Final_Project_Paper

Ben Lawrence

```
source("setup.R")
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

Introduction

In this paper we investigate the question of which student attributes affect student academic performance. The adage "Skip class you won't pass?" presents us with what might appear to be an obvious relationship. If a student does not show up to learn their class material, they will inevitably struggle when it comes time to evaluate their learning. Many state and national governments have adopted policies to ensure that students "stay in school", and parents who allow their children to be absent for enough days are punished. Clearly there is a general belief that keeping absences low will generally benefit the academic performance of students and in general benefit society.

Some educational researchers believe that there is a strong effect that failing a class has on subsequent student performance. If a student fails one class, they may fall into a pattern where

they are psychologically influenced to perceive their previous effort as wasted. In the pressure to catch up in the following grading period, students may lose the persistence and self-efficacy needed to succeed. In this way, failure can compound and become a predictor of student performance. Research by Rola Ajjawi et. al studied student attitudes and performance after failing college units and found their figures "indicate a significant contribution of academic failure to drop out".

There appears to be general consensus that prior academic grades are one of the best predictors of future academic grades of students (Alyahyan, Düştegör, 2020). However, some challenge the validity of using prior performance to predict future performance, because when data from prior academic contexts is used in can result in inaccurate predictions. For example, if a student performed poorly in high school math, that grade would likely not be a useful predictor of their success in a college level Literature class. The difference in subject and grade level takes the student's performance out of its appropriate context. A better predictor would be the student's grade in a prerequisite English or other Literature class. When proper context is considered prior grades are a strong predictor of future grades, but some also question whether grades are an accurate measure of student success. We will discuss this further in the conclusion. In the following section we shall discuss the dataset we used and conduct regression analyses to examine the predictors discussed above.

Data Description

We obtained our two datasets from the UC Irvine Machine Learning Repository (UCIMLR). These datasets were used in the article "Using Data Mining to Predict Secondary School Student Performance" by Paulo Cortez and Alice Silva from the University of Minho, Portugal. They contain student grade and background information for two core classes in the Portuguese curriculum, Portuguese and Math. The Math dataset is a 395 x 33 dataset, and the Portuguese is a 649 x 33 dataset. They were donated to the UCIMLR in 2014.

In this article, the dataset was used to train five models (naive predictor, neural network, support vector machine, decision tree, and random forest) to predict three different attributes of the data. The first attribute is a binary pass/fail classification based on a student's final grade. The second is a 5-level classification based on the Erasmus grade conversion system which ranks students from 1 to 5 based on bins of their final grade. (Table 2 in the article) The third attribute is a regression output of the final student grade, which is numeric between 0 and 20.

The main conclusion of the article is the grade of the student in the prior grading period (when available) is the most important predictor across the best performing models. When the immediate prior grade is unavailable the next prior grade becomes the most important, and when both are unavailable, other predictors such as number of past class failures, and number of absences become important.

The Portuguese school system operates on a trimester model where three distinct grading periods take place and the last trimester grade is taken as the student's final grade. We shall attempt to predict the same variable as the paper "*G3*" (the student's grade for the third trimester and final grade) using simple linear regression (SLR) and multiple linear regression (MLR) techniques. We will begin our analysis by creating simple plots for the variables that the paper identified as important. These include prior trimester grades, number of absences, and past class failures.

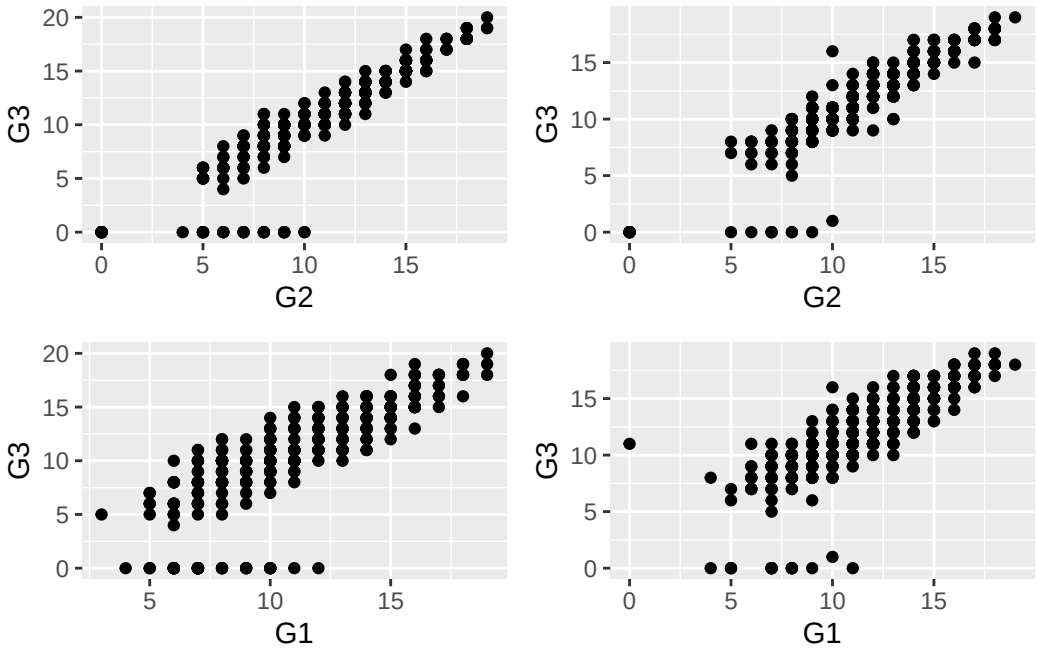
Here we transform our data to be suitable for SLR and MlR. We transform variables recorded on an ordinal scale (rate your health from 1 to 5), as factors. For descriptions of each row of the datasets, please see the Code Appendix below.

```
math_perf <- read.csv("student-mat.csv", sep=";")
port_perf <- read.csv("student-por.csv", sep=";")

# Factor these data columns
fcols <- c('school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu', 'Mjob'

# Factorize columns in both dataframes
math_perf[fcols] <- lapply(math_perf[fcols], factor)
port_perf[fcols] <- lapply(port_perf[fcols], factor)  # Corrected this line

# Transform absences - Log transformation to handle skewness
math_perf$absences <- log(math_perf$absences + 1)
port_perf$absences <- log(port_perf$absences + 1)

