Group L2C

Himari Honda	**26450779**
Lawrence Ma	**41896937**
Benedictus Harley Kartawidjaja	**77641504**
Garreth Lee	**27537224**

Introduction

Dataset Information:

The dataset utilized in this analysis originates from data collected during the 2005-2006 academic year at two public schools in Portugal's Alentejo region. Despite an observed increase in government investment in Information Technology, most public schools in Portugal continue to rely heavily on outdated paper-based systems. Therefore, this dataset was compiled by consolidating information from school reports, which primarily included period grades, school absences, and questionnaires. These questionnaires were specifically designed with predefined options covering various demographic, social/emotional, and school-related factors believed to influence student performance. Prior to administration to a larger cohort of 788 students, the questionnaires underwent validation by school professionals and were piloted with a small group of 15 students to solicit feedback. Following data collection, 111 responses lacking identification details were excluded. Subsequently, the dataset was bifurcated into two distinct sets: one focused on Mathematics (comprising 395 examples) and the other on Portuguese language classes (containing 649 records).

Variable Information:

Attribute	Description	Data Type
school	Student's school	(binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	Student's sex	(binary: 'F' - female or 'M' - male)
age	Student's age	(numeric: from 15 to 22)
address	Student's home address type	(binary: 'U' - urban or 'R' - rural)
famsize	Family size	(binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	Parent's cohabitation status	(binary: 'T' - living together or 'A' - apart)
Medu	Mother's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, or 4 – higher education)
Fedu	Father's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, or 4 – higher education)
Mjob	Mother's job	(nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob	Father's job	(nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	Reason to choose this school	(nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	Student's guardian	(nominal: 'mother', 'father' or 'other')
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)
studytime	Weekly study time	(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	Number of past class failures	(numeric: n if 1<=n<3, else 4)
schoolsup	Extra educational support	(binary: yes or no)
famsup	Family educational support	(binary: yes or no)
paid	Extra paid classes within the course subject	(binary: yes or no)
activities	Extra-curricular activities	(binary: yes or no)
nursery	Attended nursery school	(binary: yes or no)
higher	Wants to take higher education	(binary: yes or no)
internet	Internet access at home	(binary: yes or no)
romantic	With a romantic relationship	(binary: yes or no)
famrel	Quality of family relationships	(numeric: from 1 - very bad to 5 - excellent)
freetime	Free time after school	(numeric: from 1 - very low to 5 - very high)
goout	Going out with friends	(numeric: from 1 - very low to 5 - very high)
Dalc	Workday alcohol consumption	(numeric: from 1 - very low to 5 - very high)
Walc	Weekend alcohol consumption	(numeric: from 1 - very low to 5 - very high)
health	Current health status	(numeric: from 1 - very bad to 5 - very good)
absences	Number of school absences	(numeric: from 0 to 93)

Response Variable(s)

G1	First period grade	(numeric: from 0 to 20)
G2	Second period grade	(numeric: from 0 to 20)

G3	Final grade	(numeric: from 0 to 20, output target)

Motivation:

Boxplot of Final Grades vs. Weekend Alcohol Consumption:

This plot visualizes the distribution of final grades among students grouped by their reported weekend alcohol consumption levels.

Role: It helps identify any potential relationship or trend between students' weekend alcohol consumption habits and their final grades in Portuguese language studies.

Boxplot of Final Grades vs. Family Relationships:

This plot displays the distribution of final grades among students categorized by their reported perceptions of family relationship quality.

Role: It allows for the examination of any discernible patterns or associations between students' perceptions of family relationships and their final grades in Portuguese language studies.

Full Model Summary:

Provides a comprehensive summary of the regression model that includes all covariates except for other grade variables (G1 and G2).

Role: Offers insights into the overall relationship between the explanatory variables (including weekend alcohol consumption and family relationship quality) and students' final grades.

Best Model Summary (Selected via Stepwise Selection):

Summarizes the regression model selected using stepwise selection, which aims to identify the most parsimonious model with optimal predictive power.

Role: Evaluates the performance of the selected model in predicting students' final grades while considering fewer predictor variables, helping identify key predictors of academic performance.

