, col = 'red', main = "NQ Plot of Bootstrapped
Differences in Means \n of Grades based on Weekday Alcohol Consumption")
```

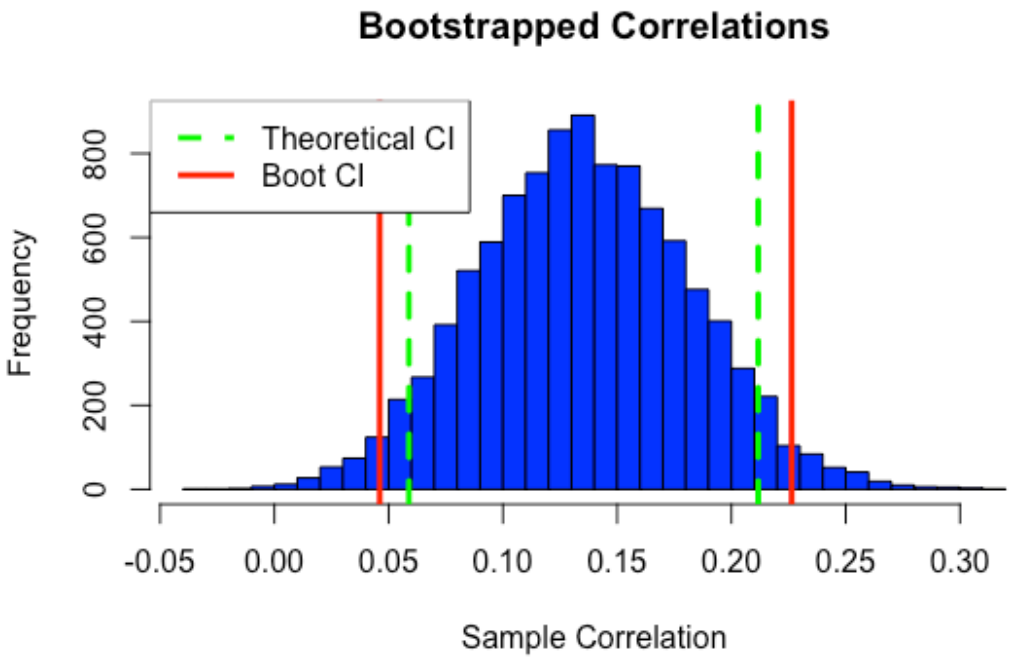



```
## [1] 3225 2008
```

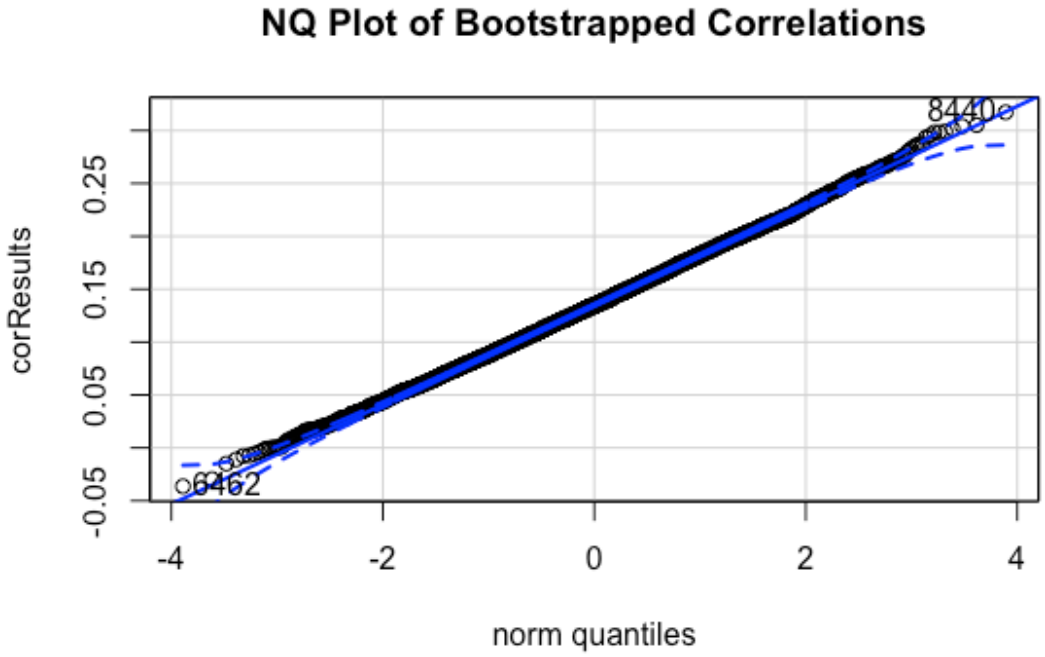
```r
#Bootstrap 2 of grades and Walc (Weekend Alcohol Consumption)
t.test(grade ~ Walc2)
boxplot(grade ~ Walc2, main = "Boxplot of Grades by Weekend Alcohol
Consumption", cex.main = 0.7, col = c(2:4), lwd = 2)
N <- 10000
diffGradeW <- rep(NA, N)
for (i in 1:N) {
  sNW <- sample(grade[Walc2 =="None"], sum(Walc2 == "None"), replace = TRUE)
  sSW <- sample(grade[Walc2 =="Some"], sum(Walc2 == "Some"), replace = TRUE)
  diffGradeW[i] <- mean(sNW) - mean(sSW)
}
ciW <- quantile(diffGradeW, c(0.025, 0.975))
round(ciW,1)
Gradet_testW <- t.test(grade ~ Walc2)$conf.int
round(Gradet_testW,2)
hist(diffGradeW, col = "blue", main = "Bootstrapped Sample Means Diff in
Grades against Weekend Alcohol Consumption", cex.main = 0.7, xlab = "Grades",
breaks = 50)
abline(v = ciW, lwd = 3, col = "red")
abline(v = Gradet_testW, lwd = 3, col = "green", lty = 2)
legend("topright", c("Original CI","Boot CI"), lwd = 3, col =
c("green","red"), lty = c(2,1))
qqPlot(diffGradeW, pch = 19, col = 'red', main = "NQ Plot of Bootstrapped
Differences in Means \n of Grades based on Weekend Alcohol Consumption")