#head(math_perf)
#head(port_perf)
```

Figure 1 below shows a scatter plot of *G3* vs. *G2* and *G3* vs. *G1* for both the Math and Portuguese classes. As we can see the relationship between them is roughly linear for all plots, meaning that *G2* and *G1* are both likely strong predictors in a linear regression model. This is consistent with other models presented in the article.

Figure 1

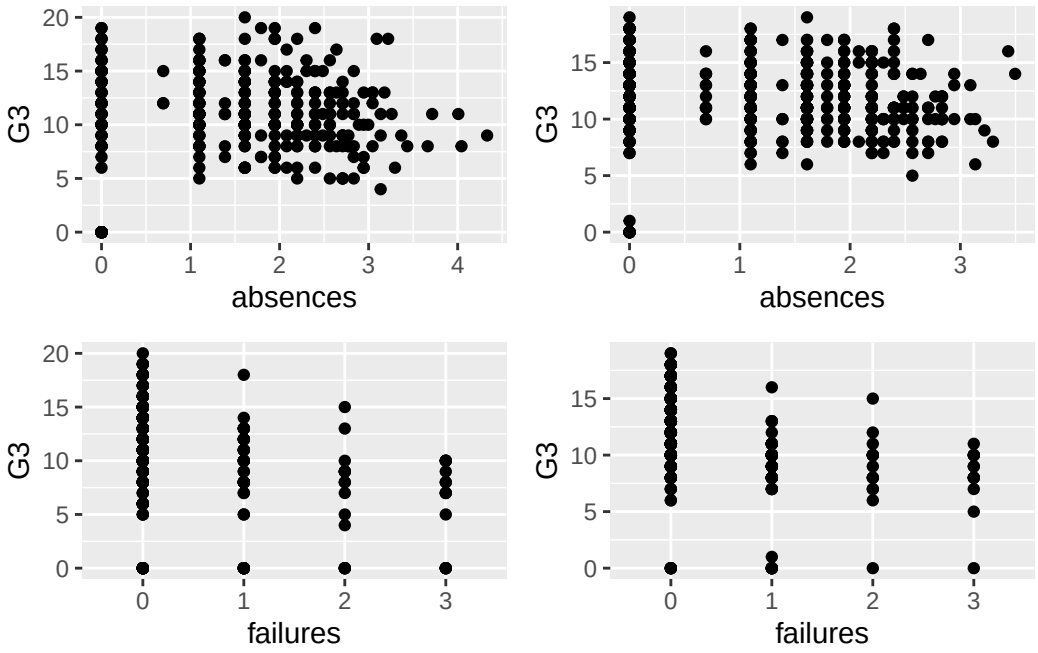
```
p1 <- gg_basic(math_perf,G2,G3)
p2 <- gg_basic(port_perf,G2,G3)
p3 <- gg_basic(math_perf,G1,G3)
p4 <- gg_basic(port_perf,G1,G3)

grid.arrange(p1, p2,p3, p4, ncol=2, nrow=2)
```



The article suggests that absences is also an important predictor. In **Figure 2** we have plotted *absences* vs. *G3* and *failures* vs. *G3* and we can see a negative correlation, but it is very slight. As we might expect, students with more absences do not perform at a high level, but we can see students who have very few absences perform both very well and very poorly. For failures, the number of failures does influence the *G3* value, but because failures is a very discrete variable (takes values 1, 2 ,3, 4), it is hard to establish a continuous linear relationship between it and *G3*.

Figure 2

```
p5 <- gg_basic(math_perf,absences,G3)
p6 <- gg_basic(port_perf,absences,G3)
p7 <- gg_basic(math_perf,failures,G3)
p8 <- gg_basic(port_perf,failures,G3)

grid.arrange(p5, p6, p7, p8, ncol=2, nrow=2)
```



This dataset contains many other variables that may affect student performance as measured by *G3.* Although given the visually weak linear relationship between variables other than prior trimester grades, it may be that the rest of the variables are weak as well or are irrelevant inputs. We shall still investigate these in a complete MLR model later.

Methods and Results (Analysis)

First we will construct SLR models for the variables we have plotted above to confirm our visual intuition. We shall also create diagnostic plots to judge model validity.

In **Figure 3**, we can see *G2* is a highly significant predictor in our models for both clases. However when we look at the diagnostic plots we can see two patterns in the fitted vs. residuals. There is a slight linear relationship with the majority of points and a clear linear relationship with the outlying lower grades. This indicates the variance between the majority grade levels and the lower levels is not constant, calling into question our model validity. The QQ plot and Leverage plot indicate that majority higher grade levels are normal, but the lower levels are not, and abnormally high and low grades tend to have high leverage in our model.

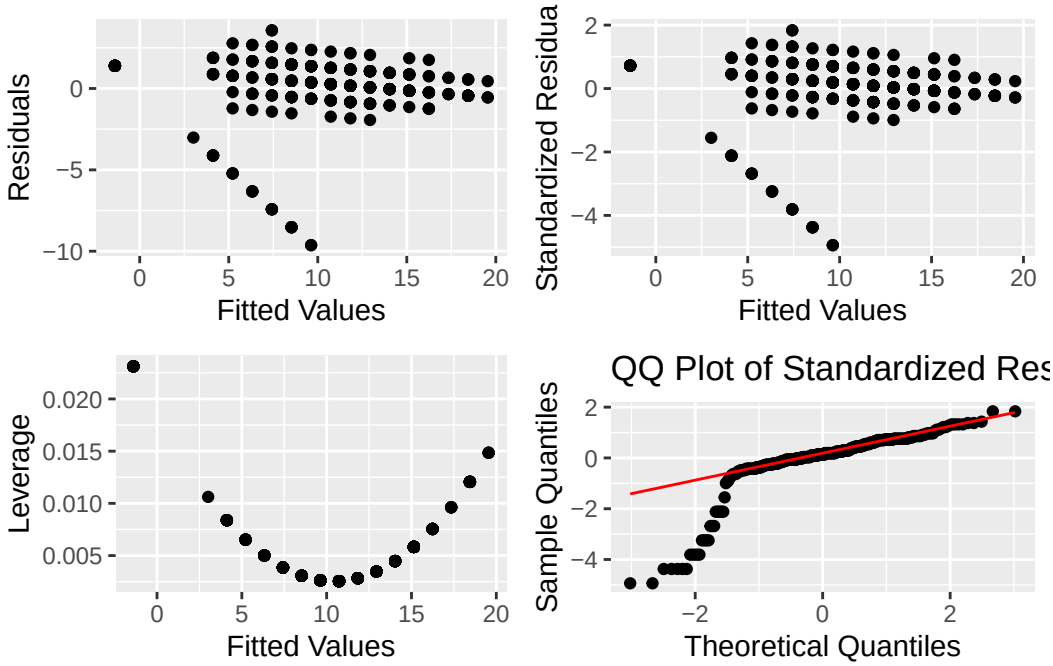
Overall, although the prior trimester grade does not seem appropriate as the sole predictor of the final grade in a linear model. This result holds with *G1* as well.

Figure 3

