# ANALYSIS OF MATHEMATICS PERFORMANCE OF STUDENTS

Pstat 126: Regression Analysis
Professor Sudeep Bapat

ANA CAKLOVIC & SUKANYA JOSHI

Section Timings: ANA: Wednesday @ 5 PM, SUKANYA: Tuesday @ 6 PM

Teaching Assistant: Jiaye Xu

Due date: December 11th, 2018

**Introduction**:

Our data interprets two schools in Portugal and looks at numerous variables which influence student performance in math. We will be examining how absences (a numerical predictor) affect a student's performance in math. We will also be testing the significance of categorical predictors, such as failures and traveltime, to see if they influence the final grades of a student. Final grades are measured from 0 to 20, absences are measured numerically from 0 to 93, and traveltime is measured from 1 to 4 hours, and failures are past class failures numerically measured from 1 to 3. Regression can be used to help us analyze data because we want to see if the number of absences, failures, and travel time can affect a student's final grade.

Questions of Interest:

1) *Do the number of absences affect the final grade of a student?*
2) *Which other aspects in a student's life have the greatest effect on the final grade?*
3) *Based on question #2, do any of those aspects have an effect on each other or with the number of absences a student has?*

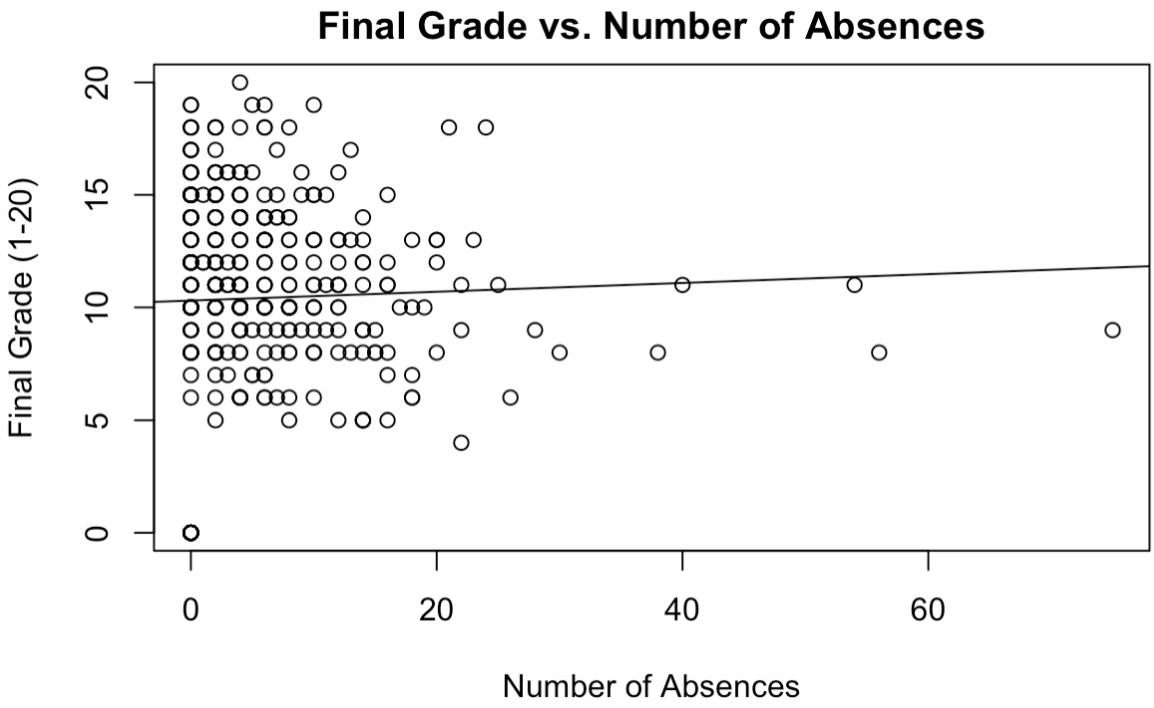
**Regression Method:**

In order to analyze our data and answer our questions of data, we will first plot the final math grade against the number of absences a student has. We will use an information criterion hypothesis tests, and test for interactions in order to find the most significant predictors of a student's final grade, and we will then use this to build the best final model.

**Regression Analysis, Results and Interpretation**:

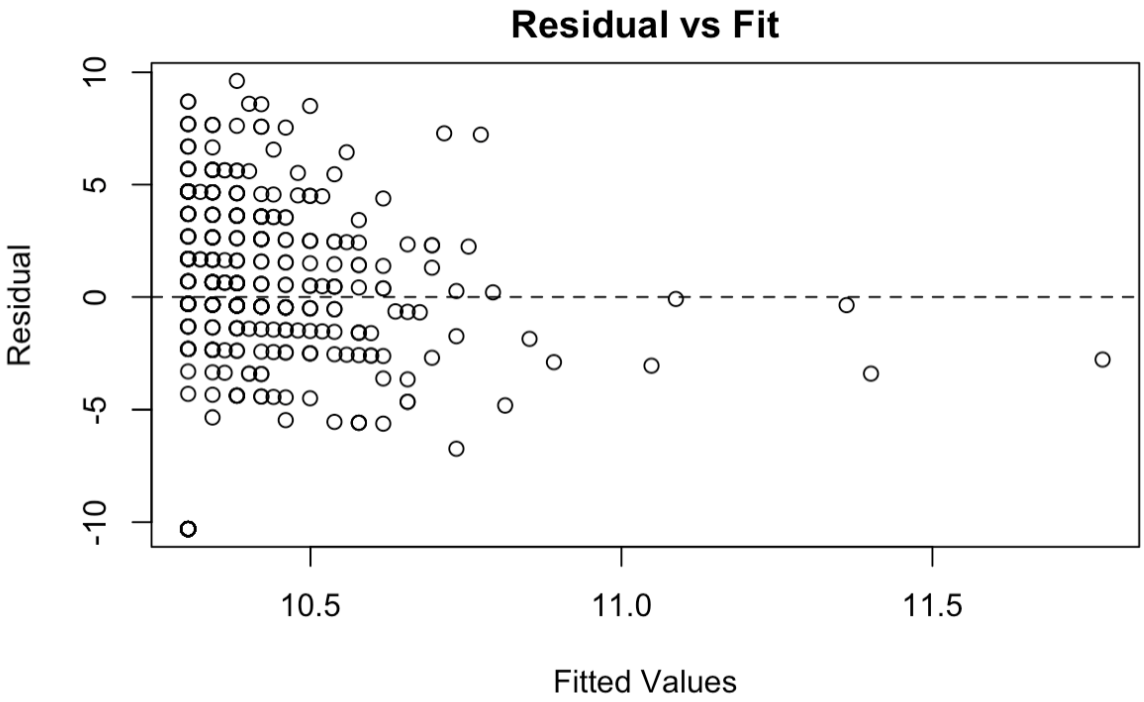
***Question 1) Do the number of absences affect the final grade of a student?***

An absence is when a student misses school and it ranges from 0 to 93 per semester. The scatterplot below shows the final math grade on absences the number of absences a student has in a school year. The goal of the scatterplot is to check the linearity of the data before creating a model where we will regress the students' final grades on their absences.



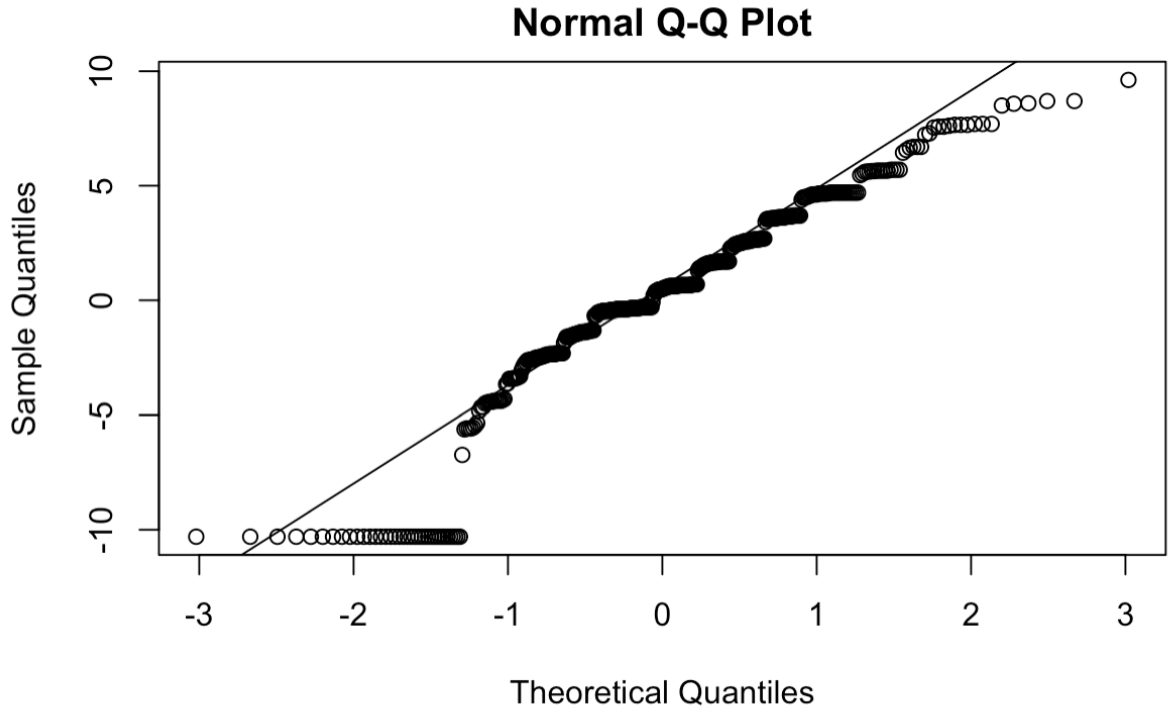

The scatterplot above has a slight linearity issue due to the cluster of observations to the left. The line of best fit shows that there is a slight positive correlation between the two variables, however due to the linearity issue the relationship between absences and final grade will be hard to determine until the linearity issue is fixed.

Below is a residual vs fit plot:



We can see that there is a funneling effect of the residual vs fit plot (where the distribution is spread out for small X values and close to zero for large X values), which means that the variances of the error terms are not equal.

