Students' Alcohol Consumption Impacts on Academic Results on the Example of a Secondary Scool Maths Class in Portugal

Group 2

13 October 2016

Content:

=======

1. Introduction

Recent events in the world economy have lead to the decline in economic performance in many countries, amongst which is the decline of economic growth in the European Union (EU). As a response to these changes, the EU has made it a priority to reverse this trend through investments that would provide new jobs and create a better educated workforce. Hence, the European Commission (EC) has set up a 'Strategy for smart, sustainable and inclusive growth' (European Comission, 2010), where a crucial part of investment is to contribute to the education of the next generations: The aim is to reduce the share of early school leavers to below 10% and ensure that at least 40% of the younger generation obtain a tertiary degree.

The implementation of this strategy would use up siginificant amounts of resources; therefore, ensuring its success is crucial. However, some factors affect the effectiveness of this strategy by lowering the academic performance of students, despite the investments going into the education system. In an effort to identify one of those factors, this paper looks into the effect of alcohol consumption on the academic performance of students. The consequences of high alcohol consumption is a field with a vast amount of literature and scientific studies, all of which point to the health, social and economic effects of alcohol consumption. For example, a study by Keng and Huffman (2010) has found a relationship between alcohol abuse and poor labor market outcomes. The choice of alcohol consumption as hindering factor the success of the EC's strategy was therefore motivated to test whether these consequences extend to academic performance as well.

The dataset used for this paper contains information about students in Portugal at the secondary school level. Portugal is among the world's highest alcohol consuming countries worldwide, ranked as the tenth highest alcohol consuming country per capita (Statitsa, 2016). In addition, Portugal is ranked amongst the top five OECD countries with the largest percentage of adults who have not attained secondary education in 2013, with a high percentage of 65% compared to an OECD average of 25%. This is coupled with the fact that unemployment affected adults without a secondary education more severely than it did those with tertiary education (OECD, 2013).

=======

2. Dataset Description and Descriptive Statistics

The dataset examined provides sample data of two secondary schools in Portugal, providing a total of 395 objects (students) and 33 variables (items on information on each student). A list of all variables and their description can be found in Appendix A. The data can be grouped into five areas of information:

a. General Information General information about the students include gender, address and age. From the descriptive statistics (section 3) we can derive that the age from the students in this dataset ranges from 15 to 22 years with an average of 16.7 years. Out of the total of 395, there are 52.7% female and 47.3% male students. Most students would live in an urban environment, while approx. 20% live in an rural environment.

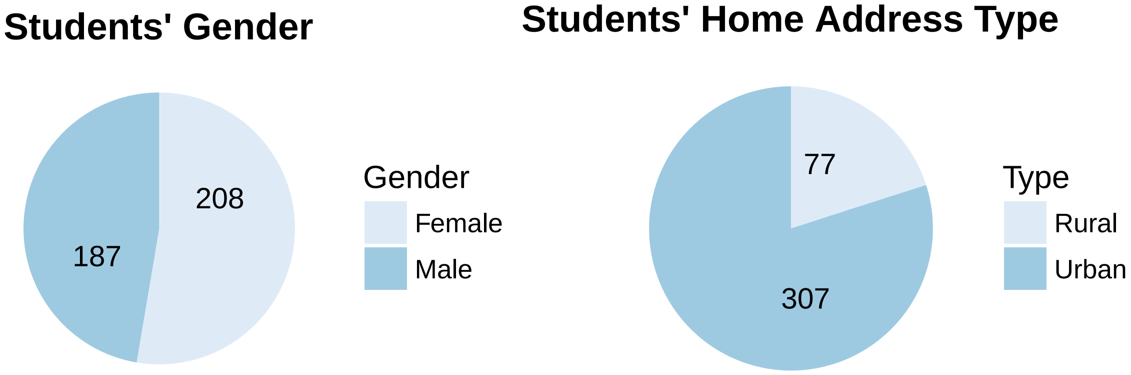


Figure 1. General Information

b. Family
Information in the students families includes the family size, educational background of their parents and their jobs, and the student's guardian.

