**Final Report: Statistical Modelling and Analysis Results**

**For Student Alcohol Consumption and Final Grade Study**

Submitted to:

Professor Marzia A Cremona

Department of Statistics

Report Prepared By:

Group Straight A

Ruxin Tong

Wenrui Cai

Ruiqi Dong

Jing Hu

Yang Zhou

December 7, 2017

**Abstract**

This report summarizes the statistical modelling and analysis results associated with the student alcohol consumption and final grades study. The dataset, obtained from Kaggle, UCL Machine Learning, is a survey of 395 students enrolled in math classes in two secondary schools in Portugal (most likely high school as observed from the age range). There are 33 variables in the raw dataset, which give the social, gender, and the study information of the students. A new predictor that we call weekly alcohol consumption (DWalc) was created by combining workday (Dalc) and weekend alcohol (Walc) consumption. Therefore, the total number of variables changed from 33 into 34. Our data analysis is based on students with positive grades, so students with zero final grade G3 were removed from the dataset. There are two objectives for the study: to find the best predictors for the response variable final grade G3 and to examine if there is a statistical relationship between the weekly alcohol consumption and the final grade G3 in both simple linear regression model and multiple linear regression models. First of all, the exploratory data analysis including scatterplot, histogram, and boxplot were used to narrow down the number of predictors from 31 to 14 with respect to the final grade G3. Then, pairwise correlation and variance inflation factor (VIF) methods were used to check if there is collinearity issue among these 14 predictors. The results indicated that there is no significant collinearity among them. Additionally, leverage, studentized residuals, and Cook’s distance were employed to identify potential influential outliers. No influential outliers were identified. Four measures of fit were used to select the potential best model subsets: adjusted coefficient of determination Radj2, Mallow’s criterion Cp, Akaike’s information criterion (AIC), and Bayesian information criterion (BIC). Lasso regression was also employed, but it was not good for this dataset because it did not produce an obvious trend where the minimum value can be easily identified. Diagnostics and VIF were then employed to evaluate models selected by four measures of fit. The model for Mallow’s criterion Cp was evaluated to be the best model, because it satisfies assumptions of linearity, equal variance, independence, and normality of errors and has low VIF among the predictors. This model explains around 17 percent of the variation of the final grade G3. In summary, predictors of sex, address, studytime, failures, goout, and absences are the best predictors for the final grade G3. For the second objective, a simple linear model was formulated and the associated p-value is 4e-04, which less than 0.05. Hence, the weekly alcohol consumption is significant in predicting the final G3 in a simple linear regression model. In addition, a multi-linear regression model was formulated, and a partial F test was performed. The p-value is 0.192, which is greater than 0.05, and thus there is not enough evidence that the weekly alcohol consumption is a significant predictor for final grade G3, given other 13 predictors.

# **Introduction**

## **PART I: Introduction to the dataset**

Raw dataset was obtained on Kaggle, UCI Machine learning in October, 2017. The dataset is a survey of students enrolled in math classes in two secondary schools (most likely high school as observed from the age range). There are 33 variables in this dataset, both quantitative and categorical, and many of them can be used for predicting students’ final grade. A new variable, DWalc, was created by combing workday and weekend alcohol consumption.

In the following list, the description of variables is provided directly from the “Content” part in the description of the dataset on the website.

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 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. *Fedu* - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – 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: fromba1 - 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)
31. *DWalc* – weekend alcohol consumption (numeric: from 2- very low to 10 – very high

These grades are related with the course subject, Math:

1. *G1* - first period grade (numeric: from 0 to 20)
2. *G2* - second period grade (numeric: from 0 to 20)
3. *G3* - final grade (numeric: from 0 to 20, output target)

## **PART II: Research questions**

The two main questions we focused on in this research project are:

1. What are the best predictors of Final Grade G3?
2. Considering weekly alcohol consumption as our interested predictor, how is weekly alcohol consumption related to the final grades using simple linear model? How is weekly alcohol consumption related to the final grades in a multiple linear model given other significant predictors?

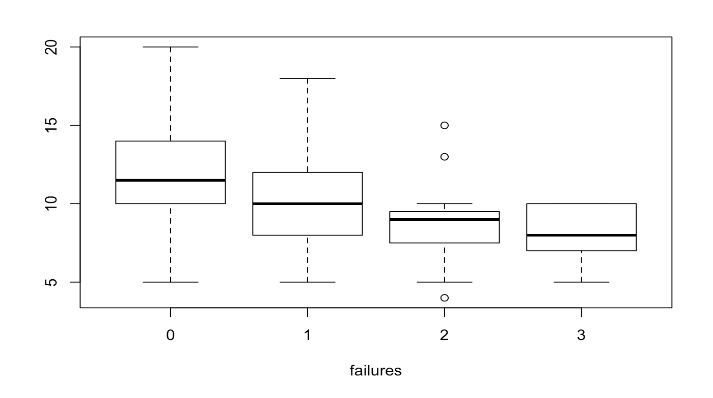
Reasons for the research questions:

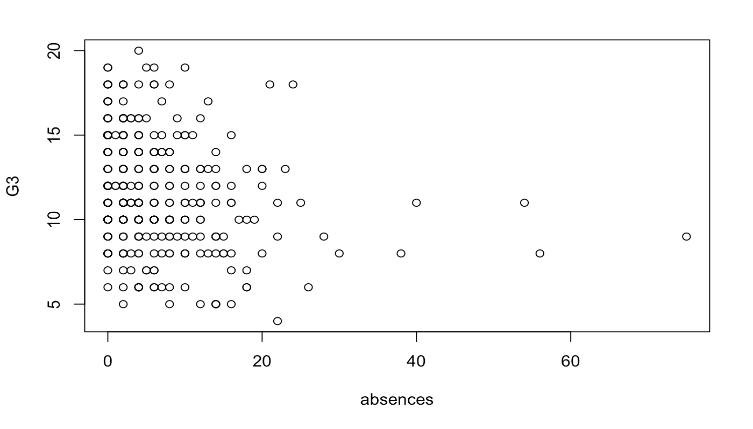
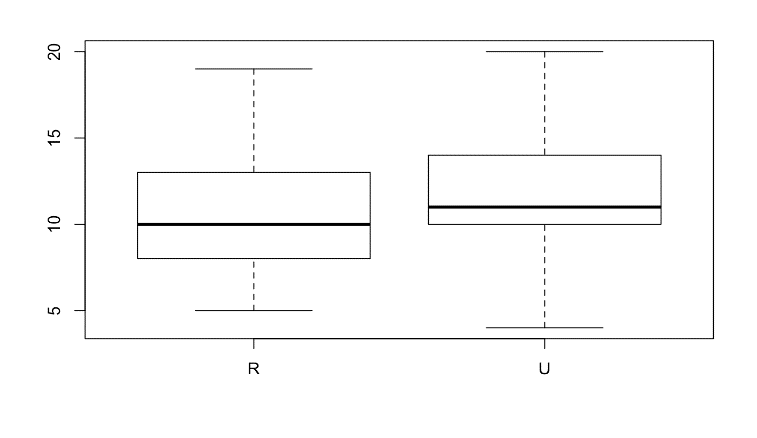
1. Final grade is usually used to evaluate how well students learn about the materials, and most colleges use high school grades to evaluate student’s qualifications. However, social, gender, and study variables may influence how well the students perform in the class. Hence, which variables significantly contribute to student’s performance is a worthy question. If some social factors do influence how students perform in the class, then it may be worthy of further exploration. If confirmed, it would be fairer and better for the college admission office to also evaluate students on the social background of the students other than focusing solely on the grade.
2. Alcohol consumption is a possible factor in determining the academic performance of the student. We analysed this specific variable to see what kind of impact, if any, this variable would have on final grade.

# **Exploratory Data Analysis**

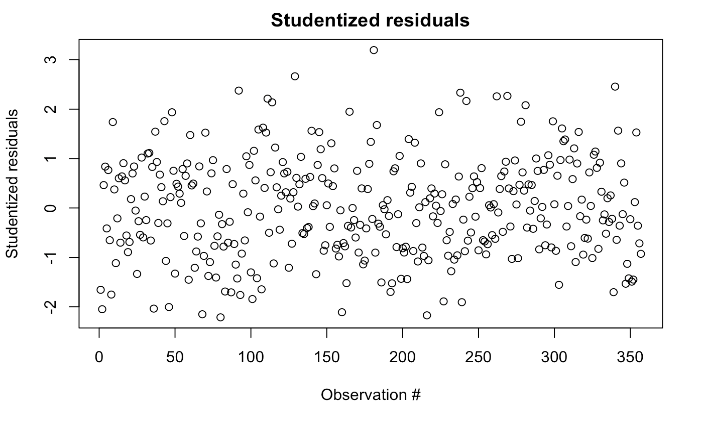
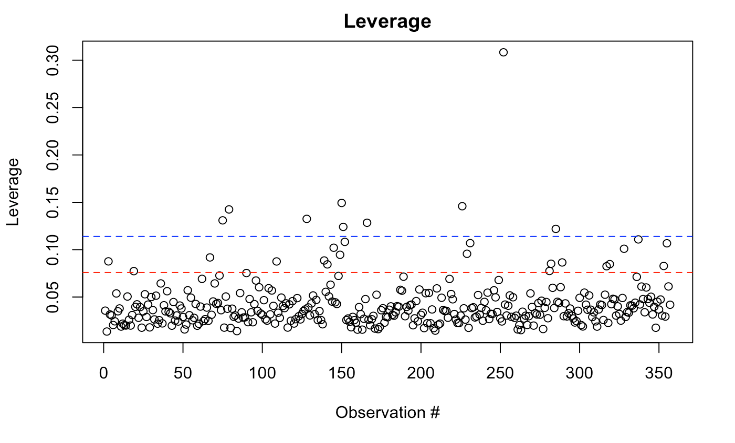
We removed 38 observations with response zero because we wanted to focus our analysis on students with positive grades. There are 34 variables (with the addition of DWalc). Exploratory data analysis including boxplot, histogram, and scatterplot exhibited distributions of the predictor from 1-26 and 29, 30, 31 in relation with G3, which is the response variable. After exploring all 29 distributions, school, famsize, Pstatus, Medu, Fedu, reason, guardian, traveltime, schoolsup, famsup, paid, nursery, and famrel are predictors that did not look significant. The remaining 14 predictors sex, age, address, studytime, failures, activities, higher, internet, romantic, freetime, goout, health, absences, and DWalc are significant. Then, leverage, studentized residuals, and Cook’s distance measurements were used to identify potential outliers. No influential outliers were identified. Also, pairwise correlation and VIF were performed to test for possible collinearity problems among the 14 predictors. No significant collinearity was identified.