#Bootstrap 3 of Dalc and age using non-binary data
(cor1 <- cor(age, Dalc))

## [1] 0.1360978

cor.test(age, Dalc)

##
##  Pearson's product-moment correlation
##
## data:  age and Dalc
## t = 3.4536, df = 632, p-value = 0.0005902
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05885448 0.21172102
## sample estimates:
##       cor
## 0.1360978

lm1 <- lm(Dalc ~ age)
summary(lm1)

##
## Call:
## lm(formula = Dalc ~ age)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9341 -0.5226 -0.4197  0.4774  3.6832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.22621    0.49931  -0.453  0.65067
## age          0.10287    0.02979   3.454  0.00059 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9094 on 632 degrees of freedom
## Multiple R-squared:  0.01852,    Adjusted R-squared:  0.01697
## F-statistic: 11.93 on 1 and 632 DF,  p-value: 0.0005902
```

```
n_samp <- 10000
corResults <- rep(NA, n_samp)
bResults <- rep(NA, n_samp)
for(i in 1:n_samp){
  s <- sample(1:634, 634, replace = T)
  fakeData <- cbind(age[s], Dalc[s])
  corResults[i] <- cor(fakeData[, 1], fakeData[, 2])
  bResults[i] <- lm(fakeData[, 2] ~ fakeData[, 1])$coef[2]
}
ci_r <- quantile(corResults, c(.025, .975))
ci_slope <- quantile(bResults, c(.025, .975))

hist(corResults, col = "blue", main = "Bootstrapped Correlations", xlab =
"Sample Correlation", breaks = 50)
abline(v = ci_r, lwd = 3, col = "red")
abline(v = cor.test(age, Dalc)$conf.int, lwd = 3, col = "green", lty = 2)
legend("topleft", c("Theoretical CI","Boot CI"), lwd = 3, col =
c("green","red"), lty = c(2, 1))
```

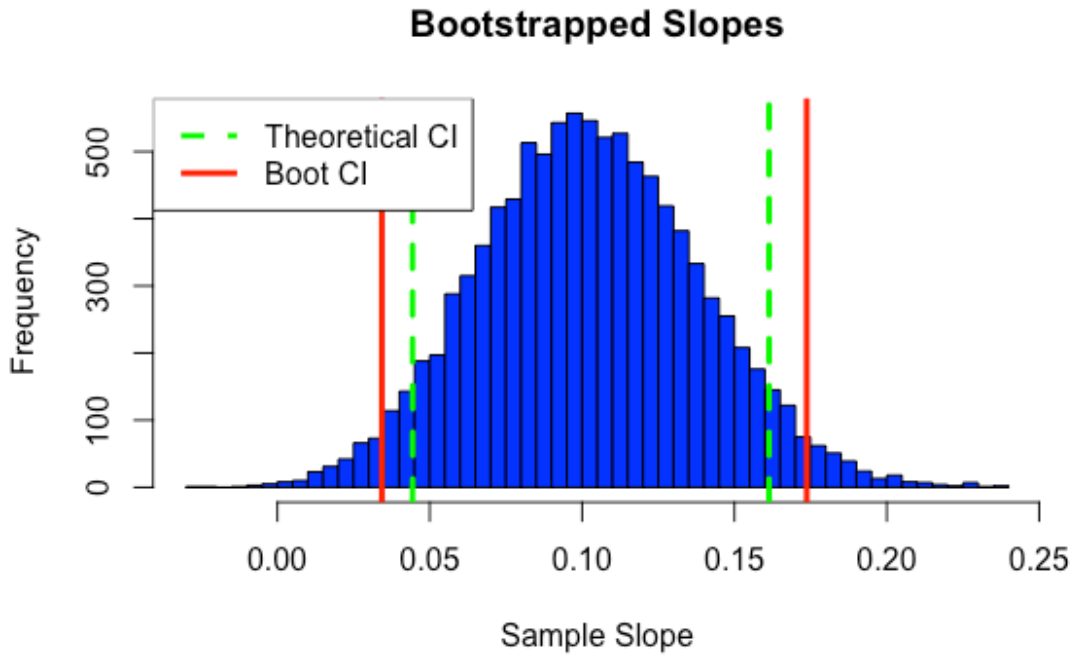



```
qqPlot(corResults, main = "NQ Plot of Bootstrapped Correlations")
```

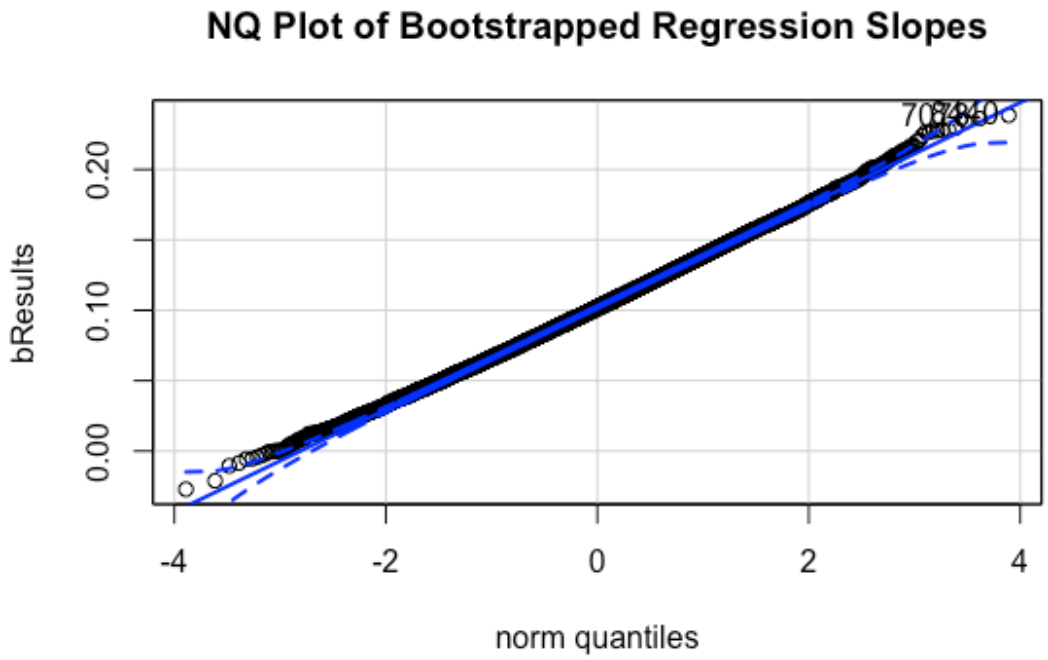