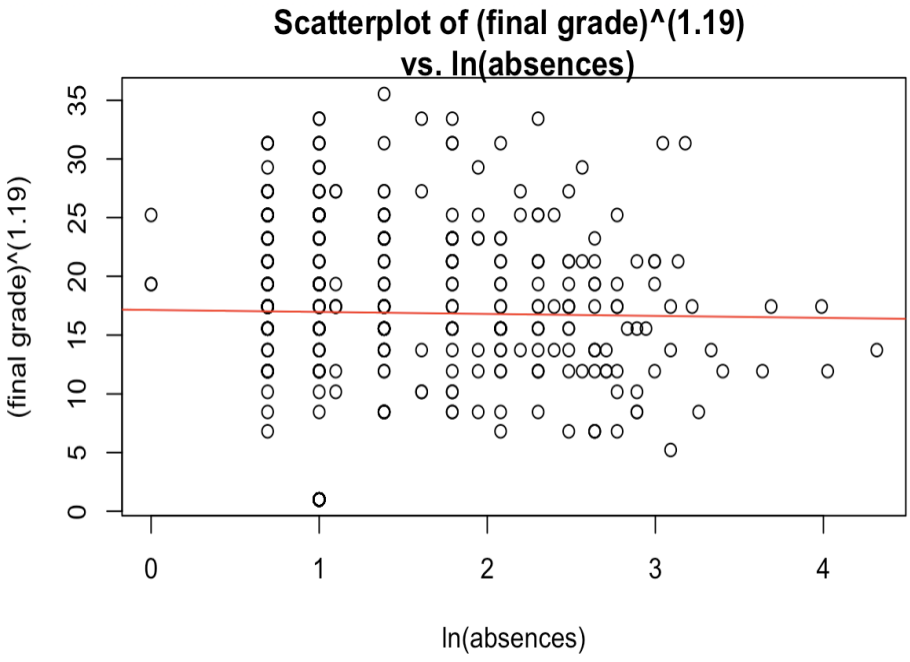
Below is a normal Q-Q plot:

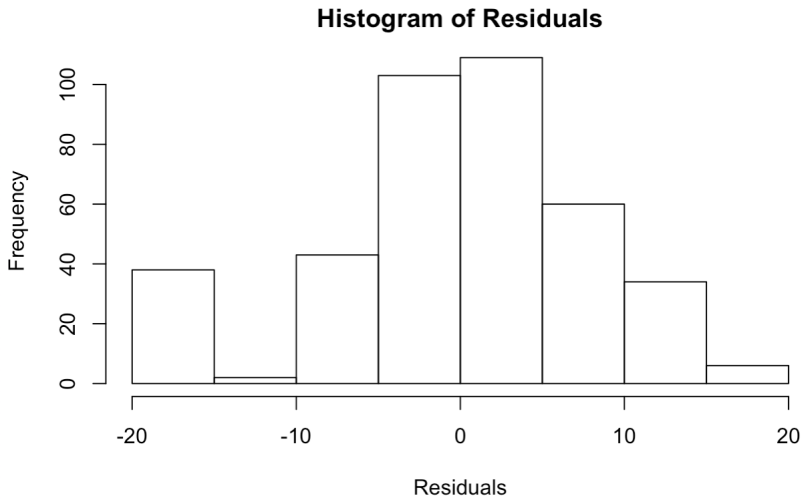



We see that the normal Q-Q plot does not follow the line of fit at the tails (it diverges by a large amount), which suggests that we have an issue with normality. We can also test the normality with the Shapiro-Wilk Normality Test., which simply allows us to confirm the visual issues we see on the normal Q-Q plot. We can see that that test returns a p-value of 4.283e-12. Since the value is less than 0.05, we can see reject the null hypothesis, H0: no normality issues, and accept the the alternative hypothesis (that there is an issue with normality).

When we plotted the residual vs fit plot and the normal Q-Q plot, we realized that regressing the response variable *final grade* on *absences* results in a model which does not fulfill linearity, normality, and equal error variances criteria. In order to resolve the linearity issue, we performed a log transformation on the predictor variable. Then, after running diagnostics, to resolve the remaining issues we used a box-cox transformation and found that the best regression is a regression of Y^1.19 on ln X.

Below is the data after a box-cox transformation:




From the box-cox transformation plotted above, we see that there is a linear relationship between the final grade and number of absences (specifically, a slight negative correlation). This suggests that the relationship of absences and final grades is that when the number of absences a student has increases, there is a slight decrease in their final grade.

The histogram of the residuals shows a fairly normal distribution, which means the normality issue has been improved. Thus, these transformations help satisfy the linearity, normality, and equal error variances criteria (LINE conditions).

Now that we have looked at how absences affect the final grade, let's see which other predictors influence final grades the most.

---

***Question 2) Which other aspects in a student's life have the greatest effect on the final grade?***

There are 30 predictor variables which affect testing score (check appendix for full list of variables). Some of these variables are related to the social, academic, and family background of the student. This includes family size, parent education level, study time, travel time, past failures, and extracurricular activities.

We grouped the variables into three different groups. The first group consists of all academic-related variables: freetime, traveltime, studytime, activities, and failures. The AIC test of the first group is shown below.

TEST 1:

```
Step:  AIC=1149.19
finalgrade ~ student$failures + student$traveltime

                     Df Sum of Sq    RSS    AIC
<none>                             7136.9 1149.2
+ student$freetime    1     14.94 7122.0 1150.4
- student$traveltime  1     58.71 7195.7 1150.4
+ student$studytime   1      6.77 7130.2 1150.8
+ student$activities  1      0.67 7136.3 1151.2
- student$failures    1   1019.48 8156.4 1199.9

Call:
lm(formula = finalgrade ~ student$failures + student$traveltime)

Coefficients:
       (Intercept)    student$failures  student$traveltime
           11.9460             -2.1723             -0.5558
```

We tested three models in total and found out that the above model has the most significant predictor variables. We first tested the model which regressed finalgrade on 1 and got an AIC value of 1203.39. Then, we tested the model which regressed finalgrade on failures and got an AIC value of 1150.43. Then, we finally tested the model above which regressed finalgrade on failures and traveltime, and got an AIC value of 1149.19 which is clearly less than the values of the other two models. Thus, we know that this model has predictors which are the most significant.

The second group consists of all family-related variables: family educational support (famsup), quality of family relationships (famrel), father's education, mother's education, father's job, and mother's job. The AIC test of the second group is shown below.

TEST 2:

```
Step:  AIC=1185.62
finalgrade ~ student$Medu + student$famsup

                 Df Sum of Sq    RSS    AIC
<none>                        7826.5 1185.6
- student$famsup  1     53.49 7880.0 1186.3
+ student$famrel  1     21.18 7805.3 1186.5
+ student$Fedu    1      7.14 7819.3 1187.3
+ student$Mjob    4     99.30 7727.2 1188.6
+ student$Fjob    4     35.97 7790.5 1191.8
- student$Medu    1    430.76 8257.2 1204.8

Call:
lm(formula = finalgrade ~ student$Medu + student$famsup)

Coefficients:
      (Intercept)       student$Medu  student$famsupyes
           8.2145             0.9717            -0.7685
```

---

We tested three models in total and found out that the above model has the most significant predictor variables. We first tested the model which regressed finalgrade on 1 and got an AIC value of 1203.39. Then, we tested the model which regressed finalgrade on mother's education (Medu) and got an AIC value of 1186.31. Then, we finally tested the model above which regressed finalgrade on mother's education and family educational support, and got an AIC value of 1185.62 which is clearly less than the values of the other two models. Thus, we know that this model has predictors which are the most significant.

The third group consists of all family-related variables: family educational support (famsup), quality of family relationships (famrel), father's education, mother's education, father's job, and mother's job. The AIC test of the third group is shown below.

TEST 3:

```
Step:  AIC=1193.62
finalgrade ~ student$goout + student$romantic

                     Df Sum of Sq    RSS    AIC
<none>                           7986.6 1193.6
+ student$freetime    1    20.248 7966.4 1194.6
+ student$activities  1     5.061 7981.6 1195.4
+ student$Dalc        1     2.736 7983.9 1195.5
+ student$Walc        1     0.046 7986.6 1195.6
- student$romantic    1   137.468 8124.1 1198.4
- student$goout       1   143.599 8130.2 1198.7

Call:
lm(formula = finalgrade ~ student$goout + student$romantic)

Coefficients:
        (Intercept)        student$goout  student$romanticyes
            12.5191              -0.5423              -1.2507
```

We tested three models in total and found out that the above model has the most significant predictor variables. We first tested the model which regressed finalgrade on 1 and got an AIC value of 1203.39. Then, we tested the model which regressed finalgrade on the number of times a student goes out with his friends (goout) and got an AIC value of 1198.36. Then, we finally

tested the model above which regressed finalgrade on goout and a student's romantic relationships (romantic) and got an AIC value of 1193.62, which is clearly less than the values of the other two models. Therefore, we know that this model has predictors which are the most significant.

Since Test #1 has an AIC value of 1149.19, which is less than the AIC values of the other two tests' final models (1185.62 and 1193.62), we know that the variables in test #1 are the most significant. Thus, the variables failures and traveltime are the two variables which are the most significant, according to the AIC tests.

This means that the aspects of a students life that affect their final grade the most are the amount of classes they have failed in their academic career, and the amount of time it takes the student to travel from home to school (measured in hour intervals where 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour).

Using the data above, we can now answer some more research questions regarding the relationship between traveltime, failures, and absences.

---

*Question 3) Based on question #2, do any of those aspects have an effect on each other or with the number of absences a student has?*

In order to answer the question of interest, we used an ANOVA table and compared reduced versus full models for traveltime and failures in relation to absences.

<u>Test 1: Testing for the significance of the interaction between absences and failures</u>
**H0:** accept the reduced model Yi = βo + βAbsences Xi + βFailures Xi + βTraveltime Xi + βAbsences,Traveltime Xi + εi
**Ha:** accept the full model Yi = βo + βAbsences Xi + βFailures Xi + βTraveltime Xi + βAbsences,Traveltime Xi + βAbsences,Failures Xi + εi

```
Analysis of Variance Table

Model 1: finalgrade ~ absences + student$failures + student$traveltime +
    absences * student$traveltime
Model 2: finalgrade ~ absences + student$failures + student$traveltime +
    absences * student$failures + absences * student$traveltime
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    390 7106.4
2    389 6927.7  1    178.71 10.035 0.001658 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table above, we can see that the failures has an interaction with the absences since its p value of 0.001658 is less than the alpha value of 0.05. Thus, we can reject the reduced model and accept the full model of the interaction between traveltime and absences.

<u>Test 2: Testing for the significance of the interaction between absences and traveltime</u>

**H0:** accept the reduced model Yi = βo + βAbsences Xi + βFailures Xi + βTraveltime Xi + βAbsences,Failures Xi + εi
**Ha:** accept the full model Yi = βo + βAbsences Xi + βFailures Xi + βTraveltime Xi + βAbsences,Traveltime Xi + βAbsences,Failures Xi + εi

```
Analysis of Variance Table

Model 1: finalgrade ~ absences + student$failures + student$traveltime +
    absences * student$failures
Model 2: finalgrade ~ absences + student$failures + student$traveltime +
    absences * student$failures + absences * student$traveltime
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    390 6934.1
2    389 6927.7  1     6.355 0.3568 0.5506
```

From the ANOVA table above, we can see that the traveltime variable does not have an interaction with the absences since its p value of 0.5506 is greater than the alpha value of 0.05. Thus, we can fail to reject the reduced model, which signifies that there is no interaction between the traveltime and absences.

Now that we have determined that there is an interaction between failures and absences and no interaction between travetime and absences, let's see if there is any interaction between traveltime and failures.

Test 3: Testing for the significance of the interaction between failures and traveltime
**H0:** accept the reduced model Yi = βo + βAbsences Xi + βFailures Xi + βTraveltime Xi + βAbsences,Failures Xi + εi
**Ha:** accept the full model Yi = βo + βAbsences Xi + βFailures Xi + βTraveltime Xi + βAbsences,Failures Xi + βFailures,Traveltime Xi + εi

```
Analysis of Variance Table

Model 1: finalgrade ~ absences + student$failures + student$traveltime +
    absences * student$failures
Model 2: finalgrade ~ absences + student$failures + student$traveltime +
    absences * student$failures + student$failures * student$traveltime
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    390 6934.1
2    389 6773.0  1     161.1 9.2527 0.002511 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table above, we can see there is an interaction between failures and traveltime, because when we compared a reduced model and a full model, the resulting p-value of the ANOVA test performed was 0.002511. This value is less than 0.05, so we can reject the reduced model and accept the full model, which means that the interaction term is significant and should be included in the model. Therefore, there is an interaction between failures and traveltime.