Q-Q Plot:

A quantile-quantile plot used to assess the normality of residuals from the regression model.

Role: Assesses whether the assumption of normality holds for the residuals of the regression model, which is crucial for the validity of regression analysis.

Kruskal-Wallis Test Results:

Conducted to test for significant differences in final grades across different levels of categorical variables (e.g., family relationship quality, weekend alcohol consumption).

Role: Helps determine whether there are statistically significant associations between categorical variables and students' final grades, providing additional insights into potential predictors of academic performance.

Each of these visualizations and analyses plays a crucial role in exploring the relationship between weekend alcohol consumption, family relationships, and students' final grades in Portuguese language studies, contributing to a comprehensive understanding of the factors influencing academic performance.

The detailed exploration of the dataset aims to shed light on the intricate relationship between students' academic performance in Portuguese language studies and several key variables, namely weekend alcohol consumption and family relationship quality. By visually analyzing boxplots of final grades against weekend alcohol consumption and family relationship quality, we can discern any observable trends or patterns that may exist. Subsequently, employing regression analysis allows for a comprehensive examination of the predictive power of these variables on final grades. The stepwise selection procedure aids in identifying the most influential predictors while ensuring model parsimony. Additionally, diagnostics such as the Q-Q plot help assess the validity of regression assumptions, ensuring the reliability of the analysis. Finally, the application of the Kruskal-Wallis test provides further insights into potential predictors of academic performance, enhancing our understanding of the factors contributing to students' final grades in Portuguese language studies. Overall, this meticulous analysis aims to provide valuable insights into the complex interplay between social, behavioral, and academic factors influencing student outcomes, thus offering practical implications for educational policies and interventions.

Analysis

Exploratory Data Analysis

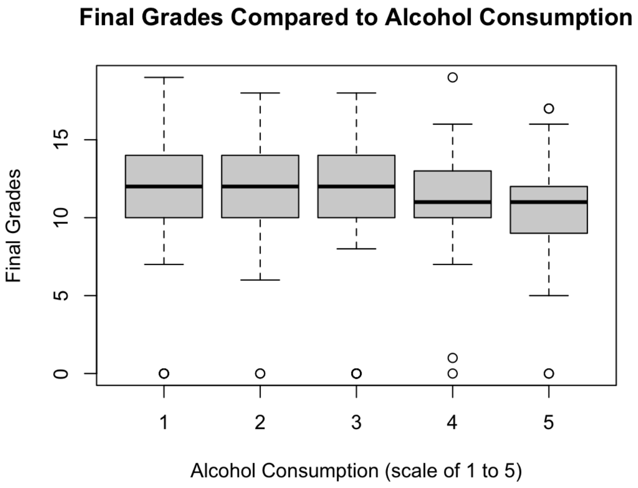


Fig. 1

In Fig. 1, we can see that the distribution of final grades across different levels of alcohol consumption does not vary much. Levels 3 and 4 are slightly more left-skewed than the other levels, but as a whole the distribution of final grades across each category is extremely similar.

