STUDENT ALCOHOL CONSUMPTION

IDS 702 MODELING AND REPRESENTATION OF DATA

Victor Nomwesigwa

Student ID: 2748775

Masters of Interdisciplinary Data Science

Duke University

December 2021

1 Summary

This analysis examines factors related to a student's final grade and how alcohol consumption influences that grade. Implementation of exploratory data analysis and stepwise selection help in determining the linear regression model. It is discovered that a student will attain a higher final grade if they have excellent first and second-period grades, excellent family relationships, or live far away from school tend to perform better than students who live less than 15 min from school. According to the analysis, female students or students who frequently go out drink a lot over the weekend. Alternatively, students who attend Mousinho da Silveira or study a lot, or have no romantic relationships drink less during the weekdays.

2 Introduction

It is no secret that students start drinking alcohol before reaching legal age. In this analysis, we aim to quantify the effect of alcohol consumption on a student's final grade. We are also interested in exploring other factors like social, gender, and study information that affect a student's class final grade.

3 Data

The data were obtained from Kaggle (1). The Student Alcohol Consumption data were collected from a survey of Math and Portuguese language courses in secondary school. We have two data sets corresponding to each course with both data sets containing a lot of interesting social, gender, and study information about students.

The social attributes include family size, parent cohabitation status, mother and father education, mother and father job information, guardian, quality of family relationships, health status, age, etc.

The study information about students contains grades from the first and second period, final grade, romantic relationships status, weekday and weekend alcohol consumption, days the student was absent from school, number of past class failures, whether student paid for extra classes, nursery school attended, school, extra-curricular activities, free time, study time information, etc.

Each data set has 33 columns and 382 students take both Math and Portuguese courses. For this project, we concentrate on the 395 students who took the Math course. Some students received 0 as a final grade and this corresponds to a 0 in both the first and second-period grades. We exclude these from the analysis. We center the means of absence, first and second-period grades to reduce multicollinearity. There is no missing data in both data sets.

3.1 Exploratory Data Analysis

Most of the variables are categorical with at least 2 levels. Age, failures, absences, and grade variables are numerical. The response variable is the final grade and has a somewhat bell-shaped curve.

Family size, parent cohabitation status, family education support, extra paid classes, nursery, romantic relationship status, extra-curricular activities predictors have the same medians.

For the numeric variables, we use scatter plots to study the relationship between the predictor variable and the response variable. The age and failures variables have a discrete distribution with values that are not randomly distributed and are not considered during modeling. The higher the first or second-period grade the higher the final grade.

For the categorical and ordered variables, we plot box plots to study the the relationship between the predictor variable and response variable. Students who live more than an hour away from school have the least median and smallest distribution compared to students who live less than 15 min away from school. Students have varying medians of family relationships with most have students having excellent family relationships. Weekday and weekend alcohol consumption have varying medians with the lowest consumption in the high category(See Figure 1).

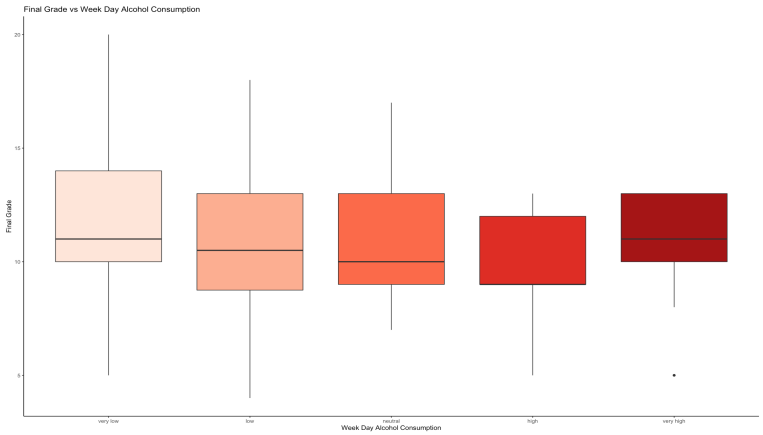

Figure 1: Final Grade vs. Weekend Alcohol Consumption

We did interactions between the response variable, predictors, and weekday and weekend alcohol consumption. There is an interaction between these predictors i.e. study time, schools, romantic relationships, weekday alcohol consumption, and the final grade. An interaction also exists between these predictors i.e. school, sex, address, a reason to choose the school, family education support, romantic relationship, going out, health, weekend alcohol consumption, and the final grade.