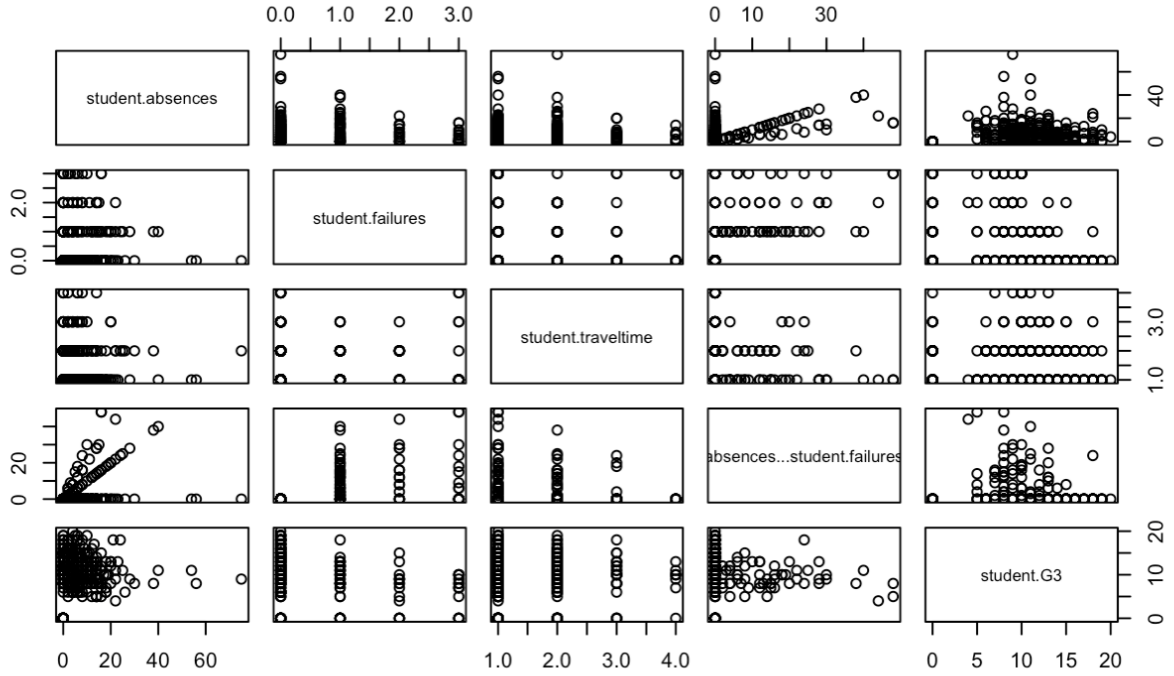
Overall, we can see that there is an interaction between the past failures of a student and his or her absences. Furthermore, we proved that there is no interaction between the travel time from a student's home to school and the number of absences that the student has. Additionally, there is also an interaction between the past failures and the travel time of a student.

The results of this research question lets us know that our final model should have the predictors absences, failures, traveltime, the interaction term between failures and traveltime, and the interaction term between absences and failures. The **final model** is: $Yi = \beta o + \beta Absences\ Xi + \beta Failures\ Xi + \beta Traveltime\ Xi + \beta Failures,Traveltime\ Xi + \beta Absences,Failures\ Xi + \varepsilon i.$

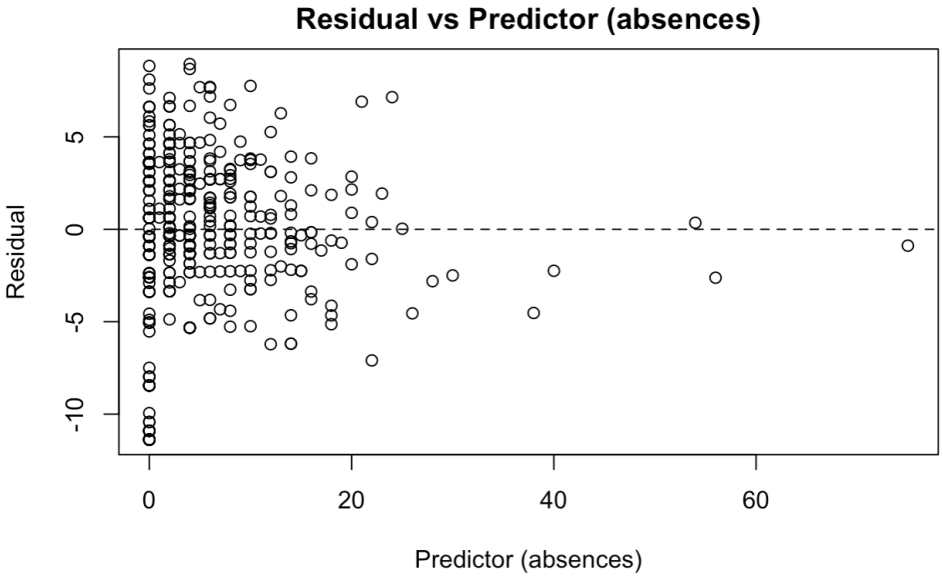
---

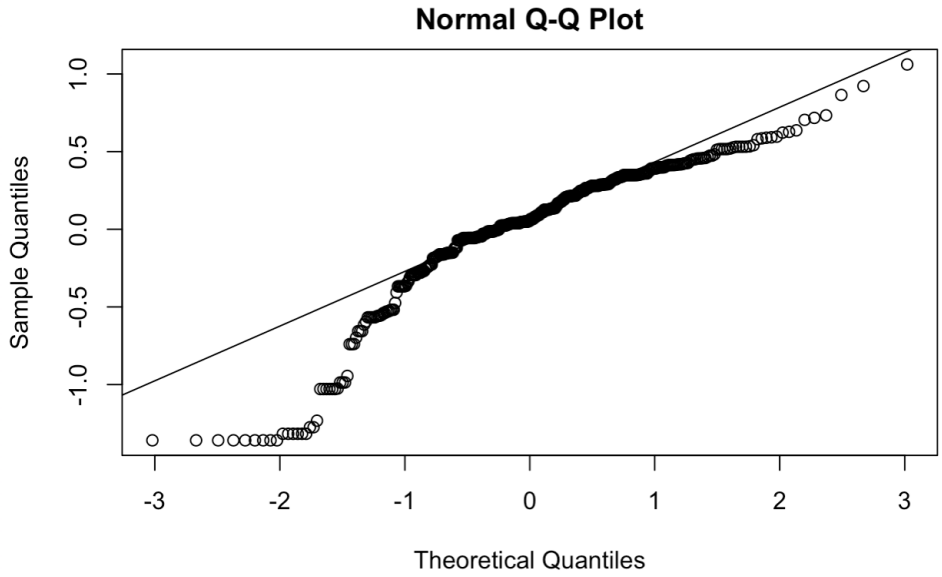
*- Diagnostics and Transformations of the Final Model:*

Here is the scatterplot matrix of our final model:

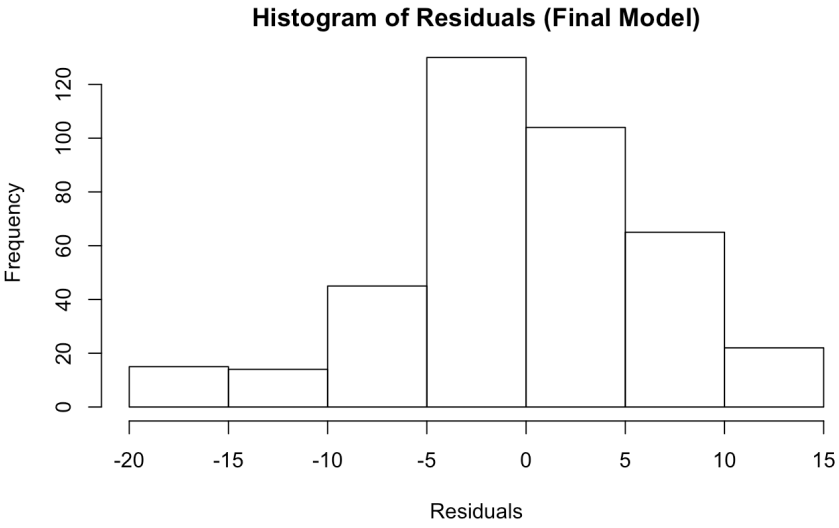


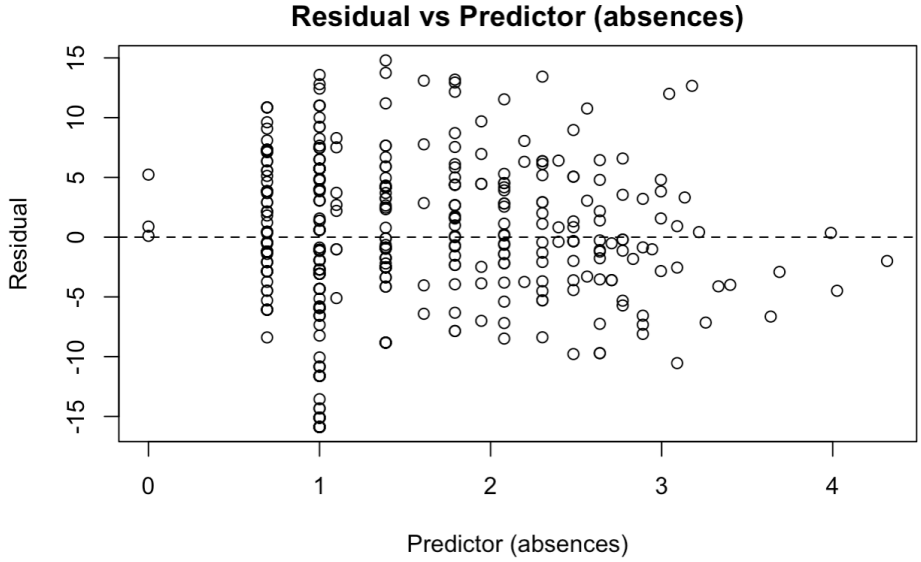

We can see that we might have a linearity issue between the predictor variable absences and the response variable final grade (noted as G3 on the scatterplot). Also, an examination of the residual vs predictor (specifically, the predictor absences) and the normal Q-Q plot shows us that the variances of the error terms are not equal (there is a clear funneling effect in the residual vs predictor plot) and that with have a normality issue (the tails of the curve diverge from the line).

