STAT 306 Group Project

Exploration on the Relationship between Students’ Performance with Grades, Demographic, Social, and School Related Features

Group #26 members:

Eric Li 12804357

Fabiola Grace 58809732

Feifei Yang 84608421

Julian Lozano 91526632

**Introduction**

**Background and motivation**

It goes without saying that education is a very important aspect in everyday life. Without it, society may fall behind in terms of development and growth, across all sorts of different sectors and industries. Although it’s not the final deciding factor by any means, we assess the quality of education that one possesses through the grades they receive.

With that being said, we decided to work with the Student Performance Data Set. This data set contains information about a math course and a Portuguese language course in two different secondary schools (Gabriel Pereira and Mousinho da Silveira) in Alentejo, Portugal. Additionally, this data was collected during the 2005-2006 school year.

For the purposes of this project, we decided to exclude the math course data, and only use information about the Portuguese course. Additionally, we decided to group the two schools together as one, and assume there is no significant difference between the two. The Portuguese language course dataset contains 649 rows (each row represents 1 unique student) with 33 different columns (also known as attributes or variables). Listing all the variables here would be too tedious, so we decided to take only a handful to use for this project, which are G1, G2, studytime, schoolsup, higher, goout, Dalc and absences. The response variable that we’re trying to predict is G3.

Since Portugal at the time was still lacking an effective school information system, most of the data collected here was through paper reports donated by the schools as well as surveys/questionnaires given to the students. More specifically, information about G1, G2, G3, and the number of absences was from the paper reports. Everything else came from the surveys given to students. The survey was designed with predetermined choices that were believed to affect student performance, mainly regarding the characteristics/behaviour of each individual student (how much they studied, whether they had extra school support, alcohol consumption levels, etc). This was information that the school couldn’t track, which is why surveys had to be given to each individual student. The final version of the survey had a total of 37 questions on single-sided A4 sheets. Lastly, some information collected from the surveys had to be discarded. One of the main reasons is simply due to the lack of responses for that particular question. For example, many students did not answer the question about family income (most likely due to privacy concerns), so that variable was ultimately omitted altogether. The other reason is due to the lack of discriminative value. For example, it was discovered that almost all students lived with their parents and had a computer. If almost all observations have something in common, there’s no point in including it in the dataset because it won’t lead to any meaningful analysis.

The motivation behind the analysis is to ultimately get a better sense of the reasons why young Portuguese adults seem to be struggling more than their European peers in terms of academics. Statistics show that Portugal has some of the highest dropout rates in all of Europe. In 2006, 40% of adults aged 18-24 dropped out of school prematurely, while the average in the EU is 15%. Hopefully through this analysis, certain patterns can be discovered, and based on the evidence of things such as correlation, swift changes can be made (whether to students or schools) in order to lower Portugal’s alarmingly high dropout rates. Having sufficient Portuguese comprehension is extremely important, especially considering the fact that it is obviously the native language in Portugal. Moreover, not only is learning proper Portuguese important for future education, but also for living everyday life within that country.

**Data Source and Description**

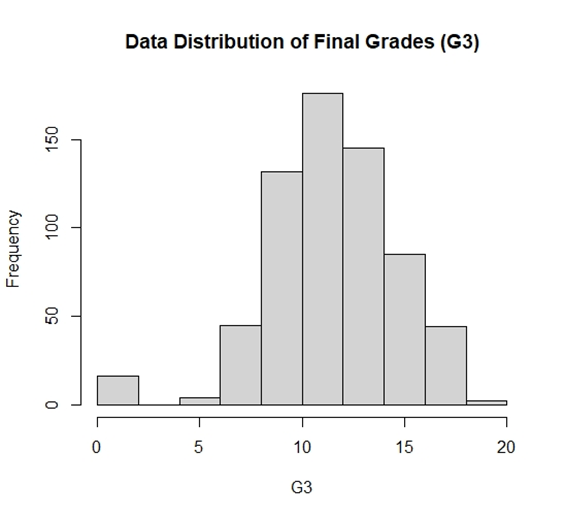
Data was taken from <https://archive.ics.uci.edu/ml/datasets/Student+Performance>.

This student performance data set predicts G3, the final grades of Portuguese language class of students in secondary education of two Portuguese schools by using student grades, demographic, social and school related features (Cortez and Silva, 2008). Data was taken through questionnaires and school reports.

For this project the response variable is G3, the final grade that is issued at the third period. There are a total of 31 predictor variables included in the dataset, however for this project we will be taking 8 predictor variables. These variables are chosen because we believe they are highly correlated with the response variable, G3, final grade of students in Portuguese class. This is especially true in the case of variables G1 and G2. It was told in the data source that the target attribute G3 has a strong correlation with attributes G2 and G1 as G3 is the final grade while G1 and G2 correspond to the first and second period grade. The variables chosen as well as their features are detailed in Table 1.