```
## [1] 8440 6462
```

```r
hist(bResults, col = "blue", main = "Bootstrapped Slopes", xlab = "Sample
Slope", breaks = 50)
abline(v = ci_slope, lwd = 3, col = "red")
abline(v = confint(lm1,'age'), lwd = 3, col = "green", lty = 2)
legend("topleft", c("Theoretical CI","Boot CI"), lwd = 3, col =
c("green","red"), lty = c(2, 1))
```




```r
qqPlot(bResults, main = "NQ Plot of Bootstrapped Regression Slopes")
```




```
## [1] 8440 7078
```

Bootstrap Discussion The t-test for the relationship between weekday alcohol consumption and a student's grades says that there is a significant difference in students' grades depending on whether they drank some or none on weekdays. The p value is extremely low in the t-test. This concurs with the bootstrap's results, which have 95% confidence intervals that almost directly overlap with the theoretical intervals. All parts of the 95% CI for differences in mean are also above 0, which show that we have 95% confidence that there is a significant positive difference between those who drink some and none based on the bootstrap. All of the above applies for the relationship between weekend alcohol consumption and a student's grades. The only difference between the two is that the range of means within the 95% confidence interval is between 0.66 and 1.56 for weekday alcohol consumption, and it is between 0.18 and 1.04 for weekend alcohol consumption. This means that on average there is a greater difference in grades between students who drink some on the weekday than on the weekends as compared to those who do not. For the third bootstrap between weekday alcohol consumption and age, there is a significant difference as well with a low p value from the linear regression. The CI for correlation values is 0.0466 and 0.228, and the CI for slopes is 0.0347 and 0.174. The theoretical CI are narrower than the bootstrapped CI for both correlations and slopes, so the real CI may be slightly larger than what the linear models suggest. Nevertheless, the CIs do not include 0, so the bootstrapped values still show that there is a significant difference.

Basic Tests - Permutation Tests

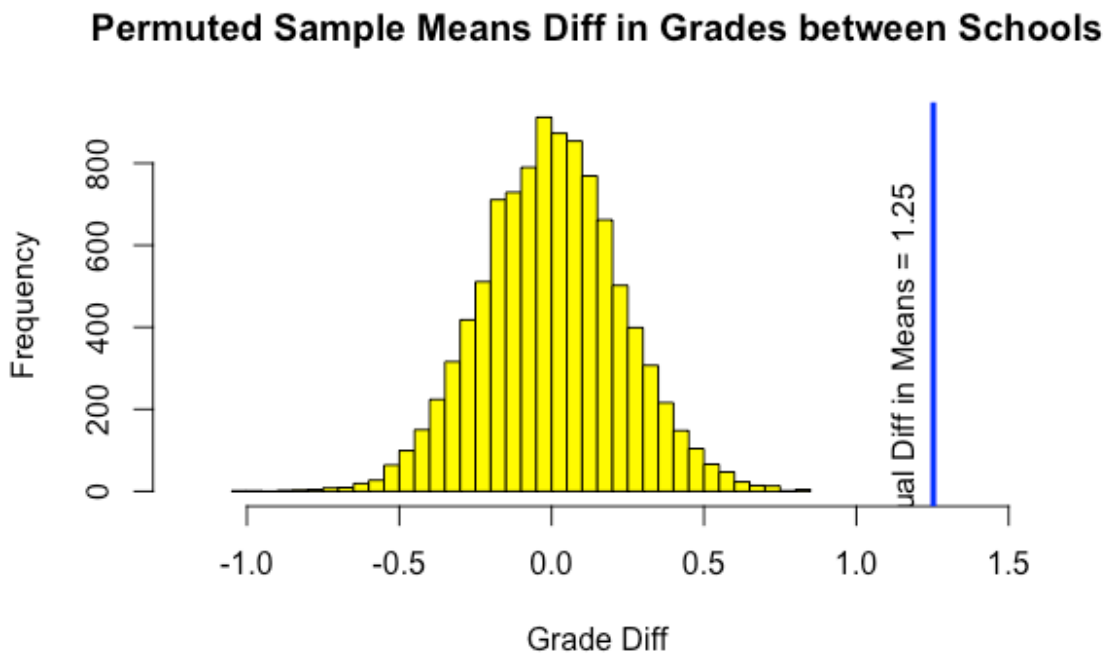
```
# Permutation test 1 (grade by school)
(actualdiff <- by(grade, school, mean))
```

```
## school: GP
## [1] 12.60664
## ------------------------------------------------------------
## school: MS
## [1] 11.35377
```