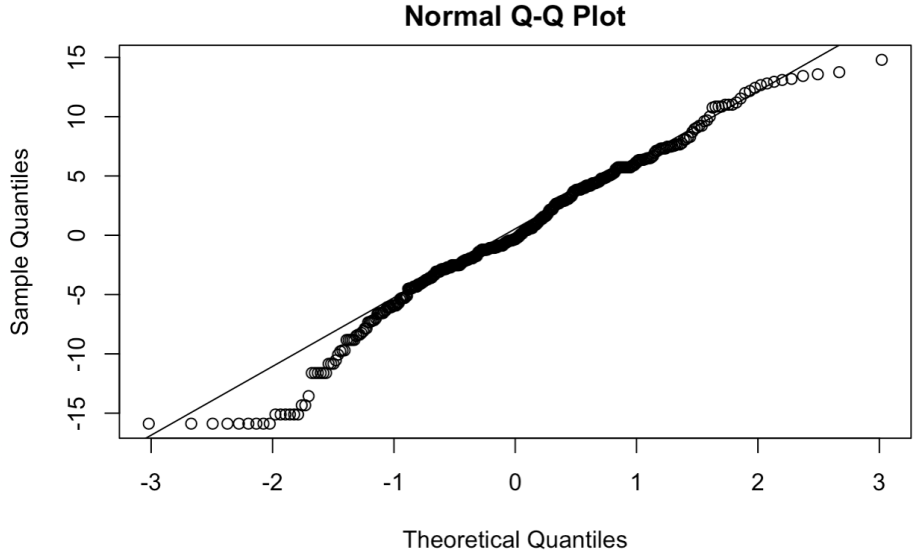



Therefore, we can see the final model requires transformations. We performed a log transformation on the only numerical predictor variable (absences). Then, after running diagnostics, to resolve the remaining issues we used a box-cox transformation and found out that the best regression is a regression of Y^1.15 on the predictors. Below are the new residual vs predictor and normality Q-Q plot after the transformations. We can see the points on the residual vs predictor plot are now well-distributed and that the tails of the curve in the normality Q-Q plot do not diverge as much. The histogram of the residuals below shows a fairly normal distribution, which also confirms the normality issue has been resolved.


Histogram of Residuals (Final Model)
Frequency
0
20
40
60
80
100
120
-20
-15
-10
-5
0
5
10
15
Residuals


Residual vs Predictor (absences)
Residual
-15
-10
-5
0
5
10
15
0
1
2
3
4
Predictor (absences)


Normal Q-Q Plot
Sample Quantiles
-15
-10
-5
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles

In all of our plots, we have noticed that there are no stark outliers in our data. In general, the traveltime of a student and the number of past failures significantly affect his or her final math grade the most. In addition, the main focus of interest, the number of absences, also has a tremendous impact on a student's final math grade. Before doing any transformations, our R^2 value for the original model ($Yi = \beta o + \beta Absences\ Xi + \varepsilon i$) was 0.00117. However, after all of the transformations, we created a final model ($Yi = \beta o + \beta Absences\ Xi + \beta Failures\ Xi + \beta Traveltime\ Xi + \beta Failures,Traveltime\ Xi + \beta Absences,Failures\ Xi + \varepsilon i$) which has an R^2 of 0.1544. This demonstrates that our new model is a better fit after adding significant predictors such as traveltime and failures. By running numerous tests, we have proven that the traveltime and failures, along with a student's absences, have the most impact on the final grade.

**Conclusion**:

Ultimately, the number of absences do affect a student's math performance. We can also see that the time that it takes a student to go to school as well as the number of past classes that they have failed significantly affect their final math grade. In addition, the number of past failed classes a student has had influences the number of absences. There is also an interaction between the traveltime and the number of absences of a student. This shows that both the traveltime and the number of classes a student has failed have a relationship with the number of absences (and consequently influence the final math grade).

There are several ways in which we could improve our final model. One way that the model could be improved is to have a larger sample size of all data. Instead of examining two schools, it would be beneficial to take the cumulative data of numerous schools so that we could have a generalized conclusions of our hypotheses. Thus, we would have more information about the traveltime, failures, number of absences, and final math grades. Furthermore, it is important to note that the data set included mostly categorical variables. Adding more numerical predictors in the data set would improve analysis and help build a better model. Including predictors such as hours of sleep and hours the students were tutored could help improve the model. Our results can not be easily generalized, and the conclusions are fairly rigid, due to the fact that the data set was only collected in two Portuguese schools. Therefore, the data cannot be used to draw general conclusions about the math performance of students outside of these two schools.

**Citations**:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
[Web Link]

## Appendix

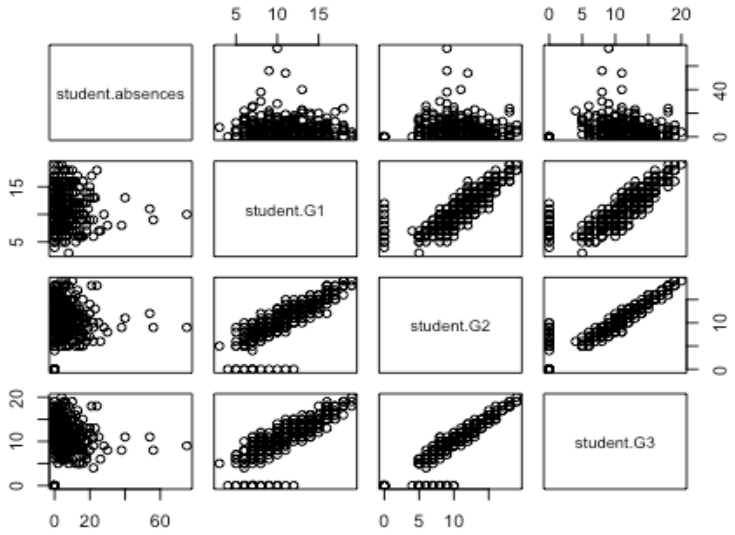
All 30 predictor variables of this data: school, sex, age, address, family size, mother's education, father's education, mother's job, father's job, parent cohabitation status, reason (to choose the school), guardian, traveltime from home to school, studytime, failures, school's extra educational support, family educational support, extra paid classes, extra-curricular activities, nursery, higher education plans, internet access at home, romantic relationships, quality of family relationships, freetime, going out with friends, workday alcohol consumption, weekend alcohol consumption, health, and number of absences.

## Code

```
#student <- read.csv2(file.choose())
student <- read.csv("student-mat(1).csv", header = TRUE)
```

create new, smaller data frames to make it easier to plot and create scatterplot matrices to check correlations

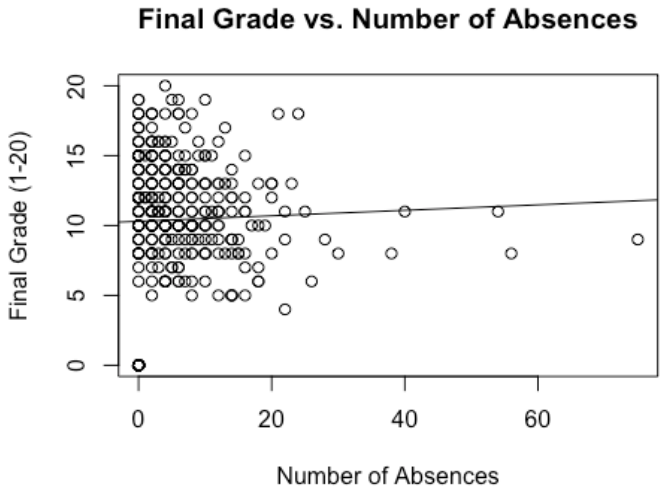
```
newstudent1 = data.frame(student$school, student$sex, student$age, student$address, student$famsize, stude
nt$Pstatus)
newstudent2 = data.frame(student$Medu, student$Fedu, student$Mjob, student$Fjob, student$reason, student$g
uardian)
newstudent3 = data.frame(student$traveltime, student$studytime, student$failures, student$schoolsup, stude
nt$famsup, student$paid)
newstudent4 = data.frame(student$activities, student$nursery, student$higher, student$internet, student$ro
mantic, student$famrel, student$freetime)
newstudent5 = data.frame(student$goout, student$Dalc, student$Walc, student$health)
newstudent6 = data.frame(student$absences, student$G1, student$G2, student$G3)

#scatter1 = pairs(newstudent1)
#scatter2 = pairs(newstudent2)
#scatter3 = pairs(newstudent3)
#scatter4 = pairs(newstudent4)
#scatter5 = pairs(newstudent5)
scatter6 = pairs(newstudent6)
```



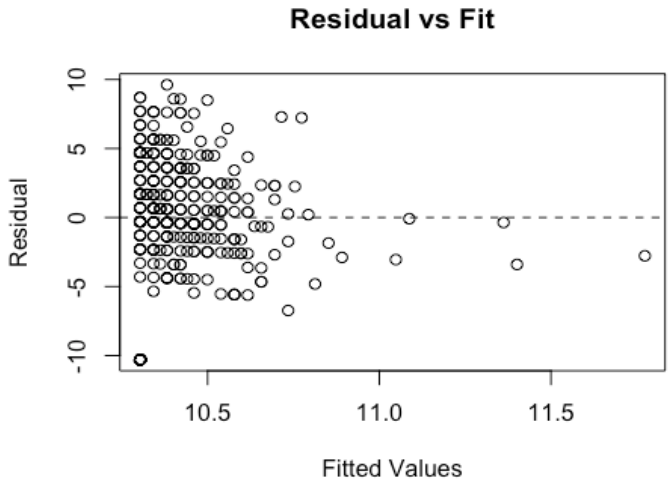

DIAGNOSTICS (OLD MODEL):

```
#par(mfrow =c(1,3))
absences = student$absences
```