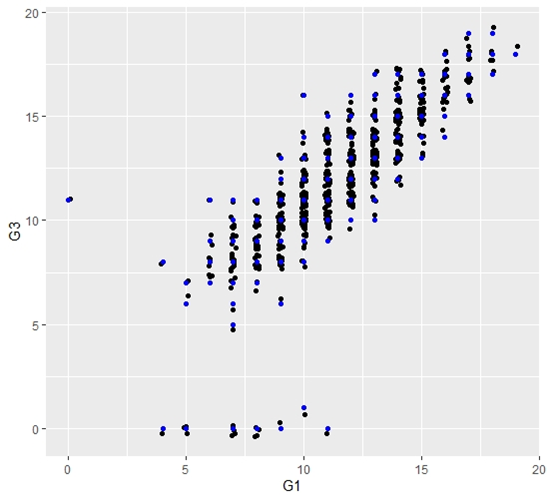
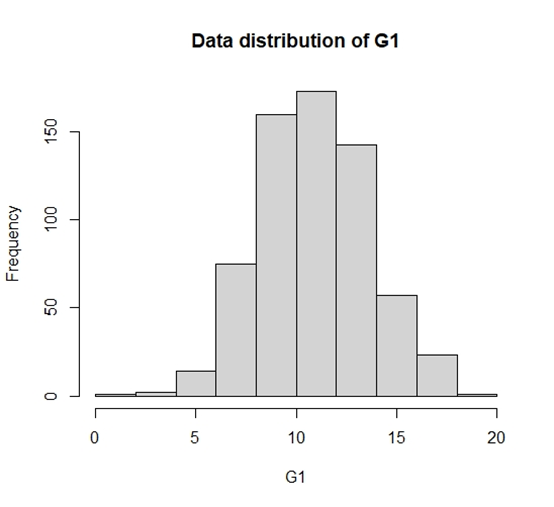
| Name | Description | Data type |
| --- | --- | --- |
| G3 | final grade, issued at the third period, our response variable | numeric: from 0 to 20 |
| studytime (1) | weekly study time | numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours |
| schoolsup (2) | extra educational support | binary: yes or no |
| higher (3) | wants to take higher education | binary: yes or no |
| goout (4) | going out with friends | numeric: from 1 - very low to 5 - very high |
| Dalc (5) | workday alcohol consumption | numeric: from 1 - very low to 5 - very high |
| absences (6) | number of school absences | numeric: from 0 to 93 |
| G1 (7) | first period grade | numeric: from 0 to 20 |
| G2 (8) | second period grade | numeric: from 0 to 20 |

Table 1: Data Description

**Preliminary Data Analysis**

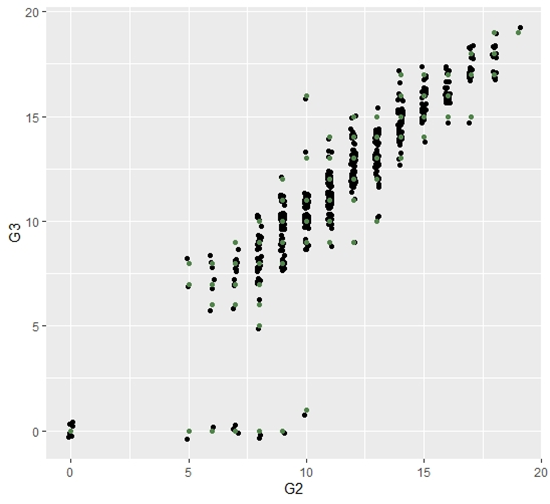
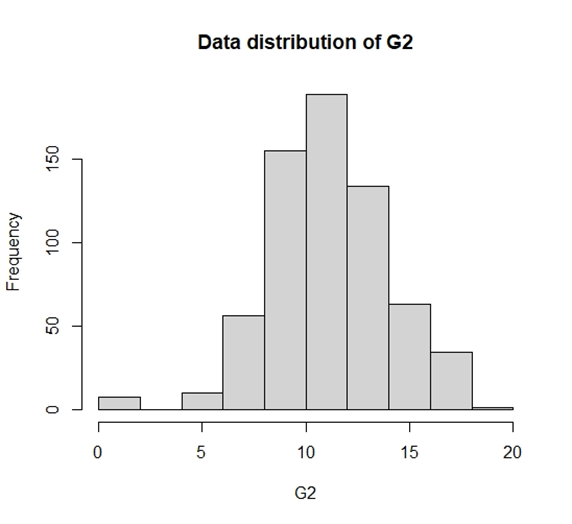
With 33 different explanatory variables in the data, we have chosen to showcase the relationship between each of 4 most significant explanatory variables and the response variable, “G3”, through visualization.

We first take a look at the data distribution of “G3”, which corresponds to the final grade. Observing the plot, we notice that there is an obvious center of distribution around 12. While there are outliers at the left end, the overall shape still suggests an approximately normal distribution. After that, we chose to plot “G3” against each of 4 predictor variables that we considered most important. Because we aimed to have a general image of the relationships between each pair, a scatterplot was employed for its efficiency.



Left: Histogram of G1 (First period grade); Right: G1 vs. G3

Analysis: Our first predictor variable is “G1”, corresponding to first period grade on a 1-20 scale. We can tell from the histogram that the distribution of “G1” is concentrated around 9-11, and that the distribution is also approximately normal as Figure 1, but with no outliers. Then, looking at the scatterplot of “G1” vs. “G3” , there seems to be a strong positive relationship, while some outliers also exist around “G1” = 5-10, “G3” = 0, and around “G1” = 0, “G3” = 11. Overall, the graph is suggesting that first period grade has significant impact on final grade.

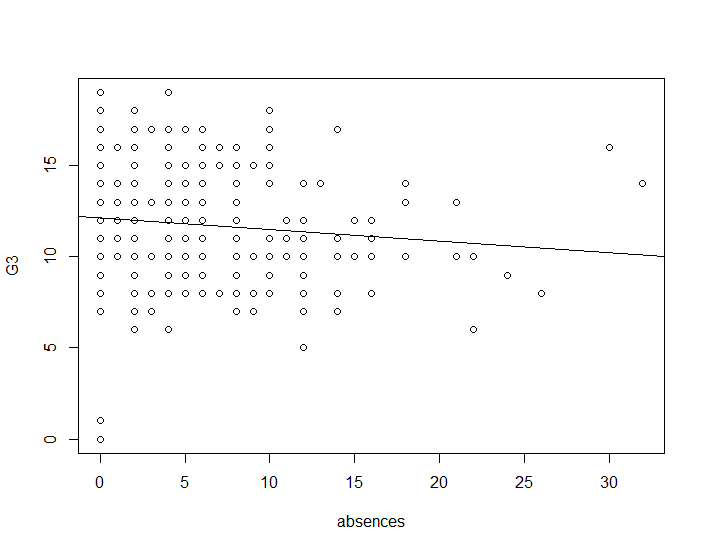
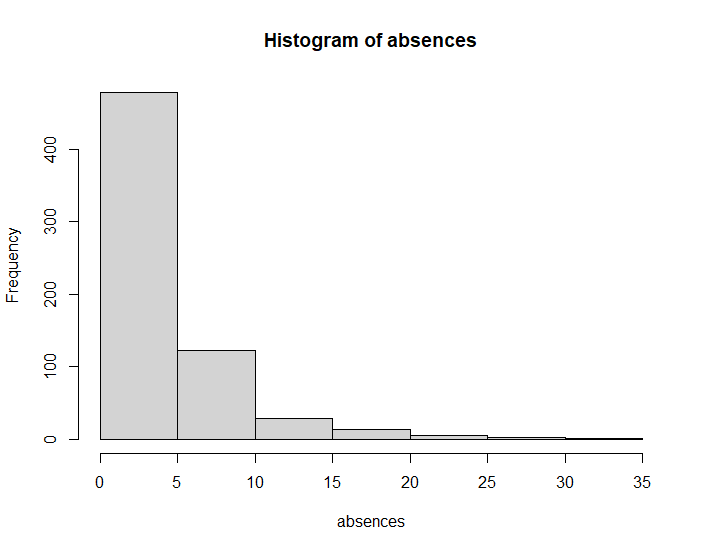