```
(actualdiff <- actualdiff[1] - actualdiff[2])
```

```
##       GP
## 1.252861
```

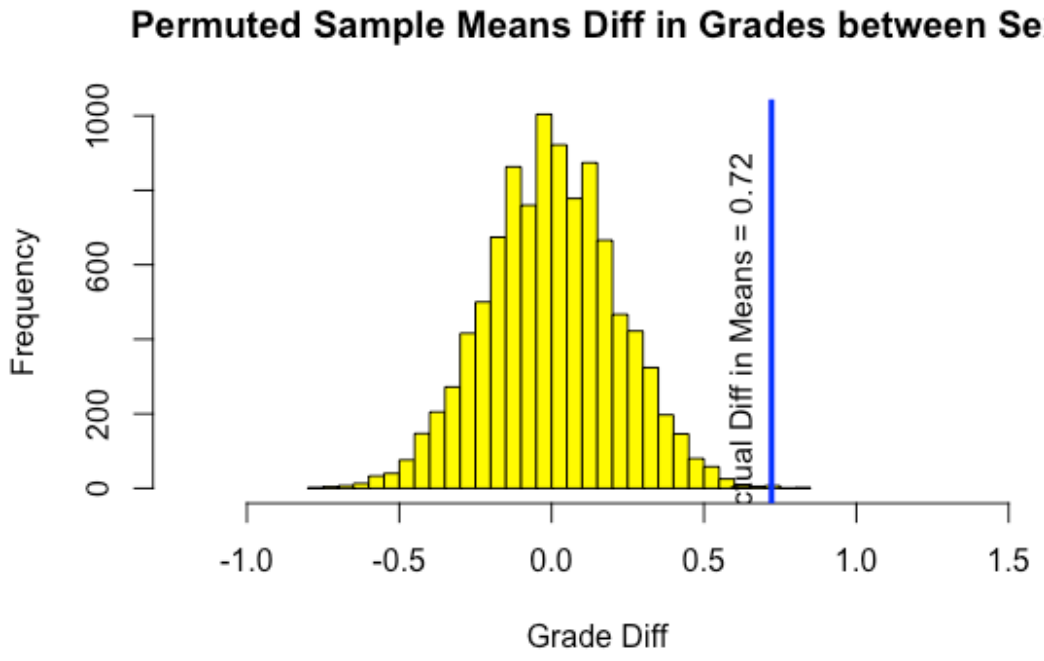
```
set.seed(1)
N <- 10000
diffvals <- rep(NA, N)
for (i in 1:N) {
  fakeschool <- sample(school)  # default is replace = FALSE
  diffvals[i] <- mean(grade[fakeschool == "GP"]) -  mean(grade[fakeschool ==
"MS"])
}
hist(diffvals, col = "yellow", main = "Permuted Sample Means Diff in Grades
between Schools", xlab = "Grade Diff", breaks = 50, xlim = c(-1.2, 1.5))
abline(v = actualdiff, col = "blue", lwd = 3)
text(actualdiff - 0.1, 300 , paste("Actual Diff in Means =",
round(actualdiff,2)),srt = 90)
```




```
(mean(abs(diffvals) >= abs(actualdiff))) ## p-value is approximately 0
```

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

```r
# Permutation test 2 (grade by sex)
(actualdiff1 <- by(grade, sex, mean))

## sex: F
## [1] 12.48138
## ------------------------------------------------------------
## sex: M
## [1] 11.75969

(actualdiff1 <- actualdiff1[1] - actualdiff1[2])

##         F
## 0.7216931

set.seed(1)
N <- 10000
diffvals1 <- rep(NA, N)
for (i in 1:N) {
  fakesex <- sample(sex)  # default is replace = FALSE
  diffvals1[i] <- mean(grade[fakesex == "F"]) -  mean(grade[fakesex == "M"])
}
hist(diffvals1, col = "yellow", main = "Permuted Sample Means Diff in Grades
between Sex", xlab = "Grade Diff", breaks = 50, xlim = c(-1.2, 1.5))
abline(v = actualdiff1, col = "blue", lwd = 3)
text(actualdiff1 - 0.1, 400 , paste("Actual Diff in Means =",
round(actualdiff1,2)),srt = 90)
```




```r
(mean(abs(diffvals1) >= abs(actualdiff1))) ## p-value is approximately 0.0012
```

```
## [1] 0.0012
```

Permutation Test Discussion The result of the first permutation test was that there is enough evidence to reject the null hypothesis that the difference in means of grades between the schools is zero. The permuted p-value was approximately 0, which is similar to the theoretical p-value calculated earlier to be 6.261e-08. These p-values are below alpha 0.05 and therefore we reject the null. The result of the second permutation test was that there is enough evidence to reject the null hypothesis that the difference in means of grades between males and females is zero. The permuted p-value was 0.0012, which is similar to the theoretical p-value calculated earlier to be 0.0009001. These p-values are below alpha 0.05 and therefore we reject the null.

Multiple Regression

Description of Plan We perform backwards step-wise multiple regression predicting grade. Our plan is to include all variables in our initial model (school, sex, age, studytime, higher education, freetime, goout, Dalc, Walc, health, absences, grade), then remove non-significant predictors in order of least significance to most significance, leaving only significant predictors. We transform Walc, Dalc, and absences as indicated and discussed in our matrix plot above. We begin with step-wise removal of interaction terms, then of the remaining predictors. Because we have categorical predictors, we use backwards stepwise regression instead of best subsets regression.

