

MS-E2112 Multivariate Statistical Analysis

Final Project Report

Student: Jeheon Kim(716954)

1. Motivation

Evolution in data analyzing technology naturally led to the active analysis of the student performance in educational research. Early prediction of student performance can be helpful for both educators and students. For example, it can allow educators to help students to achieve challenging goals and to take corrective measures for students, who are lagging behind, to mitigate their risk of failure.

For this reason, the students' performance dataset is selected and the research will focus on identifying the significant factors on students' performance and presenting good models for the prediction.

For the simplicity and brevity to meet the requirement of maximum 10 pages, the project will be focused on the math dataset and covers only few selected significant variables. Furthermore, only the linear regression (For numeric prediction) and the logistic regression (For categorical prediction) will be done. And lastly, for the better presentation of the analysis, both Python and R will be utilized.

2. Data Description

The data was obtained from the UC Irvine Machine Learning database. The data approach student achievement in secondary education of two Portuguese schools: Gabriel Pereira HS and Mousinho da Silveira HS. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. The questionnaires were answered by 788 students, in total, and 111 answers were discarded due to lack of identification to merge with the mark reports. The data was then integrated into two different datasets:

Mathematics (mat) with 395 observations and Portuguese language (por) with 649 observations. However, the project will be focused on the Mathematics data for the brevity to keep the page limit.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10
...
390	MS	M	20	U	LE3	A	2	2	services	services	...	5	5	4	4	5	4	11	9	9	9
391	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	3	14	16	16
392	MS	M	21	R	GT3	T	1	1	other	other	...	5	5	3	3	3	3	3	10	8	7
393	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	0	11	12	10
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3	2	3	3	3	5	5	8	9	9

395 rows x 33 columns

Figure 1. Top and bottom 5 rows of the student-mat dataset (out of total 395 instances with 33 predictors)

It is provided that the earlier grades (G1: 1st period grades, G2: 2nd period grades) are crucial in predicting the final grade (G3: 3rd period grades).

For this reason, two regression models, with and without “G1” and “G2”, will be compared.


```

> summary(student.mat)
school sex age address famsize Pstatus Medu Fedu Mjob
GP:349 F:208 Min. :15.0 R: 88 GT3:281 A: 41 Min. :0.000 Min. :0.000 at_home : 59
MS: 46 M:187 1st Qu.:16.0 U:307 LE3:114 T:354 1st Qu.:2.000 1st Qu.:2.000 health : 34
Median :17.0 Median :3.000 Median :2.000 other :141
Mean :16.7 Mean :2.749 Mean :2.522 services:103
3rd Qu.:18.0 3rd Qu.:4.000 3rd Qu.:3.000 teacher : 58
Max. :22.0 Max. :4.000 Max. :4.000

Fjob reason guardian traveltime studytime failures schoolsup famsup
at_home : 20 course :145 father: 90 Min. :1.000 Min. :1.000 Min. :0.0000 no :344 no :153
health : 18 home :109 mother:273 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:0.0000 yes: 51 yes:242
other :217 other : 36 other : 32 Median :1.000 Median :2.000 Median :0.0000
services:111 reputation:105 Mean :1.448 Mean :2.035 Mean :0.3342
teacher : 29 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:0.0000
Max. :4.000 Max. :4.000 Max. :3.0000

paid activities nursery higher internet romantic famrel freetime goout
no :214 no :194 no : 81 no : 20 no : 66 no :263 Min. :1.000 Min. :1.000 Min. :1.000
yes:181 yes:201 yes:314 yes:375 yes:329 yes:132 1st Qu.:4.000 1st Qu.:3.000 1st Qu.:2.000
Median :4.000 Median :3.000 Median :3.000
Mean :3.944 Mean :3.235 Mean :3.109
3rd Qu.:5.000 3rd Qu.:4.000 3rd Qu.:4.000
Max. :5.000 Max. :5.000 Max. :5.000

Dalc Walc health absences G1 G2 G3
Min. :1.000 Min. :1.000 Min. :1.000 Min. : 0.000 Min. : 3.00 Min. : 0.00 Min. : 0.00
1st Qu.:1.000 1st Qu.:1.000 1st Qu.:3.000 1st Qu.: 0.000 1st Qu.: 8.00 1st Qu.: 9.00 1st Qu.: 8.00
Median :1.000 Median :2.000 Median :4.000 Median : 4.000 Median :11.00 Median :11.00 Median :11.00
Mean :1.481 Mean :2.291 Mean :3.554 Mean : 5.709 Mean :10.91 Mean :10.71 Mean :10.42
3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:5.000 3rd Qu.: 8.000 3rd Qu.:13.00 3rd Qu.:13.00 3rd Qu.:14.00
Max. :5.000 Max. :5.000 Max. :5.000 Max. :75.000 Max. :19.00 Max. :19.00 Max. :20.00

```

Figure 4. Summary of the student-mat dataset

Important findings from the summary of the dataset:

- There are 208 females and 187 males out of total 395 respondents.
- About 10% of the students' parents were living apart. According to other studies, Mother-Father relationship have effects on academic performance of students as the love, caring and affection from the family are essential requirements for child's emotional stability. This predictor can be strongly correlated to the predictor, "famrel" (quality of family relationship).
- The mean of Mother and Father's education level is around 2.6, which lies between the middle school and the secondary education (High school or Ammattikoulu in Finland). This number can be considered relatively low and it can be backed up by to the source (The Portugal News, 2020) saying that 50% of Portuguese between 25 and 64 did not complete secondary school.
- The number of others as students' guardian is similar to the number of students' whose parents are living apart. There might be a correlation between these two predictors.
- About half of students are participating in extra-curricular activities. This is an interesting predictor as many studies have shown that the students who actively participate in extracurricular activities get a lot of benefits including higher grades, and test scores, higher educational achievements, more regularity in class attendance and higher self-confidence.
- The predictor "romantic" (With a romantic relationship), "Dalc" (Workday Alcohol Consumption), and "Walc" (Weekend Alcohol Consumptions) are interesting subjects to check whether they really affect or not the students' performances as what every parent says.
- The predictor "absence" is expected to be highly correlated with student's performance. According to the summary, the mean value of absence seems relatively low, 5.709 out of 93.
- The variables G1, G2, and G3 are in same format. (Numerical, Range)
Furthermore, the mean value of these three variables are similar at around 10.50.

4.2. Bivariate Data Analysis

First, let's take a look at the correlation heatmap.

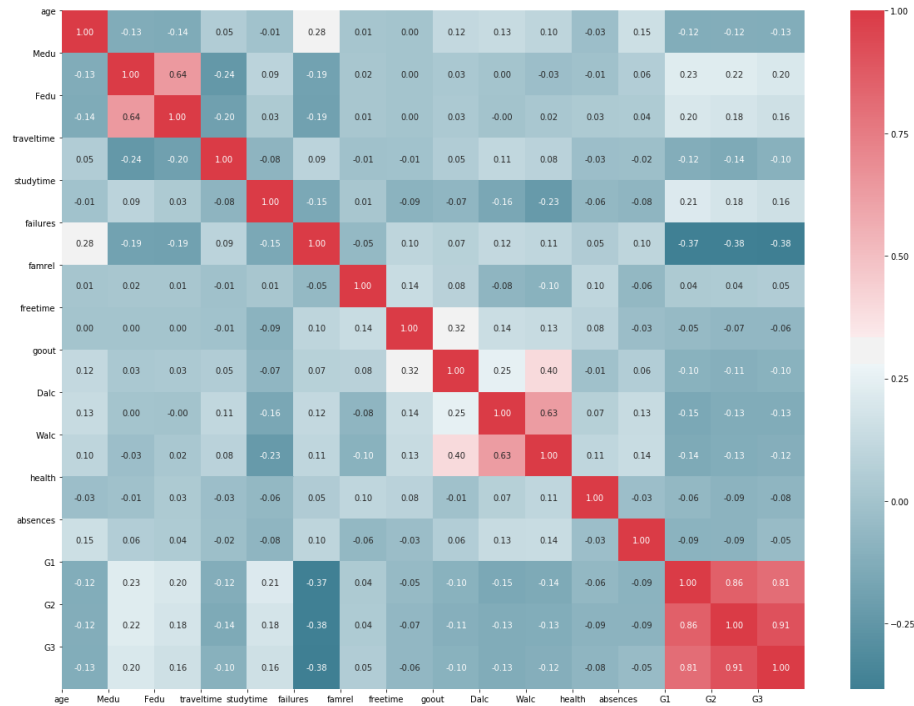


Figure 5. Correlation heatmap of the student-mat dataset

According to the heatmap, there are some high correlations between 3 sets of variables: “Medu” & “Fedu” (Parents’ education level), “Dalc” & “Walc” (Alcohol Consumption), and lastly “G1” & “G2” (Grades in first and second period). These correlations are self-evident as they are similar variables with slightly different condition.

Because there are too many explanatory variables in this dataset, the bivariate data analysis will be focused on a few selected variables that are highly correlated with the response variable, G3.

