Econ 104 - Project 2

Rebecca Zhu, Crystal Huynh, Polat Akbıyk, Tori Takeshita

6/2/2021

1. Introduction

For this project, we are analyzing the "Student Alcohol Consumption" dataset found on Kaggle (https://www.kaggle.com/uciml/student-alcohol-consumption). The data were obtained from a survey of 649 students in Portuguese classes. It contains a variety of social, gender, and study attributes about these students.

Our response variable is Talc = Dalc + Walc. Dalc and Walc are the weekday and weekend alcohol consumption of students, each on a scale of 1 (very low) to 5 (very high).

Our explanatory variables are the following:

absences

activities

age

famrel

free time

G3 (Grade 3)

health

internet

Pstatus

romantic

sex

studytime

traveltime

The reason why we selected these variables for our initial model is a combination of quantitaitive reasoning (using correlation plots) and qualitiative reasoning using our own knowledge.

2. Description of Data

Absences indicates the number of school absences (numeric: from 0 to 93)

Activities - extra-curricular activities (binary: '1' - yes or '0' -no)

Age indicates the student's age (numeric: from 15 to 22)

Famrel indicates the quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

Freetime indicates the amount of free time after school (numeric: from 1 - very low to 5 - very high)

G3 indicates the final grade (numeric: from 0 to 20)

Health indicates current health status (numeric: from 1 - very bad to 5 - very good)

Internet indicates the intrnet access at home (binary: ‘1’ - yes or ‘0’ -no)

Pstatus - parent’s cohabitation status (binary: ‘0’ - living together or ‘1’ - apart)

Sex indicates the student’s sex (binary: ‘0’- female or ‘1’ - male)

Romantic indicates students that are in a romantic relationship (binary: ‘1’ - yes or ‘0’ -no)

Studytime indicates weekly study time (numeric: 1 - 10 hours)

Traveltime indicates the home to school travel time (numeric: 1 - 4 hour)

And our response variable is Talc which indicates both workday alcohol consumption and weekend alcohol consumption (numeric: from ‘2’ - very low to ‘10’ - very high)

3. Data Analysis/Models

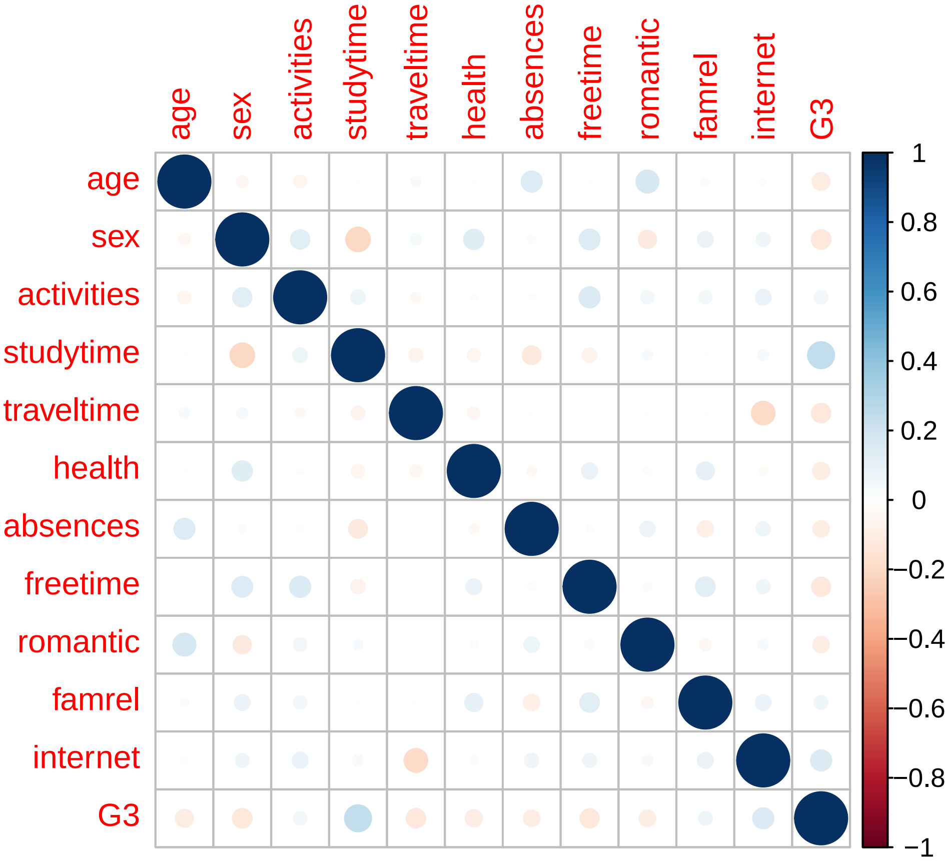
Testing/correcting for best model

First we are testing various linear regression models to determine which variables to keep and what model fits best.

Here we are doing a simple linear regression on the explantory variables as our base model.

```
library(broom)
student <- read.csv(file = ’student-por-cleaned.csv’)
attach(student)
```

```
library(corrplot)
corrplot(cor(student[, c("age", "sex", "activities", "studytime", "traveltime",
"health", "absences", "freetime",
"romantic", "famrel", "internet", "G3")] ))
```



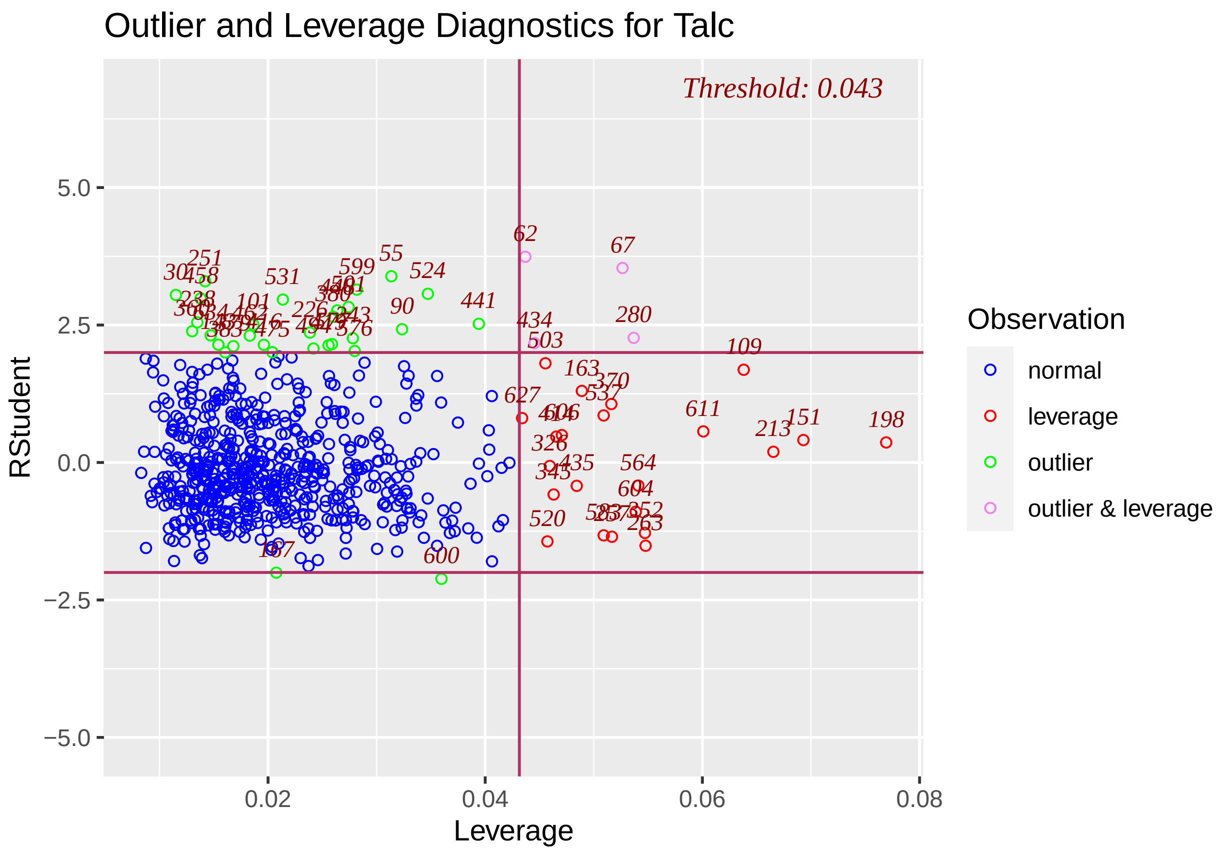
There is no significant correlation between any two variables, which is a good sign. The largest correlation is between study time and G3, but is around ~ 0.3. Next, we'll move on to creating the model and testing which variables are significant and should be included in the model (similar to as in Project 1).

```
mod1 <- lm(Talc~age+sex+activities+studytime+traveltime+health+absences+freetime
           +romantic+famrel+internet+Pstatus+G3, data=student)
summary(mod1)
```