```r
lm1 <- lm(G3 ~ G2, data = math_perf)
summary(lm1)
```

```
Call:
lm(formula = G3 ~ G2, data = math_perf)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6284 -0.3326  0.2695  1.0653  3.5759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.39276    0.29694   -4.69 3.77e-06 ***
G2           1.10211    0.02615   42.14  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.953 on 393 degrees of freedom
Multiple R-squared:  0.8188,    Adjusted R-squared:  0.8183
F-statistic:  1776 on 1 and 393 DF,  p-value: < 2.2e-16
```

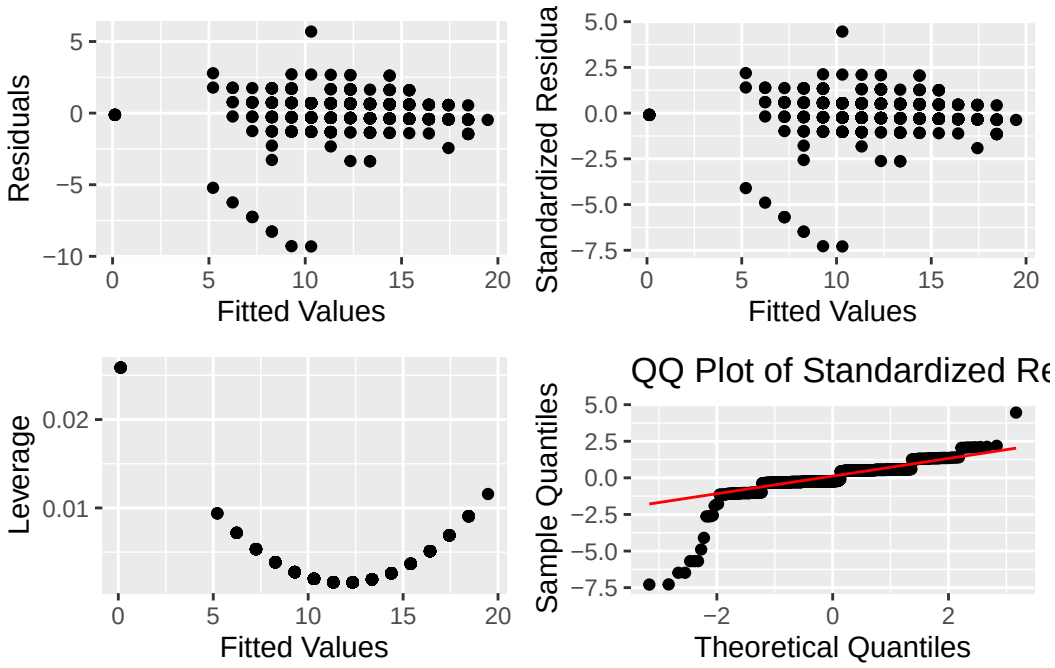
```r
lm_diag(lm1)
```




```
lm2 <- lm(G3 ~ G2, data = port_perf)
summary(lm2)
```

```
Call:
lm(formula = G3 ~ G2, data = port_perf)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3069 -0.3623 -0.2884  0.6746  5.6931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12197    0.20559   0.593    0.553
G2           1.01849    0.01723  59.104   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.278 on 647 degrees of freedom
Multiple R-squared:  0.8437,    Adjusted R-squared:  0.8435
F-statistic:  3493 on 1 and 647 DF,  p-value: < 2.2e-16
```

```
lm_diag(lm2)
```

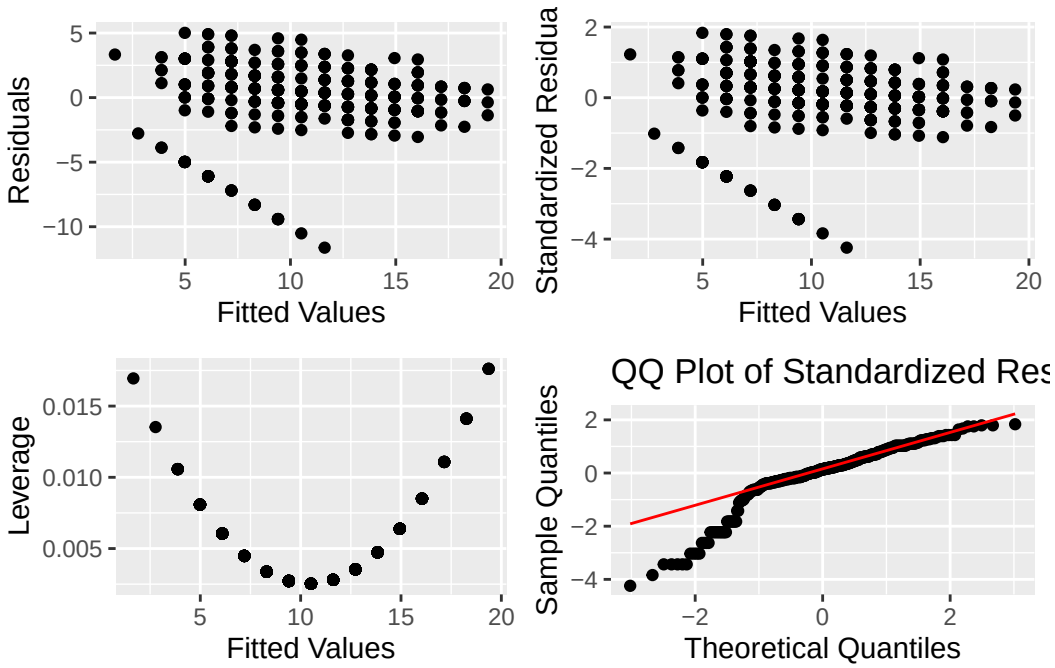



```
lm3 <- lm(G3 ~ G1, data = math_perf)
summary(lm3)
```

```
Call:
lm(formula = G3 ~ G1, data = math_perf)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6223  -0.8348   0.3777   1.6965   5.0153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.65280    0.47475  -3.481 0.000555 ***
G1           1.10626    0.04164  26.568  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.743 on 393 degrees of freedom
```

```
Multiple R-squared:  0.6424,    Adjusted R-squared:  0.6414
F-statistic: 705.8 on 1 and 393 DF,  p-value: < 2.2e-16
```