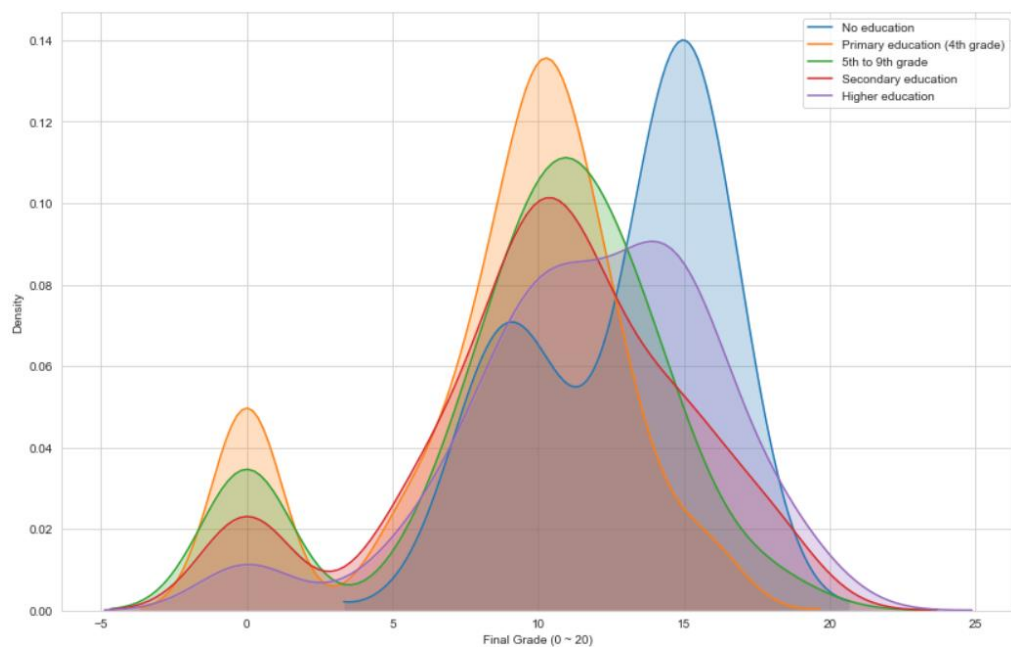


Figure 6. Density plot of Final Grades by the Mother's Education Level

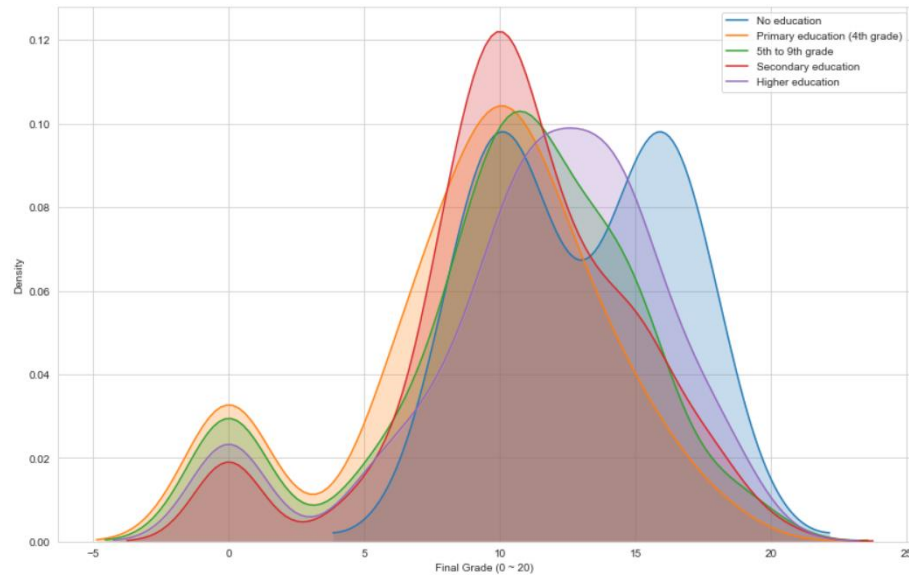


Figure 7. Density plot of Final Grades by the Father's Education Level

The result of the density plot between parents' education level and child's final grade was interesting. Although there is no clear difference in final grades between 3 attributes (Primary Education to Secondary Education), it turned out that students whose parents with extreme education level (Either No education or Higher education) got good grades than others. Good academic performance of students from highly educated parents are proven by many researches. Parents with high education tend to be more interested in their child's education and be able to help them academically. And such parental involvement in children's learning positively affects their academic performance. (Ngure, W. & Amollo, P. 2017).

However, good grades from the students whose parents have no educational backgrounds was unexpected. To interpret it, the deeper investigation is required in the context of the Portugal's cultural backgrounds, however, it is known that there is a correlation between poverty and lack of education in the capitalistic society. And the poverty can act as a double-edged sword on the child's academic performance. In one case, based on the study by Lacour, M. and Tissington, L., the poverty negatively affects child's academic achievement due to the lack of resources available for success. On the other hand, the poverty can affect positively by providing a strong motivation for child to study harder to get away from it. The latter case is, of course, rare in comparison but, surprisingly, it seems that the latter case prevailed for this dataset.

Next is the correlation between the alcohol consumption and the final grades.