Left: Histogram of G2 (Second period grade); Right: G2 vs. G3

Analysis: The second predictor variable is “G2”, corresponding to second period grade on a 1-20 scale. We can tell from the graph that the distribution of “G2” is concentrated around 11, and that the distribution is also approximately normal as Figure 1 and Figure 2, with outliers at the left end like in data distribution of “G3”. Then, looking at the scatterplot of “G2” vs. “G3”, there seems to be a strong positive relationship, while some outliers also exist around “G2” = 5-10, “G3” = 0, and around “G2” = 0, “G3” = 0, but no outliers around “G3” = 11 like Figure 3. Furthermore, the distribution appears to be more clustered together compared to Figure 3. Overall, the graph suggests that second period grade also has a significant impact on final grade, which is possibly more than that of first period grade.

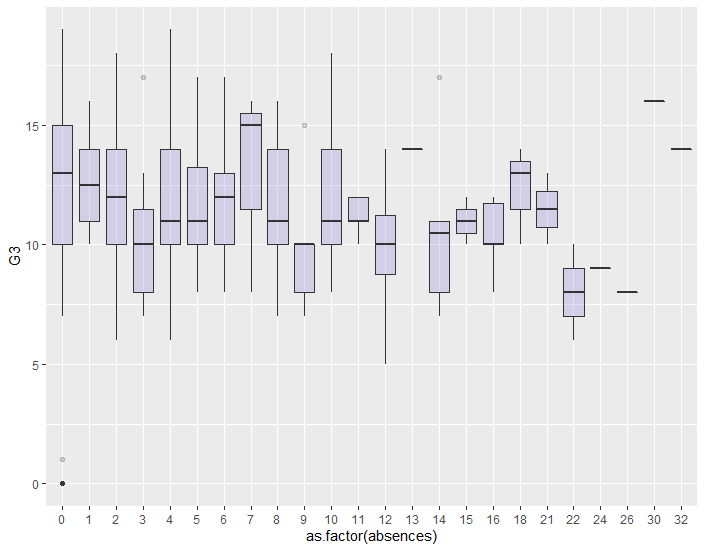
**G3 vs absences**

Left: absences histogram; Right: absences vs G3 scatterplot



Predicted G3 = 12.1388 - 0.06361\*absences

Boxplot:

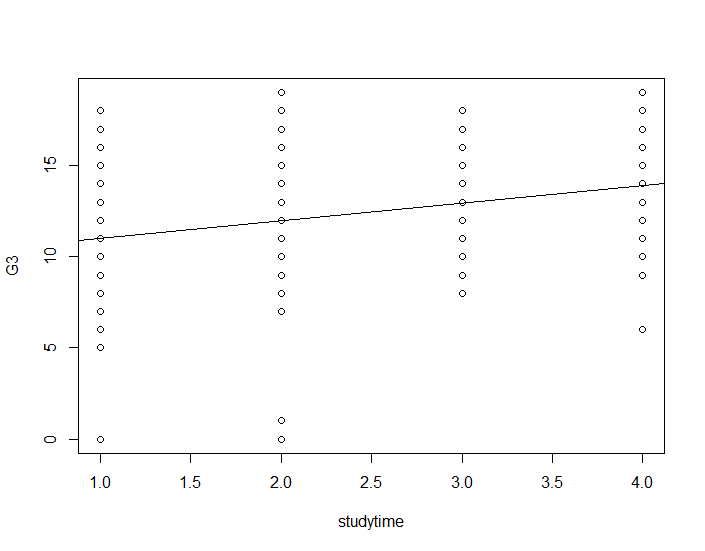
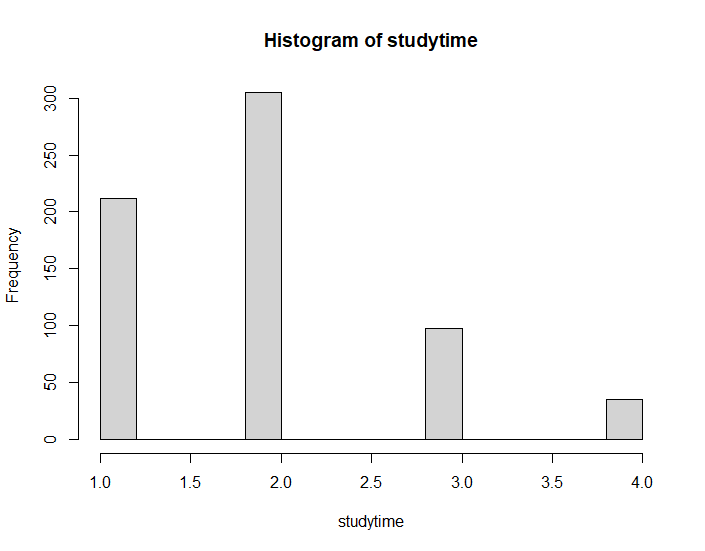


Analysis**:**

There is a slight and general downtrend for G3 as absences increase. Absence may play a part in predicting G3. There are a few outliers that scored well on the G3 despite having a significant number of absences, but for the majority of students, having more absences indicates a lower G3. The IQR is constantly shifting up and down as absences increase, which indicates that there are likely other variables that better predict G3. The other outliers that may be those who had 0 absences but also had a low G3 score. While we don’t know for certain, we speculate that those are the students who dropped the course early in the course. So they have 0 absences for the rest of the school year, but also very low overall grades. Moreover, there is also the possibility for a few students that, just because they showed up to class, does not necessarily mean they were paying attention. We decided to keep these outliers because they ultimately were still part of our data.