```
lm_diag(lm3)
```

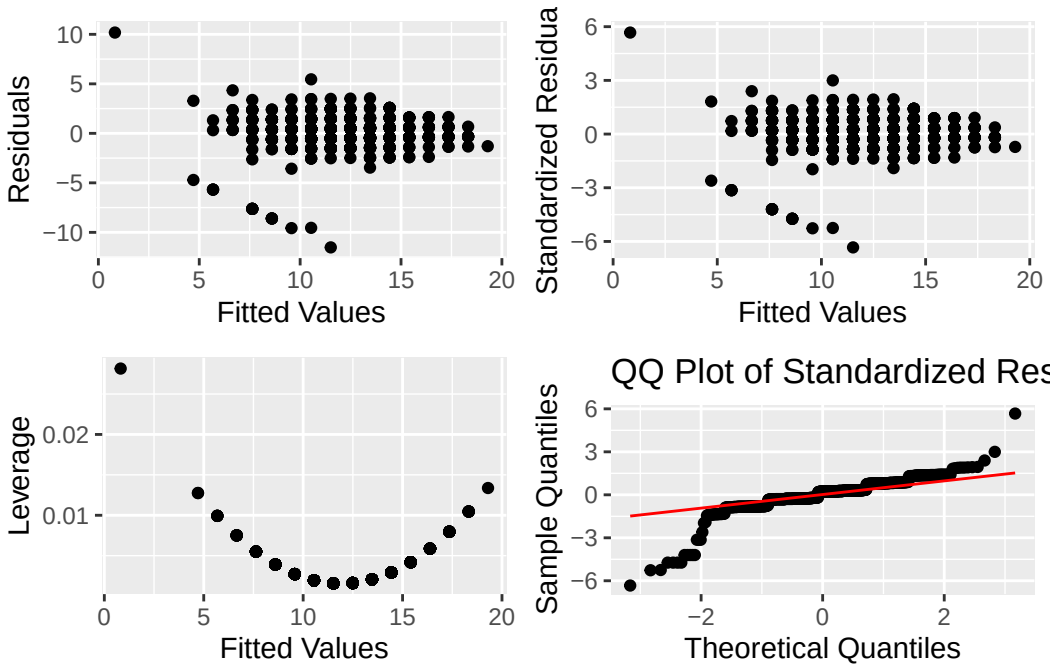



```
lm4 <- lm(G3 ~ G1, data = port_perf)
summary(lm4)
```

```
Call:
lm(formula = G3 ~ G1, data = port_perf)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5179  -0.5454   0.3996   0.6196  10.1796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.82040    0.30545   2.686  0.00742 **
G1           0.97250    0.02605  37.329  < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.821 on 647 degrees of freedom
Multiple R-squared:  0.6829,    Adjusted R-squared:  0.6824
F-statistic:  1393 on 1 and 647 DF,  p-value: < 2.2e-16
```

```
lm_diag(lm4)
```



We may observe the results of SLR models of the remaining predictors that we have discussed. We can see that the R^2 value is extremely low for these predictors. Even if this model was valid, its predictive power would be quite low, so we will omit the diagnostics for the sake of brevity.

```
lm5 <- lm(G3 ~ absences, data=math_perf)
summary(lm5)
```

```
Call:
lm(formula = G3 ~ absences, data = math_perf)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.4309 -2.2200  0.4132  3.1235  9.5691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4309     0.3730  25.283  < 2e-16 ***
absences      0.7182     0.2156   3.331 0.000948 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.524 on 393 degrees of freedom
Multiple R-squared:  0.02745,   Adjusted R-squared:  0.02498
F-statistic: 11.09 on 1 and 393 DF,  p-value: 0.000948
```

```r
lm6 <- lm(G3 ~ absences, data=port_perf)
summary(lm6)
```

```
Call:
lm(formula = G3 ~ absences, data = port_perf)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1735  -1.7332  -0.1735   2.0964   7.2220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1735     0.1910  63.747   <2e-16 ***
absences     -0.2458     0.1314  -1.871   0.0618 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.224 on 647 degrees of freedom
Multiple R-squared:  0.005381,  Adjusted R-squared:  0.003843
F-statistic:   3.5 on 1 and 647 DF,  p-value: 0.06182
```

```r
lm7 <- lm(G3 ~ failures, data=math_perf)
summary(lm7)
```

```
Call:
lm(formula = G3 ~ failures, data = math_perf)

Residuals:
    Min      1Q  Median      3Q     Max
-11.253  -2.253  -0.120   2.765   9.880

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2532     0.2418  46.539  < 2e-16 ***
failures1    -3.1332     0.6506  -4.816 2.10e-06 ***
failures2    -5.0179     1.0637  -4.717 3.33e-06 ***
failures3    -5.5657     1.0948  -5.084 5.75e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.271 on 391 degrees of freedom
Multiple R-squared:  0.1375,    Adjusted R-squared:  0.1309
F-statistic: 20.78 on 3 and 391 DF,  p-value: 1.642e-12
```

```r
lm8 <- lm(G3 ~ failures, data=port_perf)
summary(lm8)
```

```
Call:
lm(formula = G3 ~ failures, data = port_perf)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5100  -1.5100   0.3571   1.4900   7.3571

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.5100     0.1242 100.759  < 2e-16 ***
failures1    -3.8672     0.3692 -10.474  < 2e-16 ***
failures2    -3.6975     0.7378  -5.012 6.98e-07 ***
failures3    -4.4386     0.7873  -5.637 2.58e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.909 on 645 degrees of freedom
Multiple R-squared:  0.1929,    Adjusted R-squared:  0.1891
F-statistic: 51.39 on 3 and 645 DF,  p-value: < 2.2e-16
```

Clearly SLR is not an appropriate technique for creating a valid or useful model and we must expand to a MLR model that will incorporate multiple predictors. We shall begin with a full model.

```
lm_full_math <- lm(G3 ~ . , data=math_perf)
#Omitted for brevity
#summary(lm_full_math)

lm_full_port <- lm(G3 ~ ., data=port_perf)
#Omitted for brevity
#summary(lm_full_port)
```

The full model contains many insignificant predictors that should be filtered out. In **Figure 4** we can see the diagnostics of the full models appear similar to the SLR model with the prior grades as predictors.

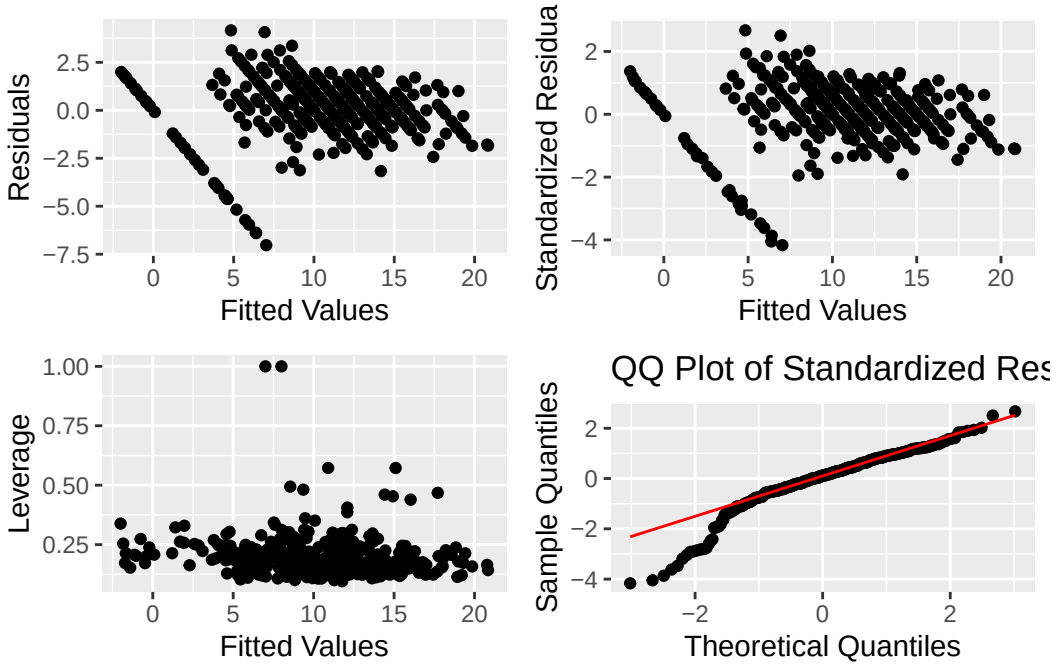
Figure 4