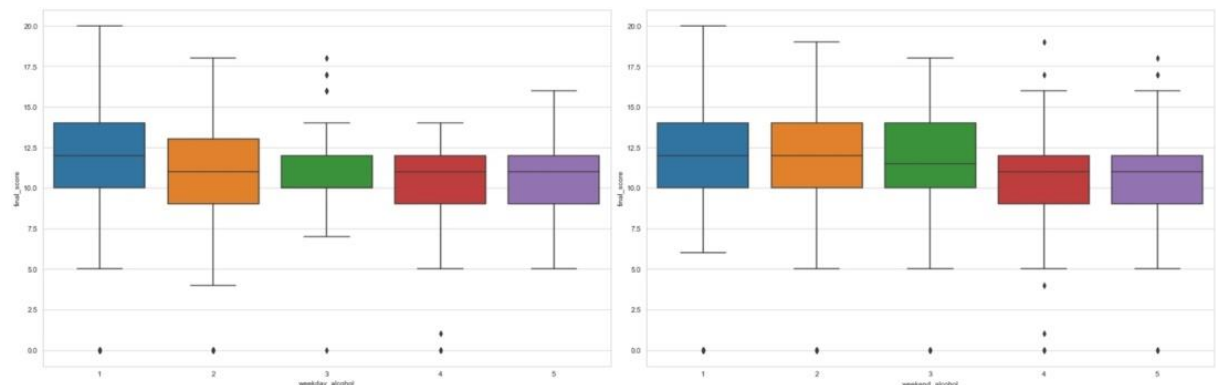


Figure 8. Boxplot of weekday and weekend alcohol consumptions, and final grades

It is clear by both boxplots that students who consumed alcohol less achieved high final grades compared to those who consumes alcohol often. Thus, based on these findings, we can say that there is a correlation between the alcohol consumption and final grades.

The last significant variable found from the heatmap is the “G1” and “G2”. However, it was already given that G1 and G2 are highly correlated with the G3 (final grades). Therefore, we will skip the pair-wise analysis of them and discuss more in the next correlation analysis.

4.3. Multivariate Data Analysis

4.3.1. Correspondence Analysis: Principal Component Analysis

In this section, we will utilize Principal Component Analysis (PCA) to visualize how the variables in the data sets relate to one another and to identify most significant variables to the response variable, G3 (final grades)

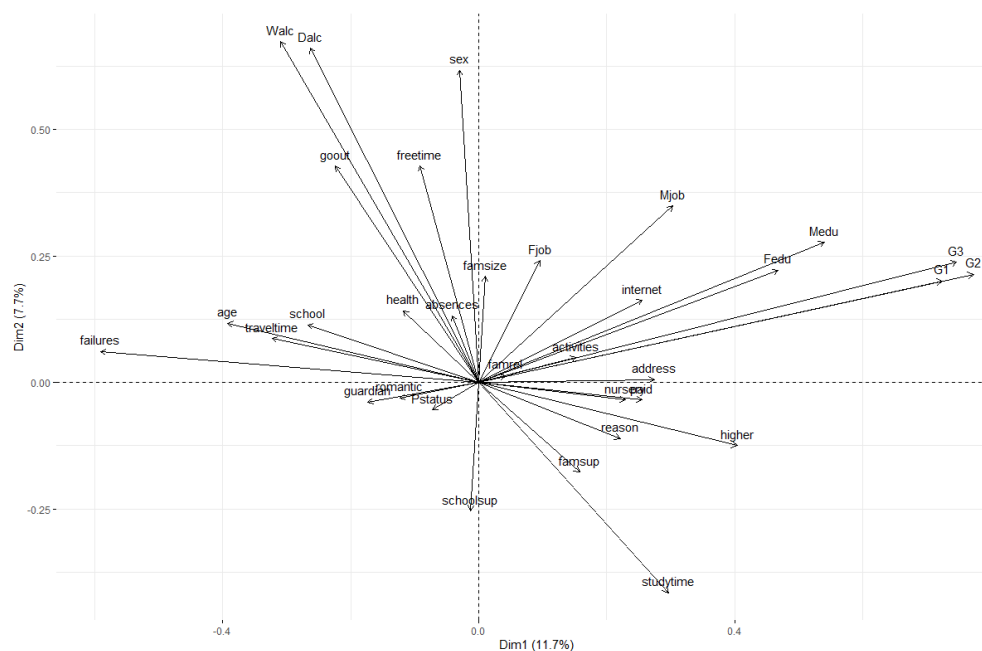


Figure 9. Correspondence Analysis with PCA for student-mat dataset

Less than 90 degrees between variables: ATTRACTION

More than 90 degrees between variables: REPULSION

Exactly 90 degrees between variables: INDEPENDENT

Based on the information on how to interpret the correlation analysis, the response variable, G3, is highly correlated with G1 and G2 as expected. Also, we can see that the education level of both father and mother are grouped together.

If we look at the other groupings, alcohol consumption level (Walc and Dalc) is grouped together with “Free time”, “Go out”, and “Absence”. This makes a lot of sense as students who consumes alcohol throughout school years tend to have a lot of free time, go out and play hook often.

Another group contains “Failures”, “Age”, “Travel times”, and “Romantic”. This group also makes a lot of sense. The older, harder to follow the class. And more time spent on travelling and romantic relationship must affect the final grades negatively.