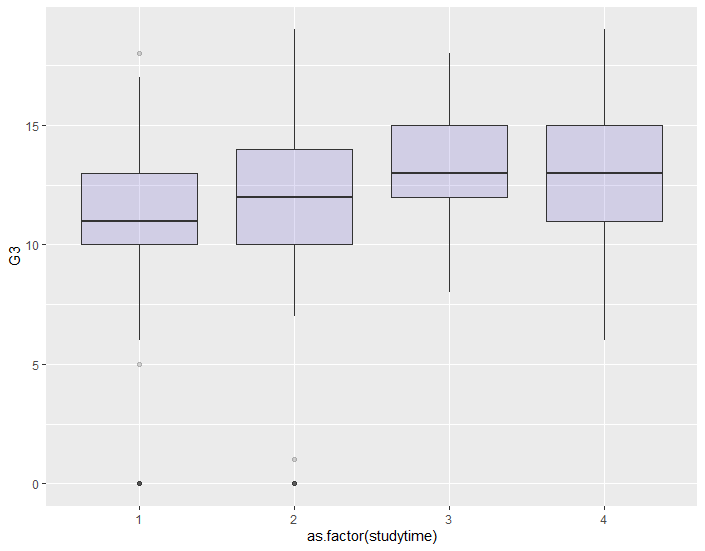
**G3 vs studytime**

Left: studytime histogram; Right: studytime vs G3 scatterplot



Predicted G3 = 10.0278 + 0.9728\*studytime

Boxplot:



Analysis:

There is an uptrend for G3 as studytime (numeric: 1: <2 hours, 2: 2-5 hours, 3: 5-10 hours, or 4: >10 hours) increases. The IQR moves smoothly upwards as studytime increases. Studytime is likely to play a larger part in predicting G3, as students probably learn at a higher concentration level studying than attending class. It is worth noting that the studytime categories do not evenly distribute time studying, i.e. “1’ accounts for (120 - 0) = 120 minutes, “2” accounts for (300 - 120) = 180, “3’ accounts for (600 - 300) = 300 minutes, and “4” accounts for >600 minutes.

**Data Analysis**

**Model Selection**

Using both exhaustive and forward methods, this is done by using the regsubset function in R. We summarize which variables are significant as we fit models with the first variable in x, the first two, the first three, and so on. (Significant variables are denoted by \*)

| Variable | studytime (1) | schoolsup (2) | higher (3) | goout (4) | Dalc (5) | absences (6) | G1 (7) | G2 (8) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model 1 |  |  |  |  |  |  |  | \* |
| Model 2 |  |  |  |  |  |  | \* | \* |
| Model 3 |  |  |  |  |  | \* | \* | \* |
| Model 4 |  |  |  |  | \* | \* | \* | \* |
| Model 5 | \* |  |  |  | \* | \* | \* | \* |
| Model 6 | \* |  | \* |  | \* | \* | \* | \* |
| Model 7 | \* | \* | \* |  | \* | \* | \* | \* |
| Model 8 | \* | \* | \* | \* | \* | \* | \* | \* |

Table 2: Significant variables for each model using forward and exhaustive method

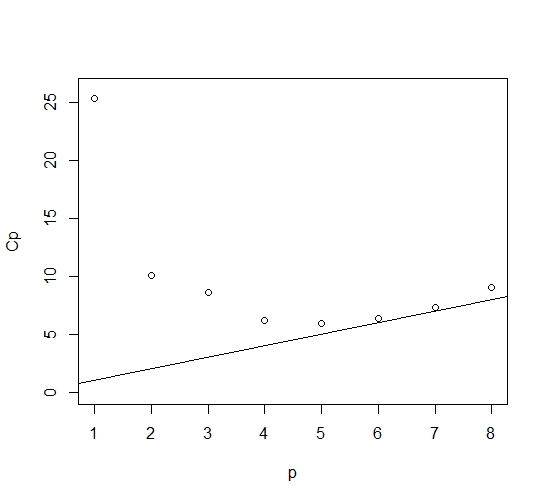
Then we want to see how many variables we should include for each model. To do so we calculated the adj R2,Cp, Residual Square and AIC through R.

|  | **Model** | **adj R2** | **Cp** | **Residual Square** | **AIC** |
| --- | --- | --- | --- | --- | --- |
| **1** |  | 0.8434889 | 25.333516 | 0.8437304 | 2164.271 |
| **2** |  | 0.8472902 | 10.041725 | 0.8477615 | 2149.31 |
| **3** |  | 0.8478684 | 8.562063 | 0.8485727 | 2147.842 |
| **4** |  | 0.8486538 | 6.206719 | 0.8495880 | 2145.476 |
| **5** |  | 0.8489532 | 5.930424 | 0.8501187 | 2145.182 |
| **6** |  | 0.8491002 | 6.305699 | 0.8504975 | 2145.54 |
| **7** |  | 0.8490958 | 7.325689 | 0.8507259 | 2146.548 |
| **8** |  | 0.8489369 | 9.000000 | 0.8508018 | 2148.2170 |

Table 3: Various different Summary Statistics for Each Model

From Table 3, arguably, model 5, is the best model. This model was chosen for several different reasons.