## **PART I: Boxplots, Histograms, Scatterplot Distributions**



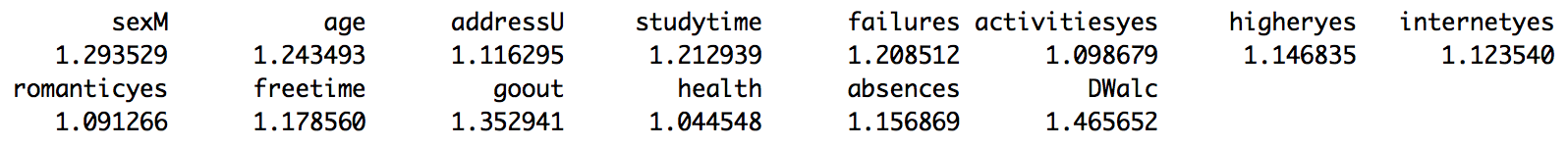


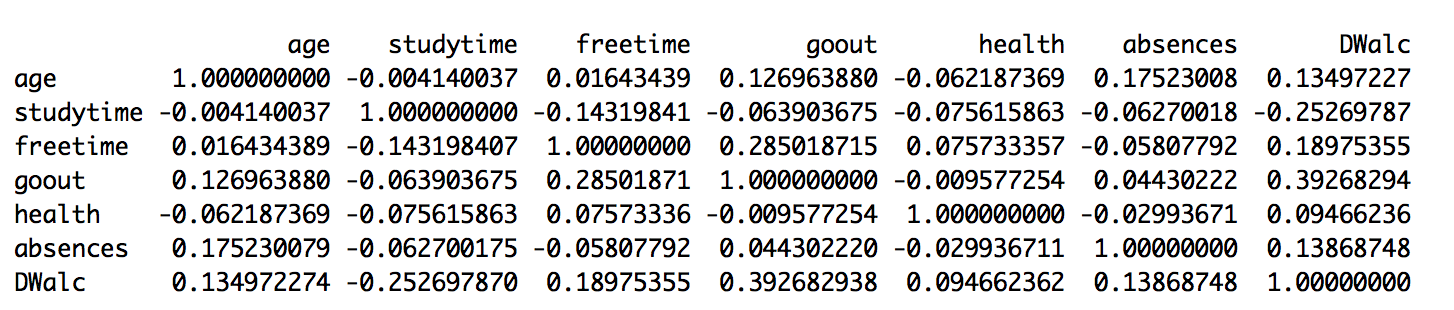
## **PART II: Leverage, Studetized Residuals, and Cook’s Distance.**



## 

## **PART III: VIF and Pairwise Corrlation**





**Method**

**PART I: Identifying the Best Predictors for the Final Grade G3**

1. The function *regsubsets* from the package *leaps* was used to select, for each p, the best model according to RSS
2. Adjusted coefficient of determination was calculated for each p, and the respective plot was created
3. Mallow’s criterion Cp was computed for each p, and the respective plot was created.
4. Akaike’s Information Criterion (AIC) was computed for each p, and then the respective plot was created
5. Bayesian Information Criterion (BIC) was computed for each p, and then the respective plot was created.
6. Lasso was employed, and the minimum was identified
7. Best subsets were chosen for each measure of fit as mentioned above
8. Diagnostics were employed for each model
9. Variance inflation factor was computed for each model
10. The four models were evaluated and compared
11. The best model was chosen

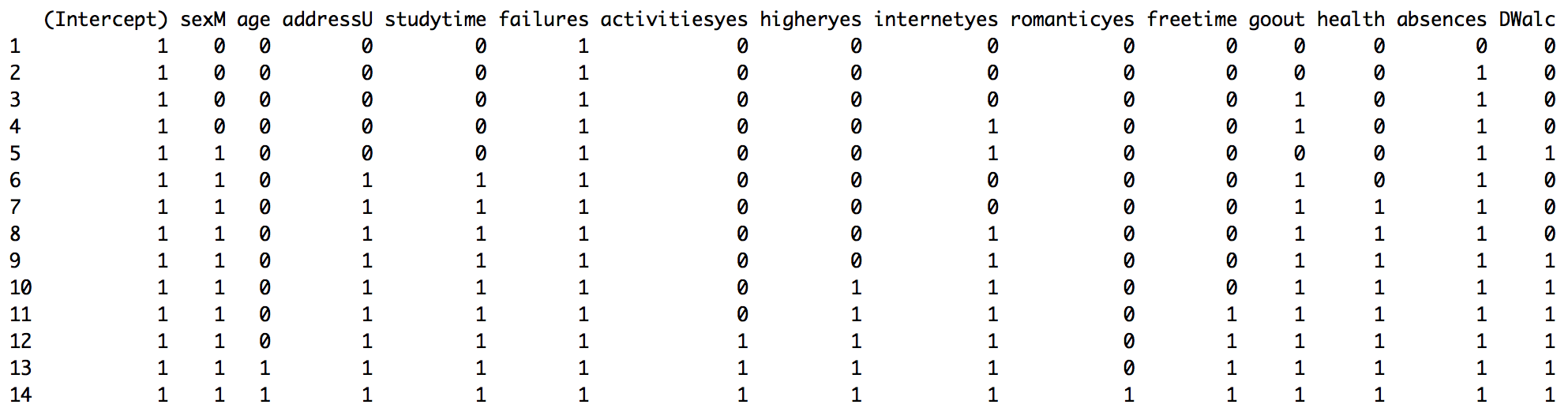
**PART II: Statistical Relationship between Weekly Alcohol Consumption (DWalc) and the Final Grade G3**