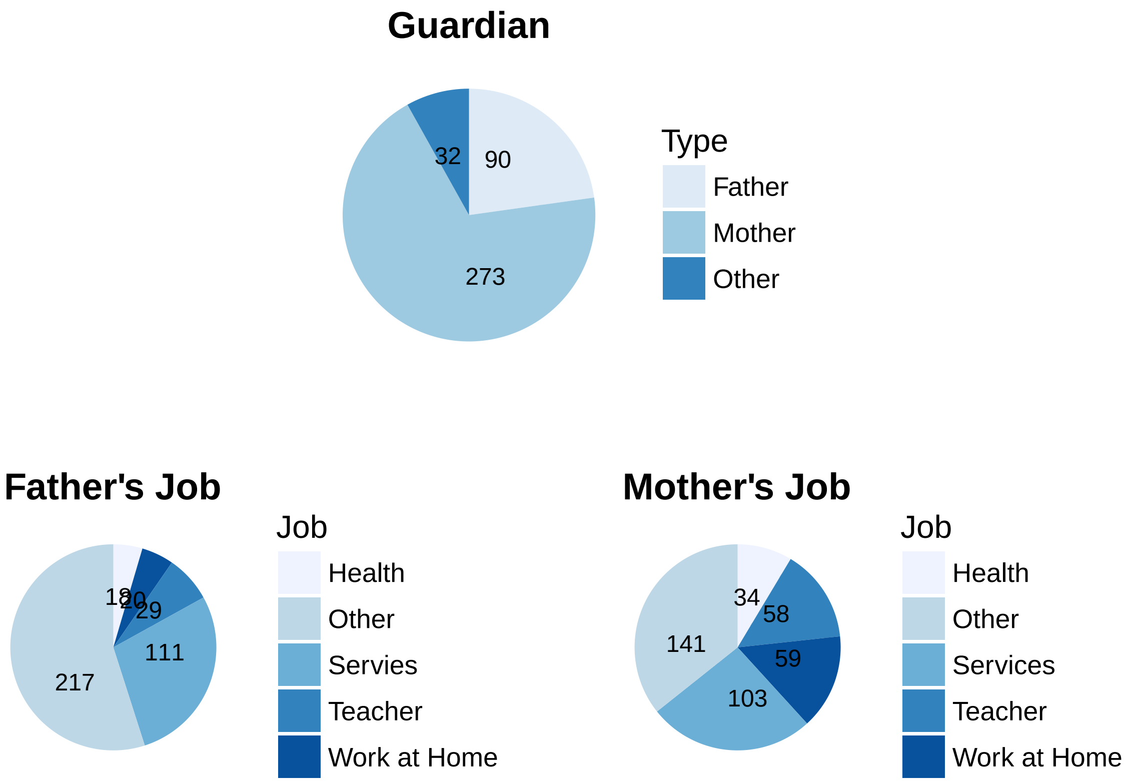


Figure 2. *Family*

c. Education The education section of this dataset covers the split between the two schools observed, why each student has chosen that particular school, and how long it takes each student to get to school. In addition, it provides information on the student's study time, extra educational support or classes taken and if the student is involved in extra curricular activities. Finally, there is data on their previous education and aspirations on whether to go into higher education.