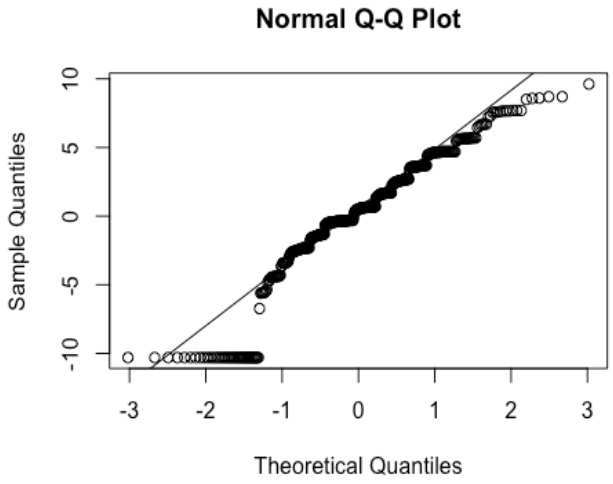
```r
finalgrade = student$G3
plot(absences, finalgrade, xlab="Number of Absences", ylab = "Final Grade (1-20)", main = "Final Grade vs. Number of Absences")

student.fit = lm(finalgrade ~ absences)
abline(student.fit) # intercept = 10.30327, slope: 0.01961
```



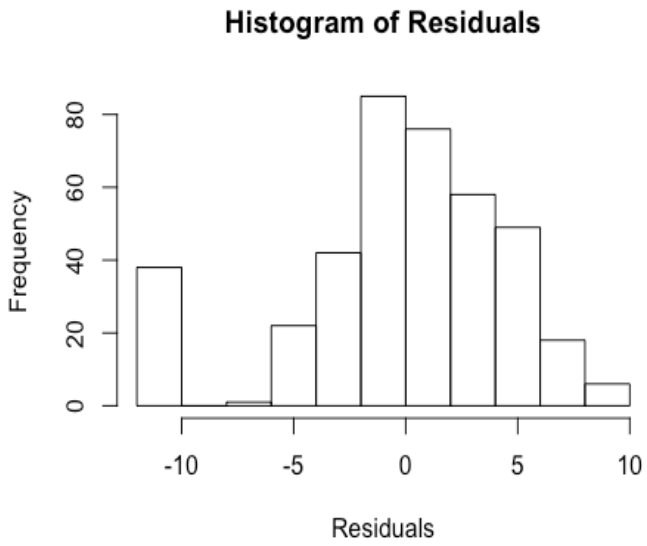

```r
# Res vs. Fit:
x = student$absences
y = student$G3
xbar = mean(x)
ybar = mean(y)
yhat = fitted(student.fit)
e = y - yhat
plot(yhat, e, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)
```




```r
# clustering on the left (clear funnel effect)

#Normal Q-Q plot
student.res = resid(student.fit)
qqnorm(student.res)
qqline(student.res)
```




```
# very skewed data, especially on the left tail
shapiro.test(student.res)

##
##  Shapiro-Wilk normality test
##
## data:  student.res
## W = 0.93511, p-value = 4.283e-12

# p-value = 4.283e-12, reject H0: data is normal --> know it has normality problem

summary(student.fit, data = "student")$r.squared

## [1] 0.001172879

hist(e, xlab = 'Residuals', main = 'Histogram of Residuals')
```



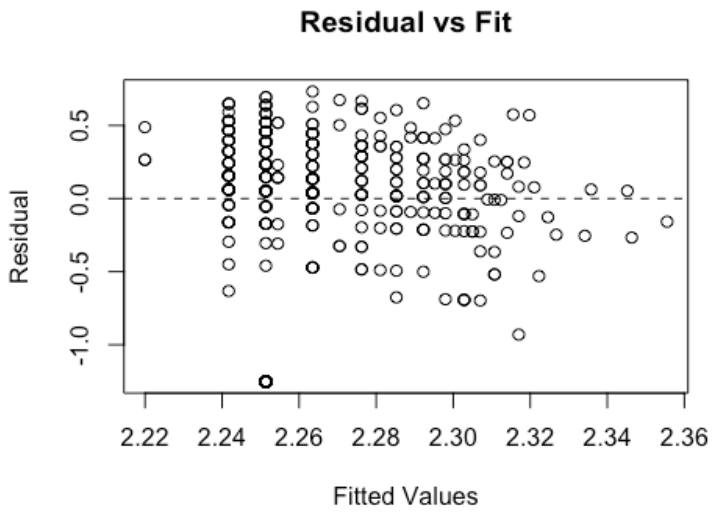

TRANSFORMATIONS: –> seem to have problems with everything (linearity, equal variances, normality) so take log of both X and Y

```
x = student$absences
y = student$G3
x.new = log(x)
#x.new
y.new = log(y)
```

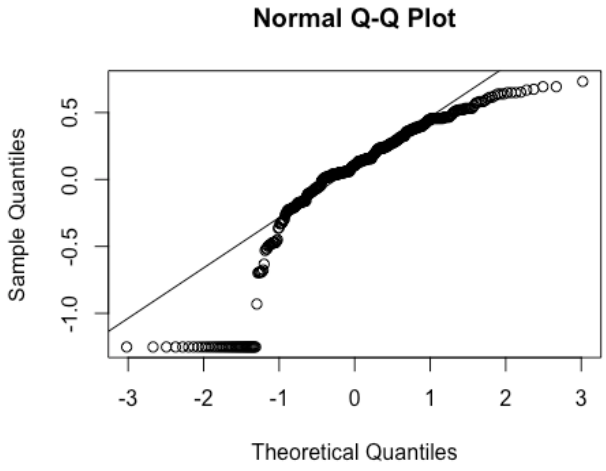
```r
#y.new
#stud.fit.new = lm(y.new ~ x.new)
# error due to log = "-Inf" bc can't do log of 0, replace "-Inf" with "1"

library(plyr)
x.new = mapvalues(x.new, from = "-Inf", to = "1")
x.new = as.numeric(as.character(x.new))
y.new = mapvalues(y.new, from = "-Inf", to = "1")
y.new = as.numeric(as.character(y.new))
stud.fit.new = lm(y.new ~ x.new)
```

DIAGNOSTICS (NEW MODEL):

```r
# res vs fit:
yhat.new = fitted(stud.fit.new)
e.new = y.new - yhat.new
plot(yhat.new, e.new, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)
```




```r
# Q-Q plot:
qqnorm(e.new)
qqline(e.new)
```

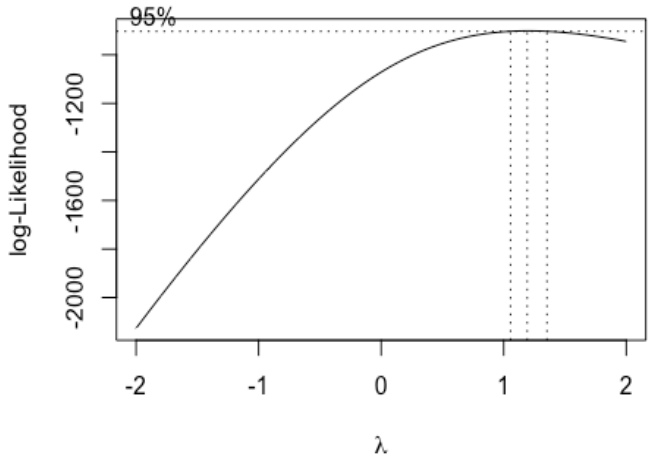



```r
# R^2
summary(stud.fit.new, data = "student")$r.squared
```

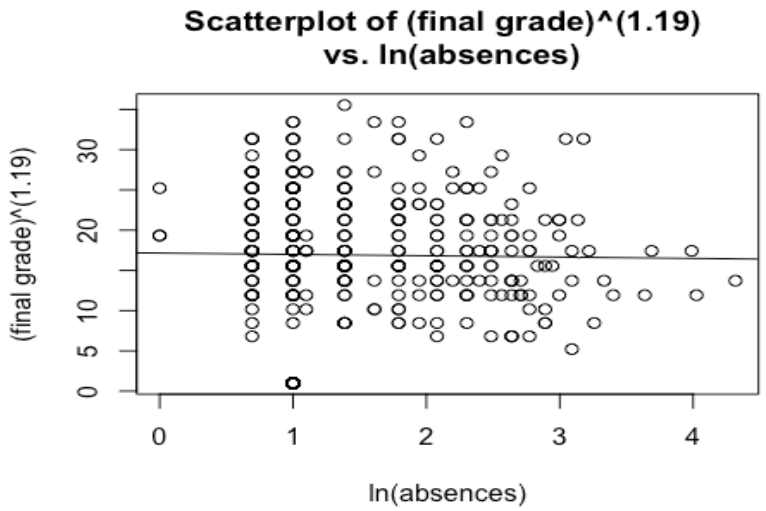
```
## [1] 0.002318529
```

BOX COX TRANSFORMATION:

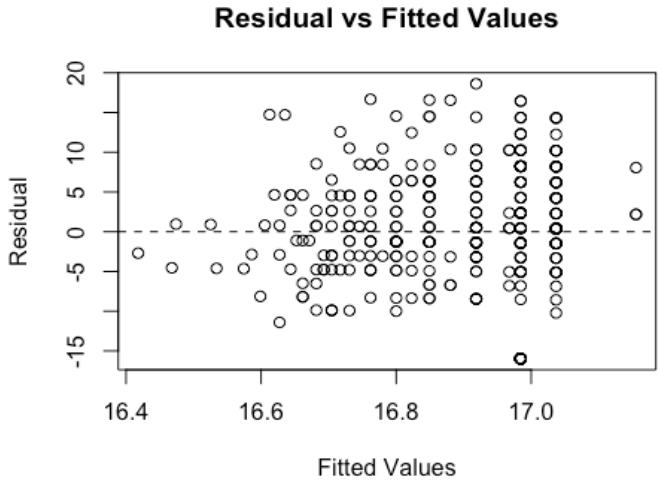
```
library(plyr)
library(SemiPar)
library(MASS)
attach(student)
y2 = student$G3
y2 = mapvalues(y2, from = "0", to = "1")
y2 = as.numeric(as.character(y2))
bc = boxcox(y2 ~ x.new)
```