4.3.2. Correspondence Analysis: Multiple Correspondence Analysis

In this section, we will utilize Multiple Correspondence Analysis (MCA) to explore deeper into the relationship between variables grouped together from the previous analysis.

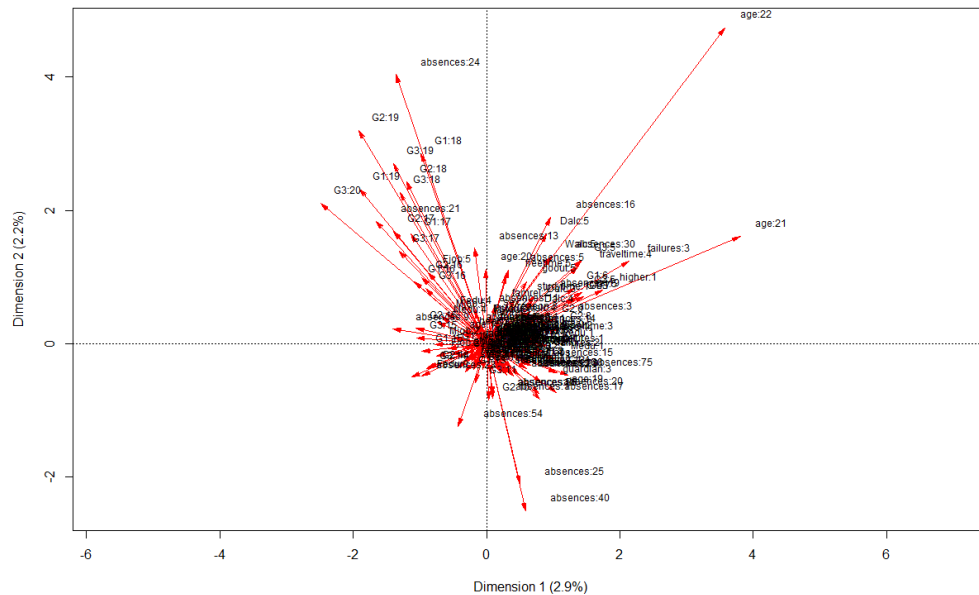


Figure 10. Correspondence Analysis with MCA for student-mat dataset

Although the cumulative proportion of variation explained by 2 components is very low, the groups are in accordance with the previous correlation analysis.

Higher grades from the G1 and G2 are grouped with the higher grades from the G3.

And the highest value of the “Failure” is grouped with maximum age of 22, the highest level of alcohol consumption, “Travel time” and the “Go out” as discovered earlier.

4.3.3. Linear regression

For determining the best linear model, the student-mat file is used as a training set and the student-por file is used as a test set. Let's fit a linear model to all of the variables.

Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.115488	2.116958	-0.527	0.598573
famrel	0.356876	0.114124	3.127	0.001912 **
freetime	0.047002	0.110209	0.426	0.670021
goout	0.012007	0.105230	0.114	0.909224
Dalc	-0.185019	0.153124	-1.208	0.227741
Walc	0.176772	0.114943	1.538	0.124966
health	0.062995	0.074800	0.842	0.400259
absences	0.045879	0.013412	3.421	0.000698 ***
G1	0.188847	0.062373	3.028	0.002645 **
G2	0.957330	0.053460	17.907	< 2e-16 ***

Figure11. Significant variables by the Linear regression model (G1 & G2 included)

From the above summary of the linear fit model, significant predictors are identified. Asterisks mark aside p-value defines significance of value. Lower the value, higher significance.

- ```
*** for the p-value between 0 ~ 0.001
** for the p-value between 0.001 ~ 0.01
* for the p-value between 0.01 ~ 0.05
The significance level is set as 5% (0.05) for this project.
```



Defined 4 most significant predictors are listed as following, in order of importance:

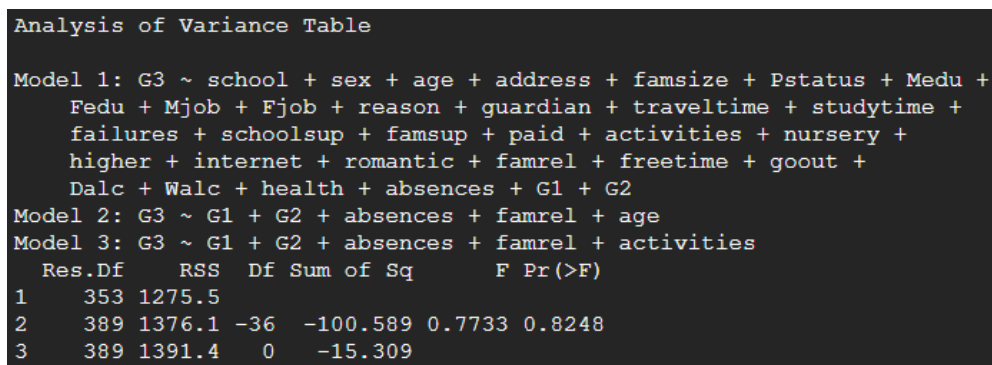
“G2” → “Absences” → “G1” → “Famrel” (quality of family relationships)

Interestingly, the study time predictor is found to be not significant with the p-value of 0.437667 and its coefficient says that the unit increase (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) of study times decreases the final grade, G3.