4 Model

$$\rho(y_i|x_i) = \mathcal{N}(11.11 + 0.11_{xiG1cent} + 0.89_{xiG2cent} - 0.13_{xitraveltime15-30-min} - 0.09_{xitraveltime30-1hr} + 0.77_{xitraveltime>1hr} + 0.24_{xihealthbad} + 0.04_{xihealthfair} + 0.06_{xihealthgood} + 0.20_{xifamrelbad} + 0.45_{xifamrelfair} + 0.33_{xifamrelgood} + 0.73_{xifamrelexcellent}, \sigma^2)$$

4.1 Model and Variable Selection

No additional transformations were done on the predictors other than the former centering of the means of the first and second-period grades and absences.

Step-wise model selection based on AIC and BIC was adopted in this study. The null model included weekday and weekend alcohol consumption as the predictors. Two full models were considered, one including all predictors minus interactions and the other was the interaction model interactions including the predictors we found worth exploring in the EDA.

Both BIC models from the two full models were the same only including the first and second period grades as the only significant predictors. The two AIC models did not differ by much except that one included more predictors than the other. We chose the latter for the final model which was significant at $\alpha = 0.1$

4.2 Model Assessment

We did a check for assumptions for a linear regression model and looking at the residuals vs. predictors plot linearity assumption holds since the points are randomly distributed. There's no clear violation of independence and equal variance assumptions since the points are randomly distributed above and below the average line with approximately equal distance toward the average line as seen in figure 3. Therefore, the observations are independent from each other and there is equal variance for all predictors on the response. From the QQ Plot shown in figure 2, normality holds since the majority of values fall at the 45-degree line.

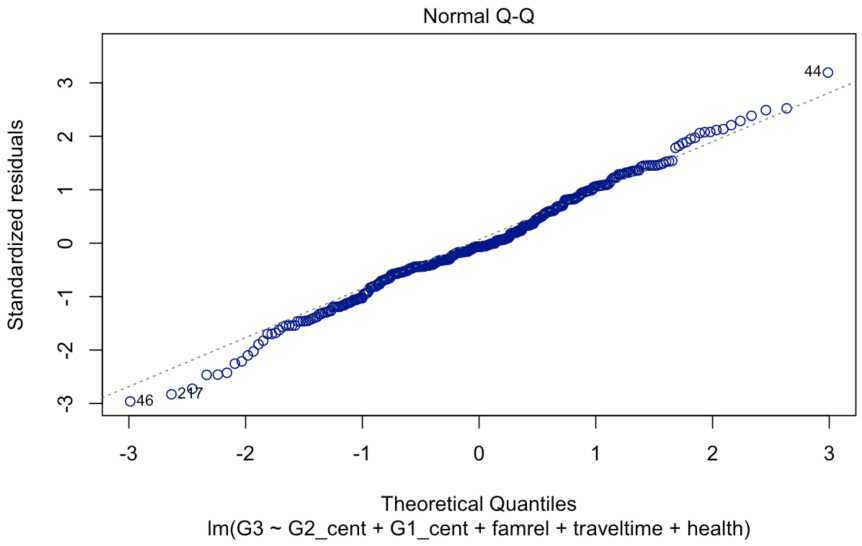

Figure 2: Normal QQ Plot

4.3 Model Validation

From the residuals vs. leverage plot, there are no presence outliers from the non-existence of high leverage or high influential points. There is a presence of multicollinearity with family relationships being highly correlated and the first and second-period grades moderately correlated. The model prediction accuracy of the final grade is 71%.