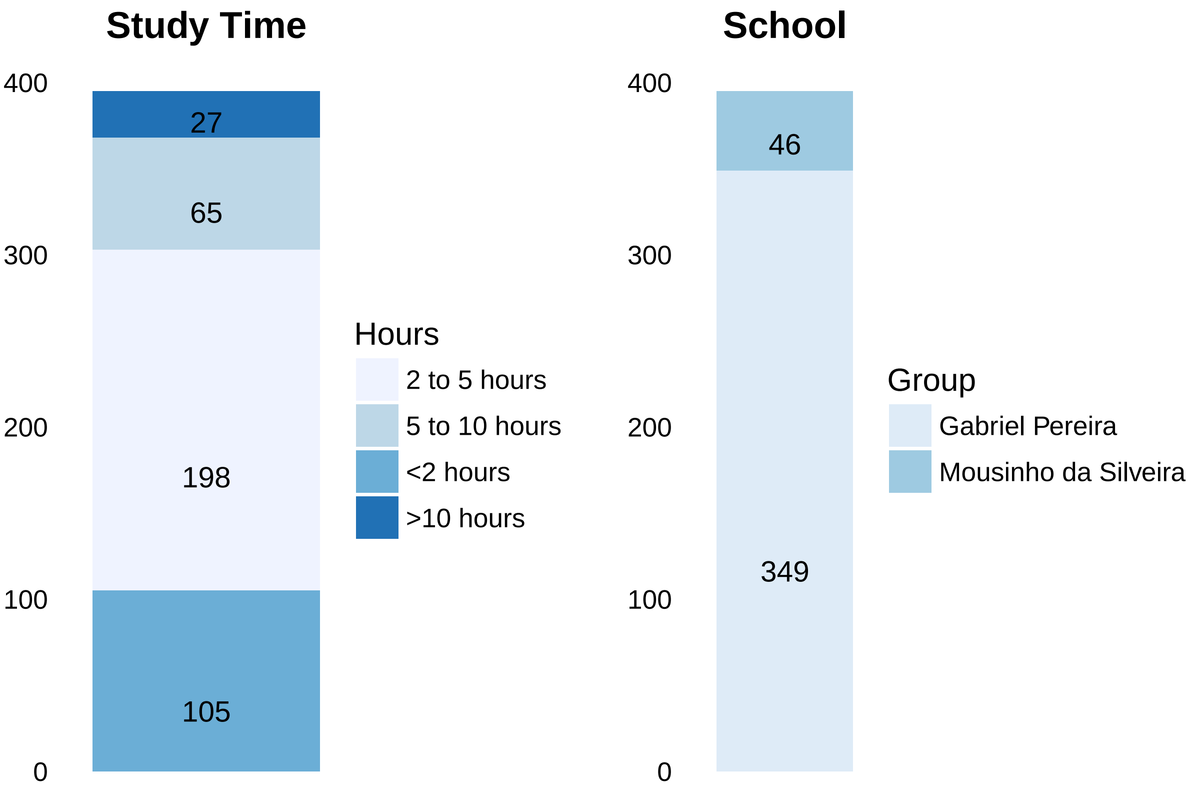


Figure 3. Education

d. Social Life and Health The social life section provides information on activities the student is involved in outside of school. This includes how much leisure time they have, whether they are in a relationship and how often they go out. Moreover, there is data on how much alcohol students consume and in what health condition they are in.

e. Grades The grades are split into three grades: the first period (G1), second period (G2) and final grade (G3). For the purpose of this analysis, students academic performance will be measured as the average of all three grades. In addition, there is information on whether they have failed to pass any classes.

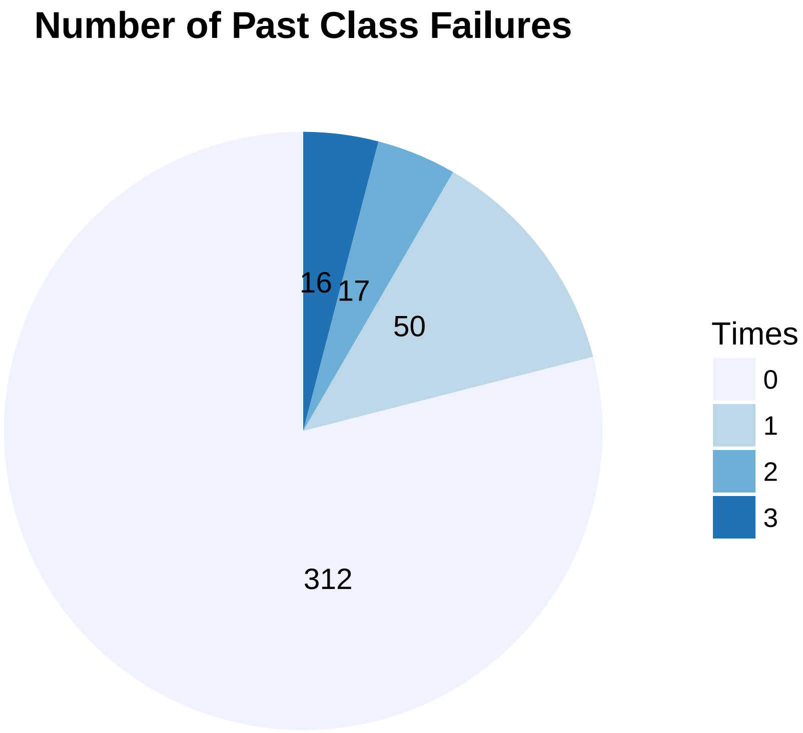


Figure 4. Past Class Failures

=======

3. Hypothesis Testing

As aforementioned, this paper looks into factors influencing academic performance. Our expectation is to see a high correlation between family background and social activities on students' performances. In particular, we are interested to see how alcohol consumption is impacting students' grades. Our basic assumption is that alcohol consumption has a negative impact on students grade. Thus, the Null Hypothesis (H0) to be checked assumes that alcohol consumption does not have an impact on academic performance. To test this hypothesis, multivariate regression analysis will be applied on the 395 samples of students described above. This will allow us to see the significant factors that contribute to academic performance, and check whether or not alcohol consumption is one of them. In this regression, grade is assigned as the dependent variable, and all other variables are treated as independent variables. If alcohol appeared as one of the significant variables in the regression, the null hypothesis will be rejected, inferring that alcohol **is** one of the variables related to academic performance.