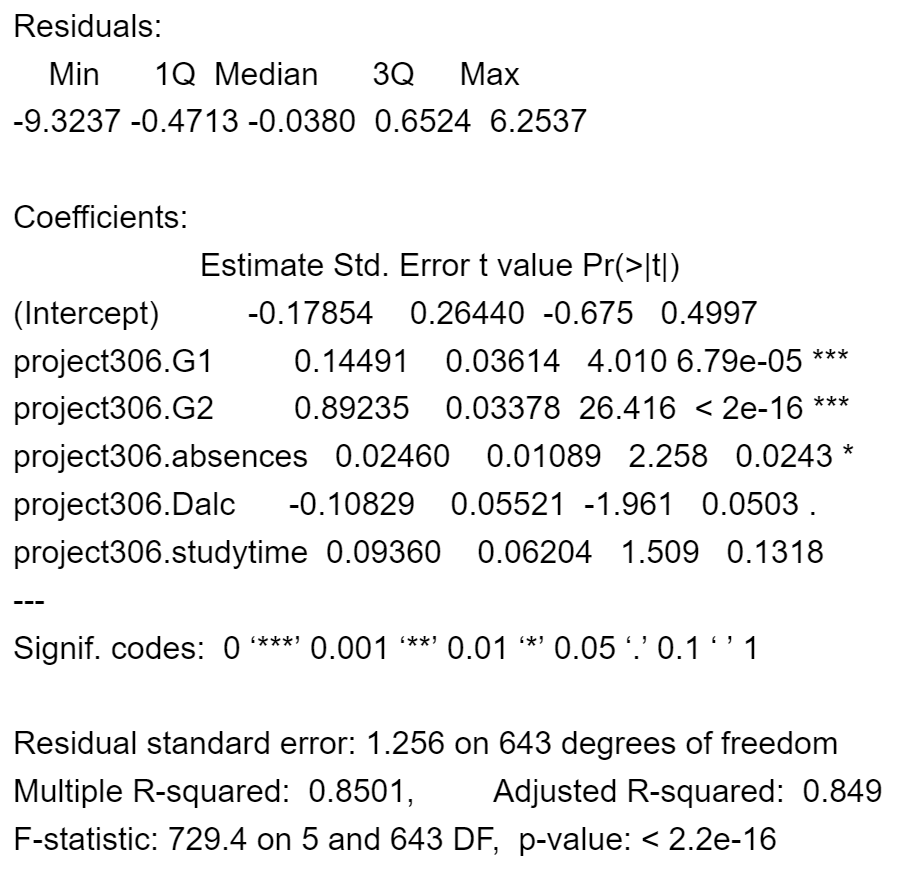
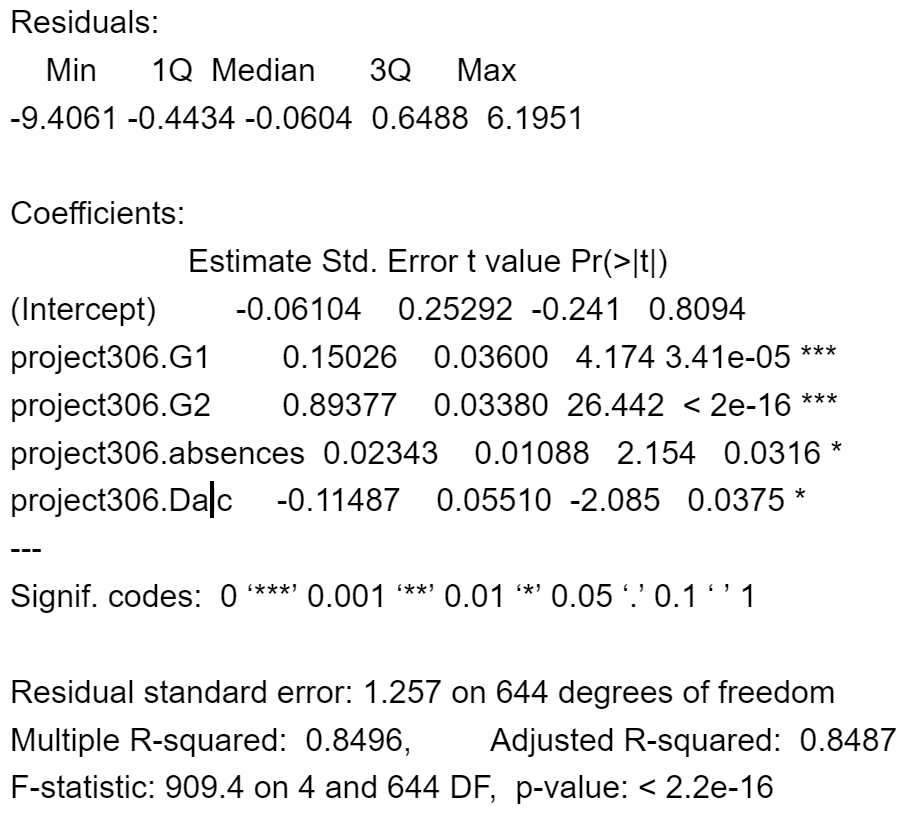
* Through Cp, by picking Cp = p, we narrowed down the better models to model 4 ((Cp) 6.21 = 5 (p)), model 5 ((Cp) 5.93 = 6 (p)), model 7 ((Cp) 7.33 = 8 (p)) and model 8 ((Cp) 9 = 9). After narrowing down, we pick the model with the lowest Cp which is model 5.
* Through AIC, model 5 has the lowest AIC which is 2145.182
* Through adj R2, model 5 rank 3rd, however there is only a difference of -0.000147 between the highest adj R2  on model 6 (adj R2=0.8491002) and model 5 adj R2(adj R2=0.8489532)
* Through residual square, it still has a very high R2 0.8501187. We know that R will always increase, hence we compare this to the full model R2 which is 0.8508018 and the difference is still very small.

After choosing that model 5 is the best, using Table 2, which compute the best variables for each model through forward and exhaustive method, we see that we should include studytime (1), Dalc (5), absences (6), G1 (7) and G2 (8). Hence our final model will be 

The Cp plot on the left shows clearly that models 4-8 have Cp values close to p values, which is the range of models that we took into consideration. We also observe that model 1 has the Cp value that’s furthest from p value, making it to appear like an outlier on the graph.