Since these 4 predictors' values are less than the significance level of 5%, we can reject the null hypothesis and consider the alternative hypothesis that there are linear correlations between these four predictors (G2, G1, Absences, Famrel) and the response variable, G3.

However, we are going to use the 5 most significant variables for the prediction. To choose the last one, we can utilize the ANOVA test. According to the p-values, either “age” or “activities” seem to be suitable for the 5<sup>th</sup> variable. Let's run the ANOVA test and compare the models with

1. All predictors
2. Most significant four + “Age”
3. Most significant 4 + “Activities”



| Analysis of Variance Table                                                                                                                                                                                                                                                                                   |        |        |     |           |        |        |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|--------|-----|-----------|--------|--------|
| Model 1: G3 ~ school + sex + age + address + famsize + Pstatus + Medu + Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime + failures + schoolsup + famsup + paid + activities + nursery + higher + internet + romantic + famrel + freetime + goout + Dalc + Walc + health + absences + G1 + G2 |        |        |     |           |        |        |
| Model 2: G3 ~ G1 + G2 + absences + famrel + age                                                                                                                                                                                                                                                              |        |        |     |           |        |        |
| Model 3: G3 ~ G1 + G2 + absences + famrel + activities                                                                                                                                                                                                                                                       |        |        |     |           |        |        |
|                                                                                                                                                                                                                                                                                                              | Res.Df | RSS    | Df  | Sum of Sq | F      | Pr(>F) |
| 1                                                                                                                                                                                                                                                                                                            | 353    | 1275.5 |     |           |        |        |
| 2                                                                                                                                                                                                                                                                                                            | 389    | 1376.1 | -36 | -100.589  | 0.7733 | 0.8248 |
| 3                                                                                                                                                                                                                                                                                                            | 389    | 1391.4 | 0   | -15.309   |        |        |

Figure 12. The result of ANOVA test of 3 different models

We can see from the result none of combination are promising for prediction. So, just four predictors obtained from the linear regression analysis will be used for the prediction.

Now let's take a look at the coefficient values of each predictors,

Noted that the response variable G3 is a numeric value: from 0 to 20.

The intercept is giving data when all the variables are 0, which means the value when all the measure done without considering any variable is: -1.115488. And we can check that a unit increase of the most significant variable, “G2” increases the G3 value by 0.957330, and a unit increase of the second most significant “absence” increases the G3 value by 0.045879.

Now, let's take a look at the R-squared value

Multiple R-squared: 0.8458, Adjusted R-squared: 0.8279

Figure 13. R-squared value of the linear regression model (G1 & G2 included)

For the Multiple R-squared, it's always between 0 to 1, high value is a better percentage of variation in the response variable that is explained by variation in the explanatory variable. This can tell how well the model is doing to explain the things. (Jaiswal, V. 2018)

Both multiple and adjusted R-squared values for our model are about 0.83, which means that that the model explains approximately 83% of the variance. This is high number and it tells us that the created linear model fits the data in question very well.



And lastly, the overall p-value on the basis of F-statistics:

**F-statistic: 61.2 on 32 and 362 DF, p-value: < 2.2e-16**

Figure 14. F-statistics result and the overall p-value of the linear regression model (G1 & G2 included)

Overall p-value is ' $2.2e - 16$ ', way lower than the 0.05, which indicates that overall model is significant. Thus, we can say our linear model well-fits the data and can well-predict the response variable, G3. However, perhaps it is because the predictors G1 and G2 are very highly correlated with the G3. Now let's take a look at the result without G1 and G2.

|          |          |         |        |          |     |
|----------|----------|---------|--------|----------|-----|
| failures | -1.55274 | 0.32158 | -4.829 | 2.03e-06 | *** |
| goout    | -0.57742 | 0.22246 | -2.596 | 0.00983  | **  |
| sex      | 1.16667  | 0.49763 | 2.344  | 0.01959  | *   |
| romantic | -1.05758 | 0.46764 | -2.262 | 0.02432  | *   |

Figure 15. Significant variables by the Linear regression model (G1 & G2 NOT included)

Without critical predictors, G1 & G2, different variables are chosen as significant. "Failures", "Sex", and "Romantic" have negative coefficients which means that the unit increase of these variables decreases the final grade by the amount of those coefficients.

**Residual standard error: 4.136 on 364 degrees of freedom  
Multiple R-squared: 0.2469, Adjusted R-squared: 0.1849  
F-statistic: 3.979 on 30 and 364 DF, p-value: 1.371e-10**

Figure 16. Summary of the linear model without G1 & G2

One thing to note is that R-squared value greatly dropped compared to the model with G1 & G2. Now the model explains only about 24% of the variance of the data. This result supports the fact that G1 and G2 are crucial predictors for the response variable, G3.