=======

4. Correlation Analysis and Regression

Prior to regression, there are several items that need to be inspected :

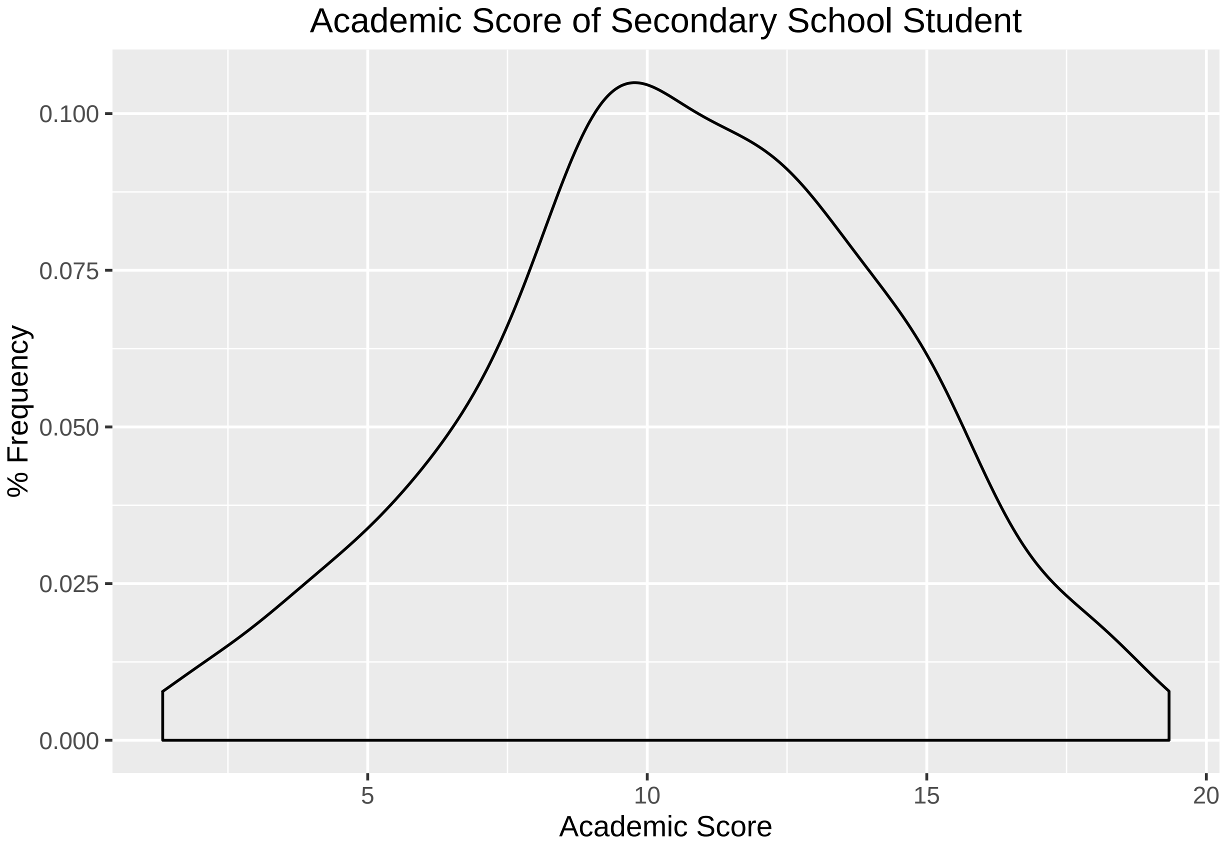
a) Distribution of Response Variable

If the distribution is highly skewed, linear regression might not give a good approximation of the response varible. Thus, variable transformation might be required, or applying another regression method might be more appropriate.