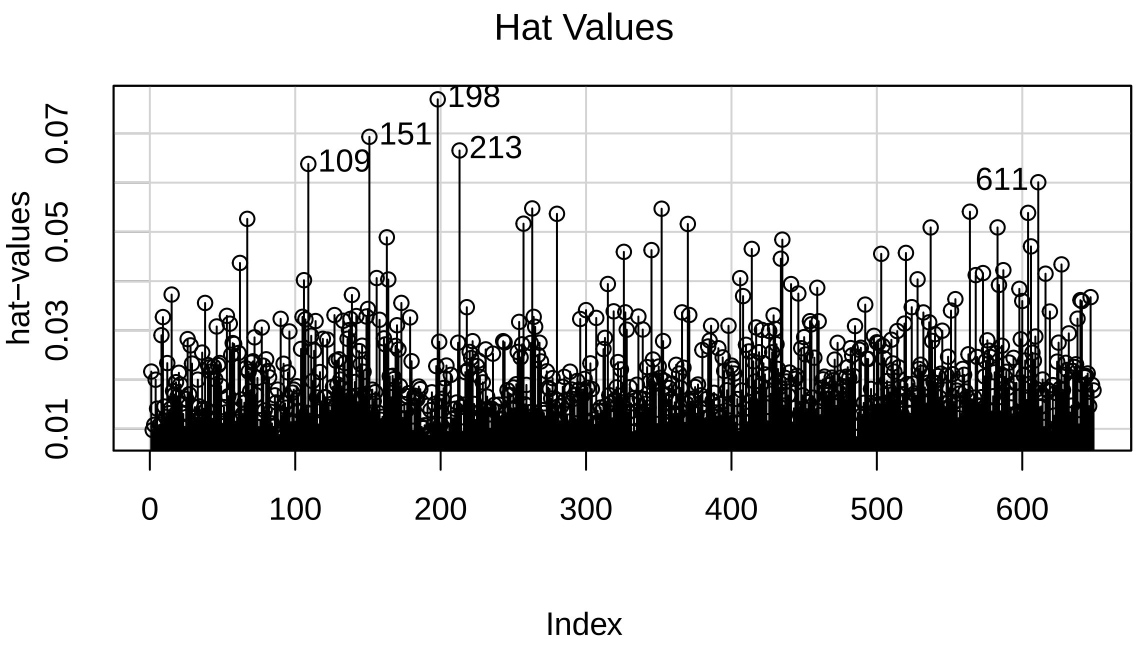
```
##
## Call:
## lm(formula = Talc ~ age + sex + activities + studytime + traveltime +
##     health + absences + freetime + romantic + famrel + internet +
##     Pstatus + G3, data = student)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6991 -1.2352 -0.3347  0.9879  6.4667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.19697    1.14632   1.044 0.296797
## age           0.15216    0.05949   2.558 0.010769 *
## sex           1.16674    0.15175   7.688 5.69e-14 ***
## activities    0.00337    0.14528   0.023 0.981500
## studytime    -0.21688    0.08974  -2.417 0.015938 *
## traveltime    0.15487    0.09649   1.605 0.108981
## health       0.08479    0.04965   1.708 0.088144 .
## absences     0.05817    0.01567   3.711 0.000224 ***
## freetime     0.14088    0.06934   2.032 0.042612 *
## romantic     0.03284    0.15044   0.218 0.827263
## famrel      -0.25610    0.07533  -3.400 0.000717 ***
## internet     0.29608    0.17354   1.706 0.088471 .
## Pstatus      0.35671    0.21724   1.642 0.101073
## G3          -0.06485    0.02347  -2.763 0.005886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.785 on 635 degrees of freedom
## Multiple R-squared:  0.2131, Adjusted R-squared:  0.1969
## F-statistic: 13.22 on 13 and 635 DF,  p-value: < 2.2e-16
```

From what we see in mod1, not all variables are statistically significant and the R squared could be better as it is only 0.19. We will then test the model for outliers, leverages, and influential points to see if we can better the statistical significance and R squared value.

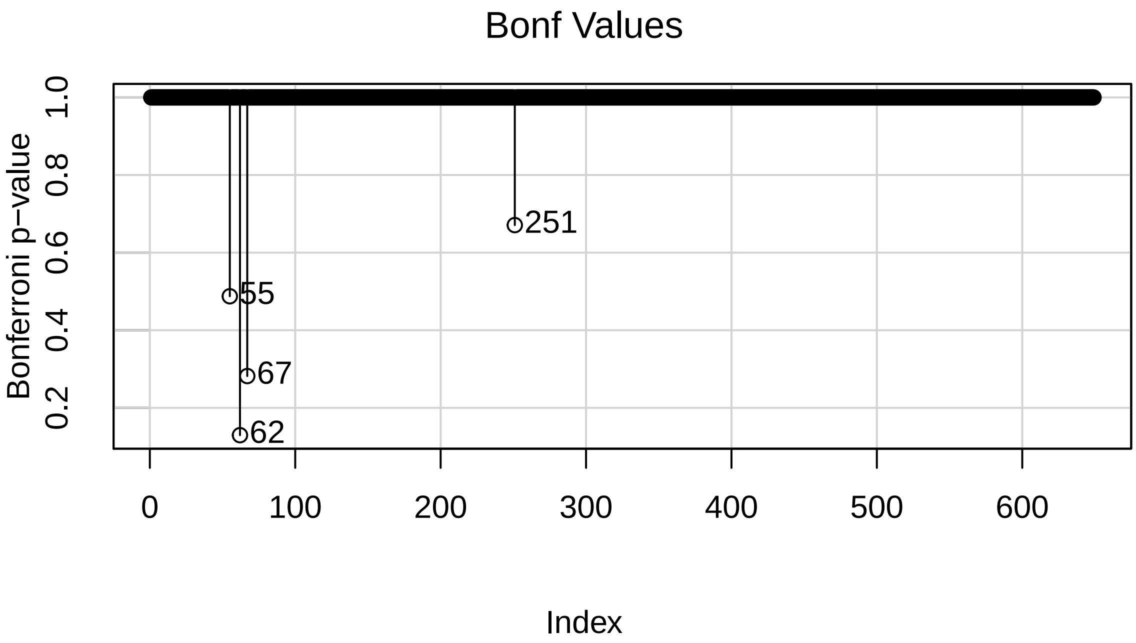
```
library(olsrr)
library(car)

ols_plot_resid_lev(mod1)
```



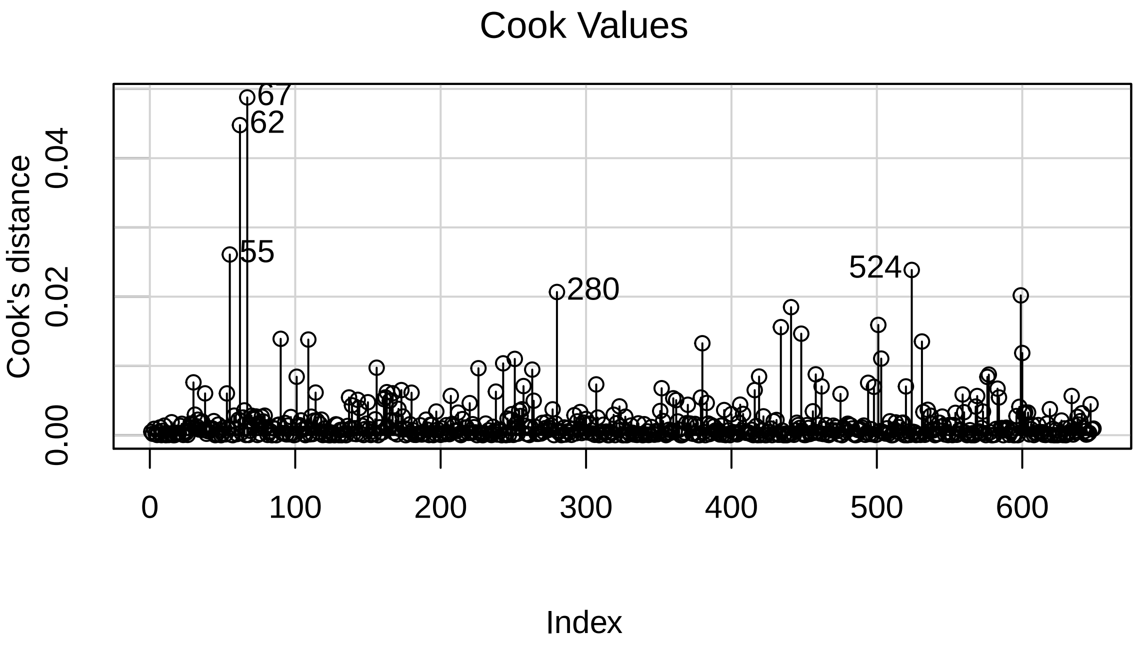
```
#leverages
influenceIndexPlot(mod1, id=list(n=5), vars="hat", main="Hat Values")
```



```
# outliers
influenceIndexPlot(mod1, id=list(n=4), vars="Bonf", main="Bonf Values")
```



```r
# influential
influenceIndexPlot(mod1, id=list(n=5), vars="cook", main="Cook Values")
```



Leverage points: 109, 151, 198, 213, 611 Outliers: 55, 67, 62, 251 Influential points: 55, 62, 67, 280, 524 (3 of the points are within both the top 5 outliers and top 5 influential points).

```
# Do again without the leverages, outliers, influential points
mod2 <- lm(Talc~age+sex+activities+studytime+traveltime+health+absences+freetime
           +romantic+famrel+internet+Pstatus+G3, data=student,
           subset=-c(109, 151, 198, 213, 611, 55, 67, 62, 251, 280, 524))
summary(mod2)
```

```
##
## Call:
## lm(formula = Talc ~ age + sex + activities + studytime + traveltime +
##     health + absences + freetime + romantic + famrel + internet +
##     Pstatus + G3, data = student, subset = -c(109, 151, 198,
##     213, 611, 55, 67, 62, 251, 280, 524))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3921 -1.2214 -0.3182  1.0141  5.5718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.05977    1.12386   0.943 0.346058
## age          0.16063    0.05829   2.756 0.006026 **
## sex          1.08875    0.14694   7.409 4.14e-13 ***
## activities  -0.03580    0.13953  -0.257 0.797573
## studytime   -0.26366    0.08795  -2.998 0.002827 **
## traveltime   0.12516    0.09436   1.326 0.185203
## health       0.10514    0.04825   2.179 0.029704 *
## absences     0.06150    0.01628   3.776 0.000174 ***
## freetime     0.10353    0.06659   1.555 0.120524
## romantic    -0.09509    0.14513  -0.655 0.512573
## famrel      -0.23850    0.07389  -3.228 0.001314 **
## internet     0.23084    0.16673   1.385 0.166689
## Pstatus      0.51346    0.21235   2.418 0.015891 *
## G3          -0.06112    0.02295  -2.663 0.007947 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.704 on 624 degrees of freedom
## Multiple R-squared:  0.2237, Adjusted R-squared:  0.2076
## F-statistic: 13.83 on 13 and 624 DF,  p-value: < 2.2e-16
```