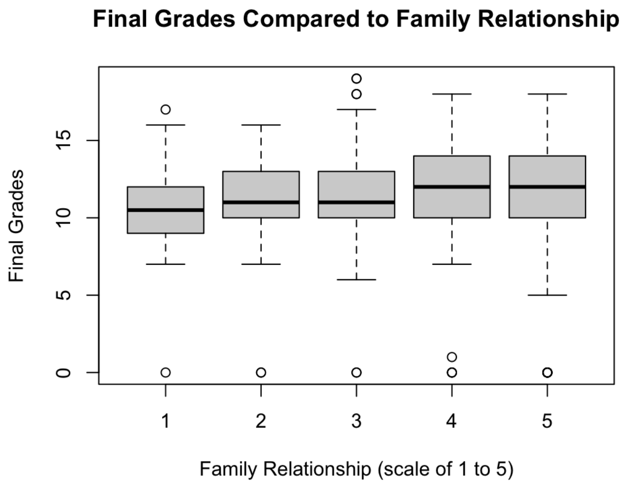
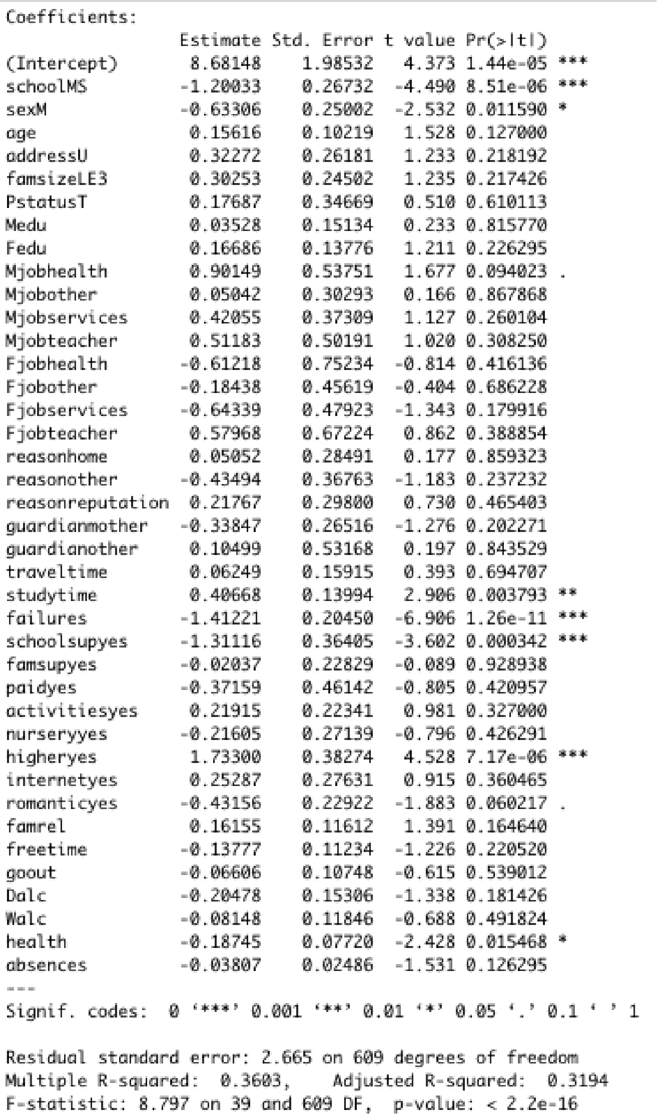


Fig. 2

Fig. 2 shows that better family relationships are associated with slightly higher grades, though the overall distribution once again does not differ much. Good family relationships display a higher median final grade as well as a higher 25th and 75th percentile for grades, though these are the only notable differences.