b) Correlation Analysis

Correlation between response and explanatory variables is required to get a general view of the variables relationships. In addition, correlation between explanatory variables will also be assessed to avoid a multicollinearity problem, as well as to provide other insights from the data.

a. Distribution of Response Variable



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.333   8.333  10.670  10.680  13.330  19.330
```

As seen from the graph, the distribution is slightly right-skewed, but very close to normal distribution. The deviation from the normal distribution could be due to a small sample size. Thus, applying a multivariate regression method is appropriate to test the hypothesis, and no variable transformation is required.

b. Correlation Analysis

After performing correlation of all explanatory variables against grade (response variable), four variables appear to have high a correlation to grade :

1) Parent's Education

```
##
##  Pearson's product-moment correlation
##
## data:  sm[, "f_or_medu"] and sm[, "Score"]
## t = 4.7011, df = 393, p-value = 3.586e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1351481 0.3220802
## sample estimates:
##       cor
## 0.2307422
```

The correlation between parent's education and grade is 0.23. With 95% confidence interval and a p-value of 3.586e-06, we can reject the null hypothesis, which states that there is no correlation between these two variables. This means that parent's education **has a positive correlation** with the student's grade, and that students who have parents with university degrees are more likely to have higher scores than those who do not.

2) Want to have higher education

```
## 
##  Pearson's product-moment correlation
## 
## data:  sm[, "higher_ed"] and sm[, "Score"]
## t = 3.8257, df = 393, p-value = 0.0001517
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0925427 0.2828659
## sample estimates:
##       cor
## 0.1894835
```

The correlation between students' aspirations to get higher education and grade is 0.189. With 95% confidence interval and a p-value of 0.0001517, we can reject the null hypothesis, which states that theres is no correlation between these two variables, and that grades do in fact correlate positively with their aspiration to take higher education.

3) Number of past class failure

```
## 
##  Pearson's product-moment correlation
## 
## data:  sm$Score and sm$failures
## t = -8.0382, df = 393, p-value = 1.08e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4574686 -0.2877569
## sample estimates:
##        cor
## -0.3757589
```

The correlation between number of past class failure and grade is -0.375. With 95% confidence interval and a p-value of 1.08e-14, we can reject the null hypothesis, which states that there is no correlation between these two variables, and conclude that indeed, the number of past class failures is negatively correlated with grades. Students who have failed more classes tend to have lower grades.

4) Frequency of going out with friends

```
## 
##  Pearson's product-moment correlation
## 
## data:  sm$Score and sm$goout
## t = -3.1003, df = 393, p-value = 0.002073
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.24938036 -0.05670482
## sample estimates:
##        cor
## -0.1545113
```

The correlation between the frequency of going out with friends and grade is -0.154. With 95% confidence interval and a p-value of 0.002073, we can reject the null hypothesis, which states that there is no correlation between these two variables, and conclude the frequency of going out with friends are negatively correlated with grades. Students who go out more often are more likely to have lower grades. In the following section, regression analysis will be performed to identify a fitting model which explains the students' academic performances and influencing factors.

Correlation between Alcohol Consumption and Students' Performances

```
##
##  Pearson's product-moment correlation
##
## data:  sm$Score and sm$Walc
## t = -1.7518, df = 393, p-value = 0.08058
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.18508812  0.01073964
## sample estimates:
##         cor
## -0.08802467
```

As shown above, alcohol consumption does not have a significant correlation to students' grades. The test result shows a correlation of -0.08. With 95% confidence interval and a p-value of 8%, we can **not** reject the null hypothesis, which states that there is no correlation between these two variables. Therefore, we can conclude alcohol consumption does not have a significant impact on students' grades, which will also be verified in the following multivariate regression analysis.

c. Regression

After assessing the correlations between all the variables, multivariate regression will be applied. Firstly, all variables are put in the regression model. Secondly, the variable with the highest p-value is eliminated from the model. This iterative process is repeated until a model with a satisfying R2 value can be derived, which consists of significant explanatory variables.

Initial Regression Model:

```r
#Initial Regression

reg <- lm(Score ~ 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 +
                f_or_mteach + f_or_medu
                , data = sm)
```

Final Regression Model

The results of the iterative process are shown in the figure below.

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Score
## -----------------------------------------------
## sexM                          0.94***
##                               (0.35)
##
## addressU                      0.80**
##                               (0.40)
##
## studytime                     0.56***
##                               (0.21)
##
## failures                     -1.51***
##                               (0.23)
##
## schoolsupyes                 -1.32***
##                               (0.50)
##
## famsupyes                     -0.70**
##                               (0.35)
##
## goout                        -0.45***
##                               (0.15)
##
## f_or_medu                     1.19***
##                               (0.35)
##
## Constant                     10.50***
##                               (0.80)
##
## -----------------------------------------------
## Observations                    395
## R2                             0.23
## Adjusted R2                    0.22
## Residual Std. Error       3.27 (df = 386)
## F Statistic           14.68*** (df = 8; 386)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Figure 5. Final Regression Model

These results verfify that alcohol consumption is not one of the significant variables that explains student academic performance. Instead, the dependent variables which best explain the model are following:

1) Sex
2) Address
3) Study time
4) Number of past class failure
5) Extra educational school support
6) Family educational support
7) Frequency of going out with friends
8) Parents level of education

The adjusted R2 is 21.74%, which means that the model can explain 21.74% of the variance in students' academic scores. The residual plot shown below (figure 6) does not show a randomly dispersed points around the x axis, which means that the residual does not sum up to zero. For predicted academic scores around 10, the model is closer to the observed values. For lower predicted scores, the model overestimates, and for higher scores, the predicted values are too low. There are other significant factors which are neither captured in the dataset nor the model, that can contribute to an explanation of students' academic scores. General knowledge indicates that IQ score, EQ score, nutrition, or time spent for physical exercise can have an impact on students' academic performance.

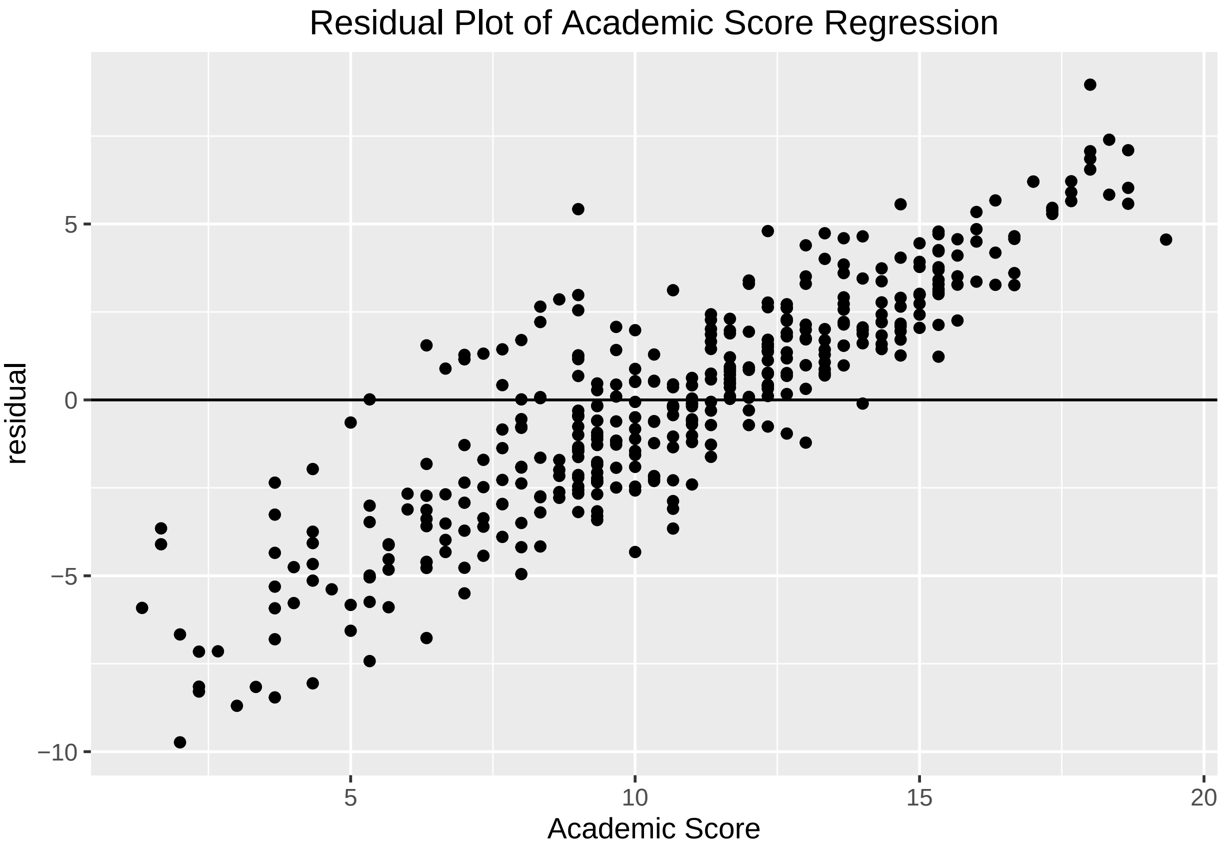


Figure 6. Residual Plot

Grades and Alcohol Consumption

Both the correlation test and regression model show that alcohol consumption does not directly affect students' grades. However, examining the correlation between alcohol consumption and the frequency of going out as well as students' study time, which have shown to have an impact on students' grades, demonstrates a high correlation.

1) Frequency of going out with friends