```
lm_diag(lm_full_math)
```

```
Warning: Removed 2 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 2 rows containing non-finite values (`stat_qq()`).
```

```
Warning: Removed 2 rows containing non-finite values (`stat_qq_line()`).
```

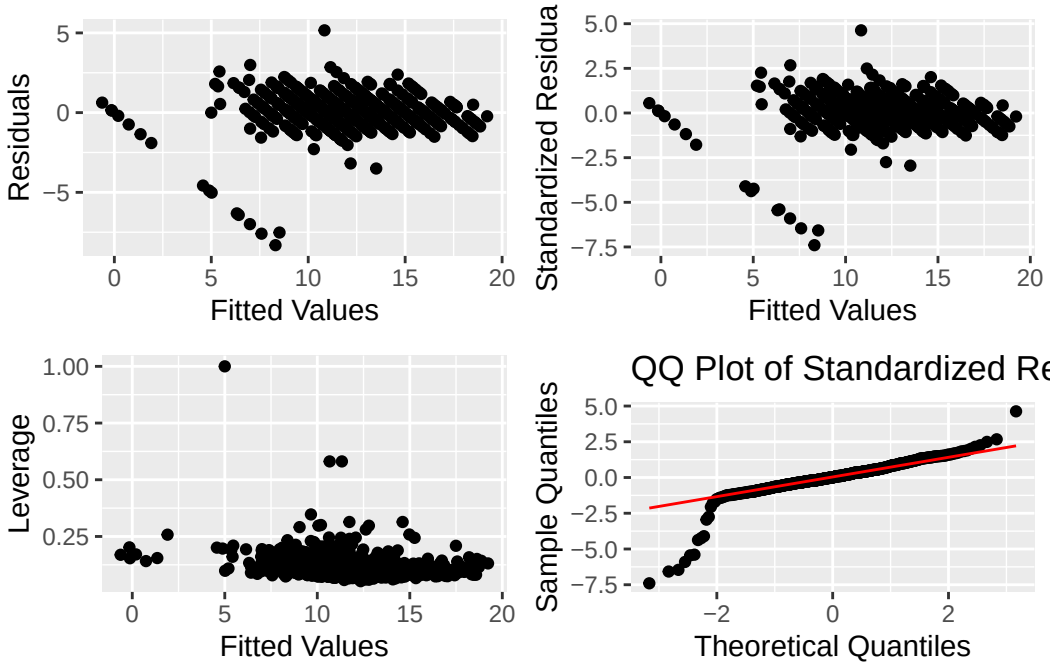



```
lm_diag(lm_full_port)
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 1 rows containing non-finite values (`stat_qq()`).
```

```
Warning: Removed 1 rows containing non-finite values (`stat_qq_line()`).
```



First we will cull our models to obtain only significant predictors using the step function.

```
#Commented Out for Brevity
#lm9 <- step(lm_full_math)

#Significant predictors
#schoolsup, romantic, health
#activities, famrel, failures
#G1, absences, G2

#Commented Out for Brevity
#lm10 <- step(lm_full_port)

#Significant predictors
#goout, address, reason
#Dalc, sex, traveltime
#failures, absences, G1, G2
```

Above we have recorded the significant predictors found using the best AIC values for both datasets. It is interesting to note some differences between the two, and they both share *G1*, *G2*, *absences*, and *failures* as significant predictors. We shall create MLR models using these predictors and continue to cull based on significance values.

```r
lm11 <- lm(G3 ~ schoolsup + romantic + health + activities + famrel + failures + G1 + abse
summary(lm11)
```

```
Call:
lm(formula = G3 ~ schoolsup + romantic + health + activities +
    famrel + failures + G1 + absences + G2, data = math_perf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2496 -0.6727  0.2524  0.9769  3.9876

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.899035   0.755638  -3.837 0.000146 ***
schoolsupyes   0.456193   0.275938   1.653 0.099113 .
romanticyes   -0.354339   0.194854  -1.818 0.069783 .
health2       -0.655167   0.372936  -1.757 0.079767 .
health3        0.129031   0.325080   0.397 0.691651
health4        0.007215   0.343320   0.021 0.983245
health5        0.264746   0.302006   0.877 0.381249
activitiesyes -0.353627   0.179389  -1.971 0.049422 *
famrel2       -0.569281   0.765967  -0.743 0.457813
famrel3        0.407781   0.674320   0.605 0.545723
famrel4        0.637127   0.652312   0.977 0.329334
famrel5        1.016422   0.663314   1.532 0.126277
failures1     -1.078230   0.279764  -3.854 0.000137 ***
failures2     -0.600999   0.457972  -1.312 0.190215
failures3     -0.042205   0.476252  -0.089 0.929432
G1             0.186631   0.053408   3.494 0.000532 ***
absences       0.666584   0.085901   7.760 8.06e-14 ***
G2             0.941075   0.046699  20.152  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.76 on 377 degrees of freedom
Multiple R-squared:  0.8588,    Adjusted R-squared:  0.8525
F-statistic: 134.9 on 17 and 377 DF,  p-value: < 2.2e-16
```

```r
lm12 <- lm(G3 ~ goout + address + reason + Dalc + sex + traveltime + failures + absences +
summary(lm12)
```

```
Call:
lm(formula = G3 ~ goout + address + reason + Dalc + sex + traveltime +
    failures + absences + G1 + G2, data = port_perf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6665 -0.5319 -0.0095  0.6421  5.0162