FULL MODEL - ALL COVARIATES MINUS OTHER GRADES

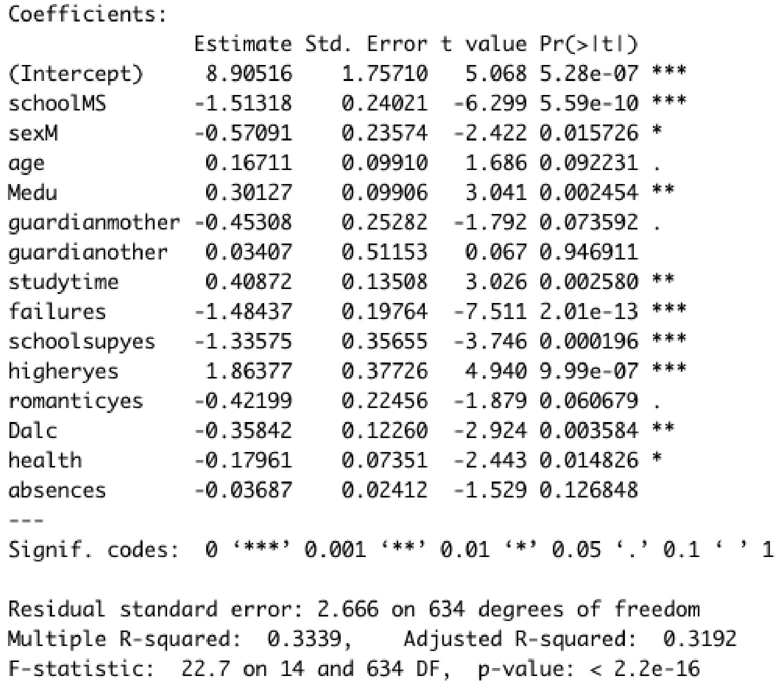


```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       8.68148    1.98532   4.373 1.44e-05 ***
schoolMS         -1.20033    0.26732  -4.490 8.51e-06 ***
sexM             -0.63306    0.25002  -2.532 0.011590 *  
age               0.15616    0.10219   1.528 0.127000    
addressU          0.32272    0.26181   1.233 0.218192    
famsizeLE3        0.30253    0.24502   1.235 0.217426    
PstatusT          0.17687    0.34669   0.510 0.610113    
Medu              0.03528    0.15134   0.233 0.815770    
Fedu              0.16686    0.13776   1.211 0.226295    
Mjobhealth        0.90149    0.53751   1.677 0.094023 .  
Mjobother         0.05042    0.30293   0.166 0.867868    
Mjobservices      0.42055    0.37309   1.127 0.260104    
Mjobteacher       0.51183    0.50191   1.020 0.308250    
Fjobhealth       -0.61218    0.75234  -0.814 0.416136    
Fjobother        -0.18438    0.45619  -0.404 0.686228    
Fjobservices     -0.64339    0.47923  -1.343 0.179916    
Fjobteacher       0.57968    0.67224   0.862 0.388854    
reasonhome        0.05052    0.28491   0.177 0.859323    
reasonother      -0.43494    0.36763  -1.183 0.237232    
reasonreputation  0.21767    0.29800   0.730 0.465403    
guardianmother   -0.33847    0.26516  -1.276 0.202271    
guardianother     0.10499    0.53168   0.197 0.843529    
traveltime        0.06249    0.15915   0.393 0.694707    
studytime         0.40668    0.13994   2.906 0.003793 ** 
failures         -1.41221    0.20450  -6.906 1.26e-11 ***
schoolsupyes     -1.31116    0.36405  -3.602 0.000342 ***
famsupyes        -0.02037    0.22829  -0.089 0.928938    
paidyes          -0.37159    0.46142  -0.805 0.420957    
activitiesyes     0.21915    0.22341   0.981 0.327000    
nurseryyes       -0.21605    0.27139  -0.796 0.426291    
higheryes         1.73300    0.38274   4.528 7.17e-06 ***
internetyes       0.25287    0.27631   0.915 0.360465    
romanticyes      -0.43156    0.22922  -1.883 0.060217 .  
famrel            0.16155    0.11612   1.391 0.164640    
freetime         -0.13777    0.11234  -1.226 0.220520    
goout            -0.06606    0.10748  -0.615 0.539012    
Dalc             -0.20478    0.15306  -1.338 0.181426    
Walc             -0.08148    0.11846  -0.688 0.491824    
health           -0.18745    0.07720  -2.428 0.015468 *  
absences         -0.03807    0.02486  -1.531 0.126295    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.665 on 609 degrees of freedom
Multiple R-squared:  0.3603,	Adjusted R-squared:  0.3194 
F-statistic: 8.797 on 39 and 609 DF,  p-value: < 2.2e-16
```

Fig. 3

BEST MODEL - USING BACK & FORWARD SELECTION (based on AIC)