```
##
##  Pearson's product-moment correlation
##
## data:  sm$Walc and sm$goout
## t = 9.1848, df = 393, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3356369 0.4983839
## sample estimates:
```

```
##       cor
## 0.4203857
```

```
##
##  Pearson's product-moment correlation
##
## data:  sm$Dalc and sm$goout
## t = 5.4923, df = 393, p-value = 7.137e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1728772 0.3562789
## sample estimates:
##       cor
## 0.2669938
```

As shown above, the frequency of going out with friends has a 0.42 correlation with weekend alcohol consumption and 0.26 with workday alcohol consumption. The higher frequency of going out, the higher the students' alcohol consumption during weekend and workday.

2) Study time

```
##
##  Pearson's product-moment correlation
##
## data:  sm$Walc and sm$studytime
## t = -5.2014, df = 393, p-value = 3.191e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3438455 -0.1590977
## sample estimates:
##        cor
## -0.2537847
```

```
##
##  Pearson's product-moment correlation
##
## data:  sm$Dalc and sm$studytime
## t = -3.9628, df = 393, p-value = 8.797e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.28909871 -0.09926822
## sample estimates:
##        cor
## -0.1960193
```

Study time also shows a high correlation of -0.25 and -0.19 for weekend and workday alcohol consumption, respectively. This implies that the more students' consume alcohol during weekend or weekday, the less study time they have.

=======

5. Conclusion

The study done by Keng and Huffman (2010) outlined that alcohol abuse has a negative impact labor market outcomes. However, this correlation could not be replicated for secondary school students: the result of the analysis show no direct correlation between alcohol consumption and secondary school students' academic performance. Alcohol consumption amongst secondary school students is highly correlated to two factors, which do in fact impact students academic performance, namely frequency of going out with friends as well as study time. Further research needs to be done to identify reasons for discrepancy between both studies, which could be due to a variety of different factors. Firstly, the values given in the data (from 1 - very low to 5- very high) do not indicate whether there is abusive alcohol consumption and leave room for personal interpretation. Furthermore, small sample size and age restriction for drinking alcohol could also impact the result. It could also be researched whether the alcohol's impact on people's concentration and performance takes time to show effects and thus would not necessarily impact students' grades at this stage. In conclusion, a more specific study needs to be designed to investigate the impact of these factors and other aspects to draw further conclusions on alcohol consumption in relation to academic achievements.

=======

6. References

European Commission (2010). Europe 2020 – A Strategy for smart, sustainable and inclusive growth. [online] Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF [Accessed: 11.10.2016]

Keng, S.H. and Huffman, W.E., 2010. Binge drinking and labor market success: a longitudinal study on young people. Journal of Population Economics, 23(1), pp.303-322.

OECD (2013). Education at a Glance 2013. Available at https://www.oecd.org/edu/Portugal_EAG2013%20Country%20Note.pdf. [Accessed 11.10.2016]

Statista (2016). Countries with the highest per capita consumption of alcohol in 2015 (in liters of pure alcohol per year). Available at https://www.statista.com/statistics/280139/countries-with-highest-per-capita-consumption-of-alcohol/ [Accessed 11.10.2016]

=======

7. Appendix

Attributes in the Dataset:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2 sex - student's sex (binary: 'F' - female or 'M' - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: 'U' - urban or 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) 6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student’s guardian (nominal: ‘mother’, ‘father’ or ‘other’)

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

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

30 absences - number of school absences (numeric: from 0 to 93)

Grades Observed in the Dataset:

31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)