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.530e-02  3.185e-01   0.111  0.91180
goout2            1.067e-01  2.090e-01   0.510  0.61001
goout3            1.534e-01  2.001e-01   0.767  0.44363
goout4            1.891e-01  2.101e-01   0.900  0.36848
goout5           -2.143e-01  2.181e-01  -0.983  0.32610
addressU          2.075e-01  1.152e-01   1.801  0.07213 .
reasonhome       -1.069e-01  1.291e-01  -0.828  0.40804
reasonother      -4.502e-01  1.646e-01  -2.735  0.00642 **
reasonreputation -1.220e-01  1.300e-01  -0.938  0.34843
Dalc2            -1.517e-01  1.302e-01  -1.164  0.24468
Dalc3             2.178e-01  2.070e-01   1.052  0.29317
Dalc4            -7.993e-01  3.129e-01  -2.554  0.01087 *
Dalc5            -4.424e-06  3.208e-01   0.000  0.99999
sexM             -2.002e-01  1.037e-01  -1.931  0.05391 .
traveltime2       5.160e-02  1.106e-01   0.467  0.64102
traveltime3       2.721e-01  1.913e-01   1.422  0.15547
traveltime4       8.214e-01  3.246e-01   2.530  0.01164 *
failures1        -5.048e-01  1.720e-01  -2.935  0.00346 **
failures2        -4.330e-01  3.243e-01  -1.335  0.18224
failures3        -4.602e-01  3.453e-01  -1.333  0.18309
absences          1.684e-01  5.222e-02   3.225  0.00132 **
G1                1.512e-01  3.593e-02   4.208 2.96e-05 ***
G2                8.622e-01  3.402e-02  25.345  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.226 on 626 degrees of freedom
Multiple R-squared:  0.861, Adjusted R-squared:  0.8561
F-statistic: 176.2 on 22 and 626 DF,  p-value: < 2.2e-16
```

Based on the significance values above we will cull any predictor with p-value more than 0.05.

If a categorical variable value has significance less than 0.05, all values of that variable will be included.

```
lm13 <- lm(G3 ~ activities + failures + absences + G1 + G2, data = math_perf)
summary(lm13)
```

```
Call:
lm(formula = G3 ~ activities + failures + absences + G1 + G2,
    data = math_perf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5946 -0.5430  0.2333  0.9944  4.1407

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.04200    0.37636  -5.426 1.02e-07 ***
activitiesyes -0.32737    0.18289  -1.790 0.074231 .
failures1    -1.09101    0.28409  -3.840 0.000143 ***
failures2    -0.87716    0.46316  -1.894 0.058992 .
failures3    -0.09449    0.48454  -0.195 0.845488
absences      0.62079    0.08744   7.100 6.02e-12 ***
G1            0.16263    0.05288   3.075 0.002253 **
G2            0.95004    0.04677  20.311  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.809 on 387 degrees of freedom
Multiple R-squared:  0.8469,    Adjusted R-squared:  0.8442
F-statistic: 305.9 on 7 and 387 DF,  p-value: < 2.2e-16
```

```
lm14 <- lm(G3 ~ reason + Dalc + failures + absences + G1 + G2, data = port_perf)
summary(lm13)
```

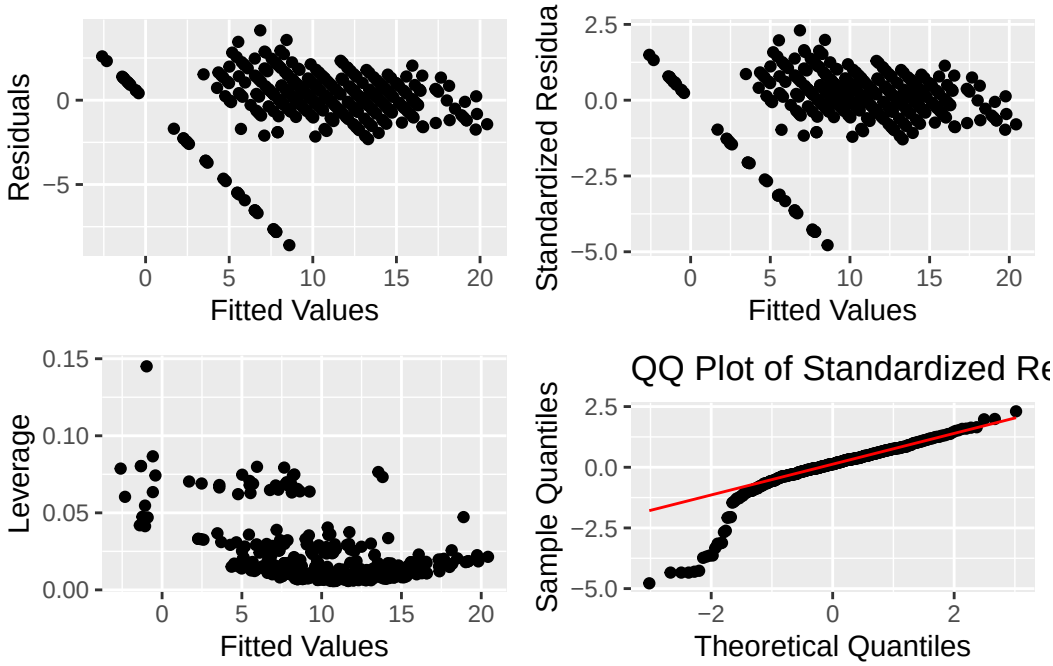
```
Call:
lm(formula = G3 ~ activities + failures + absences + G1 + G2,
    data = math_perf)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.5946 -0.5430  0.2333  0.9944  4.1407

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.04200    0.37636  -5.426 1.02e-07 ***
activitiesyes -0.32737    0.18289  -1.790 0.074231 .
failures1    -1.09101    0.28409  -3.840 0.000143 ***
failures2    -0.87716    0.46316  -1.894 0.058992 .
failures3    -0.09449    0.48454  -0.195 0.845488
absences      0.62079    0.08744   7.100 6.02e-12 ***
G1            0.16263    0.05288   3.075 0.002253 **
G2            0.95004    0.04677  20.311  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.809 on 387 degrees of freedom
Multiple R-squared:  0.8469,    Adjusted R-squared:  0.8442
F-statistic: 305.9 on 7 and 387 DF,  p-value: < 2.2e-16
```

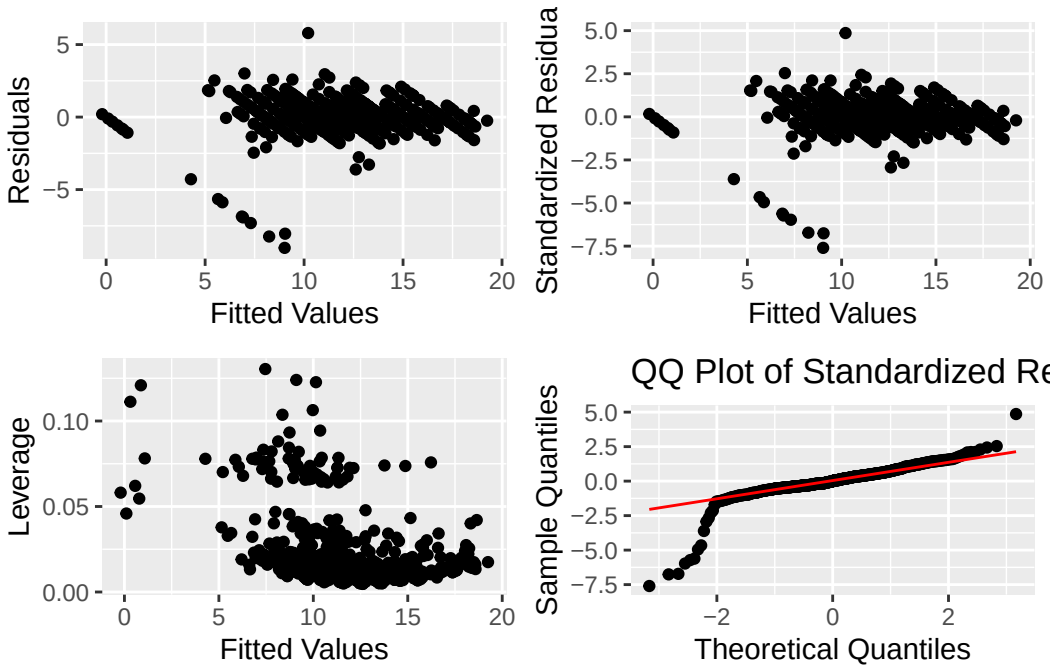
In Figure 5 we examine the diagnostics of both models.

Figure 5