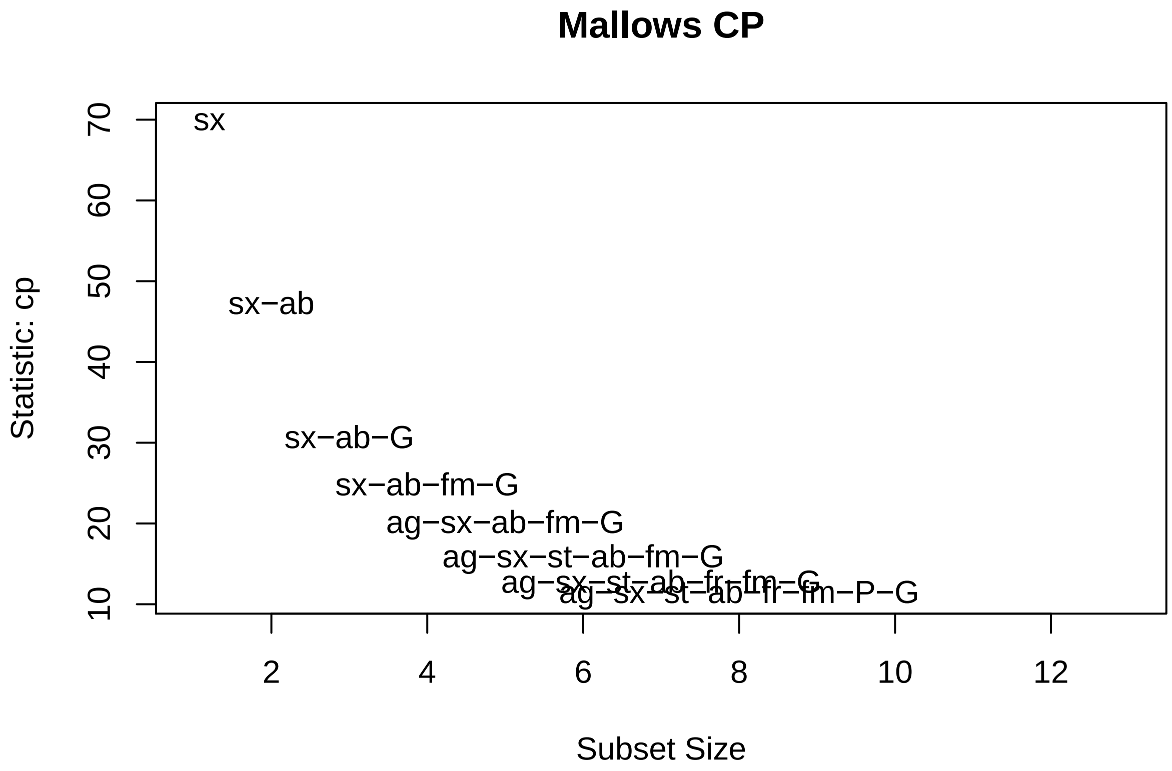
Age, studytime, health, and PstatusT got more statistically significant after removing the leverage, outlier, and influential data points. The variables freetime, famrel, and internet are now less statistically significant. As a whole, the model, mod2, is a bit better than mod1 because the R squared value increased from 0.197 to 0.208.

Next we will test what variables should be keep and which should be removed from the model with the Mallows CP test and Boruta Test.

```r
library(AER)
library(leaps)
ols_mallows_cp(mod2, mod1)
```

```
## [1] -41.53815
```

```r
ss = regsubsets(Talc~age+sex+activities+studytime+traveltime+health+absences+
                  freetime+romantic+famrel+internet+Pstatus+G3, data=student)
subsets(ss, statistic="cp", legend = F, main="Mallows CP", col="steelblue4")
```

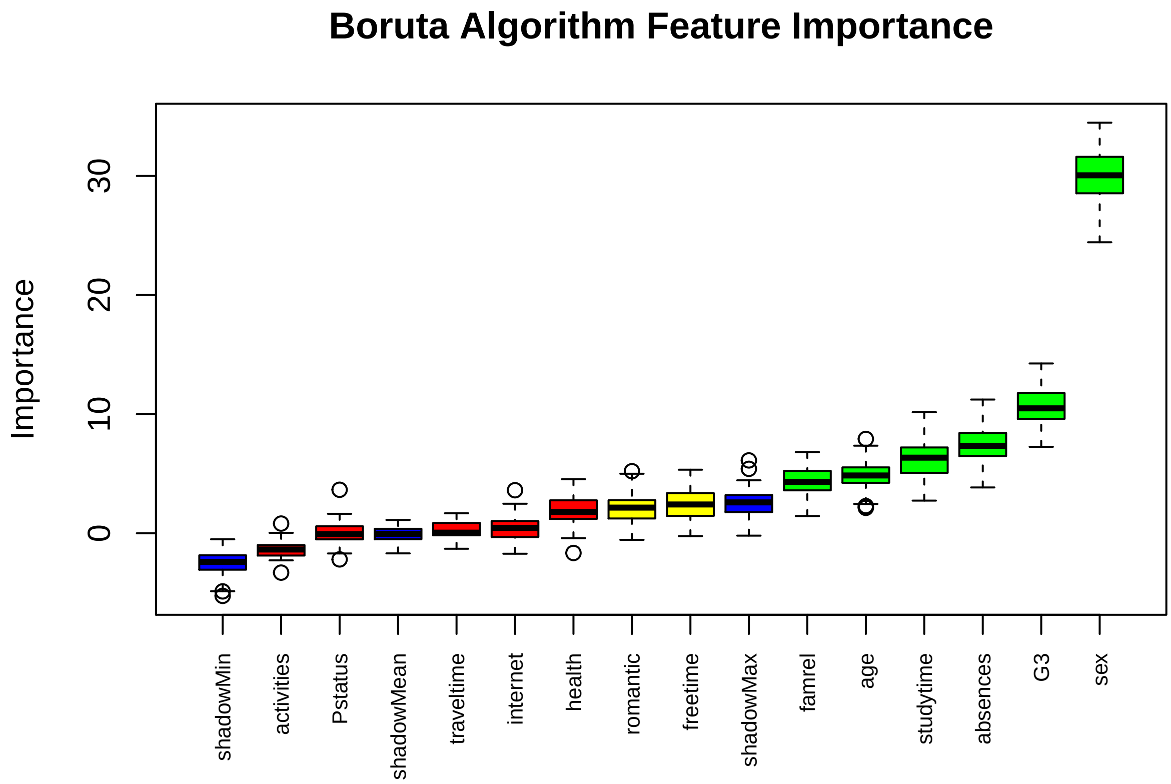


```
##             Abbreviation
## age                   ag
## sex                   sx
## activities            ac
## studytime             st
## traveltime             t
## health                 h
## absences              ab
## freetime              fr
## romantic               r
## famrel                fm
## internet               i
## Pstatus                P
## G3                     G
```

The output of Mallows CP tells us that the most significant variables are: age, sex, studytime, absences, freetime, famrel, Pstatus, and G3. This excludes activities, traveltime, health, romantic, and internet.

Next, we will also perform the Boruta test.

```
library(Boruta)
mod.Bor <- Boruta(Talc~age+sex+activities+studytime+traveltime+health+absences+freetime
                  +romantic+famrel+internet+Pstatus+G3, data=student, doTrace=2)
plot(mod.Bor, xlab="", xaxt= "n", main="Boruta Algorithm Feature Importance")

lz<-lapply(1:ncol(mod.Bor$ImpHistory),function(i)
mod.Bor$ImpHistory[is.finite(mod.Bor$ImpHistory[,i]),i])
names(lz) <- colnames(mod.Bor$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
at = 1:ncol(mod.Bor$ImpHistory), cex.axis = 0.7)
```



The output of Boruta corresponds with the Mallows CP test - most of the variables are green, with freetime being yellow.

Next, we will create a new model based on the most significant variables (determined by these past few tests)

```
mod3 <- lm(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3,
           data=student, subset=-c(109, 151, 198, 213, 611, 55, 67, 62, 251, 280, 524))
summary(mod3)
```

```
##
## Call:
## lm(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3, data = student, subset = -c(109, 151,
##     198, 213, 611, 55, 67, 62, 251, 280, 524))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.589 -1.221 -0.320  1.077  5.583
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.67053    1.09977   1.519 0.129270
## age          0.15955    0.05754   2.773 0.005723 **
## sex          1.14700    0.14383   7.975 7.23e-15 ***
## studytime   -0.26830    0.08785  -3.054 0.002353 **
## absences     0.06286    0.01622   3.875 0.000118 ***
## freetime     0.11049    0.06596   1.675 0.094406 .
## famrel      -0.21746    0.07356  -2.956 0.003232 **
## Pstatus      0.52801    0.21112   2.501 0.012638 *
## G3          -0.06402    0.02239  -2.860 0.004377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.708 on 629 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.2041
## F-statistic: 21.42 on 8 and 629 DF,  p-value: < 2.2e-16
```

Looking at the statistical significance of the variables, we find that the statisitical significance of freetime increased. The others variables have the same statistical significance. However the standard errors of all the variables in mod3 is less than that of mod2 (though my very small amounts ~0.01 to 0.001) However the R squared decreased by ~ 0.02 for mod3, which is not ideal.

We will now test for multicollinearity for mod3.

```
# testing for multicollinearity
tidy(vif(mod3))
```

```
## # A tibble: 8 x 2
##   names         x
##   <chr>     <dbl>
## 1 age        1.04
## 2 sex        1.09
## 3 studytime  1.14
## 4 absences   1.07
## 5 freetime   1.05
## 6 famrel     1.04
## 7 Pstatus    1.02
## 8 G3         1.11
```

All of our VIF scores are less than 4, so we do not need to remove any variables.