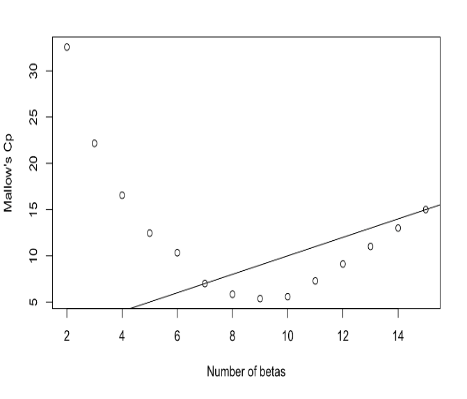
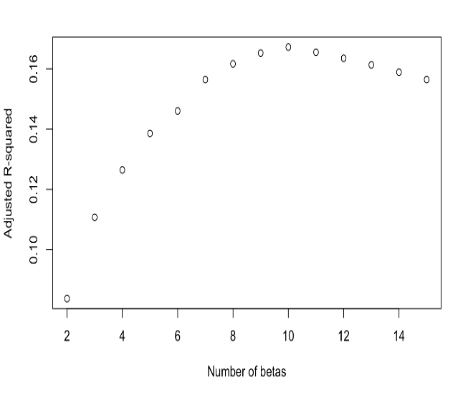
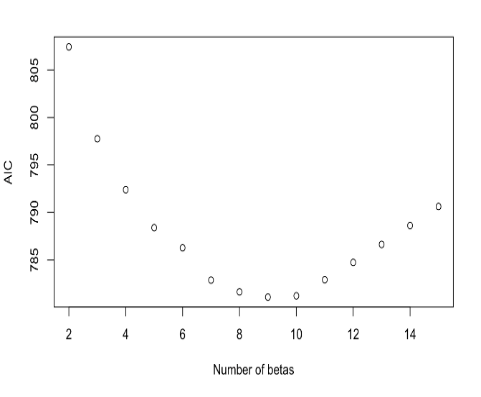
1. Simple Linear Regression Model
2. A null hypothesis was formed: H0:B1=0, Ha: B1≠ 0
3. A simple linear regression model lm(G3~DWalc) was created, and the respective p-value was computed
4. The T-test was computed by comparing the p-value with the predetermined significance level of α=0.05
5. Based on whether the p-value was larger or smaller than the significance level, the conclusion was deduced (FIX)
6. Multilinear Regression Model
7. A null hypothesis was formed: H0:B15=0, Ha: B15≠ 0
8. Regression models were defined under the null and alternative hypothesis
9. Analysis of variance was employed
10. A partial F-test was computed by comparing the p-value to the predetermined significance level of α=0.05
11. Based on whether the p-value was larger or smaller than the significance level, the conclusion was deduced (FIX)

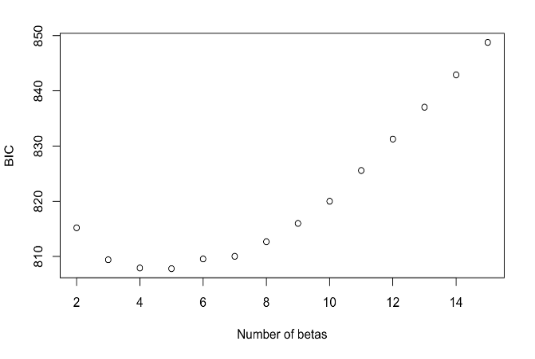
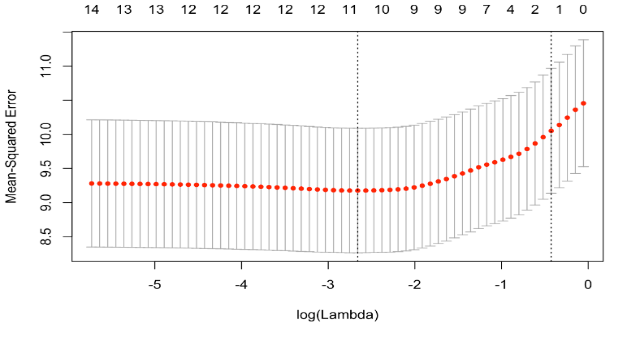
**Results**

## **PART I: Best Predictors for the Final Grade G3**

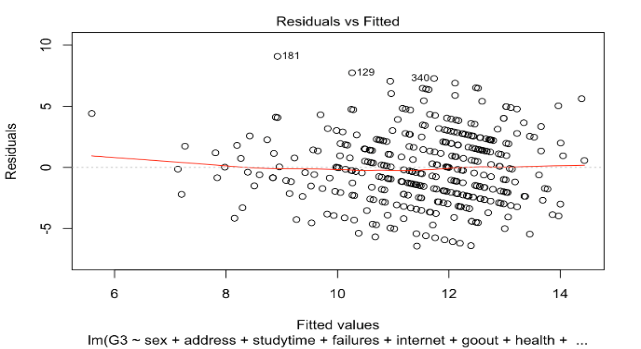
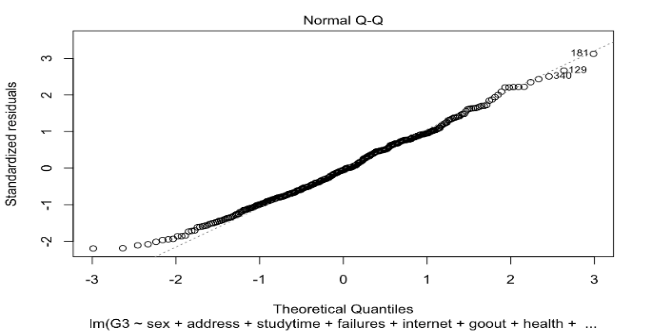
1. Best Model Subsets Selection