```
lm_diag(lm13)
```


Residuals
Fitted Values
Standardized Residua
Fitted Values
Leverage
Fitted Values
QQ Plot of Standardized Re
Sample Quantiles
Theoretical Quantiles

```
lm_diag(lm14)
```


Residuals
Fitted Values
Standardized Residua
Fitted Values
Leverage
Fitted Values
QQ Plot of Standardized Re
Sample Quantiles
Theoretical Quantiles

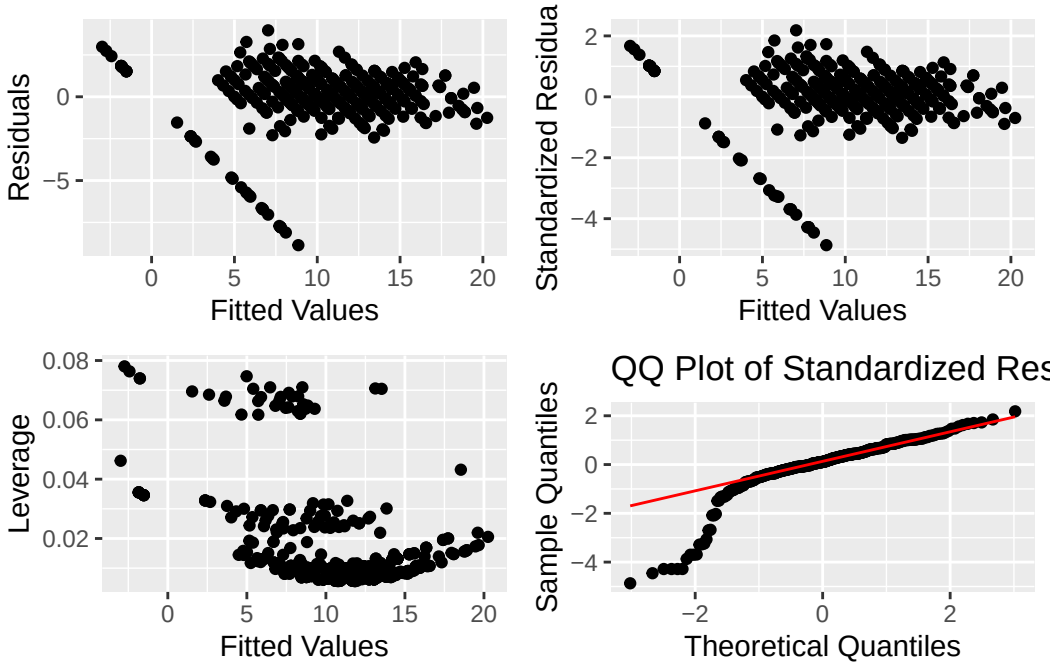
We still have similar problems with the residuals that we had with prior models. We shall investigate if there is multicollinearity present among the predictors by creating a correlation matrix. We can see that *G1* and *G2* are highly correlated so it would appropriate to only include 1 in our next models.

```
cor_matrix(lm13)
```

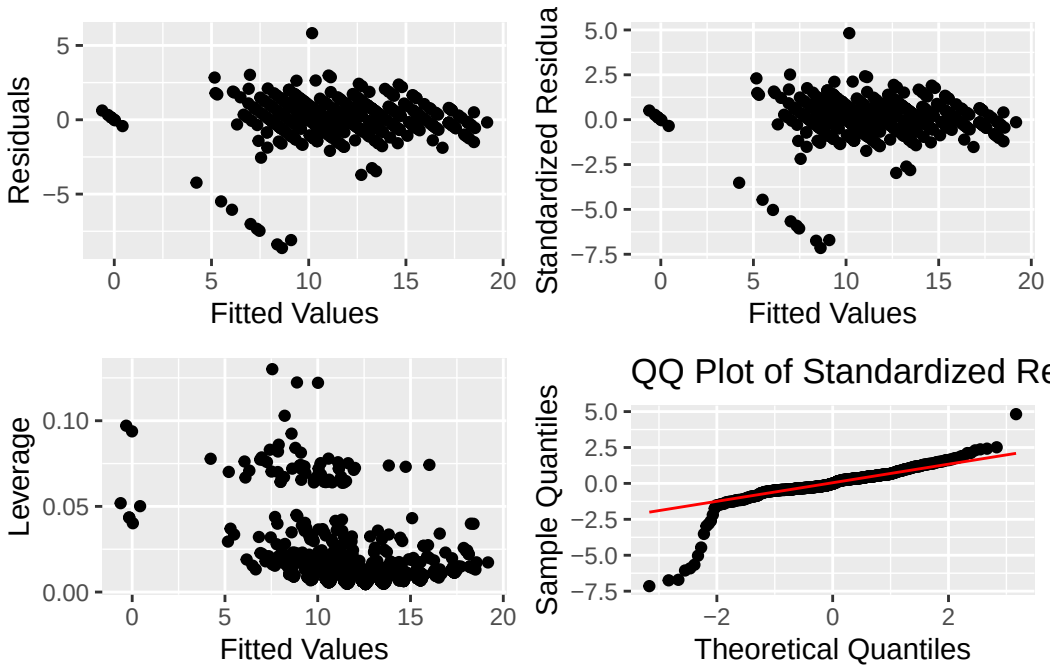
```
                   G3  activities    failures    absences           G1
G3          1.0000000  0.01609970 -0.36041494 0.165693126  0.801467932
activities  0.0160997  1.00000000 -0.06934053 0.037927975  0.057009689
failures   -0.3604149 -0.06934053  1.00000000 0.035366779 -0.354717613
absences    0.1656931  0.03792798  0.03536678 1.000000000  0.009645126
G1          0.8014679  0.05700969 -0.35471761 0.009645126  1.000000000
G2          0.9048680  0.05055171 -0.35589563 0.043675293  0.852118066
                   G2
G3          0.90486799
activities  0.05055171
failures   -0.35589563
absences    0.04367529
G1          0.85211807
G2          1.00000000
```

The models below are likely the "best" we will obtain. The problem with a pattern in the residuals is still present, however in the majority of fitted values there is not a visually obvious pattern. There is a set of outliers that have clearly linear residuals. This may indicate there are a subset of students (who generally get graded lower) which our models may not adequately address. That being said, the Leverage plots indicate few outliers that are also high leverage, providing evidence our models are sound for the average student. The QQ plot further supports this idea, given that the lower outliers fall short of the normal line.