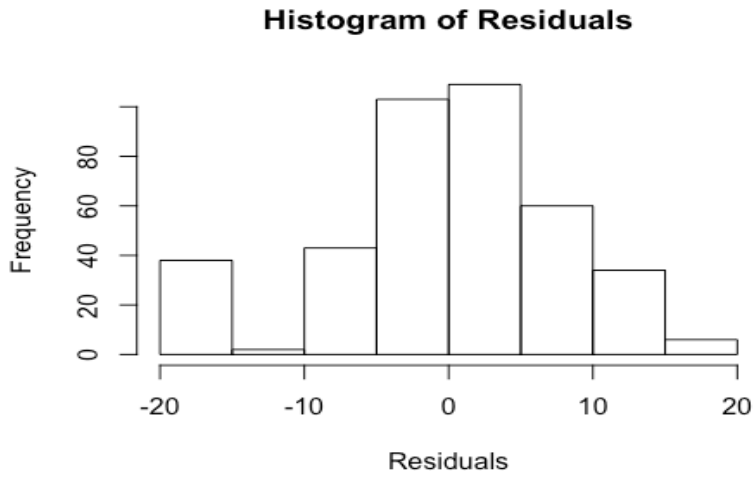

```
lambda = bc$x # lambda values
lik = bc$y # likelihood values for SSE
bc.df = cbind(lambda, lik)
sorted_bc = bc.df[order(-lik)] # values are sorted to identify the lambda value for the maximum log likeli
hood for obtaining minimum SSE
head(sorted_bc, n=10)

##  [1] 1.191919 1.232323 1.151515 1.272727 1.111111 1.313131 1.070707
##  [8] 1.353535 1.030303 1.393939

y_lambda = y2^(1.191919)
fit.new = lm(y_lambda ~ x.new)
summary(fit.new)$r.squared

## [1] 0.0002818789

yhat.new = fitted(fit.new)
e.new = y_lambda - yhat.new

# new scatterplot:
plot(x.new, y_lambda, xlab = 'ln(absences)', ylab = '(final grade)^(1.19)', main = 'Scatterplot of (final
grade)^(1.19) \n vs. ln(absences)')
abline(fit.new)
```



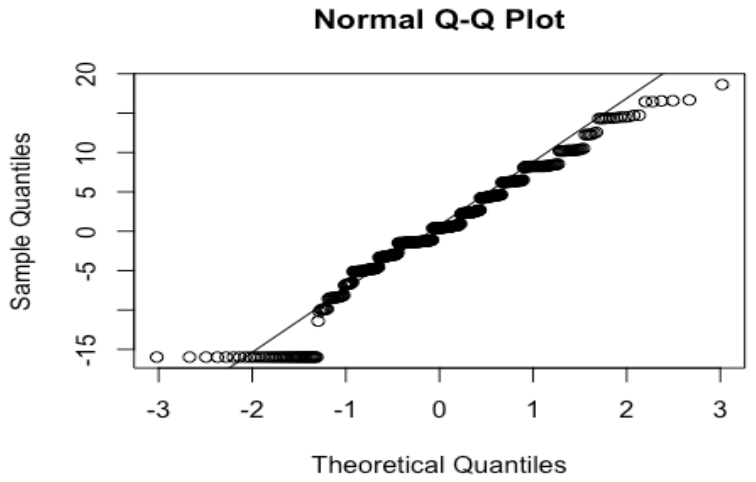

```r
# res vs. fit
plot(yhat.new, e.new, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fitted Values')
abline(h = 0, lty = 2)
```




```r
hist(e.new, xlab = 'Residuals', main = 'Histogram of Residuals')
```




```
#normal Q-Q plot
qqnorm(e.new)
qqline(e.new)
```




SIGNIFICANCE OF OTHER PREDICTORS:

```
### AIC - choosing bw our categorical predictors: (TEST 1)
mod0 = lm(finalgrade ~ 1)
mod.upper = lm(finalgrade ~ student$traveltime + student$studytime + student$failures + student$activities
+ student$freetime)
step(mod0, scope = list(mod0, upper = mod.upper))

## Start:  AIC=1203.39
## finalgrade ~ 1
##
##                       Df Sum of Sq    RSS    AIC
## + student$failures     1   1074.25 7195.7 1150.4
## + student$traveltime   1    113.48 8156.4 1199.9
## + student$studytime    1     79.13 8190.8 1201.6
## <none>                             8269.9 1203.4
## + student$activities   1      2.14 8267.8 1205.3
## + student$freetime     1      1.06 8268.9 1205.3
##
## Step:  AIC=1150.43
## finalgrade ~ student$failures
##
##                       Df Sum of Sq    RSS    AIC
```

```
## + student$traveltime  1     58.71 7136.9 1149.2
## <none>                            7195.7 1150.4
## + student$freetime    1     16.49 7179.2 1151.5
## + student$studytime   1     10.60 7185.1 1151.8
## + student$activities  1      0.66 7195.0 1152.4
## - student$failures    1   1074.25 8269.9 1203.4
##
## Step:  AIC=1149.19
## finalgrade ~ student$failures + student$traveltime
##
##                      Df Sum of Sq    RSS    AIC
## <none>                            7136.9 1149.2
## + student$freetime    1     14.94 7122.0 1150.4
## - student$traveltime  1     58.71 7195.7 1150.4
## + student$studytime   1      6.77 7130.2 1150.8
## + student$activities  1      0.67 7136.3 1151.2
## - student$failures    1   1019.48 8156.4 1199.9

##
## Call:
## lm(formula = finalgrade ~ student$failures + student$traveltime)
##
## Coefficients:
##        (Intercept)    student$failures  student$traveltime
##            11.9460             -2.1723             -0.5558

# TEST 2
mod02 = lm(finalgrade ~ 1)
mod.upper2 = lm(finalgrade ~ student$Medu + student$Fedu + student$Mjob + student$Fjob + student$famsup +
student$famrel)
step(mod02, scope = list(mod02, upper = mod.upper2))

## Start:  AIC=1203.39
## finalgrade ~ 1
##
##                  Df Sum of Sq    RSS    AIC
## + student$Medu    1    389.95 7880.0 1186.3
## + student$Fedu    1    192.22 8077.7 1196.1
## + student$Mjob    4    306.64 7963.3 1196.5
## <none>                        8269.9 1203.4
## + student$famrel  1     21.82 8248.1 1204.3
## + student$famsup  1     12.68 8257.2 1204.8
## + student$Fjob    4    109.06 8160.9 1206.2
##
## Step:  AIC=1186.31
## finalgrade ~ student$Medu
##
##                  Df Sum of Sq    RSS    AIC
## + student$famsup  1     53.49 7826.5 1185.6
## <none>                        7880.0 1186.3
## + student$famrel  1     22.55 7857.4 1187.2
## + student$Fedu    1      3.94 7876.0 1188.1
## + student$Mjob    4     89.33 7790.6 1189.8
## + student$Fjob    4     29.22 7850.7 1192.8
## - student$Medu    1    389.95 8269.9 1203.4
##
## Step:  AIC=1185.62
## finalgrade ~ student$Medu + student$famsup
##
##                  Df Sum of Sq    RSS    AIC
## <none>                        7826.5 1185.6
## - student$famsup  1     53.49 7880.0 1186.3
## + student$famrel  1     21.18 7805.3 1186.5
## + student$Fedu    1      7.14 7819.3 1187.3
## + student$Mjob    4     99.30 7727.2 1188.6
## + student$Fjob    4     35.97 7790.5 1191.8
## - student$Medu    1    430.76 8257.2 1204.8

##
## Call:
## lm(formula = finalgrade ~ student$Medu + student$famsup)
##
```

```
## Coefficients:
##       (Intercept)       student$Medu  student$famsupyes
##            8.2145             0.9717            -0.7685
```

The suggested model includes the Medu (mother's education) and famsup (family educational support) predictors

Now, let's focus on predictors that look at social life:

```
# TEST 3
mod03 = lm(finalgrade ~ 1)
mod.upper3 = lm(finalgrade ~ student$goout + student$Dalc + student$Walc + student$romantic + student$acti
vities + student$freetime)
step(mod03, scope = list(mod03, upper = mod.upper3))

## Start:  AIC=1203.39
## finalgrade ~ 1
##
##                       Df Sum of Sq    RSS    AIC
## + student$goout        1   145.828 8124.1 1198.4
## + student$romantic     1   139.697 8130.2 1198.7
## <none>                             8269.9 1203.4
## + student$Dalc         1    24.708 8245.2 1204.2
## + student$Walc         1    22.310 8247.6 1204.3
## + student$activities   1     2.144 8267.8 1205.3
## + student$freetime     1     1.057 8268.9 1205.3
##
## Step:  AIC=1198.36
## finalgrade ~ student$goout
##
##                       Df Sum of Sq    RSS    AIC
## + student$romantic     1   137.468 7986.6 1193.6
## <none>                             8124.1 1198.4
## + student$freetime     1    21.749 8102.3 1199.3
## + student$activities   1     4.092 8120.0 1200.2
## + student$Dalc         1     3.285 8120.8 1200.2
## + student$Walc         1     0.152 8123.9 1200.3
## - student$goout        1   145.828 8269.9 1203.4
##
## Step:  AIC=1193.62
## finalgrade ~ student$goout + student$romantic
##
##                       Df Sum of Sq    RSS    AIC
## <none>                             7986.6 1193.6
## + student$freetime     1    20.248 7966.4 1194.6
## + student$activities   1     5.061 7981.6 1195.4
## + student$Dalc         1     2.736 7983.9 1195.5
## + student$Walc         1     0.046 7986.6 1195.6
## - student$romantic     1   137.468 8124.1 1198.4
## - student$goout        1   143.599 8130.2 1198.7

##
## Call:
## lm(formula = finalgrade ~ student$goout + student$romantic)
##
## Coefficients:
##         (Intercept)        student$goout  student$romanticyes
##             12.5191              -0.5423              -1.2507
```

We can see the best model includes the predictors goout and romantic.

```
summary(lm(finalgrade ~ student$absences + student$traveltime))

##
## Call:
## lm(formula = finalgrade ~ student$absences + student$traveltime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.652  -1.918   0.292   3.236   9.273
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        11.41838    0.55579  20.544   <2e-16 ***
## student$absences    0.01874    0.02870   0.653   0.5142
## student$traveltime -0.76664    0.32932  -2.328   0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.559 on 392 degrees of freedom
## Multiple R-squared:  0.01479,    Adjusted R-squared:  0.009767
## F-statistic: 2.943 on 2 and 392 DF,  p-value: 0.05387
```

```
summary(lm(finalgrade ~ student$absences + student$failures))
```

```
##
## Call:
## lm(formula = finalgrade ~ student$absences + student$failures)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.977  -1.980   0.023   2.980   8.892
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.97700    0.27842  39.426  < 2e-16 ***
## student$absences  0.03289    0.02697   1.219    0.224
## student$failures -2.24298    0.29029  -7.727 9.35e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.276 on 392 degrees of freedom
## Multiple R-squared:  0.1332, Adjusted R-squared:  0.1288
## F-statistic: 30.12 on 2 and 392 DF,  p-value: 6.815e-13
```