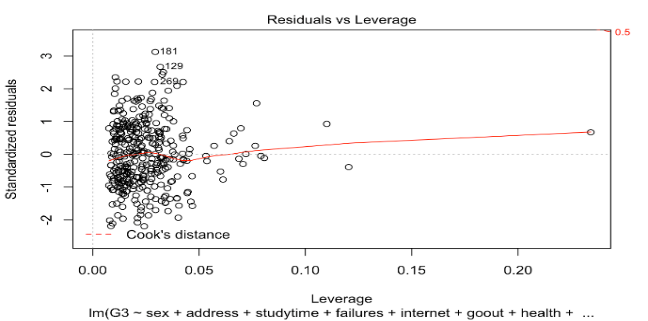
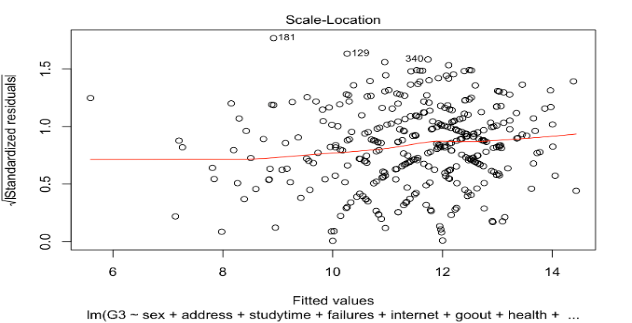


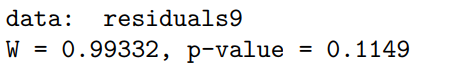
1. Measures of Fit and Lasso Regression



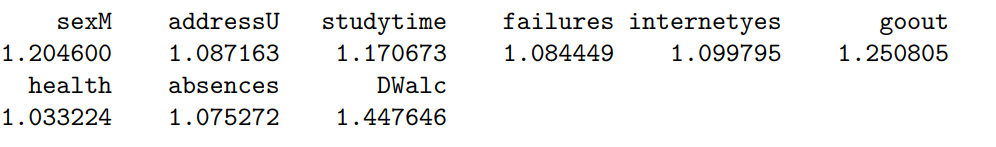
1. Diagnostic Plot and Variation Inflation Plot
2. Radj2 :



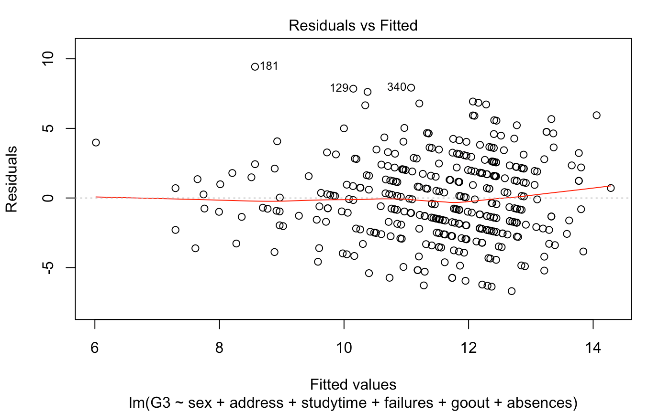
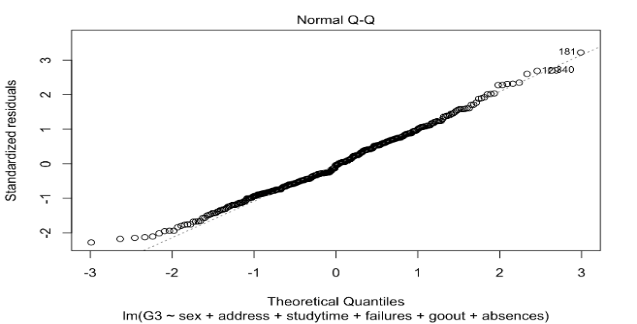


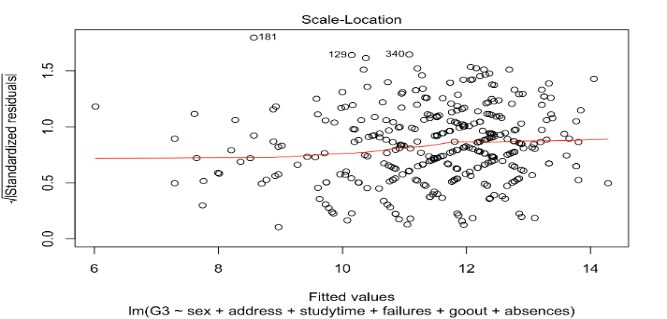
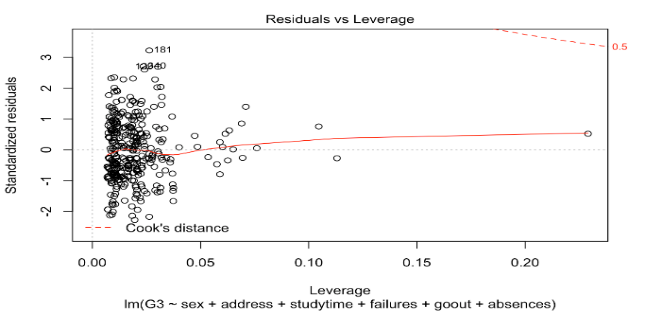