Now we will predict the response variable, G3 with these linear regression models.

Noted that the observations are filtered by  $G3 \neq 0$ , as the prediction model with all observation turned out to be heavily clustered when G3 is 0.

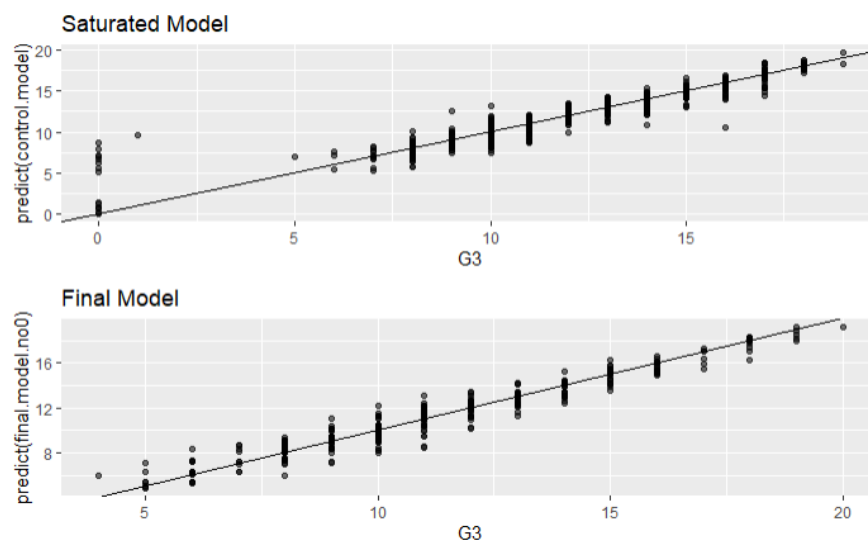


Figure 17. Visualization of the linear regression models: All predictors vs Most significant 5 predictors

The saturated model is in which all 33 predictors are included, and the final model is which only 5 most significant variables are included. In the graph, the straight line represents a perfect model and it is clear that our model is in accordance with it. Thus, we can say that our model is considerably accurate and did a great prediction of the test-set.

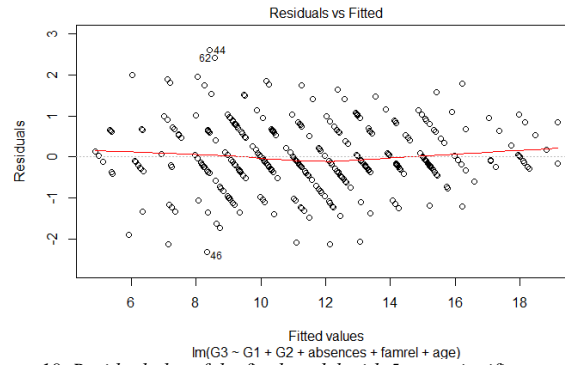


Figure 18. Residual plot of the final model with 5 most significant predictors

Also, the residual plot looks great. Firstly, no pattern. Secondly, the variance does not seem to change with the change of x. And lastly, residuals are evenly distributed on both sides of zero.

And now, let's take a quick look at the prediction result without G1 & G2 predictors.

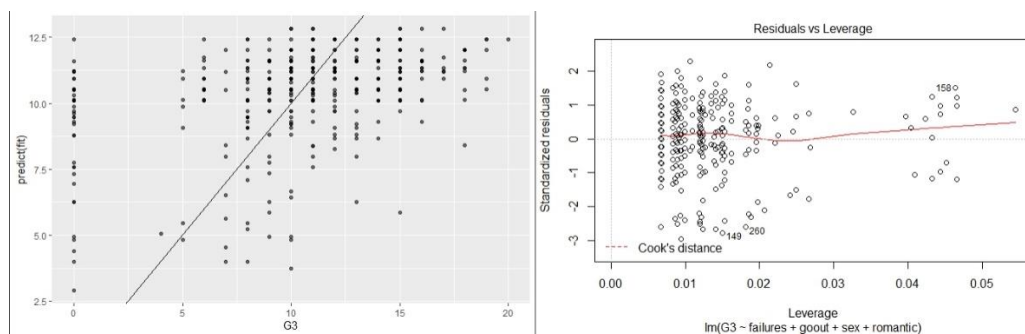


Figure 19. Prediction result of the model without critical predictors G1 & G2

It is clear that the model does not fit the data well. It corresponds to its low R-squared value which indicated that the model explains only about 24% of the variance of the data. From this observation, we can conclude that two variables G1 and G2 are critical to predict G3.

#### 4.3.4. Logistic regression

Because the linear regression assumes all independent variables are numerical, for categorical variables, we need to use a different model. In this project, we will use the Logistic Regression.