```
# Multiple Regression to predict grade (Backwards Stepwise Regression)
# transformations indicated in matrix plot
logWalc <- log(Walc + 1)
logDalc <- log(Dalc + 1)
sqrtabs <- sqrt(absences)

mod2 <- lm(grade ~ ., data = selected_data)

# with transformed variables
mod2 <- lm(grade ~ logWalc + logDalc + school + sex + age + studytime +
higher + freetime + health + sqrtabs)

# final model
mod2 <- lm(grade ~ logDalc +  + school + age + studytime + higher + health +
sqrtabs)
```

Multiple Regression - Results and Discussion

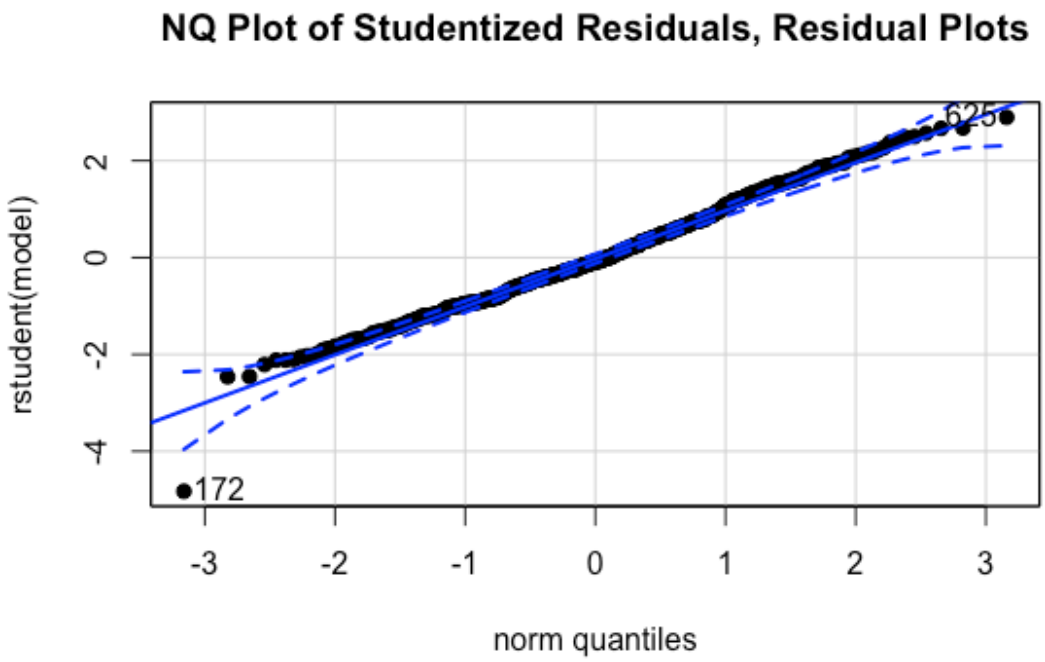
```
summary(mod2)

##
## Call:
## lm(formula = grade ~ logDalc + +school + age + studytime + higher +
##     health + sqrtabs)
##
## Residuals:
```

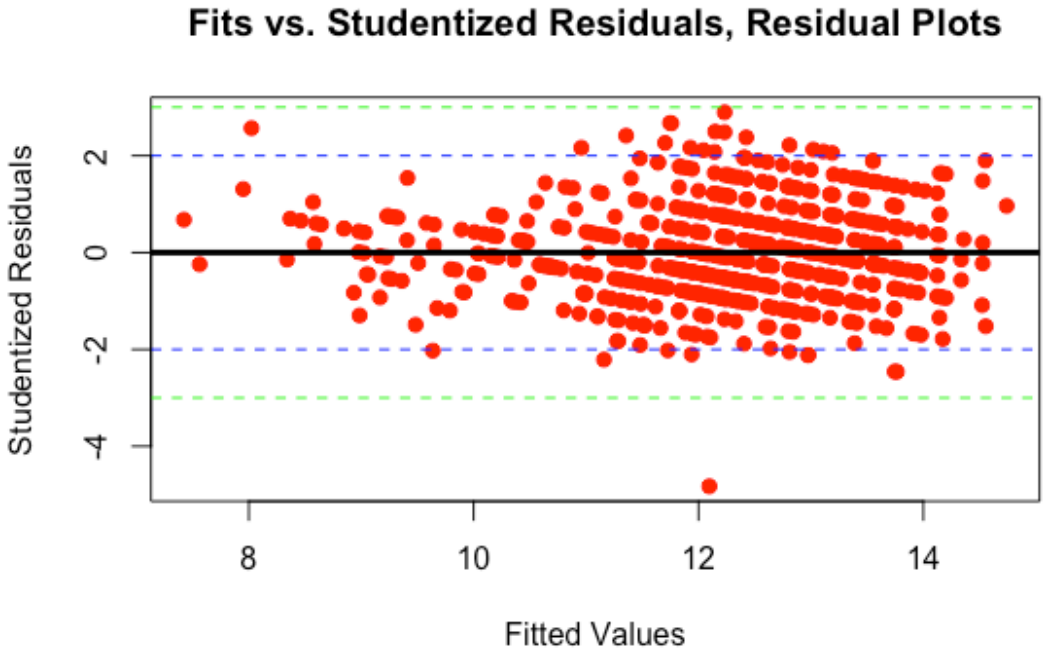
```
##      Min       1Q   Median       3Q      Max
## -11.0965  -1.6251  -0.1998   1.5121   6.7678
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.67312    1.47869   5.865 7.26e-09 ***
## logDalc     -1.17089    0.33288  -3.517 0.000467 ***
## schoolMS    -1.12540    0.20499  -5.490 5.85e-08 ***
## age          0.18970    0.08130   2.333 0.019942 *
## studytime    0.40482    0.11741   3.448 0.000603 ***
## higheryes    2.41994    0.33088   7.314 7.98e-13 ***
## health      -0.19139    0.06528  -2.932 0.003492 **
## sqrtabs     -0.38266    0.07634  -5.012 7.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.362 on 626 degrees of freedom
## Multiple R-squared:  0.2385, Adjusted R-squared:   0.23
## F-statistic: 28.01 on 7 and 626 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value of the model is 0.23, which means our model accounts for 23% of the variation in our response variable, grade. Our model shows that plans to attend higher education and studytime were major significant positive predictors of grades. We also found that logDalc was a major significant negative predictor of grades. The p-value of the model was < 2.2e-16, which is below alpha 0.05, therefore our model is statistically significant. The criteria used to choose our final model was that each predictor was a statistically significant predictor of our response variable grade.