Next, we will test for heteroskedasticity.

```
library(car)
ncvTest(mod3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 73.44337, Df = 1, p = < 2.22e-16
```

```
bptest(mod3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 68.619, df = 8, p-value = 9.251e-12
```

The p-value is low for both the NCV test and the BP test, meaning we reject the null hypothesis that there is no heteroskedasticity, and accept the alternative hypothesis that heteroskedasticity is present. Therefore, we must correct for that.

```
#correcting heteroskedasticity with white standard errors
cov1 <- hccm(mod3, type="hc1")
mod4 <- coeftest(mod3, vcov.=cov1)
summary(mod3)
```

```
##
## Call:
## lm(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3, data = student, subset = -c(109, 151,
##     198, 213, 611, 55, 67, 62, 251, 280, 524))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.589 -1.221 -0.320  1.077  5.583
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.67053    1.09977   1.519 0.129270
## age          0.15955    0.05754   2.773 0.005723 **
## sex          1.14700    0.14383   7.975 7.23e-15 ***
## studytime   -0.26830    0.08785  -3.054 0.002353 **
## absences     0.06286    0.01622   3.875 0.000118 ***
## freetime     0.11049    0.06596   1.675 0.094406 .
## famrel      -0.21746    0.07356  -2.956 0.003232 **
## Pstatus      0.52801    0.21112   2.501 0.012638 *
## G3          -0.06402    0.02239  -2.860 0.004377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.708 on 629 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.2041
## F-statistic: 21.42 on 8 and 629 DF,  p-value: < 2.2e-16
```

```
mod4
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1.670528   1.057569  1.5796 0.1147032
## age          0.159546   0.057291  2.7849 0.0055164 **
## sex          1.146997   0.148366  7.7309 4.249e-14 ***
## studytime   -0.268299   0.087034 -3.0827 0.0021411 **
## absences     0.062864   0.018268  3.4412 0.0006176 ***
## freetime     0.110487   0.071648  1.5421 0.1235540
## famrel      -0.217464   0.076412 -2.8459 0.0045726 **
## Pstatus      0.528012   0.193307  2.7315 0.0064822 **
## G3          -0.064023   0.023410 -2.7349 0.0064161 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#correcting heteroskedasticity with fgls
student2 <- student[-c(109, 151, 198, 213, 611, 55, 67, 62, 251, 280, 524)]
ehatsq <- resid(mod3)^2
sighatsq.ols <- lm(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3,
                   data=student2)
vari <- exp(fitted(sighatsq.ols))
mod.fgls <- lm(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3,
               data=student2, weights = 1/vari)
summary(mod.fgls)
```

```
##
## Call:
## lm(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3, data = student2, weights = 1/vari)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32559 -0.22411 -0.04898  0.14518  1.63241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.55394    0.96647   2.643 0.008430 **
## age          0.07778    0.05326   1.461 0.144634
## sex          0.87229    0.16598   5.255 2.02e-07 ***
## studytime   -0.25601    0.07120  -3.595 0.000349 ***
## absences     0.02862    0.01772   1.615 0.106707
## freetime     0.20280    0.05983   3.390 0.000743 ***
## famrel      -0.21815    0.07083  -3.080 0.002158 **
## Pstatus      0.21552    0.16413   1.313 0.189603
## G3          -0.01053    0.02158  -0.488 0.625815
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2696 on 640 degrees of freedom
## Multiple R-squared:  0.1113, Adjusted R-squared:  0.1002
## F-statistic: 10.02 on 8 and 640 DF,  p-value: 3.325e-13
```

Now we will use AIC and BIC to test which model is the best fit.

```
AIC(mod1, mod2, mod3, mod4, mod.fgls)
```

```
##          df      AIC
## mod1     15 2610.050
## mod2     15 2506.606
## mod3     10 2504.472
## mod4      10 2504.472
## mod.fgls 10 2606.053
```

```
BIC(mod1, mod2, mod3, mod4, mod.fgls)
```

```
##           df      BIC
## mod1      15 2677.182
## mod2      15 2573.481
## mod3      10 2549.056
## mod4      10 2549.056
## mod.fgls 10 2650.807
```

The best model is mod3 / mod4 as they are the same value. While mod4 is corrected for heteroskedasticity, the difference between the two models is minimal. Thus we will go with mod3 for flexibility as it is a linear model and easier to compare with other (future) models.

We are cross validating the mod4 model to measure the performance of the model.

```
library(lmvar)
fit= lm(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3, data=student2,
        weights = 1/vari, x = TRUE, y = TRUE)

summary(Talc) # give us range of 2-10
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   3.000   3.783   5.000  10.000
```

```
cv.lm(fit, k = 3)
```

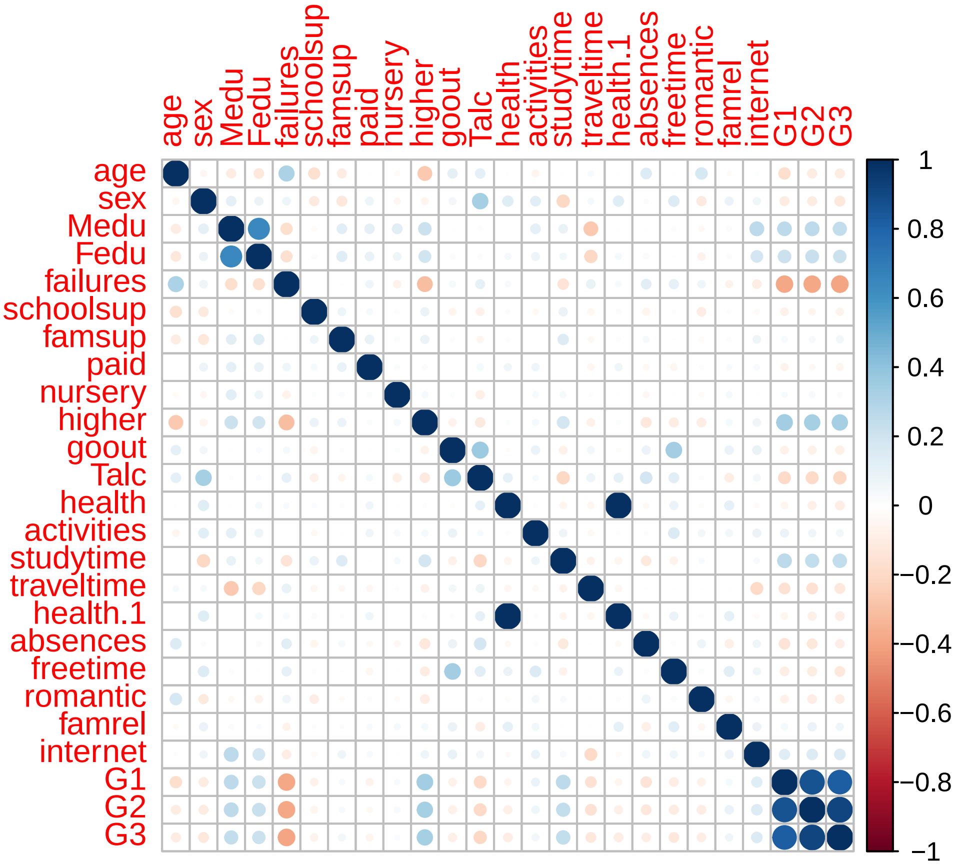
```
## Mean absolute error        :  1.433384
## Sample standard deviation  :  0.06692037
##
## Mean squared error         :  3.348622
## Sample standard deviation  :  0.412192
##
## Root mean squared error    :  1.827686
## Sample standard deviation  :  0.1108241
```

Given that the RMSE is 1.805 and Talc varies from 2-10, then 1.8/8 is about 22.5%. This means that the model is off by about 22.5%. This is not ideal, given that the range that we would like to be in is 1% to 8%.

Probit Models

Here we are trying to test the probability of a student's gender (male or female). The explanatory variables include: studytime, Talc, and activities.

```
# using corrplot to see what variables are correlated with sex:
corrplot(cor(student[, c("age", "sex", "Medu", "Fedu", "failures", "schoolsup",
                         "famsup", "paid", "nursery", "higher", "goout", "Talc",
                         "health", "activities", "studytime", "traveltime",
                         "health", "absences", "freetime", "romantic", "famrel",
                         "internet", "G1", "G2", "G3")] ))
```



```
#baseline model
reg.mod2 <- lm(sex~studytime+Talc, data=student2)
    # using studytime, Talc, and activities because
    #they are the most correlated with sex
summary(reg.mod2)
```

```
##
## Call:
## lm(formula = sex ~ studytime + Talc, data = student2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9637 -0.3532 -0.2590  0.4942  0.9026
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.285779   0.063125   4.527 7.12e-06 ***
## studytime   -0.085254   0.022183  -3.843 0.000133 ***
## Talc         0.076315   0.009236   8.263 8.06e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4587 on 646 degrees of freedom
## Multiple R-squared:  0.1341, Adjusted R-squared:  0.1314
## F-statistic:    50 on 2 and 646 DF,  p-value: < 2.2e-16
```

```r
# probit model
probit.mod2 = glm(sex~studytime+Talc, family=binomial(link="probit"), data=student2)
summary(probit.mod2)
```

```
##
## Call:
## glm(formula = sex ~ studytime + Talc, family = binomial(link = "probit"),
##     data = student2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1682  -0.9244  -0.7655   1.1548   2.0116
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.57521    0.18086  -3.180  0.00147 **
## studytime   -0.24146    0.06554  -3.684  0.00023 ***
## Talc         0.21254    0.02787   7.627  2.4e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 878.50  on 648  degrees of freedom
## Residual deviance: 788.02  on 646  degrees of freedom
## AIC: 794.02
##
## Number of Fisher Scoring iterations: 4
```

```r
# we created a function to output confusion matrix, sensitivity, specificity, PPV and NPV
confusion <- function(model, threshold) {
  pred.classes <- ifelse(fitted(model) > threshold, 1, 0)
  table <- table(pred.classes, student2$sex)
  print(table)
  cat("Sensitivity: ", table[1]/(table[1] + table[2]), "\n")
  cat("Specificity: ", table[4]/(table[3] + table[4]), "\n")
  cat("PPV: ", table[1]/(table[1] + table[3]), "\n")
  cat("NPV: ", table[4]/(table[2] + table[4]), "\n")
}


# default threshold = 0.5
confusion(reg.mod2, 0.5)
```

```
##
## pred.classes   0   1
##            0 330 148
##            1  53 118
## Sensitivity:  0.8616188
## Specificity:  0.443609
## PPV:  0.6903766
## NPV:  0.6900585
```

```
confusion(probit.mod2, 0.5)
```

```
##
## pred.classes   0   1
##            0 305 134
##            1  78 132
## Sensitivity:  0.7963446
## Specificity:  0.4962406
## PPV:  0.6947608
## NPV:  0.6285714
```

```
# higher threshold
confusion(reg.mod2, 0.6)
```

```
##
## pred.classes   0   1
##            0 369 194
##            1  14  72
## Sensitivity:  0.9634465
## Specificity:  0.2706767
## PPV:  0.6554174
## NPV:  0.8372093
```

```
confusion(probit.mod2, 0.6)
```

```
##
## pred.classes   0   1
##            0 369 194
##            1  14  72
## Sensitivity:  0.9634465
## Specificity:  0.2706767
## PPV:  0.6554174
## NPV:  0.8372093
```

```
# lower threshold
confusion(reg.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 252  94
##            1 131 172
## Sensitivity:  0.6579634
## Specificity:  0.6466165
## PPV:  0.7283237
## NPV:  0.5676568
```

```
confusion(probit.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 252  95
##            1 131 171
## Sensitivity:  0.6579634
## Specificity:  0.6428571
## PPV:  0.7262248
## NPV:  0.5662252
```

Result of confusion matrices: probit gave slightly higher sensitivity/true negative rate for 0.5, but adjusting up and down had no impact.