**Regression analysis**

To closely examine how well our model fits the data, we created linear regression for the 5-variable model, as well as for the 4-variable model as comparison. Looking at the summary of regressions, we observe that the 5-variable model has a relatively higher Adjusted R-Squared value of 0.849, and a relatively smaller Residual Standard Error of 1.256. However, we also notice that the explanatory variable “Studytime” is not significant under the 5-variable model, while the “Dalc” variable could also be seen as insignificant. Overall, a model of 5 variables seems to be a better fit compared to 4 variables. From our model, we can also see that “Absences” has a positive estimate. This is not feasible, since this implies that more absences will increase your grade. This can be caused either by collinearity or that “Absences” have some interaction with another variable. However, since we have excluded some variables from the model, we do not know which variable this may be. In this case, for the completeness of the project, we will include “Absences” as one of our variables.



Left: Regression summary on 5 variables; Right: Regression summary on 4 variables

From our regression summary, the final model is determine to be

**Residual plots and Normal QQ Plots**

Ideally, the residual plot is patternless around zero. Moreover, the Normal QQ plot should follow a straight line.

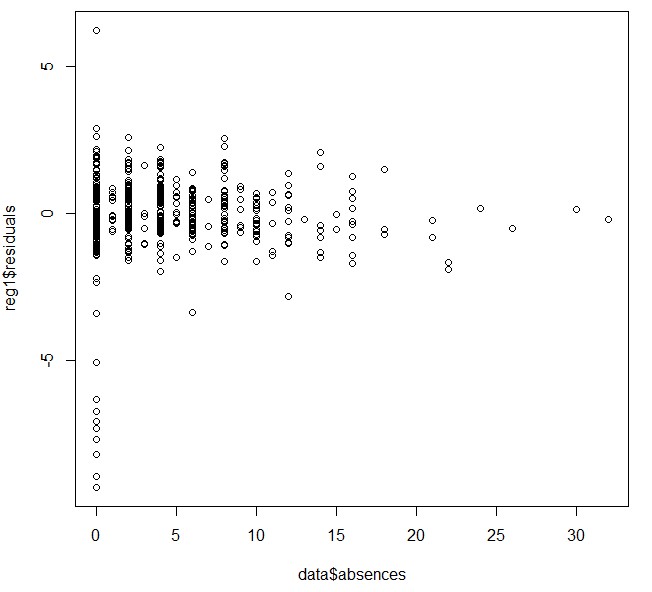
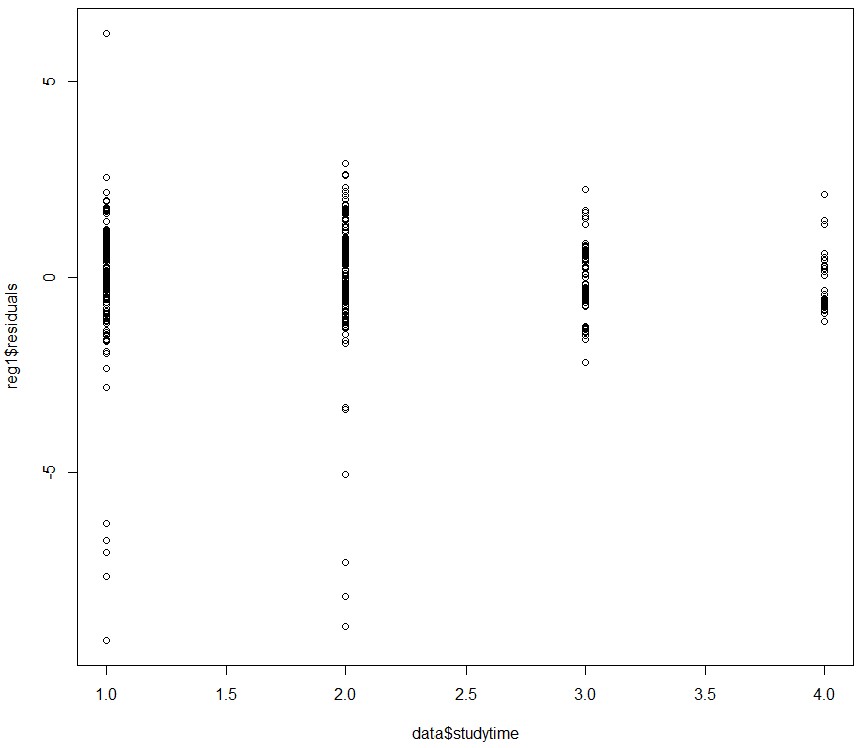
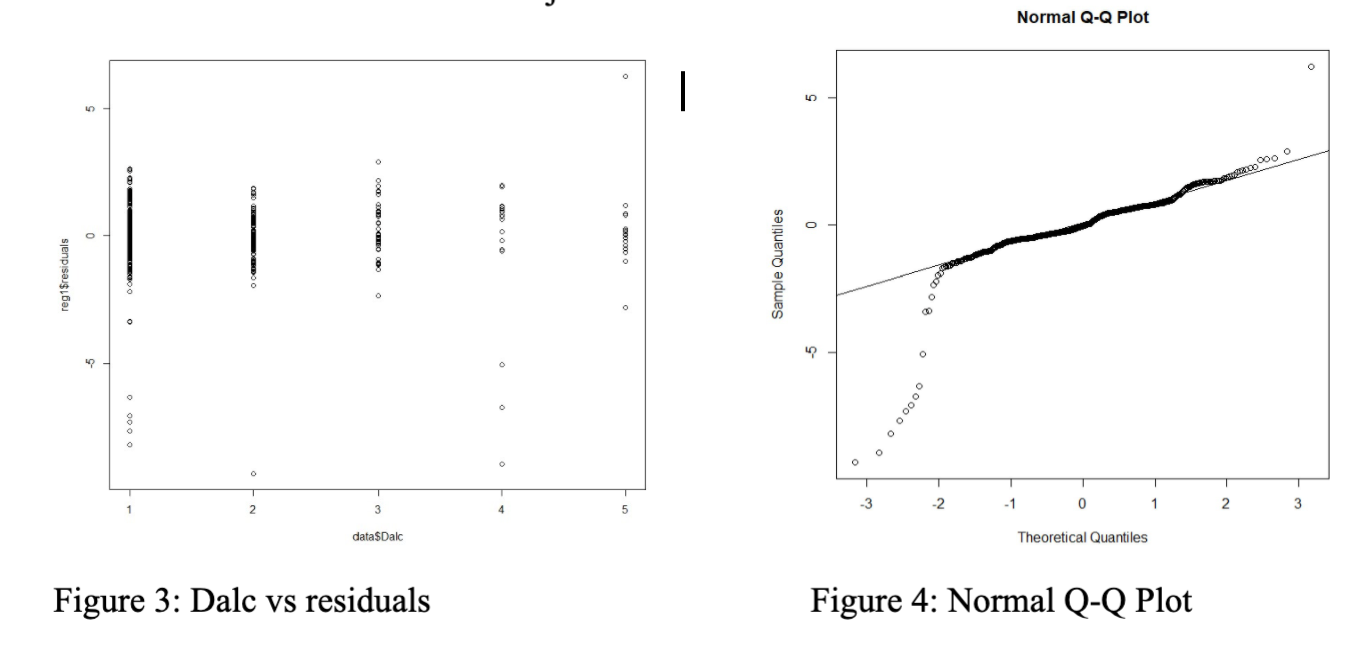
Figure 1: studytime vs residuals Figure 2: absences vs residuals