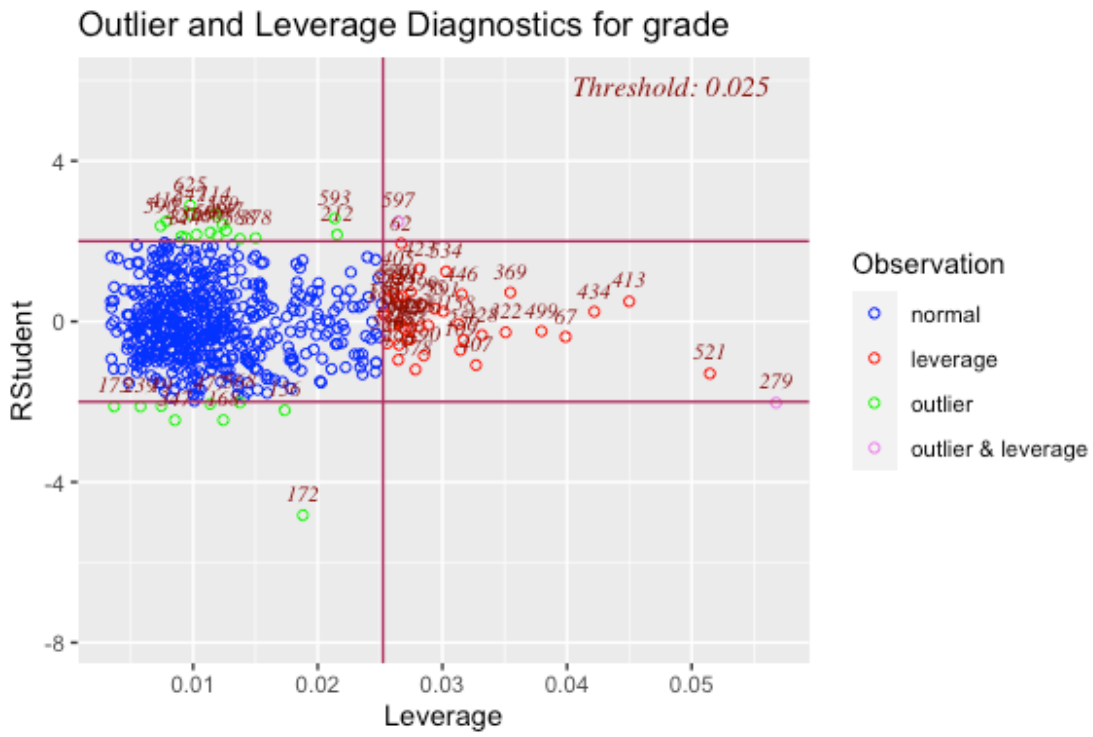
Multiple Regression - Residual Plots

```
myResPlots2(mod2)
```





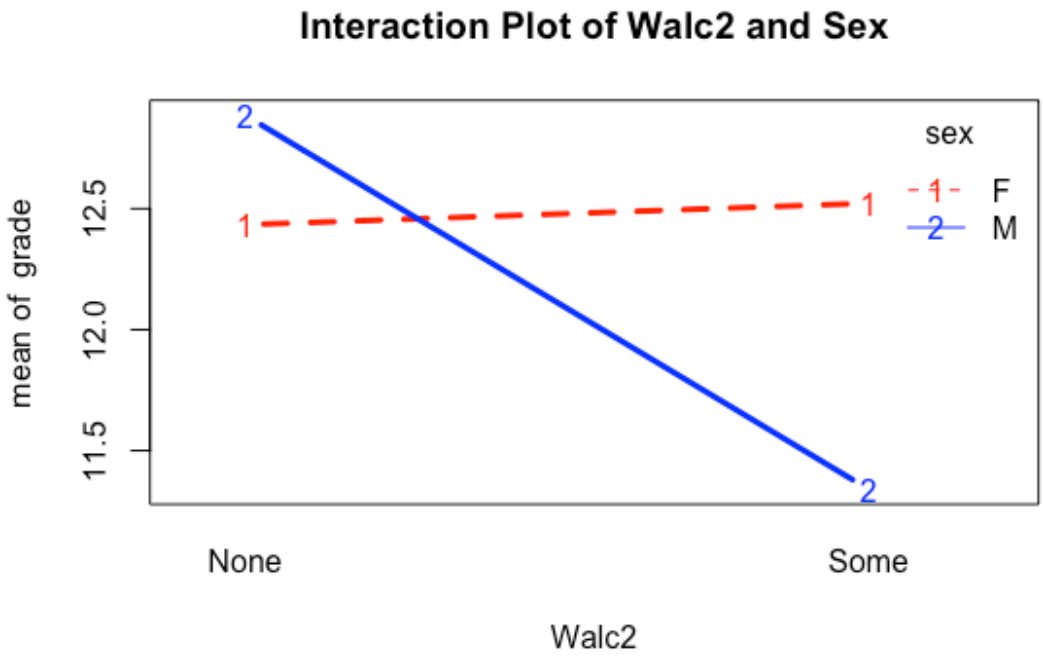

```
ols_plot_resid_lev(mod2)
```



Our normal quantile plot of residuals of our model shows that the residuals are approximately normally distributed. The plot of fits vs. residuals does not show heteroskedasticity or signs of collinearity. This indicates that the assumptions of our multiple regression model are sufficiently met. An outlier and Leverage Diagnostics graph shows that there are not many outliers to this model.