```
lm15 <- lm(G3 ~ activities + failures + absences + G2, data = math_perf)
lm_diag(lm15)
```




```
lm16 <- lm(G3 ~ reason + Dalc + failures + absences + G2, data = port_perf)
lm_diag(lm16)
```



Conclusion

Our models largely agree with the models in the original article. For both classes, the most significant predictors are *G1*, *G2*, *absences*, and *failures*. For our models each class had one additional significant predictor different from the other, *activities* for Math and *reason* for Portuguese. Since these predictors are not highly significant with their respective datasets, it is likely this is not suggestive of a difference between performance in Math and Portuguese and they both may be discarded in pursuit of a more general model of student performance without much concern.

Our results and the results of the article suggest several interesting questions of which interventions may be effective in increasing student performance. Since number of class failures and number of student absences are significant across models and class subjects, these would be useful starting points for interventions. Strategies which reduce student absenteeism and allow students additional opportunities (extra credit, etc.) to avoid class failure could improve student performance.

However it should be noted that when considered individually as SLR models, these predictors offered very little predictive power according to their R^2 values. This could be a result of the how the data was formatted in the two datasets. Failures, in particular was partitioned into 4 discrete values, making it a poor candidate to have a continuous linear relationship with *G3*. The classification and Random Forest models in the article may be better ways to assess the importance of these two predictors than SLR or MLR. All models agree however, when present, *G2* is the most significant predictor of *G3* (Cortez, Silva, 2008).

This suggests that prior student grades are strong predictors of future grades. This is not particularly useful information when designing academic interventions because interventions that improve grades should perform the same across time in the same class. It also may suggest that academic performance is a skill that is independent of other sociological factors attributed to students. Student performance may not be as closely tied to factors that we reasonably suspect influence academic ability (such as support outside of class). Professors who experience a spike in office hour attendance near the end of the semester likely have suspected as much.

There has been debate in the educational community on whether grades are an accurate measure of educational achievement. Studies have shown that secondary student grades do not always correlate with other measures of achievement like standardized test scores (Pollio, Hochbein, 2015). Teachers have reported giving grade credit for nonacademic factors such as participation, behavior, and attitude. Moreover, teachers often differentiate assignments between students based on academic aptitude. Students who are high performing receive assignments tailored to the lesson content, while low performing students may receive assignments tailored to cover material in the previous grade level to remediate perceived deficiencies. Both groups would receive the same amount of grade credit for completing the assignment,

yet the low performing group had not achieved mastery of the lesson content. For any concerned parties, it should follow that interventions designed to improve student grades may not improve actual content mastery.

Care should be taken to examine the complex array of sociological factors and nuances involved in promoting the academic success of students. The datasets used in this paper are from a particular education level from a particular country. Conclusions derived from this data should bear in mind that there is great variation between teachers, grade levels, schools, and nations in how student performance is evaluated. This presents a problem in generalizing conclusions from this dataset outside of a Portuguese secondary education context and without an expert sociological analysis of our results.

In conclusion we present some additional questions for further research.

- Do the conclusions of our models hold in a Portuguese post-secondary context? Primary?
- Do the conclusions hold in analyses of the American education system? British? Chinese? What effects do cultural evaluations of student performance have?
- What sociological factors not present in our dataset reasonably affect student performance or academic ability? Are they highly correlated with prior student grades?
- Are there superior models to SR, MLR, Decision Trees, or Random Forests when predicting student performance?

Our findings help frame the dynamic topic of improving student educational performance and underscore the importance of making data driven decisions in education while keeping broader cultural and sociological factors in mind.

Code Appendix

```
#Relevant libraries
library(ggplot2)
library(plotly)
library(grid)
library(gridExtra)

#Custom functions to avoid redundant coding
#Creates basic scatter plots
gg_basic <- function(data, x, y) {
  plot <- ggplot(data = data, aes(x = {{x}}, y = {{y}})) +
          geom_point()
  return(plot)
}
```

```r
#Creates a 2x2 grid of basic diagnostic plots of a lm object
lm_diag <- function(lm_model) {
  p1 <- gg_basic(lm_model$model, fitted(lm_model), residuals(lm_model)) +
    labs(x="Fitted Values", y="Residuals")
  p2 <- gg_basic(lm_model$model, fitted(lm_model), rstandard(lm_model)) +
    labs(x="Fitted Values", y="Standardized Residuals")
  p3 <- gg_basic(lm_model$model, fitted(lm_model), hatvalues(lm_model)) +
    labs(x="Fitted Values", y="Leverage")

  # Create QQ plot of standardized residuals
  p4 <- ggplot(lm_model$model, aes(sample = rstandard(lm_model))) +
        stat_qq() +
        stat_qq_line(colour="red") +
        labs(x ="Theoretical Quantiles", y="Sample Quantiles") +
        ggtitle("QQ Plot of Standardized Residuals")

  # Arrange all plots into a grid
  return(grid.arrange(p1, p2, p3, p4, ncol=2, nrow=2))
}


#Creates a correlation matrix for predictors in a lm model
cor_matrix_from_lm <- function(lm_model) {
  model_data <- lm_model$model

  # Get names of all variables from the model formula
  all_vars <- all.vars(formula(lm_model))

  # Get names of predictors from the model formula (exclude the response variable)
  predictors <- all_vars[all_vars != response_var]
  pred_data <- model_data[predictors]

  # Attempt to convert factors to numeric if not already
  for(col in colnames(pred_data)) {
    if(is.factor(pred_data[[col]])) {
      pred_data[[col]] <- as.numeric(as.factor(pred_data[[col]]))
    }
  }

  cor_matrix <- cor(pred_data, use = "complete.obs")  # using complete cases
  return(cor_matrix)
```

```
}
```

References

Ajjawi, R., Dracup, M., Zacharias, N., Bennett, S., & Boud, D. (2020). Persisting students' explanations of and emotional responses to academic failure. *Higher Education Research & Development*, *39*(2), 185–199. https://doi.org/10.1080/07294360.2019.1664999

Alyahyan, E., Düştegör, D. Predicting academic success in higher education: literature review and best practices. *Int J Educ Technol High Educ* **17**, 3 (2020). https://doi.org/10.1186/s41239-020-0177-7

Cortez, P., & Silva, A.M. (2008). Using data mining to predict secondary school student performance.

Pollio, M., & Hochbein, C. (2015). The Association between Standards-Based Grading and Standardized Test Scores as an Element of a High School Reform Model. Teachers College Record, 117(11), 1-28. https://doi.org/10.1177/016146811511701106