INTERACTIONS Now, we want to check if any of these predictors interact with absences:

```
### AIC Model 1: failures and traveltime
mod.reduced = lm(finalgrade ~ absences + student$failures + student$traveltime)
mod.full = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
absences*student$traveltime)
anova(mod.reduced, mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$failures + student$traveltime
## Model 2: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures + absences * student$traveltime
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    391 7111.2
## 2    389 6927.7  2    183.51 5.1523 0.006188 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value = 0.006 < 0.05 --> can reject null
```

```
### AIC Model 2: Medu and famsup
mod.reduced = lm(finalgrade ~ absences + student$Medu + student$famsup)
mod.full = lm(finalgrade ~ absences + + student$Medu + student$famsup + absences*student$Medu + absences*s
tudent$famsup)
anova(mod.reduced, mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$Medu + student$famsup
## Model 2: finalgrade ~ absences + +student$Medu + student$famsup + absences *
##     student$Medu + absences * student$famsup
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    391 7825.1
## 2    389 7737.8  2    87.301 2.1944 0.1128
```

```
# p-value = 0.11 > 0.05 --> fail to reject H0 (can't use full)
```

```
### AIC Model 3: goout and romantic
mod.reduced = lm(finalgrade ~ absences + student$goout + student$romantic)
```

```
mod.full = lm(finalgrade ~ absences + student$goout + student$romantic + absences*student$goout + absences
*student$romantic)
anova(mod.reduced, mod.full)

## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$goout + student$romantic
## Model 2: finalgrade ~ absences + student$goout + student$romantic + absences *
##     student$goout + absences * student$romantic
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    391 7956.2
## 2    389 7919.8  2    36.415 0.8943 0.4097

# p-value = 0.4097 > 0.05 --> fail to reject H0 (can't use full)



summary(lm(finalgrade ~ student$failures))

##
## Call:
## lm(formula = finalgrade ~ student$failures)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1572  -2.1572  -0.1572   2.8428   9.0632
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.1572     0.2361   47.26  < 2e-16 ***
## student$failures  -2.2204     0.2899   -7.66 1.47e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.279 on 393 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1277
## F-statistic: 58.67 on 1 and 393 DF,  p-value: 1.466e-13

summary(lm(finalgrade ~ student$traveltime))

##
## Call:
## lm(formula = finalgrade ~ student$traveltime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.76  -1.76   0.24   3.24   9.24
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         11.5294     0.5288  21.805   <2e-16 ***
## student$traveltime  -0.7694     0.3290  -2.338   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.556 on 393 degrees of freedom
## Multiple R-squared:  0.01372,    Adjusted R-squared:  0.01121
## F-statistic: 5.468 on 1 and 393 DF,  p-value: 0.01987
```

1) For every amount of absences, is there a difference in the mean effect for the other two predictors? We need to test the null hypothesis: H0: B2 = B3 = B12 = B13 = 0 vs H1: at least one of these slope parameters is not 0.

```
# using Model 1 (failures and traveltime)
mod.reduced = lm(finalgrade ~ absences)
mod.full = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
absences*student$traveltime)
anova(mod.reduced, mod.full)

## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences
## Model 2: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures + absences * student$traveltime
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    393 8260.2
## 2    389 6927.7  4    1332.5 18.705 4.471e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# there's a significant difference in all the terms

# testing models for interactions with absences
### AIC Model 1: failures and traveltime
mod.reduced = lm(finalgrade ~ absences + student$failures + student$traveltime)
mod.full = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
absences*student$traveltime)
anova(mod.reduced, mod.full)

## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$failures + student$traveltime
## Model 2: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures + absences * student$traveltime
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    391 7111.2
## 2    389 6927.7  2    183.51 5.1523 0.006188 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# p-value = 0.006 < 0.05 --> can reject null
# at least one of the interaction parameters is not zero
```

Now let us see which interaction term is significant:

```
# Testing significance of absences*student$failures:
mod.reduced = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$travelti
me)
mod.full = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
absences*student$traveltime)
anova(mod.reduced, mod.full)

## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$traveltime
## Model 2: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures + absences * student$traveltime
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    390 7106.4
## 2    389 6927.7  1    178.71 10.035 0.001658 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Testing significance of absences*student$traveltime:
mod.reduced = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures
)
mod.full = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
absences*student$traveltime)
anova(mod.reduced, mod.full)

## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures
## Model 2: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures + absences * student$traveltime
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    390 6934.1
## 2    389 6927.7  1     6.355 0.3568 0.5506

# the interaction term that is significant is: absences*student$failures
# so the failures predictor has an interaction with absences; the traveltime predictor does not

# interaction between failures and traveltime?
mod.reduced = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures
```

```
)
mod.full = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
student$failures*student$traveltime)
anova(mod.reduced, mod.full)

## Analysis of Variance Table
##
## Model 1: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures
## Model 2: finalgrade ~ absences + student$failures + student$traveltime +
##     absences * student$failures + student$failures * student$traveltime
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    390 6934.1
## 2    389 6773.0  1     161.1 9.2527 0.002511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, let us test the new model to ensure the predictors and interaction terms are all significant:

```
#### AIC of new model:#####
mod0 = lm(finalgrade ~ absences)
mod.upper = lm(finalgrade ~ absences + student$failures + student$traveltime + absences*student$failures +
absences*student$traveltime)
step(mod0, scope = list(mod0, upper = mod.upper))

## Start:  AIC=1204.93
## finalgrade ~ absences
##
##                       Df Sum of Sq    RSS    AIC
## + student$failures     1   1091.74 7168.5 1150.9
## + student$traveltime   1    112.64 8147.6 1201.5
## - absences             1      9.70 8269.9 1203.4
## <none>                              8260.2 1204.9
##
## Step:  AIC=1150.93
## finalgrade ~ absences + student$failures
##
##                              Df Sum of Sq    RSS    AIC
## + absences:student$failures   1    191.16 6977.3 1142.3
## + student$traveltime          1     57.23 7111.2 1149.8
## - absences                    1     27.18 7195.7 1150.4
## <none>                                   7168.5 1150.9
## - student$failures            1   1091.74 8260.2 1204.9
##
## Step:  AIC=1142.26
## finalgrade ~ absences + student$failures + absences:student$failures
##
##                              Df Sum of Sq    RSS    AIC
## + student$traveltime          1    43.227 6934.1 1141.8
## <none>                                   6977.3 1142.3
## - absences:student$failures   1   191.160 7168.5 1150.9
##
## Step:  AIC=1141.8
## finalgrade ~ absences + student$failures + student$traveltime +
##     absences:student$failures
##
##                                Df Sum of Sq    RSS    AIC
## <none>                                      6934.1 1141.8
## - student$traveltime            1    43.227 6977.3 1142.3
## + absences:student$traveltime   1     6.355 6927.7 1143.4
## - absences:student$failures     1   177.160 7111.2 1149.8

##
## Call:
## lm(formula = finalgrade ~ absences + student$failures + student$traveltime +
##     absences:student$failures)
##
## Coefficients:
##               (Intercept)                   absences
##                  11.85838                   -0.01355
##          student$failures          student$traveltime
##                  -2.92712                   -0.47827
```

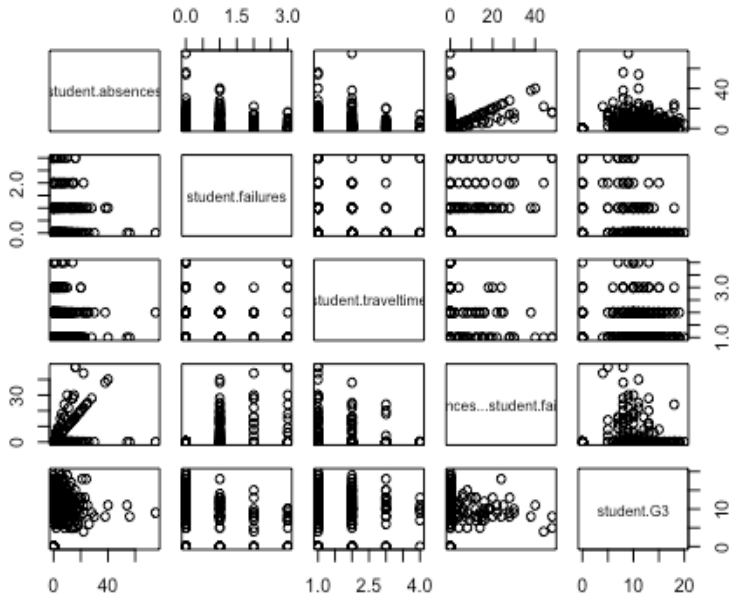
```
## absences:student$failures
##                   0.13338
```

FINAL MODEL: absences + failures + traveltime + absences x failures + failures x traveltime

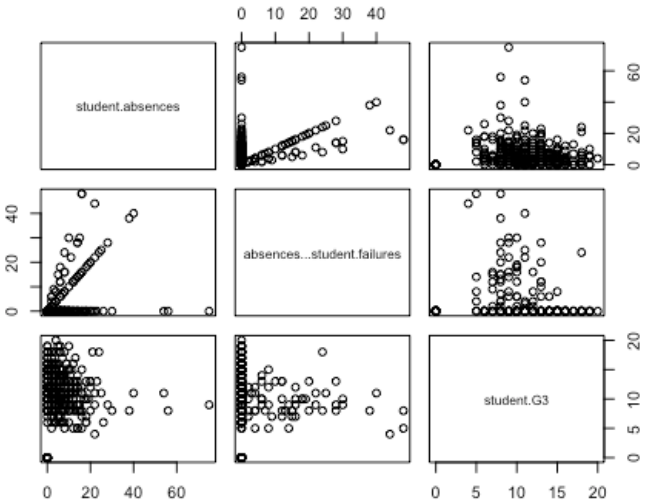
---

Checking (running diagnostics on) the New/Final Model:

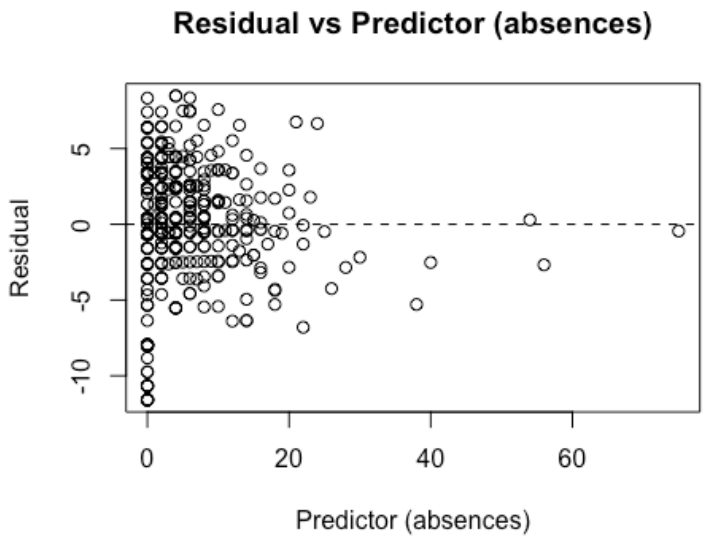
```
### Scatterplot matrix:
finalstudent = data.frame(student$absences, student$failures, student$traveltime, absences*student$failure
s, student$G3)
pairs(finalstudent)
```



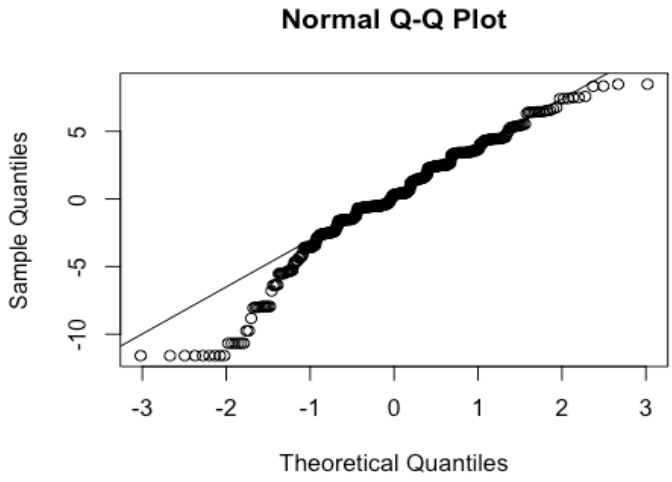