4.4 Model Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.1119	0.3139	35.40	0.0000
G2_cent	0.8947	0.0319	28.07	0.0000
G1_cent	0.1054	0.0308	3.42	0.0007
famrelbad	0.2025	0.3756	0.54	0.5901
famrelfair	0.4467	0.3299	1.35	0.1767
famrelgood	0.3277	0.3184	1.03	0.3041
famrelexcellent	0.7283	0.3224	2.26	0.0245
traveltime15 to 30 min.	-0.1307	0.0983	-1.33	0.1844
traveltime30 min. to 1 hour	-0.0967	0.1926	-0.50	0.6160
traveltime>1 hour	0.7667	0.3129	2.45	0.0148
healthbad	0.2385	0.1796	1.33	0.1851
healthfair	0.0444	0.1526	0.29	0.7710
healthgood	0.0603	0.1627	0.37	0.7112
healthvery good	-0.1417	0.1408	-1.01	0.3147

The adjusted R-squared(R^2) of this model is 0.94 meaning about 94% variability in the final grade is explained by our model. In addition, the residual standard error is 0.801 with 343 degrees of freedom.

We expect the average final grade to increase by about 0.89 at the average second-period grade points, holding all other variables constant with 95% CI (0.83, 0.96).
We expect the average final grade to increase by about 0.11 at the average first-period grade points, holding all other variables constant with 95% CI (0.04, 0.17).

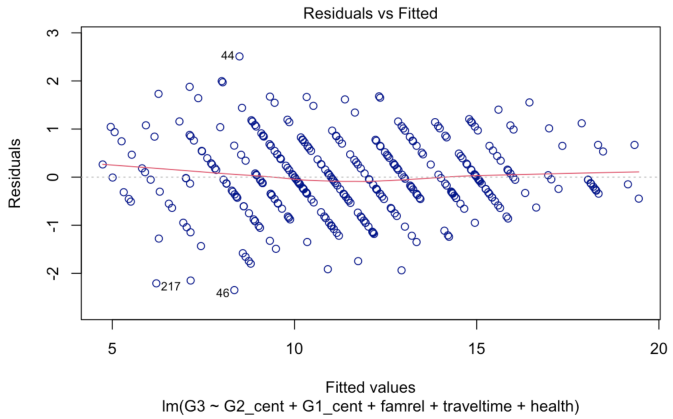

Figure 3: Residual vs. Fitted values

Holding all other factors constant, we expect students with excellent family relationships to have a final grade around 0.73 higher than students with very bad family relationships at a 95% CI (0.09, 1.36).
Holding all other factors constant, we expect students with a travel time greater than 1 hour to have a final grade of 0.76 higher than students with a travel time of less than 15 min at a 95% CI (0.15, 1.38).

Regarding how alcohol consumption by a student affects their grade, we discovered that females consume a lot of alcohol over the weekend in comparison to males. The Gabriel Pereira school has a higher weekday alcohol consumption compared to Mousinho da Silveira. Weekday alcohol consumption is low during increased study times and in students with no romantic relationships. More so, students who go out a lot drink more over the weekend than those who do not go out.

5 Conclusions

This study is beneficial in educating us on what drives high alcohol consumption in teenagers. The results of which lead us in finding ways to reduce alcohol usage among young adults. By understanding which factors affect a child's final grade, parents and teachers can help their children.

Although the model performs relatively well in predicting the final grade, we found multicollinearity in some predictors when not resolved limits our ability to interpret those coefficients or identify hidden statistically significant independent predictors.
In the future, we can continue the analysis study by exploring data in Portuguese language students and compare the results between students who took both courses.

References

[1] Student Alcohol Consumption. (2016, October 19). Kaggle. https://www.kaggle.com/uciml/student-alcohol-consumption