Figure 1 shows the studytime variable plotted against residuals. It appears that studytime values of 1 and 2 have a higher spread. Figure 2 depicts a residual plot where absences are plotted against residuals, and it appears absences of 0 notably have a much larger residual spread. These could be explained through the presence of outliers. From these two plots, we can see that the data recorded has some issues that need to be addressed. They do not necessarily correlate to each other, and we can come to the conclusion that the data recorded has some major flaws.

In figure 3, where daily alcohol consumption is plotted against residuals, it appears the lowest and 2nd highest alcohol consumption levels have the most spread. Figure 4 suggests the fitted model may not be the most ideal, as there is a presence of a tail at the start of the plot. This introduces the possibility of outliers being present in the data.

**Conclusion**

After careful and thorough analysis based on the data provided, the final grade for the Portuguese course is best predicted by the time students study, how much students consume alcohol on workdays, students’ absences, students’ first period grades, and student’s second period grades. This model was a tough choice to make, as two other models with more predictors had slightly higher adjusted R-squared values, but the chosen model had the lowest Mallow’s Cp and AIC. Furthermore, this model only includes 5 of the 33 possible predictors included in the data set, which should reduce the risk of overfitting in comparison to the two aforementioned models. However, since we are only fitting 5 out of 33 predictors, we see that fitting 4 or 5 variables doesn’t make much of a difference in their adjusted R-squared, Mallow’s Cp, and AIC.

There are a few factors or limitations that may challenge this model. Since the sample size has 649 students, this can introduce significant sampling bias, since there are likely at least 600,000 students currently in secondary school in Portugal[2][3]. This sample accounts for about a thousandth of the secondary school population, thus this model may not be accurate for predicting grades for students outside this school. Other secondary schools in Portugal may apply different grading schemes or are geographically situated in more or lesser affluent regions. It is also possible that students may not have answered truthfully or accurately. It is plausible that students do not accurately recall on the spot how much free time they have or how long they study while answering a questionnaire, so it follows that they formulate their responses from guesstimates. Students can also feel self-conscious about some questions such as alcohol consumption or income, and may overstate, understate or leave out their response(s).

Further analysis also needs to be done to show the effect of collinearity between the explanatory variables as well as identifying outliers. One major problem with one of our explanatory variables, “absences”. As the paper didn’t specify how the data was recorded, we do not know if this variable was recorded as well as it should. Hence, there was a problem with this variable having a positive estimate even though more absences shouldn’t increase G3 final grade. The issue with this variable persists when we plot the residual plot. Zero absences have a higher residual spread which may indicate that even though the students show up in class it doesn’t necessarily mean that their grade would increase. Regarding this variable, we reach the conclusion that this isn’t as informative as it could’ve been.

In conclusion, the model chosen seems reasonable in predicting final grades in Portuguese but remains largely unknown how well it predicts for the entire Portuguese secondary student population.

**References**

1. 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]](http://www3.dsi.uminho.pt/pcortez/student.pdf)

1. <http://uis.unesco.org/country/PT> lists >600,000 portuguese students (~2019)
2. <https://en.wikipedia.org/wiki/Education_in_Portugal> lists >700,000 (unknown yr)

**R Code**

R code for Preliminary Data Analysis

hist(datapor$G3, main = "Data Distribution of Final Grades (G3)", xlab = "G3" ) #plot1

hist(datapor$G1, main = "Data distribution of G1", xlab = "G1" ) #plot2

ggplot(datapor, aes(G1, G3)) +

+ geom\_jitter(position = position\_jitter(width = .1)) + geom\_point(color = "blue") #plot3

hist(datapor$G2, main = "Data distribution of G2", xlab = "G2" ) #plot4

ggplot(datapor, aes(G2, G3)) +

+ geom\_jitter(position = position\_jitter(width = .1)) + geom\_point(color = "#52854C") #plot5

R code for Data Analysis

**# regsubset for model selection**

> s <- regsubsets(project306.G3 ~., data=project306\_subset, method="exhaustive")

> ss = (summary(s))

> ss$adjr2

[1] 0.8434889 0.8472902 0.8478684 0.8486538 0.8489532 0.8491002 0.8490958 0.8489369