ANOVA

```
# interaction plots
interaction.plot(Walc2, sex, grade, type = 'b', lwd = 3, col =
c('red','blue'), main = "Interaction Plot of Walc2 and Sex")
```



Interaction Plot Discussion An interaction plot of sex and plot indicates an interaction between these variables on grade.

```
# ANOVA Model (predicting grade with Walc2 (binary weekend alcohol
consumption) and sex
gradeaov <- aov(grade ~ Walc2 + sex + Walc2*sex)
summary(gradeaov)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## Walc2         1     56   55.73   8.008 0.004805 **
## sex           1     58   58.42   8.395 0.003894 **
## Walc2:sex     1     88   88.08  12.656 0.000403 ***
## Residuals   630   4384    6.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Comparing Pairs of Means in Two-Way ANOVA
combo <- as.factor(paste(Walc2, sex))
gradeaov2 <- aov(grade ~ combo)
summary(gradeaov2)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## combo         3    202   67.41   9.686 2.95e-06 ***
## Residuals   630   4384    6.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary.lm(gradeaov2)
```

```
##
## Call:
## aov(formula = grade ~ combo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3404  -1.5222  -0.3404   1.6596   7.6596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.43353    0.20057  61.991  < 2e-16 ***
## comboNone M  0.45219    0.37370   1.210    0.227
## comboSome F  0.08864    0.27297   0.325    0.745
## comboSome M -1.09310    0.27793  -3.933 9.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.638 on 630 degrees of freedom
## Multiple R-squared:  0.04409,    Adjusted R-squared:  0.03954
## F-statistic: 9.686 on 3 and 630 DF,  p-value: 2.949e-06
```

```
# Bartlett Test
bartlett.test(grade, combo)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  grade and combo
## Bartlett's K-squared = 0.4204, df = 3, p-value = 0.936
```

```
# Ratio of Max/Min Sample SD
sds <- tapply(grade, combo, sd)
print("Ratio of Max/Min Sample SD")
```

```
## [1] "Ratio of Max/Min Sample SD"
```

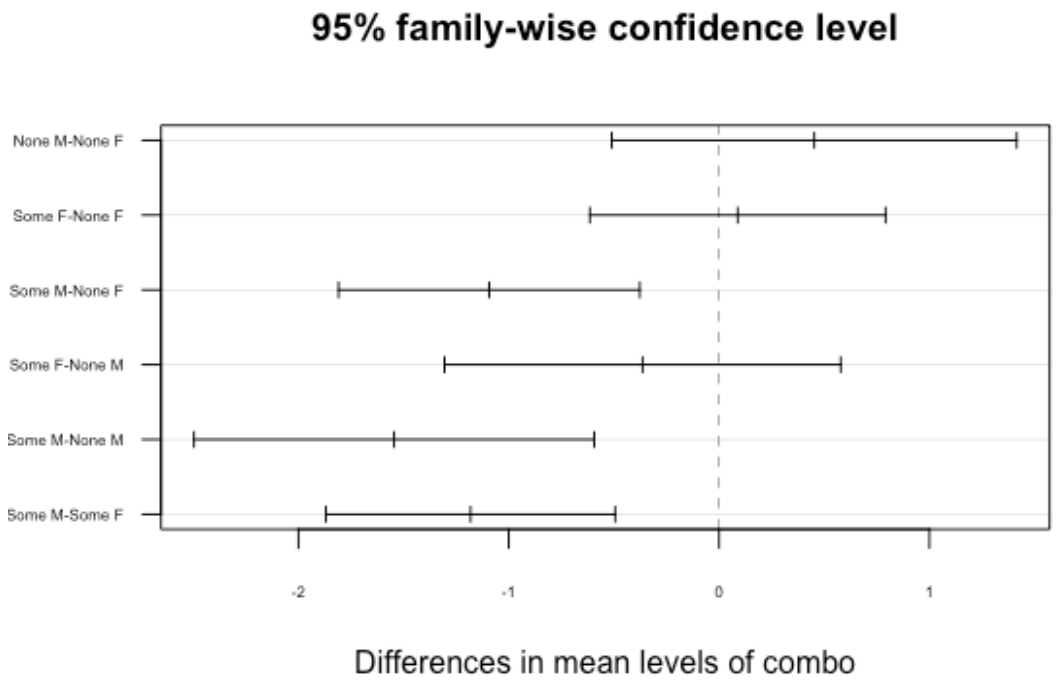
```
round(max(sds)/min(sds), 1)
```

```
## [1] 1.1
```