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8.90516    1.75710   5.068 5.28e-07 ***
schoolMS       -1.51318    0.24021  -6.299 5.59e-10 ***
sexM           -0.57091    0.23574  -2.422 0.015726 *  
age             0.16711    0.09910   1.686 0.092231 .  
Medu            0.30127    0.09906   3.041 0.002454 ** 
guardianmother -0.45308    0.25282  -1.792 0.073592 .  
guardianother   0.03407    0.51153   0.067 0.946911    
studytime       0.40872    0.13508   3.026 0.002580 ** 
failures       -1.48437    0.19764  -7.511 2.01e-13 ***
schoolsupyes   -1.33575    0.35655  -3.746 0.000196 ***
higheryes       1.86377    0.37726   4.940 9.99e-07 ***
romanticyes    -0.42199    0.22456  -1.879 0.060679 .  
Dalc           -0.35842    0.12260  -2.924 0.003584 ** 
health         -0.17961    0.07351  -2.443 0.014826 *  
absences       -0.03687    0.02412  -1.529 0.126848    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.666 on 634 degrees of freedom
Multiple R-squared:  0.3339,	Adjusted R-squared:  0.3192 
F-statistic:  22.7 on 14 and 634 DF,  p-value: < 2.2e-16
```

Fig. 4

Using back and forward selection, we found that a model with half the number of variables produces the same adjR2 as the full model. We decided to use this new model to avoid the extra computational requirements the full model would require.

QQPLOT

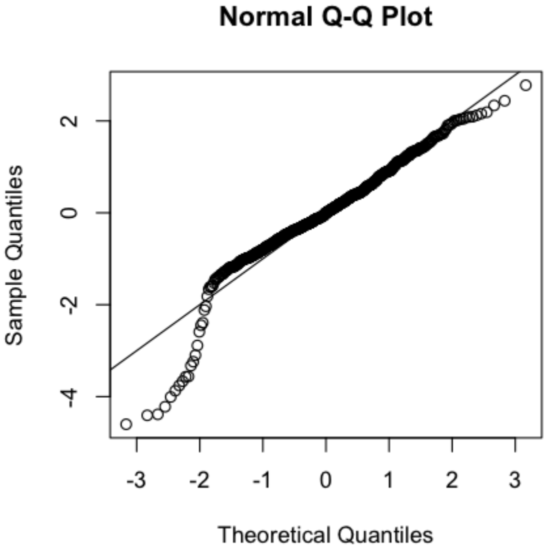


Fig. 5

The QQ plot (Fig. 5) shows a left-skew, indicating that the theoretical quantiles are much lower than the actual quantiles. This violates the normal distribution condition of linear regression and invalidates the regression analysis.

For future work, we can do the Kruskal Wallis test (since it has no distributional assumptions) to test for any significant differences between scores of students from different family backgrounds

```
kruskal.test(G3 ~ famrel, data = student)       # p-val: 0.0144 (significant at 5%)
kruskal.test(G3 ~ Walc, data = student)         # p-val: 6.963e-05 (significant at 5%)
kruskal.test(G3 ~ traveltime, data = student)    # p-val: 0.002043 (significant at 5%)
```

If we get a significant p-value, we can conclude that at least one of the groups (of families/alcohol consumption / travel time, etc.) is significantly different from the other groups.

Conclusion

Our overarching goal with this analysis was to explore the intricate relationships between academic performance of Portuguese language students and several other factors related to their lifestyle, academic habits, familial relationships, and background. We took special interest in examining the effects of weekend alcohol consumption and level of familial support on final grades, and aimed to implement the model selection method with the most appropriate linear fit.

To address these objectives, we first compared two box plots displaying impacts of alcohol consumption and familial relationship (Fig. 1, 2). Students who consumed more alcohol on the weekends tend to do slightly worse, while students receiving the best support from family performed slightly better academically. After this initial analysis, we implemented a backward and forward stepwise elimination method to select the model with the lowest AIC.

In model selection, models of smaller size are preferred compared to complicated ones especially if their predictive powers are roughly equal (compare adjusted R-squared between full and “best” model). A parsimonious model with the smallest prediction error is attained (Fig. 4). To check for assumptions of the linear regression model, we derive the QQ-plot which indicates a left skew (Fig. 5). Since the distribution is skewed, a non-parametric test that makes no distributional assumption would be appropriate to implement for further investigation on each group. In replacement of the ANOVA test, the Kruskal-Wallis test allows for the comparison of more than two discrete groups. In this case, we conclude that at least two groups are meaningfully different from each other.

Despite its usefulness, model selection statistics such as the R-squared values and p-values do not provide concrete evidence to support or reject a model without context. With real-life data comes practical issues related to sampling methodology, especially when many covariate measures are qualitative (scales of 1 to 5). The robustness of some statistical measures can be a double-edged sword; it can provide an informative framework for parameter selections, but also risk removing influential covariates.