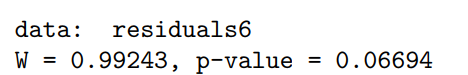
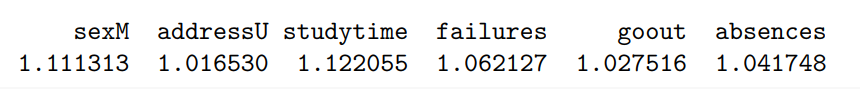


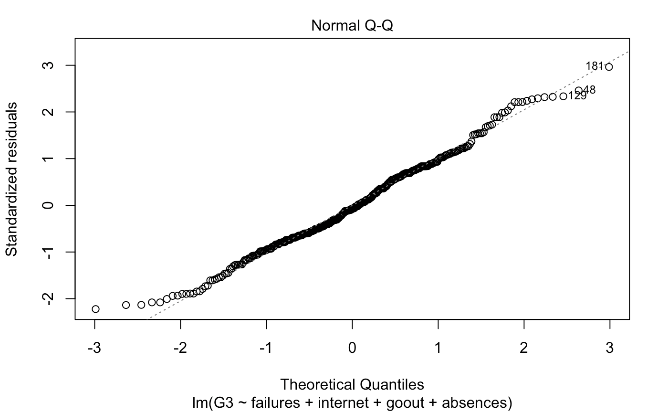
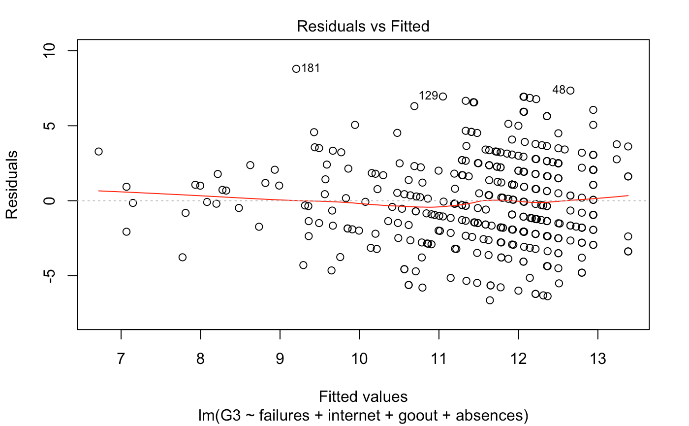


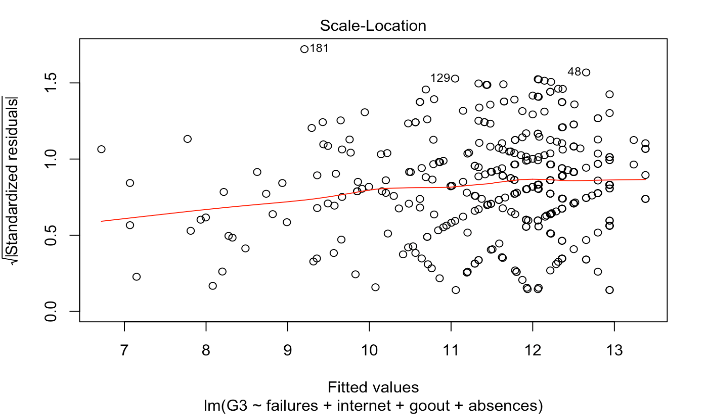
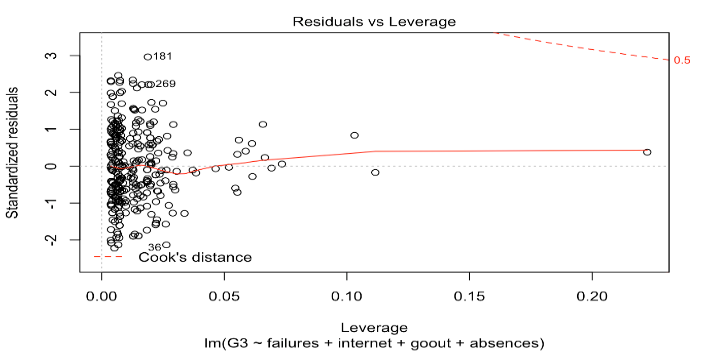
1. Mallow’s Criterion Cp:

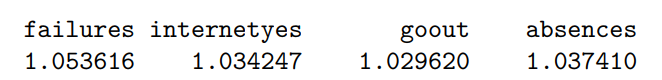




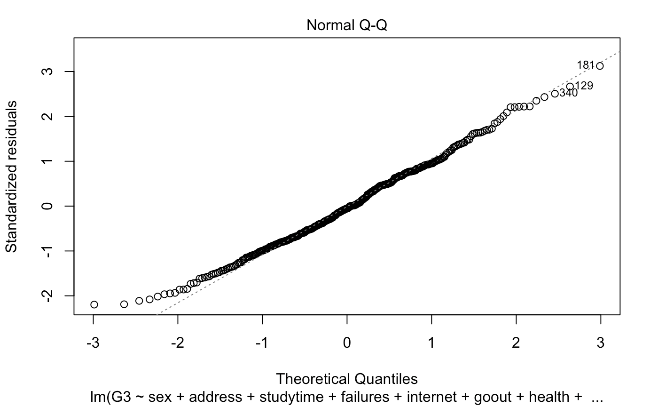
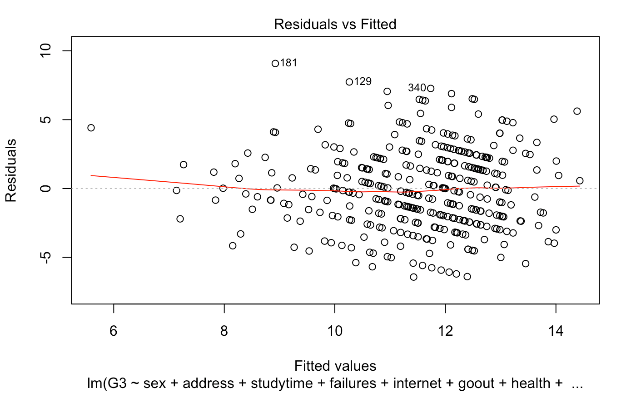


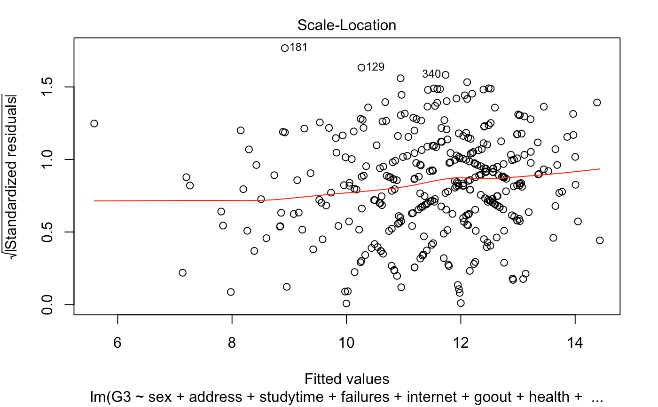
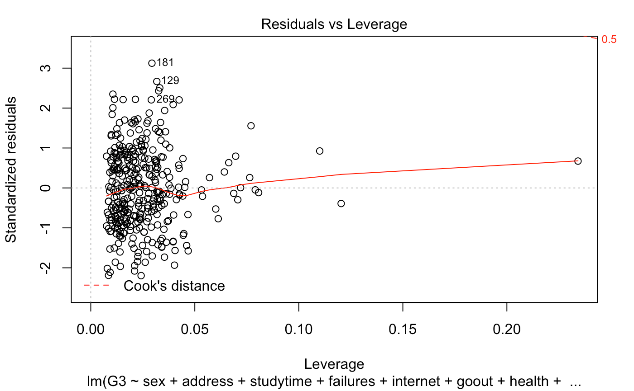
1. BIC:

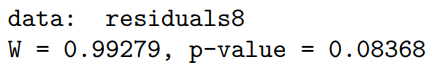
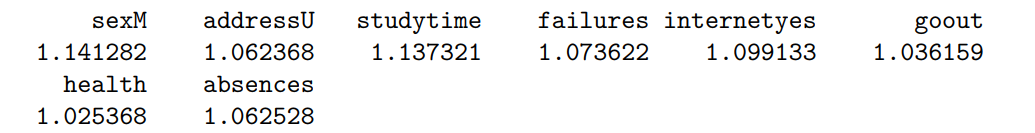