In reg.mod2 vs. probit.mod2, when using 0.5, probit has a higher sensitivity rating (0.496) than the regular (0.4436).

Logit Models

```r
#same baseline model
summary(reg.mod2)
```

```
##
## Call:
## lm(formula = sex ~ studytime + Talc, data = student2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9637 -0.3532 -0.2590  0.4942  0.9026
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.285779   0.063125   4.527 7.12e-06 ***
## studytime   -0.085254   0.022183  -3.843 0.000133 ***
## Talc         0.076315   0.009236   8.263 8.06e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4587 on 646 degrees of freedom
## Multiple R-squared:  0.1341, Adjusted R-squared:  0.1314
## F-statistic:    50 on 2 and 646 DF,  p-value: < 2.2e-16
```

```r
#logit model
logit.mod2 = glm(sex~studytime+Talc, family=binomial(link="logit"), data=student2)
summary(logit.mod2)
```

```
##
## Call:
## glm(formula = sex ~ studytime + Talc, family = binomial(link = "logit"),
##     data = student2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1287  -0.9263  -0.7550   1.1475   2.0080
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.90292    0.29749  -3.035 0.002404 ** 
## studytime   -0.41628    0.11018  -3.778 0.000158 ***
## Talc         0.34752    0.04714   7.372 1.68e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 878.50  on 648  degrees of freedom
## Residual deviance: 787.42  on 646  degrees of freedom
## AIC: 793.42
## 
## Number of Fisher Scoring iterations: 4
```

```
# default threshold = 0.5
confusion(reg.mod2, 0.5)
```

```
## 
## pred.classes   0   1
##            0 330 148
##            1  53 118
## Sensitivity:  0.8616188
## Specificity:  0.443609
## PPV:  0.6903766
## NPV:  0.6900585
```

```
confusion(logit.mod2, 0.5)
```

```
## 
## pred.classes   0   1
##            0 305 134
##            1  78 132
## Sensitivity:  0.7963446
## Specificity:  0.4962406
## PPV:  0.6947608
## NPV:  0.6285714
```

```
# higher threshold
confusion(reg.mod2, 0.6)
```

```
## 
## pred.classes   0   1
##            0 369 194
##            1  14  72
## Sensitivity:  0.9634465
## Specificity:  0.2706767
## PPV:  0.6554174
## NPV:  0.8372093
```

```
confusion(logit.mod2, 0.6)
```

```
##
## pred.classes   0   1
##            0 358 182
##            1  25  84
## Sensitivity:  0.9347258
## Specificity:  0.3157895
## PPV:  0.662963
## NPV:  0.7706422
```

```
# lower threshold
confusion(reg.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 252  94
##            1 131 172
## Sensitivity:  0.6579634
## Specificity:  0.6466165
## PPV:  0.7283237
## NPV:  0.5676568
```

```
confusion(logit.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 260  96
##            1 123 170
## Sensitivity:  0.6788512
## Specificity:  0.6390977
## PPV:  0.7303371
## NPV:  0.5802048
```

Here we are comparing the sensitivity, specificity, positive and negative predicited values for the OLS, probit, and logit models.

```
# regular - probit and logit are same, better than OLS for spec/PPV, but worse for sens/NPV
confusion(reg.mod2, 0.5)
```

```
##
## pred.classes   0   1
##            0 330 148
##            1  53 118
## Sensitivity:  0.8616188
## Specificity:  0.443609
## PPV:  0.6903766
## NPV:  0.6900585
```

```
confusion(probit.mod2, 0.5)
```

```
##
## pred.classes   0   1
##            0 305 134
##            1  78 132
## Sensitivity:  0.7963446
## Specificity:  0.4962406
## PPV:  0.6947608
## NPV:  0.6285714
```

```
confusion(logit.mod2, 0.5)
```

```
##
## pred.classes   0   1
##            0 305 134
##            1  78 132
## Sensitivity:  0.7963446
## Specificity:  0.4962406
## PPV:  0.6947608
## NPV:  0.6285714
```

```
# move up - OLS and probit become the same, logit is better for spec/PPV, worse for sens/NPV
confusion(reg.mod2, 0.6)
```

```
##
## pred.classes   0   1
##            0 369 194
##            1  14  72
## Sensitivity:  0.9634465
## Specificity:  0.2706767
## PPV:  0.6554174
## NPV:  0.8372093
```

```
confusion(probit.mod2, 0.6)
```

```
##
## pred.classes   0   1
##            0 369 194
##            1  14  72
## Sensitivity:  0.9634465
## Specificity:  0.2706767
## PPV:  0.6554174
## NPV:  0.8372093
```

```
confusion(logit.mod2, 0.6)
```

```
##
## pred.classes   0   1
##            0 358 182
##            1  25  84
## Sensitivity:  0.9347258
## Specificity:  0.3157895
## PPV:  0.662963
## NPV:  0.7706422
```

```
# move down - logit better for everything except specificity
confusion(reg.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 252  94
##            1 131 172
## Sensitivity:  0.6579634
## Specificity:  0.6466165
## PPV:  0.7283237
## NPV:  0.5676568
```

```
confusion(probit.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 252  95
##            1 131 171
## Sensitivity:  0.6579634
## Specificity:  0.6428571
## PPV:  0.7262248
## NPV:  0.5662252
```

```
confusion(logit.mod2, 0.4)
```

```
##
## pred.classes   0   1
##            0 260  96
##            1 123 170
## Sensitivity:  0.6788512
## Specificity:  0.6390977
## PPV:  0.7303371
## NPV:  0.5802048
```

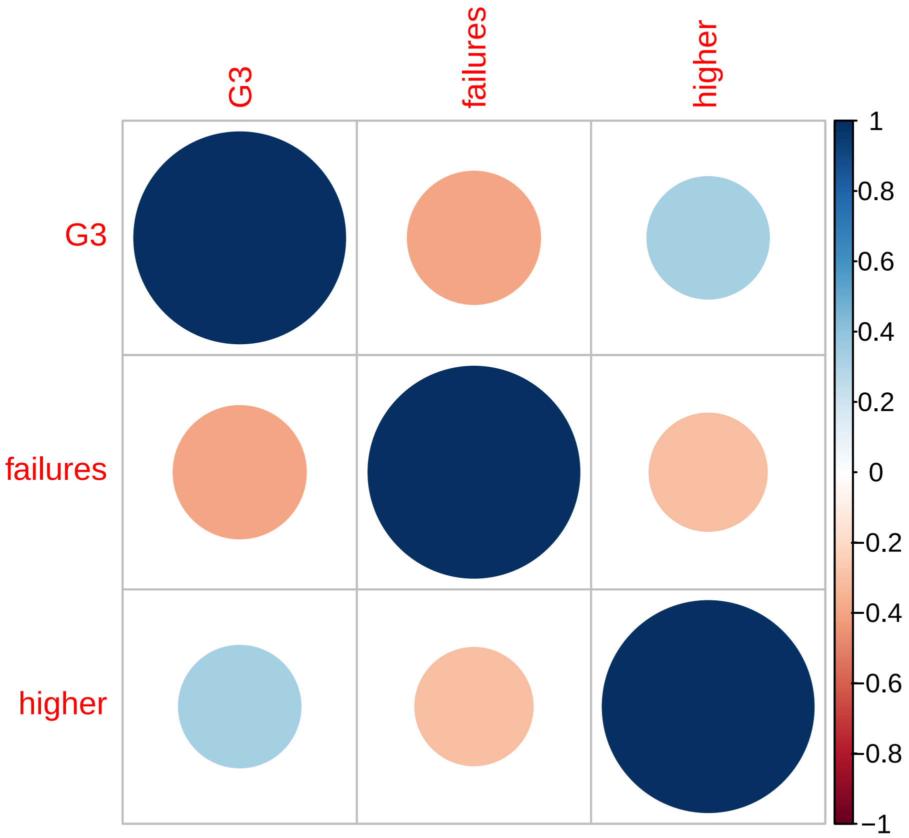
By comparing the models, we see that there is a variability of results. When comparing the 0.5 thresholds, the probit and logit model have the same results. They are better than the OLS model for specificity and PPV, but not for sensitivity or NPV. However when looking at that a higher threshold, we find that the OLS and probit model have the same results. Logit is better for specificity and PPV but worse for sensitivity and NPV. When we lower the threshold, we find that the logit is better for all the test except for specificity.

Instrumental Variables

```
mod.test <- lm(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+failures,
               data=student2)
summary(mod.test)
```

```
##
## Call:
## lm(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + failures, data = student2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.954 -1.243 -0.323  1.010  6.730
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.67671    1.11028   0.609 0.542417
## age          0.17982    0.06196   2.902 0.003835 **
## sex          1.25077    0.14918   8.384 3.25e-16 ***
## studytime   -0.27729    0.08850  -3.133 0.001809 **
## absences     0.06165    0.01572   3.922 9.74e-05 ***
## freetime     0.17597    0.06900   2.550 0.010998 *
## famrel      -0.25510    0.07533  -3.387 0.000751 ***
## Pstatus      0.38896    0.21716   1.791 0.073755 .
## failures    -0.02551    0.12885  -0.198 0.843104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 640 degrees of freedom
## Multiple R-squared:  0.1938, Adjusted R-squared:  0.1837
## F-statistic: 19.23 on 8 and 640 DF,  p-value: < 2.2e-16
```

```
corrplot(cor(student2[, c("G3", "failures", "higher")] ))
```



```r
mod.test2 <- lm(G3~failures+higher, data=student2)
summary(mod.test2)
```

```
##
## Call:
## lm(formula = G3 ~ failures + higher, data = student2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5534  -1.5534   0.1963   1.4466   6.4466
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.1157     0.3792  26.677  < 2e-16 ***
## failures     -1.7497     0.2011  -8.702  < 2e-16 ***
## higher        2.4377     0.3867   6.304 5.37e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.887 on 646 degrees of freedom
## Multiple R-squared:  0.2037, Adjusted R-squared:  0.2012
## F-statistic: 82.62 on 2 and 646 DF,  p-value: < 2.2e-16
```

The mod.test2 shows that the variables failure and higher both explain G3, as they are all highly statistically significant. Looking at the R squared, it explains about 20% of G3, which is also evident in the correlation plot.