```
# just to take a closer look at correlations:
finalstudent2 = data.frame(student$absences, absences*student$failures, student$G3)
pairs(finalstudent2)
```



```
### FINAL MODEL:
finalmod = lm(finalgrade ~ absences + failures + traveltime + absences*failures + traveltime*failures, dat
a= student)

### Residual vs. predictor:
# Residual vs. Absences
x2 = student$absences
y = student$G3
xbar = mean(x2)
ybar = mean(y)
yhat = fitted(finalmod)
e = y - yhat
plot(x2, e, xlab = 'Predictor (absences)', ylab = 'Residual', main = 'Residual vs Predictor (absences)')
abline(h = 0, lty = 2)
```



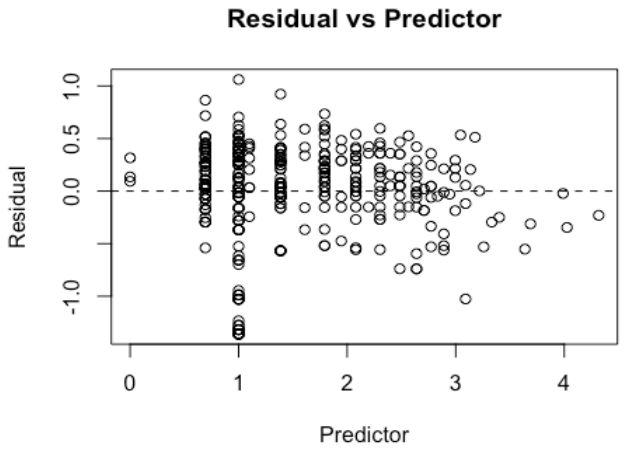

```
# clustering on the left (clear the fanning effect)

### Normal Q-Q plot
student.res = resid(finalmod)
finalstudent.res = resid(finalmod)
qqnorm(finalstudent.res)
qqline(finalstudent.res)
```



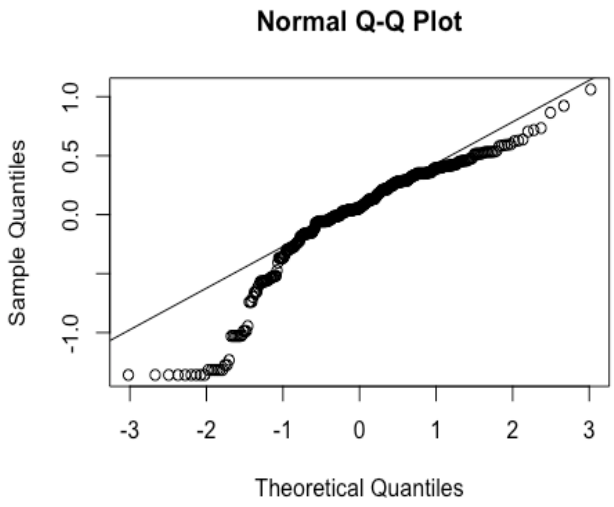

```
# normality issue (diverges on the tails)
```

TRANSFORMATIONS ON FINAL MODEL:

```
y = student$G3
x.new = log(student$absences) # can I just take the log of the numerical predictor???
#x.new
y.new = log(y)
#y.new
#stud.fit.new = lm(y.new ~ x.new)
# error due to log = "-Inf" bc can't do log of 0, replace "-Inf" with "1"

library(plyr)
x.new = mapvalues(x.new, from = "-Inf", to = "1")
x.new = as.numeric(as.character(x.new))
y.new = mapvalues(y.new, from = "-Inf", to = "1")
y.new = as.numeric(as.character(y.new))
stud.fit2.new = lm(y.new ~ x.new + failures + traveltime + absences*failures, data= student)

### DIAGNOTICS OF FINAL MODEL:

# res vs predictor:
yhat.new = fitted(stud.fit2.new)
e.new = y.new - yhat.new
plot(x.new, e.new, xlab = 'Predictor', ylab = 'Residual', main = 'Residual vs Predictor')
abline(h = 0, lty = 2)
```

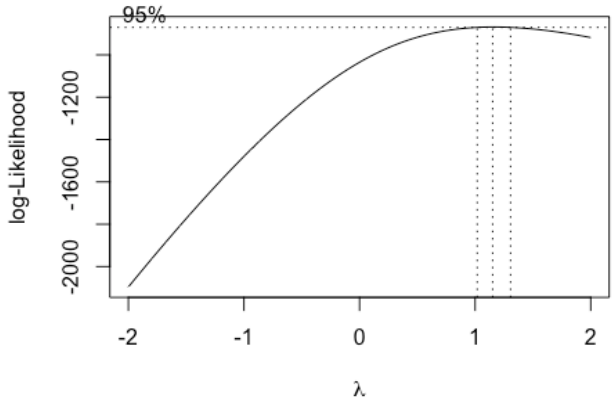



```
# Q-Q plot:
qqnorm(e.new)
qqline(e.new)
```



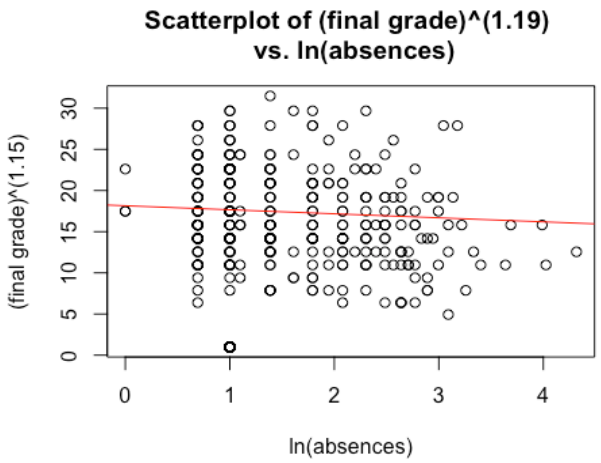

BOX COX TRANSFORMATION:

```
library(plyr)
library(SemiPar)
library(MASS)
attach(student)
y2 = student$G3
y2 = mapvalues(y2, from = "0", to = "1")
y2 = as.numeric(as.character(y2))
bc2 = boxcox(y2 ~ x.new + failures + traveltime + absences*failures, data= student)
```

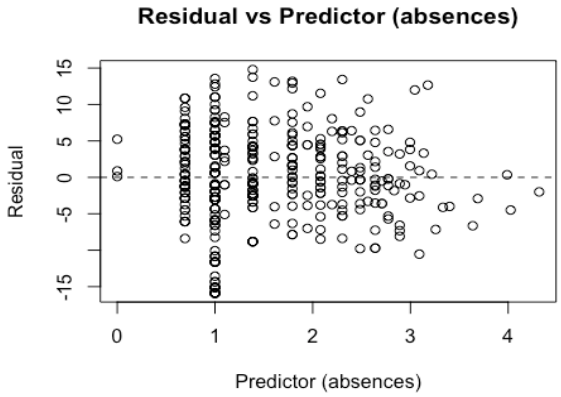



```
lambda = bc2$x # lambda values
lik = bc2$y # likelihood values for SSE
bc.df = cbind(lambda, lik)
sorted_bc = bc.df[order(-lik)] # values are sorted to identify the lambda value for the maximum log likeli
hood for obtaining minimum SSE
head(sorted_bc, n=10)

##  [1] 1.151515 1.191919 1.111111 1.232323 1.070707 1.272727 1.030303
##  [8] 1.313131 0.989899 1.353535

y_lambda = y2^(1.151515)
fit.new = lm(y_lambda ~ x.new + failures + traveltime + absences*failures, data= student)
summary(fit.new)$r.squared

## [1] 0.1544031
```

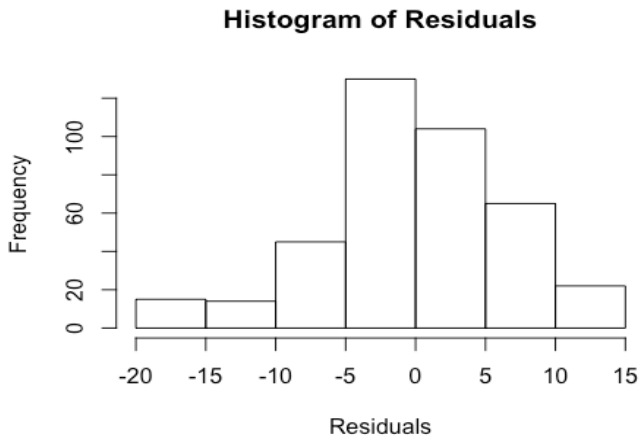
```r
yhat.new = fitted(fit.new)
e.new = y_lambda - yhat.new

# new scatterplot:
plot(x.new, y_lambda, xlab = 'ln(absences)', ylab = '(final grade)^(1.15)', main = 'Scatterplot of (final
grade)^(1.19) \n vs. ln(absences)')
abline(fit.new, col = 2)
```




```r
# res vs. predictor
plot(x.new, e.new, xlab = 'Predictor (absences)', ylab = 'Residual', main = 'Residual vs Predictor (absenc
es)')
abline(h = 0, lty = 2)
```




```r
hist(e.new, xlab = 'Residuals', main = 'Histogram of Residuals')
```




```r
#normal Q-Q plot
qqnorm(e.new)
qqline(e.new)
```