1. AIC:



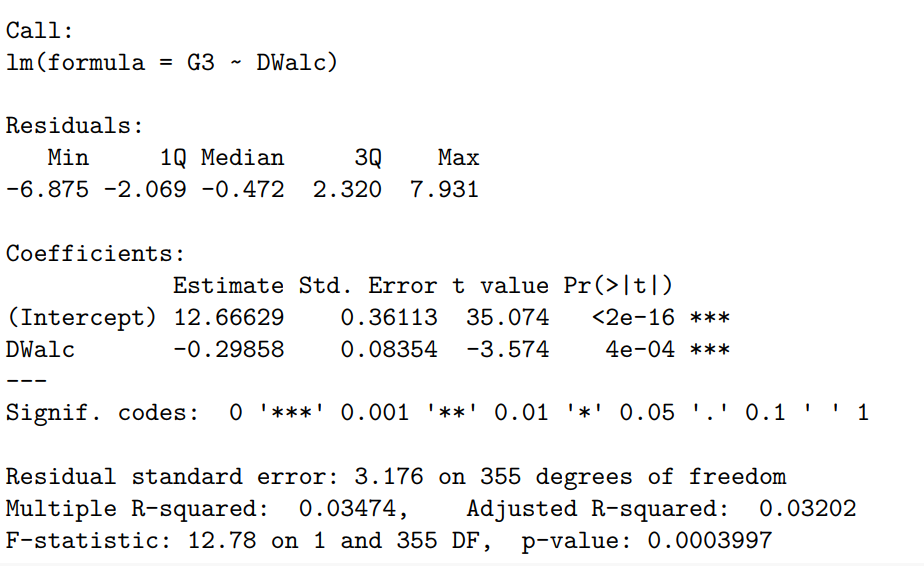




The maximum Radj2 was considered, which gave nine predictors; Mallow’s criterion Cp should have small p and the Cp value close to p. For this model, six predictors were chosen. As for AIC and BIC, the number of predictors that minimize their values were considered, and the values were 8 predictors and 4 predictors respectively. For the lasso regression, the minimum value was considered, which corresponds to 11 betas. However, the lasso regression was not explored in this analysis because it was not useful in this dataset as it did not produce a significant shape where the minimum can be easily identified. Hence, only four models were formed according to the first four measures of fit and then evaluated in terms of diagnostics and VIF; all these four models are proper in terms of diagnostics and VIF. However, model 2 was preferred as it satisfies all the assumptions of errors. In additional, this model performed better in terms of the plot of leverage, diagnostics, and Cook’s distance. In fact, the points were really close to the left of the graph, which showed low leverage, and were clustered around the 0, which demonstrated relatively low studentized residuals. Its VIF indicated the absence of collinearity problems among the predictors.

## **PART II: Statistical Relationship between Weekly Alcohol Consumption (DWalc) and the Final Grade G3**

1. **Simple linear Regression Model and T Test**

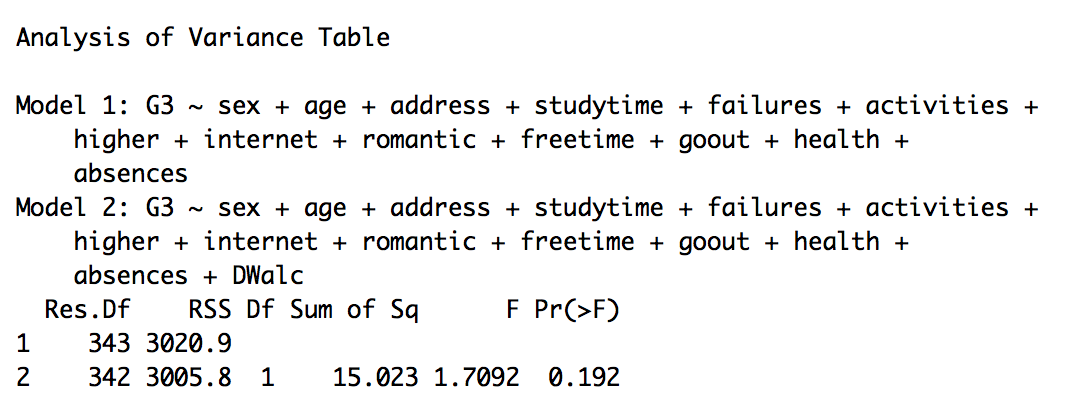


H0: β1=0

Ha: β1 ≠ 0

α=0.05

1. **Multilinear Regression Model and Partial F Test**

Partial F test:

H0: β15=0

Ha: β15 ≠ 0

α=0.05

For part II, weekly alcohol consumption is a significant predictor in the simple linear regression model. Its p-value is 4e-04, which is less than 0.05, hence, there is enough evidence to reject the null hypothesis. Our interpretation in this model is: for every increase in unit of alcohol consumption, there is a 0.3 decrease of final grade score. However, this is not a satisfactory model because the R2 is very low (0.03). In terms of the multilinear model, partial f-test indicates that weekly alcohol consumption is not significant in predicting the final grade G3, when other predictors are given. Its p-value is 0.192, which is higher than 0.05.

# **Conclusion**

Statistical analysis indicates that sex, address, studytime, failures, goout, and absences are best predictors in predicting the final grade G3. Weekly Alcohol consumption a significant predictor for the final grade in a simple linear regression model but it is not significant if considered in our multilinear regression model. In retrospect, the dataset has some limitations. For example, it does not take into account on the family’s income, which could be a significant factor in the final grade. In addition, variables such as the quality of previous education, family issues, student’s psychological state, safety of school district, safety of neighborhood could also be taken into consideration as they may have significant impact. Furthermore, the data with response as zero were deleted in the analysis, some of which have studied as many as 10 hours per week. These observations are definitely strange, therefore further exploration could be conducted to find reasons behind. Potential reason may be low quality of study, drop out from the course, or violation of academic integrity.

# **Team Member Contributions**

Ruxin Tong: designed the questions, performed the model selection, created diagnostic plots, performed lasso, presented the research, and contributed to writing the final report and presentation

Wenrui Cai: selected the dataset, designed the questions, performed the model selection, created diagnostic plots, created the PowerPoint presentation, presented the research, and contributed to writing the final report

Ruiqi Dong: designed the questions, performed and interpreted the exploratory data analysis, and fitted the initial model, presented the research, and contributed to writing the final report.

Jing Hu: performed and interpreted the exploratory data analysis, prepared the Power-point presentation, performed the outlier’s analysis, and contributed to writing the final report.

Yang Zhou: performed and interpreted the exploratory data analysis, presented the research, prepared for the figures and tables for the final report, and contributed to writing the final report.

# **Reference**

* + - 1. UCL Learning Machine. (2016, October 19). Student Alcohol Consumption. Retrieved October 10, 2017, from https://www.kaggle.com/uciml/student-alcohol-consumption
      2. Linear Models with R, Julian J. Faraway, Taylor & Francis