Given the results above, we are going to test failures an higher as an instrumental variable for G3.

```r
# IV hypothesis #1: failures as IV for G3
mod.instr1 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    |age+sex+studytime+absences+freetime+famrel+Pstatus+failures,
                    data=student2)
summary(mod.instr1)
```

```
## 
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime + 
##     famrel + Pstatus + G3 | age + sex + studytime + absences + 
##     freetime + famrel + Pstatus + failures, data = student2)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9976 -1.2462 -0.3262  1.0005  6.6728 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.53297    1.47878   0.360 0.718656    
## age          0.17949    0.06164   2.912 0.003717 ** 
## sex          1.25598    0.15332   8.192  1.4e-15 ***
## studytime   -0.28654    0.10545  -2.717 0.006757 ** 
## absences     0.06187    0.01587   3.899 0.000107 ***
## freetime     0.17886    0.07231   2.474 0.013634 *  
## famrel      -0.25755    0.07745  -3.325 0.000934 ***
## Pstatus      0.38918    0.21779   1.787 0.074414 .  
## G3           0.01333    0.06749   0.197 0.843550    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.805 on 640 degrees of freedom
## Multiple R-Squared: 0.1892,  Adjusted R-squared: 0.179 
## Wald test: 19.12 on 8 and 640 DF,  p-value: < 2.2e-16
```

```r
# IV hypothesis #2: higher as IV for G3
mod.instr2 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    |age+sex+studytime+absences+freetime+famrel+Pstatus+higher,
                    data=student2)
summary(mod.instr2)
```

```
## 
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime + 
##     famrel + Pstatus + G3 | age + sex + studytime + absences + 
##     freetime + famrel + Pstatus + higher, data = student2)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.7542 -1.2895 -0.3169  1.0388  6.9164 
## 
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31348    1.66903   0.787 0.431591
## age          0.16566    0.06274   2.640 0.008489 **
## sex          1.22942    0.15446   7.959 7.89e-15 ***
## studytime   -0.24191    0.11407  -2.121 0.034333 *
## absences     0.06030    0.01582   3.811 0.000152 ***
## freetime     0.16169    0.07383   2.190 0.028886 *
## famrel      -0.24373    0.07810  -3.121 0.001885 **
## Pstatus      0.38946    0.21601   1.803 0.071857 .
## G3          -0.03861    0.08538  -0.452 0.651300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.791 on 640 degrees of freedom
## Multiple R-Squared: 0.2024,  Adjusted R-squared: 0.1924
## Wald test: 19.46 on 8 and 640 DF,  p-value: < 2.2e-16
```

```
# IV hypothesis #3: using failures and higher as IVs for G3
mod.instr3 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3 | age+sex+studytime+abser
summary(mod.instr3)
```

```
##
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3 | age + sex + studytime + absences +
##     freetime + famrel + Pstatus + failures + higher, data = student2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9122 -1.2468 -0.3251  1.0248  6.7582
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.806682   1.380855   0.584 0.559298
## age          0.174639   0.060730   2.876 0.004166 **
## sex          1.246665   0.151742   8.216 1.17e-15 ***
## studytime   -0.270890   0.100862  -2.686 0.007425 **
## absences     0.061320   0.015776   3.887 0.000112 ***
## freetime     0.172840   0.071146   2.429 0.015399 *
## famrel      -0.252704   0.076623  -3.298 0.001028 **
## Pstatus      0.389279   0.216967   1.794 0.073256 .
## G3          -0.004886   0.057915  -0.084 0.932786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.798 on 640 degrees of freedom
## Multiple R-Squared: 0.1952,  Adjusted R-squared: 0.1852
## Wald test: 19.26 on 8 and 640 DF,  p-value: < 2.2e-16
```

```
summary(mod3)
```

```
##
## Call:
```

```
## lm(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3, data = student, subset = -c(109, 151,
##     198, 213, 611, 55, 67, 62, 251, 280, 524))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.589 -1.221 -0.320  1.077  5.583
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.67053    1.09977   1.519 0.129270
## age          0.15955    0.05754   2.773 0.005723 **
## sex          1.14700    0.14383   7.975 7.23e-15 ***
## studytime   -0.26830    0.08785  -3.054 0.002353 **
## absences     0.06286    0.01622   3.875 0.000118 ***
## freetime     0.11049    0.06596   1.675 0.094406 .
## famrel      -0.21746    0.07356  -2.956 0.003232 **
## Pstatus      0.52801    0.21112   2.501 0.012638 *
## G3          -0.06402    0.02239  -2.860 0.004377 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.708 on 629 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.2041
## F-statistic: 21.42 on 8 and 629 DF,  p-value: < 2.2e-16
```

Though there is some correlation between failure and higher with G3, when using it as an instrumental variable, it decrease the statistical significance of G3. Out of the three models, we found that using higher gave the highest R squared out of the the three options. Howvever the R squared is lower than the mod3 model, which we had deemed best from the various tests above. This is not surprising as the correlation plot revealed little correlation between the variables as a whole.

Here we tested to see if there were any other variables were instrumental variables. We made our hypothesis based on the correlation chart (pictured above) using instrumental variables highly correlated with one of the explanatory variables but not highly correlated with the response variable. Though the models were not as good as the models above (and to the OLS model), we decided to include the work to show our thought process.

```r
# IV hypothesis #4: goout as IV for freetime
# Really good for increasing significance of freetime, but bad for everything else / R-squared
mod.instr4 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    |age+sex+studytime+absences+goout+famrel+Pstatus+G3, data=student2)
summary(mod.instr4)
```

```
##
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3 | age + sex + studytime + absences +
##     goout + famrel + Pstatus + G3, data = student2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.04407 -1.58446 -0.08958  1.63379  9.30876
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.544541   1.783928  -1.987 0.047356 *
## age          0.167947   0.082557   2.034 0.042331 *
## sex          0.789680   0.219369   3.600 0.000343 ***
## studytime   -0.178830   0.125851  -1.421 0.155813
## absences     0.066903   0.021967   3.046 0.002418 **
## freetime     1.865462   0.289011   6.455 2.14e-10 ***
## famrel      -0.469380   0.111590  -4.206 2.97e-05 ***
## Pstatus      0.273712   0.303811   0.901 0.367965
## G3          -0.003004   0.033718  -0.089 0.929026
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.514 on 640 degrees of freedom
## Multiple R-Squared: -0.5721, Adjusted R-squared: -0.5917
## Wald test: 15.28 on 8 and 640 DF,  p-value: < 2.2e-16
```

```r
# IV hypothesis #5: higher as IV for age
# Really good for increasing significance of freetime, but bad for everything else / R-squared
mod.instr5 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    |higher+ failures +sex+freetime+absences+studytime+famrel+
                      Pstatus+G3, data=student2)
summary(mod.instr5)
```

```
##
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3 | higher + failures + sex + freetime +
##     absences + studytime + famrel + Pstatus + G3, data = student2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6928 -1.2566 -0.3667  1.0444  6.9635
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.77426    2.89329   1.650  0.09941 .
## age         -0.02056    0.16726  -0.123  0.90219
## sex          1.19047    0.15110   7.879 1.43e-14 ***
## studytime   -0.21192    0.09024  -2.348  0.01916 *
## absences     0.06633    0.01683   3.942 8.98e-05 ***
## freetime     0.15118    0.06906   2.189  0.02895 *
## famrel      -0.23543    0.07543  -3.121  0.00188 **
## Pstatus      0.39978    0.21750   1.838  0.06652 .
## G3          -0.07346    0.02392  -3.071  0.00222 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.801 on 640 degrees of freedom
## Multiple R-Squared: 0.1926,  Adjusted R-squared: 0.1825
## Wald test: 19.34 on 8 and 640 DF,  p-value: < 2.2e-16
```

```r
# IV hypothesis #6: Medu and Fedu as IV for G3
# really good for increasing significance of freetime, but bad for everything else / R-squared
mod.instr6 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    |age +sex+studytime+absences+freetime+famrel+Pstatus+Medu +
                      Fedu, data=student2)
summary(mod.instr6)
```

```
##
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3 | age + sex + studytime + absences +
##     freetime + famrel + Pstatus + Medu + Fedu, data = student2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8572 -1.2719 -0.3313  1.0101  6.8133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98310    1.77735   0.553 0.580369
## age          0.17151    0.06379   2.689 0.007363 **
## sex          1.24066    0.15619   7.943 8.88e-15 ***
## studytime   -0.26080    0.11939  -2.185 0.029285 *
## absences     0.06097    0.01591   3.833 0.000139 ***
## freetime     0.16896    0.07518   2.247 0.024961 *
## famrel      -0.24958    0.07901  -3.159 0.001659 **
## Pstatus      0.38934    0.21655   1.798 0.072660 .
## G3          -0.01663    0.09443  -0.176 0.860308
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.795 on 640 degrees of freedom
## Multiple R-Squared: 0.1983,  Adjusted R-squared: 0.1883
## Wald test: 19.34 on 8 and 640 DF,  p-value: < 2.2e-16
```

```r
# IV hypothesis #7: romantic as IV for age
# really good for increasing significance of freetime, but bad for everything else / R-squared
mod.instr7 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    | romantic +sex+studytime+absences+freetime+famrel+Pstatus+G3,
                    data=student2)
summary(mod.instr7)
```

```
##
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3 | romantic + sex + studytime + absences +
##     freetime + famrel + Pstatus + G3, data = student2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8145 -1.2771 -0.3222  1.0038  7.0964
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.13339    6.40205  -0.021  0.98338
## age          0.26763    0.37478   0.714  0.47542
## sex          1.23049    0.15748   7.814 2.29e-14 ***
## studytime   -0.22201    0.09061  -2.450  0.01454 *
## absences     0.05527    0.02114   2.615  0.00914 **
## freetime     0.15339    0.06880   2.230  0.02613 *
## famrel      -0.23697    0.07512  -3.155  0.00168 **
## Pstatus      0.38340    0.21736   1.764  0.07823 .
## G3          -0.06179    0.02742  -2.253  0.02458 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.793 on 640 degrees of freedom
## Multiple R-Squared: 0.1999,  Adjusted R-squared: 0.1899
## Wald test: 19.57 on 8 and 640 DF,  p-value: < 2.2e-16
```

```
# IV hypothesis #8:  higher and romantic as IV for age
# really good for increasing significance of freetime, but bad for everything else / R-squared
mod.instr8 <- ivreg(Talc~age+sex+studytime+absences+freetime+famrel+Pstatus+G3
                    |higher+ romantic+sex+studytime+absences+freetime+famrel+
                      Pstatus+G3, data=student2)
summary(mod.instr8)
```

```
##
## Call:
## ivreg(formula = Talc ~ age + sex + studytime + absences + freetime +
##     famrel + Pstatus + G3 | higher + romantic + sex + studytime +
##     absences + freetime + famrel + Pstatus + G3, data = student2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.689 -1.256 -0.341  1.077  7.034
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.16706    3.62488   0.598 0.550166
## age          0.13254    0.21079   0.629 0.529723
## sex          1.21173    0.15112   8.018 5.11e-15 ***
## studytime   -0.21728    0.08973  -2.422 0.015728 *
## absences     0.06045    0.01744   3.467 0.000562 ***
## freetime     0.15236    0.06858   2.221 0.026671 *
## famrel      -0.23625    0.07491  -3.154 0.001687 **
## Pstatus      0.39108    0.21610   1.810 0.070811 .
## G3          -0.06726    0.02433  -2.765 0.005858 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.789 on 640 degrees of freedom
## Multiple R-Squared: 0.2039,  Adjusted R-squared: 0.194
## Wald test: 19.66 on 8 and 640 DF,  p-value: < 2.2e-16
```