Since the response variable G3 is in numerical format. We first need to transform it into the categorical variables. To do this, we divide the G3 into two groups. First group with binary value 0 contains students with G3 (final grades) value less than 10. And another group with binary value 1 contains students with G3 value more than or equal to 10.

```

testing_result
model_pred_result 0 1
0 57 3
1 16 122
> mean(model_pred_result == testing_result)
[1] 0.9040404

testing_result
model_pred_result 0 1
0 15 6
1 49 128
> mean(model_pred_result == testing_result)
[1] 0.7222222

```

Figure 20. The result of logistic regression of two models in confusion matrix

Same as the linear regression analysis, one model with 5 most significant predictors (including G1 & G2), and another model with 4 most significant predictors (without G1 & G2) were used.

As we can confirm from the confusion matrix, the model with G1 & G2 predictors made a better prediction with the accuracy of 90%, while the other model without G1 & G2 resulted in lower accuracy of 72%. However, the model without critical predictors G1 & G2 predicted relatively well, compared to the result of the linear regression on the numerical variables.

## 5. *Conclusion*

From the analysis, the project can provide following answers to the predefined research questions:

1. As expected, the predictors G1 & G2 (First and Second period grades) are proved crucial to predict the response variable, G3 (Third period [Final] grade). In addition, the absence (the number of absence), famrel (quality of family relationships), and the age were selected as significant. However, it was different when G1 & G2 were excluded. Instead, failures (number of past class failures), goout (going out with friends), sex (student's sex), and romantic (with a romantic relationship) were chosen as significant. One thing to note is that most of the latter (without G1 & G2) selected variables had negative coefficients which means their unit increase decreases the final grades. On the other hand, most of the former (with G1 & G2) selected variables had positive coefficients.
2. 3. For the prediction of G3 in numerical format, the linear regression model predicted well with good accuracy. Although the accuracy and the amount of explained variance dropped significantly when two key predictors (G1 & G2) were not included. Therefore, we can say that G1 & G2 are crucial to predict the numerical value of G3. It is perhaps natural as they are very similar. (Range of the values, Mean value at around 10.50 and etc.) For the logistic regression, we divided the G3 numerical value into two categorical groups (less than 10 points, and more than or equal to 10 points) and both models (with and without G2 & G3) made a good prediction. With G2 & G3, the accuracy was almost 90% and even without them the categorical prediction accuracy was about 72%, which is fairly good number.

## 6. *Critical Evaluation of the Analysis*

Although created models turned out to be quite accurate and fit the data well, there clearly are things to be improved in this project. Especially, the limited number of prediction model (Linear and Logistics) must be compensated by other models such as, Ridge and Rasso regression, the K-means clustering or the random forest, to find an optimal model for the data.

In addition, besides the findings listed above, the research was able to identify many interesting relationships between some predictors and the response variable. Although all of them couldn't be covered in this project due to the limitation, but they worth further analysis.

Though the author had to consult lecture materials and many other sources online, the interpretation of the results was done by himself and the author could learn a lot from this project. Especially, it was great opportunity to review the lecture coming to an end and for us to try out what we've learnt.

The one problem of the project worth mentioning is the violation of the maximum pages by 2.

It is caused by the relatively big font size (12) and the figures' sizes. However, for the better readability and understanding of the project, it was inevitable to set them large. Hope it is not a big problem.

## 7. *References*

Jaiswal, V. 2018. Interpret R Linear/Multiple Regression output. Medium.

<https://medium.com/analytics-vidhya/interpret-r-linear-multiple-regression-output-lm-output-point-by-point-also-with-python-8e53b2ee2a40>

The Portugal News. 2020. 50% of Portuguese between 25 and 64 did not complete secondary school

<https://www.theportugalnews.com/news/50-of-portuguese-between-25-and-64-did-not-complete-secondary-school/53517>

Ngure, W., Amollo, P. 2017. Influence of Parents Education Level on Academic Achievement of Unity Preschool Children in Embakasi, Nairobi County. Kenyatta University.

<http://www.researchpublish.com/download.php?file=Influence%20of%20Parents%20Education%20Level-4466.pdf&act=book>

Lacour, M., Tissington, L. 2011. The effects of poverty on academic achievement, Southern Arkansas University.

[https://academicjournals.org/article/article1379765941\\_Lacour%20and%20Tissington.pdf](https://academicjournals.org/article/article1379765941_Lacour%20and%20Tissington.pdf)

Frigaard, M. 2019. Diagnosing the accuracy of your linear regression in R. Storybench.

<https://www.storybench.org/diagnosing-the-accuracy-of-your-linear-regression-in-r/>

Mehta, R. 2019. Analyzing Student Performance Through Data Mining Model.

<https://github.com/raghavm23/Analyzing-Student-Performance-Using-Data-Mining-Techniques/blob/master/Analyzing%20Student%20Performance.pdf>

Alice, M. 2015. How to perform a Logistic Regression in R

<https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/>

Prabhakaran, S. Logistic Regression

<http://r-statistics.co/Logistic-Regression-With-R.html>