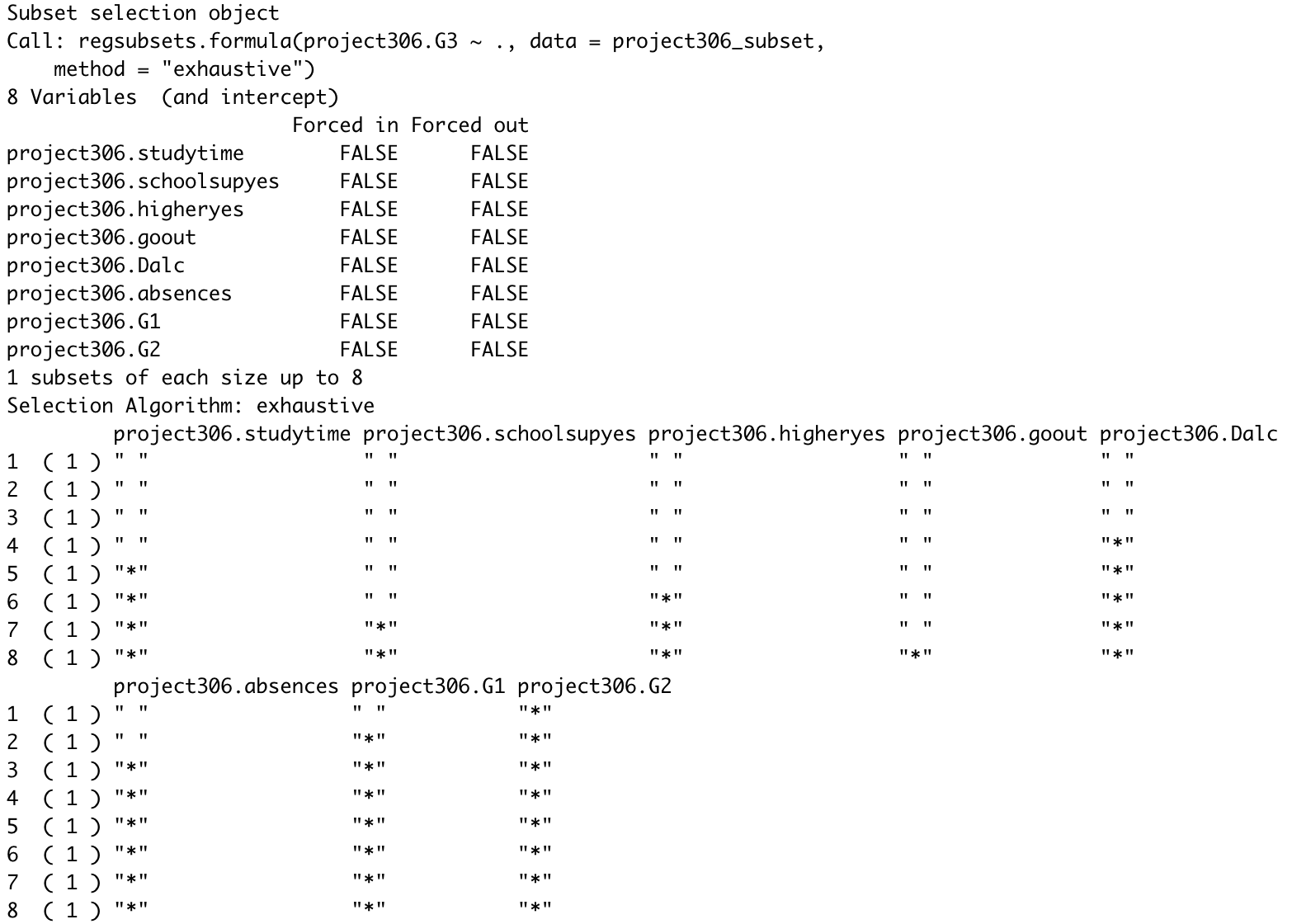
> ss$cp

[1] 25.333516 10.041725 8.562063 6.206719 5.930424 6.305699 7.325689 9.000000

> ss$rsq

[1] 0.8437304 0.8477615 0.8485727 0.8495880 0.8501187 0.8504975 0.8507259 0.8508018

> ss



**# to get AIC for each model**

> AIC(lm(project306.G3 ~ project306.G2))

[1] 2164.271

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2, data=project306\_subset))

[1] 2149.31

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences, data=project306\_subset))

[1] 2147.842

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc, data=project306\_subset))

[1] 2145.476

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc + project306.studytime, data=project306\_subset))

[1] 2145.182

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc + project306.studytime + project306.higher, data=project306\_subset))

[1] 2145.54

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc + project306.studytime + project306.higher + project306.schoolsup, data=project306\_subset))

[1] 2146.548

> AIC(lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc + project306.studytime + project306.higher + project306.schoolsup + project306.goout, data=project306\_subset))

[1] 2148.217

**# Cp plot codes**

> size <-as.numeric(rownames(ss$which))

> plot(size, ss$cp, xlab = "p", ylab = "Cp", ylim=c(0,26))

> abline(0,1)

**# regression codes**

> project306\_reg = lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc + project306.studytime, data=project306\_subset)

> summary(project306\_reg)

> project306\_reg2 = lm(project306.G3 ~ project306.G1 + project306.G2 + project306.absences + project306.Dalc, data=project306\_subset)

> summary(project306\_reg2)

**# Graphing G3 vs absences and G3 vs studytime**

> por <- student.por

> absences <- por$absences

> G3 < por$G3

> library(ggplot2)

> library(dplyr)

> library(gapminder)

> ggplot(por, aes(x=as.factor(studytime), y=G3)) + geom\_boxplot(fill="slateblue", alpha=0.2)

> ggplot(por, aes(x=as.factor(absences), y=G3)) + geom\_boxplot(fill="slateblue", alpha=0.2)

> plot(absences, G3)

> plot(studytime, G3)

> lm(G3 ~ absences)

> lm(G3 ~ studytime)

**# model 5 residuals**

> plot(fitted(model5),resid(model5))

> summary(model5)

Call:

lm(formula = G3 ~ G1 + G2 + studytime + Dalc + absences, data = student.por)

Residuals:

Min 1Q Median 3Q Max

-9.3237 -0.4713 -0.0380 0.6524 6.2537

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.17854 0.26440 -0.675 0.4997

G1 0.14491 0.03614 4.010 6.79e-05 \*\*\*

G2 0.89235 0.03378 26.416 < 2e-16 \*\*\*

studytime 0.09360 0.06204 1.509 0.1318

Dalc -0.10829 0.05521 -1.961 0.0503 .

absences 0.02460 0.01089 2.258 0.0243 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.256 on 643 degrees of freedom

Multiple R-squared: 0.8501, Adjusted R-squared: 0.849

F-statistic: 729.4 on 5 and 643 DF, p-value: < 2.2e-16