Given that our previous models still had low predictive ability, we tried a different approach of starting with a model that used all the numerical variable from the data set as explanatory variable. From the new model (mod5), we gradually removed the insignificant variables. However, it was not as good as our best model (mod3), where the R squared was 0.2041 while mod11 was 0.157 and contained many coefficients has high P-values. However, we wanted to include our work to show our attempt at finding a better model.

```
#new model
mod.5 <- lm(Dalc~age + sex + Pstatus + Medu + Fedu + traveltime + studytime +
              failures + schoolsup + famsup + paid + activities + nursery + higher
            + internet + romantic + famrel + freetime + goout + Walc + Talc +
              health + absences + G1 + G2 + G3, data=student)
summary(mod.5)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + Medu + Fedu + traveltime +
##     studytime + failures + schoolsup + famsup + paid + activities +
##     nursery + higher + internet + romantic + famrel + freetime +
##     goout + Walc + Talc + health + absences + G1 + G2 + G3, data = student)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.681e-14 -5.950e-16  8.900e-17  6.650e-16  5.089e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.636e-14  2.082e-15  7.857e+00 1.73e-14 ***
## age         -6.331e-16  1.068e-16 -5.925e+00 5.16e-09 ***
## sex         -1.204e-15  2.657e-16 -4.530e+00 7.08e-06 ***
## Pstatus     -2.040e-15  3.535e-16 -5.772e+00 1.24e-08 ***
## Medu         3.430e-17  1.405e-16  2.440e-01  0.80726
## Fedu         2.084e-16  1.378e-16  1.512e+00  0.13103
## traveltime   3.464e-16  1.613e-16  2.148e+00  0.03213 *
## studytime    1.340e-17  1.494e-16  9.000e-02  0.92858
## failures    -2.539e-16  2.249e-16 -1.129e+00  0.25938
## schoolsup    1.051e-15  3.886e-16  2.705e+00  0.00702 **
## famsup      -5.507e-16  2.439e-16 -2.258e+00  0.02431 *
## paid         4.398e-16  4.910e-16  8.960e-01  0.37072
## activities  -1.671e-16  2.367e-16 -7.060e-01  0.48055
## nursery      9.040e-18  2.897e-16  3.100e-02  0.97512
## higher       1.476e-17  4.173e-16  3.500e-02  0.97180
## internet    -6.619e-16  2.878e-16 -2.300e+00  0.02178 *
## romantic    -2.235e-16  2.457e-16 -9.100e-01  0.36328
## famrel       7.735e-17  1.245e-16  6.210e-01  0.53458
## freetime     7.664e-16  1.200e-16  6.385e+00 3.35e-10 ***
## goout       -7.099e-16  1.141e-16 -6.219e+00 9.18e-10 ***
## Walc        -1.000e+00  2.505e-16 -3.992e+15  < 2e-16 ***
## Talc         1.000e+00  1.614e-16  6.197e+15  < 2e-16 ***
## health      -2.792e-17  8.130e-17 -3.430e-01  0.73143
## absences    -8.362e-19  2.599e-17 -3.200e-02  0.97434
## G1          -5.069e-16  8.643e-17 -5.865e+00 7.30e-09 ***
## G2           3.195e-16  1.138e-16  2.808e+00  0.00514 **
## G3           7.148e-17  9.201e-17  7.770e-01  0.43753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.883e-15 on 622 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.565e+30 on 26 and 622 DF,  p-value: < 2.2e-16
```

```
# remove Medu and Fedu - very insignificant
mod.6 <- lm(Dalc~age + sex + Pstatus + traveltime + studytime + failures +
              schoolsup + famsup + paid + activities + nursery + higher +
              internet + romantic + famrel + freetime + goout + Walc + Talc +
              health + absences + G1 + G2 + G3, data=student)
summary(mod.6)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Walc + Talc + health + absences + G1 + G2 + G3, data = student)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.459e-14 -6.340e-16  2.900e-17  6.320e-16  5.015e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.861e-14  1.944e-15  9.575e+00  < 2e-16 ***
## age         -8.403e-16  1.002e-16 -8.385e+00 3.38e-16 ***
## sex         -1.395e-15  2.463e-16 -5.665e+00 2.25e-08 ***
## Pstatus     -2.018e-15  3.306e-16 -6.103e+00 1.83e-09 ***
## traveltime  -6.354e-16  1.481e-16 -4.290e+00 2.07e-05 ***
## studytime   -3.303e-17  1.402e-16 -2.360e-01 0.813782
## failures     2.435e-17  2.108e-16  1.160e-01 0.908043
## schoolsup    1.007e-15  3.642e-16  2.764e+00 0.005871 **
## famsup      -1.674e-16  2.274e-16 -7.360e-01 0.461905
## paid        -2.450e-16  4.588e-16 -5.340e-01 0.593498
## activities  -2.318e-16  2.217e-16 -1.046e+00 0.296138
## nursery      4.483e-16  2.699e-16  1.661e+00 0.097199 .
## higher       6.353e-16  3.892e-16  1.632e+00 0.103135
## internet    -5.090e-16  2.653e-16 -1.918e+00 0.055511 .
## romantic    -5.015e-17  2.304e-16 -2.180e-01 0.827799
## famrel      -2.354e-16  1.168e-16 -2.016e+00 0.044202 *
## freetime     3.373e-16  1.126e-16  2.996e+00 0.002840 **
## goout        2.580e-16  1.071e-16  2.409e+00 0.016267 *
## Walc        -1.000e+00  2.345e-16 -4.265e+15  < 2e-16 ***
## Talc         1.000e+00  1.513e-16  6.610e+15  < 2e-16 ***
## health       9.396e-18  7.623e-17  1.230e-01 0.901942
## absences    -1.411e-17  2.435e-17 -5.790e-01 0.562580
## G1          -5.444e-16  8.106e-17 -6.716e+00 4.22e-11 ***
## G2           3.528e-16  1.066e-16  3.308e+00 0.000992 ***
## G3           7.458e-17  8.634e-17  8.640e-01 0.388030
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.706e-15 on 624 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.154e+30 on 24 and 624 DF,  p-value: < 2.2e-16
```

```
# remove those with P-value above 0.8
mod.7 <- lm(Dalc~age + sex + Pstatus + traveltime + schoolsup + famsup + paid +
              activities + nursery + higher + internet + famrel + freetime +
              goout + Walc + Talc + absences + G1 + G2 + G3, data=student)
summary(mod.7)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + traveltime + schoolsup +
##     famsup + paid + activities + nursery + higher + internet +
##     famrel + freetime + goout + Walc + Talc + absences + G1 +
##     G2 + G3, data = student)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.421e-14 -6.690e-16 -7.500e-17  4.790e-16  5.049e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.668e-14  2.049e-15  8.138e+00 2.17e-15 ***
## age         -7.405e-16  1.013e-16 -7.314e+00 7.95e-13 ***
## sex         -4.706e-16  2.575e-16 -1.828e+00  0.06806 .
## Pstatus     -1.968e-15  3.519e-16 -5.592e+00 3.35e-08 ***
## traveltime  -4.335e-16  1.575e-16 -2.753e+00  0.00607 **
## schoolsup    1.155e-15  3.850e-16  3.001e+00  0.00280 **
## famsup       5.983e-16  2.401e-16  2.492e+00  0.01295 *
## paid        -3.455e-17  4.875e-16 -7.100e-02  0.94353
## activities  -2.174e-16  2.344e-16 -9.280e-01  0.35401
## nursery      7.403e-16  2.871e-16  2.579e+00  0.01015 *
## higher      -1.769e-16  4.091e-16 -4.320e-01  0.66561
## internet    -1.752e-16  2.818e-16 -6.220e-01  0.53428
## famrel      -3.588e-17  1.234e-16 -2.910e-01  0.77134
## freetime    -1.084e-16  1.192e-16 -9.100e-01  0.36327
## goout        9.001e-16  1.135e-16  7.931e+00 9.99e-15 ***
## Walc        -1.000e+00  2.477e-16 -4.037e+15  < 2e-16 ***
## Talc         1.000e+00  1.605e-16  6.231e+15  < 2e-16 ***
## absences     1.832e-17  2.581e-17  7.100e-01  0.47815
## G1          -4.854e-16  8.583e-17 -5.655e+00 2.36e-08 ***
## G2           3.743e-16  1.135e-16  3.299e+00  0.00102 **
## G3           3.962e-17  9.149e-17  4.330e-01  0.66513
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.885e-15 on 628 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.331e+30 on 20 and 628 DF,  p-value: < 2.2e-16
```

```
# remove G3
mod.8 <- lm(Dalc~age + sex + Pstatus + traveltime + schoolsup + famsup + paid +
              activities + nursery + higher + internet + famrel + freetime +
              goout + Walc + Talc + absences + G1 + G2, data=student)
summary(mod.8)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + traveltime + schoolsup +
##     famsup + paid + activities + nursery + higher + internet +
##     famrel + freetime + goout + Walc + Talc + absences + G1 +
##     G2, data = student)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.422e-14 -6.480e-16 -7.900e-17  4.630e-16  5.054e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.692e-14  2.048e-15  8.264e+00 8.38e-16 ***
## age         -7.551e-16  1.012e-16 -7.463e+00 2.83e-13 ***
## sex         -5.988e-16  2.570e-16 -2.330e+00  0.02013 *
## Pstatus     -2.042e-15  3.515e-16 -5.810e+00 9.95e-09 ***
## traveltime  -4.446e-16  1.571e-16 -2.831e+00  0.00479 **
## schoolsup    1.169e-15  3.844e-16  3.042e+00  0.00245 **
## famsup       6.268e-16  2.396e-16  2.616e+00  0.00910 **
## paid        -5.889e-17  4.867e-16 -1.210e-01  0.90372
## activities  -1.965e-16  2.342e-16 -8.390e-01  0.40176
## nursery      7.903e-16  2.868e-16  2.755e+00  0.00604 **
## higher      -1.897e-16  4.082e-16 -4.650e-01  0.64229
## internet    -1.706e-16  2.812e-16 -6.070e-01  0.54422
## famrel      -8.906e-18  1.233e-16 -7.200e-02  0.94243
## freetime    -7.817e-17  1.190e-16 -6.570e-01  0.51150
## goout        7.715e-16  1.134e-16  6.803e+00 2.40e-11 ***
## Walc        -1.000e+00  2.475e-16 -4.041e+15  < 2e-16 ***
## Talc         1.000e+00  1.602e-16  6.241e+15  < 2e-16 ***
## absences     1.908e-17  2.574e-17  7.410e-01  0.45877
## G1          -4.800e-16  8.485e-17 -5.656e+00 2.35e-08 ***
## G2           4.097e-16  7.884e-17  5.196e+00 2.75e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.883e-15 on 629 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.51e+30 on 19 and 629 DF,  p-value: < 2.2e-16
```

```r
# remove those with P-value above 0.8 (famrel and paid)
mod.9 <- lm(Dalc~age + sex + Pstatus + traveltime + schoolsup + famsup +
              activities + nursery + higher + internet + freetime + goout +
              Walc + Talc + absences + G1 + G2, data=student)
summary(mod.9)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + traveltime + schoolsup +
##     famsup + activities + nursery + higher + internet + freetime +
##     goout + Walc + Talc + absences + G1 + G2, data = student)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.061e-14 -7.260e-16 -1.390e-16  4.600e-16  5.887e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  2.025e-14  2.421e-15  8.367e+00 3.81e-16 ***
## age         -6.743e-16  1.217e-16 -5.540e+00 4.44e-08 ***
## sex          7.193e-16  3.063e-16  2.349e+00 0.019154 *
## Pstatus     -1.660e-15  4.226e-16 -3.929e+00 9.46e-05 ***
## traveltime  -4.594e-16  1.887e-16 -2.435e+00 0.015187 *
## schoolsup    6.084e-16  4.622e-16  1.316e+00 0.188578
## famsup       5.227e-16  2.866e-16  1.824e+00 0.068665 .
## activities  -3.635e-16  2.811e-16 -1.293e+00 0.196434
## nursery     -1.116e-16  3.449e-16 -3.240e-01 0.746245
## higher      -1.748e-16  4.907e-16 -3.560e-01 0.721723
## internet    -5.082e-16  3.377e-16 -1.505e+00 0.132848
## freetime    -5.139e-16  1.422e-16 -3.615e+00 0.000325 ***
## goout        9.635e-16  1.357e-16  7.102e+00 3.32e-12 ***
## Walc        -1.000e+00  2.975e-16 -3.362e+15  < 2e-16 ***
## Talc         1.000e+00  1.926e-16  5.193e+15  < 2e-16 ***
## absences    -1.401e-17  3.085e-17 -4.540e-01 0.649961
## G1          -5.137e-16  1.016e-16 -5.057e+00 5.58e-07 ***
## G2           4.223e-16  9.440e-17  4.474e+00 9.09e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.468e-15 on 631 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.711e+30 on 17 and 631 DF,  p-value: < 2.2e-16
```

```
# remove those with P-value above 0.6
mod.10 <- lm(Dalc~age + sex + Pstatus + traveltime + schoolsup + famsup +
               activities + internet + freetime + goout + Walc + Talc
             + G1 + G2, data=student)
summary(mod.10)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + traveltime + schoolsup +
##     famsup + activities + internet + freetime + goout + Walc +
##     Talc + G1 + G2, data = student)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.645e-14 -6.180e-16  3.200e-17  6.390e-16  5.181e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.954e-14  1.941e-15  1.007e+01  < 2e-16 ***
## age         -7.819e-16  9.978e-17 -7.836e+00 1.97e-14 ***
## sex         -1.262e-15  2.573e-16 -4.904e+00 1.20e-06 ***
## Pstatus     -2.007e-15  3.519e-16 -5.703e+00 1.80e-08 ***
## traveltime  -4.990e-16  1.586e-16 -3.147e+00  0.00173 **
## schoolsup    6.930e-16  3.870e-16  1.791e+00  0.07382 .
## famsup       6.114e-16  2.401e-16  2.547e+00  0.01111 *
## activities  -2.301e-16  2.359e-16 -9.750e-01  0.32970
## internet    -1.593e-16  2.829e-16 -5.630e-01  0.57350
## freetime     3.397e-16  1.189e-16  2.857e+00  0.00442 **
## goout       -7.705e-16  1.138e-16 -6.769e+00 2.96e-11 ***
## Walc        -1.000e+00  2.497e-16 -4.004e+15  < 2e-16 ***
## Talc         1.000e+00  1.611e-16  6.206e+15  < 2e-16 ***
## G1          -5.188e-16  8.483e-17 -6.116e+00 1.68e-09 ***
## G2           4.349e-16  7.908e-17  5.499e+00 5.54e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.915e-15 on 634 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.66e+30 on 14 and 634 DF,  p-value: < 2.2e-16
```

```r
# remove Walc/Talc, activities, internet
mod.11 <- lm(Dalc~age + sex + Pstatus + traveltime + schoolsup + famsup +
               freetime + goout + G1 + G2, data=student)
summary(mod.11)
```

```
##
## Call:
## lm(formula = Dalc ~ age + sex + Pstatus + traveltime + schoolsup +
##     famsup + freetime + goout + G1 + G2, data = student)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4216 -0.5183 -0.2158  0.1916  3.6921
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.24034    0.56393  -0.426  0.67012
## age          0.08290    0.02883   2.876  0.00417 **
## sex          0.50106    0.07026   7.131 2.71e-12 ***
## Pstatus      0.05424    0.10181   0.533  0.59439
## traveltime   0.05740    0.04534   1.266  0.20600
## schoolsup    0.06973    0.11238   0.621  0.53514
## famsup       0.05760    0.06961   0.827  0.40833
## freetime    -0.01367    0.03431  -0.398  0.69047
## goout        0.16495    0.03058   5.395 9.67e-08 ***
## G1          -0.02163    0.02467  -0.877  0.38093
## G2          -0.02302    0.02297  -1.002  0.31663
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8491 on 638 degrees of freedom
## Multiple R-squared:   0.17,  Adjusted R-squared:  0.157
## F-statistic: 13.07 on 10 and 638 DF,  p-value: < 2.2e-16
```