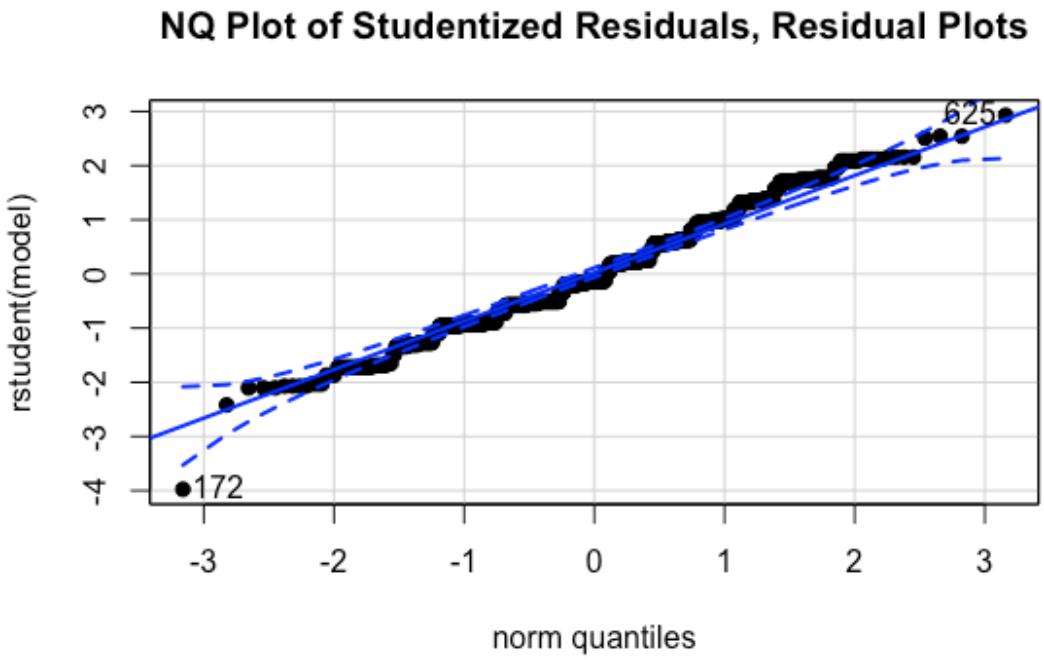
```
oneway.test(grade ~ combo)
```

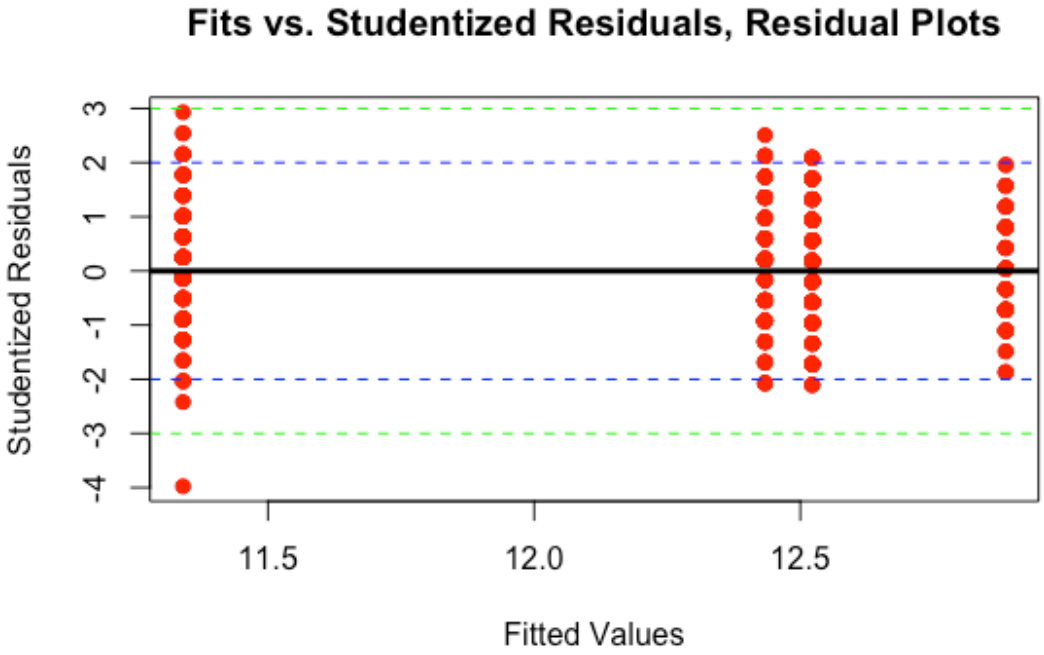
```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  grade and combo
## F = 9.822, num df = 3.00, denom df = 259.37, p-value = 3.699e-06
```

```
# Tukey comparisons
plot(TukeyHSD(gradeaov2), las = 1, cex.axis=.5)
```




```
# Residual plots
myResPlots2(gradeaov)
```





ANOVA Discussion We model grade with predictors Walc2 (binary weekend alcohol consumption) and sex with a two-way ANOVA model. We use Walc2 in the binary because we use it as a categorical variable and because many students selected an option that they consumed no alcohol. Because the ratio of max/min sample SD for our predictors is 1.1, we can reasonably pass the assumption of normal distributed and constant variance across

groups. In the model summary, we can see that all predictors (Walc2, sex, and the interaction between the two) are significant. We conclude that there are differences in grade due to none or some alcohol consumption and male or female sex and there is a statistically signifcant interaction between these two factors. Our two-way anova model had an adjusted R-squared value of 0.040 which indicates that it accounts for approximately only 4% of the variation in grades and shows that in males, weekend alcohol consumption had a negative effect on grades. This was not true for females who indicated that they drink some alcohol on weekends. Tukey comparisons indicate that there is a true difference in the mean grades between males who drink some and females who drink some, males who drink some and males who drink none, and males who drink some and females who drink none. A normal quantile plot of residuals shows that they are approximately normally distributed because they fall linearly. A fits vs. residuals plot does not indicate heteroskedasticity. These residual plots show that the assumptions of the model were valid.

Conclusion

Our project identified and analyzed predictors of grades earned by students in two secondary schools in Portugal. We found that both weekday and weekend alcohol consumption were both statistically significantly negatively correlated with grades. Permutation tests confirmed theoretical t-test results that found that there was a statistically significant difference in grades between males and females and between the two schools. The bootstrap for grades and Dalc was consistent with the t-test which showed significant difference in mean grade between groups who consumed none and some alcohol on weekdays, and the bootstrap for Dalc and age indicated that they are highly correlated, affirming theoretical correlations done above. For the second bootstrap, the bootstrapped CIs are wider than the theoretical CIs. Our multiple regression model is statistically significant and accounts for 23% of the variation in the response variable, grade. Plans to attend higher education was the biggest significant positive predictors of grades. Log of weekday consumption was a major significant negative predictor of grades. Our two-way anova model accounts for approximately only 4% of the variation in grades and our interaction term shows that in males, weekend alcohol consumption had a negative effect on grades. Future analysis can seek to compare the effect of alcohol consumption in a larger number of schools and in schools in other countries, such as the United States. Additionally, future analysis may investigate interactions of school and variables such as studytime and higher education on grade to see how their effects might be different between the schools. A related dataset includes data about students in math classes, which could be used to investigate interesting questions about the predictive power of these variables for grades in portugese vs. math classes. One weakness of our analysis is that our variables, such as weekday and weekend consumption are not metric, and are not standardized. This may have distorted the results of our analysis. However, our recoding of these variables to binary “none” or “some” in our ANOVA attempted to account for this distortion.