4. Conclusions

In this project, we wanted to predicted the probability that the sex of a student is male and the total weekly alcohol consumption. To test the probability that the sex of a student is male, we created probit and logit models. Our conclusion when comparing these models was that there was no one best model. Rather it depends on the preference of the study in regard to the predictions. If the researchers would prefer higher specificity meaning the model is better at predicting true positives (fewer false positives), then the logit model is better for the higher threshold (0.6) and baseline threshold (0.5) than the probit model. However, when looking at the lower threshold (0.4), the probit and OLS models have a higher specificity. If researchers are looking for higher sensitivity, at the regular threshold, the OLS model is better than the logit and probit model. For the higher threshold, the OLS and probit model are the same, and thus both are better than the logit model. For the lower threshold, the logit model is better for sensitivity than the other two models.

To predict total weekly alcohol consumption, the best model is the OLS model (mod3) as it has the highest R squared and most of its coefficients are statistically significant compared to the instrumental models that were predicted. It also had the lowest value when comparing all the models with both the AIC and BIC test. When predicting if failures and higher could be a instrumental variable for G3, we found that though the R square was similar to the mod3 model, the variable G3 was not significant, making the model worse than the original OLS model. This was the case for the other instrumental variable hypotheses.

5. Future Work

The data set and model that we selected did not present much success with instrumental variables. For future work, we can attempt all possible combinations of instrumental variables. Although this would take significant computational power, this could give better results. In addition, we can experiment with interaction terms and higher order terms for the OLS model. To accurately do this would require more advanced data science models, e.g. polynomial regressions, linear/cubic/smoothing splines, and general additive models. Given more time, we can look at the non-numerical variables by converting them and seeing if they have any effect on our models. We could also increase the number of observations by conducting more surveys.

For the future, we can attempt to find a data set is compatible with instrumental variables, though for this project we wanted to focus on probit/logit and instrumental variable models. This made it difficult to source a data set that was perfectly compatible with both models.

6. References

As mentioned, our data set came from https://www.kaggle.com/uciml/student-alcohol-consumption. We referenced the website http://www.alexdeforge.com/phil-155-reasons-and-arguments to help with interpreting specificity and sensitivity.