Final Report

due November 16, 2021 by 11:59 PM

Danielle Mensah, Haby Sow, Colin Lee

11/12/2021

Abstract

This report outlines the effects of alcohol consumption on students' efficacy of learning and whether family circumstances play a role in the amount of alcohol consumed by these students. We used data from a dataset we found on Kaggle that was posted by UC Irvine. The dataset collected data on student self-reported alcohol consumption in Portuguese math class in two secondary schools, Gabriel Pereira and Moushino de Silveira during the 2005-2006 school year.

To focus on the effect of alcohol consumption on student learning we used the variables relating to alcohol consumption. On the other hand, to focus on the family circumstances in comparison to alcohol consumption we used the variables related to alcohol consumption. With these variables, we built regression models to decide which attributes were significant to our findings. In addition, we used proper visual models to display our findings in a visual manner.

Our model shows that having extra paid classes within the course subject of math plays a significant factor in whether or not one drinks more often. In addition, our data showed that increased alcohol consumption also showed an increase in the amount of failed courses.

Introduction

There are many beliefs about the prevalence of student drinking and what sort of factors impact or are indicative of high levels of drinking as well as how drinking may impact individual academic performance. For our project, we will be studying two main questions with regard to alcohol consumption.

The first question is what family or external circumstances are drivers for high student alcohol consumption. Insight into the impacts of these factor may allow for a better understanding of how family education, support, size, and others may be associated with binge drinking.

The second question we seek to answer is how student alcohol consumption correlates with student life whether it be through absences, class failures, marks received on exams, activities outside of school, or desire to attend university.

In order to conduct our study of these two questions, we will be working with a dataset we found on Kaggle. The data was collected for a research paper called "Using Data Mining to Predict Secondary School Student Performance." The authors describe how the surveys were developed and reviewed by school staff and students (Cortez et al 2). The surveys were then delivered to the students through paper questionnaires with predefined answer choices for 37 different questions, 33 of which are variables in the dataset (Cortez et a 2). The questionnaires were answered by all 395 math students in the Gabriel Pereira and Moushino da Silveira secondary public schools (Cortez et al 2). The secondary school students are what would be considered as "high school" in the United States, with age ranges from 15 - 19 years old, but with a few older individuals up to 22 years old. Additionally different from the United States is that the legal drinking age in Portugal is 18 years old, which is on the upper spectrum for secondary students. Thus, the drinking prevalence in this

data set may not correlate to prevalence found in American high schools. However, this data may mirror drinking at the university level better, as to the 21 year age requirement in the US is in the upper age range for university students as is 18 in the upper age range for Portuguese secondary school students.

Overall Methodology

In order to assess the first question, we modeled a logistic regression with binge drinking being the response variable and familial and external variables serving as the predictor variables. Through this analysis, we intend to understand the significance of certain family and external factors in impacting whether or not a student is a workday binge drinker and interpret the severity of impact of those variables.

In order to answer the second question, we modeled a multiple linear regression with workday alcohol consumption as the response variable and more internal factors, such as free time, study time, quality of relationships, extracurriculars, and academic performance as explanatory variables. We additionally modeled linear regressions to understand how workday alcohol consumption impacted academic performance.

With regard to the data availability, there are no missing values and all questions in the survey were filled out by all 395 respondents. Further data wrangling and discussion of the the specific variables utilized for each of the above analyses will be described in their specific sections below.

Logistic Regression Methodology:

We conducted a logistic regression to determine what sort of family or external factors may be drivers for student alcohol consumption. We decided to pursue a logistic regression in this case due to the categorical nature of many of the variables. Furthermore, we wanted to do a logistic regression on top of a linear regression, where instead of quantifying the amount of alcoholic drinks, we can assess how these indicators are involved with impacting whether or not a student can be considered an "alcoholic" or "binge drinker."

Thus, in order to set up this logistic regression, we had to begin by transforming the daily alcoholic consumption value into a binary variable.According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA), binge drinking is defined as 4 or more drinks is for females and 5 or more drinks is for males (NIAAA). Furthermore, for youths, which is where we will categorize the students here, the drinks are 3 for girls and 3 to 5 for boys (NIAAA). However, for this dataset, the values 1-5 do not indicate the amount of drinks, but rather, a self-reported drinking level, where 1 is "very low" drinking and 5 is "very high" drinking. This is very arbitrary, but in order to create a more binary definition, we defined drinking levels of 4 and 5 (high and very high drinking) on workdays. We additionally focus on workday binge drinking as opposed to the other variable, which is weekend binge drinking, because workday binge drinking is more negatively stigmatized, and is a larger factor in a student's academic and family life than weekend binge drinking.

Thus, in order to create this binary variable, we created an additional variable called "binger" with the mutate() function, where 1 represents someone who self-reported a workday drinking level of 4 or 5 (high and very high drinking level) and 0 represented moderate, low, or very low drinking levels. Furthermore, we factored the data to "Yes" and "No" for binge drinker and not binge drinker, respectively, and then we releveled the data to make non-binge drinkers the referent group. We made non-binge drinkers the referent group because that should be the norm, and we want to observe how binge drinkers compare with non-binge drinkers.

Next, we had to select the predictor variables for our regression. Because we wanted to answer what sort of external or familial factors impact whether a student is a binge drinker, and to accomplish this, we chose variables that were more representative of the circumstances surrounding them than something they could control themselves. Thus, we used the following variables: residence, family size, parental separation, mother and father education level, educational support, family educational support, extra paid tutoring classes, whether or not the student attended nursery school.

Thus, we developed the following logistic regression model:

logit(pii) = log(pii/(1-pii)) = B0 + B1(urbani) + B2(famsmalli) + B3(parents_togetheri) + B4(mother_secondaryi) + B5(father_secondaryi) + B6(school_supporti) + B7(family_supporti) + B8(extra_tutoringi) + B9(attended_nurseryi) + B10(internet_accessi) + B11(traveltimei)

such that

urbani = 1 if the respondent lives in an urban environment, famsmalli = 1 if the family size is less than 3, parents_togetheri = 1 if the parents are together, mother_secondaryi = 1 if the mother's education level is secondary or above, father_secondaryi = 1 if the father's education level is secondary or above, school_supporti = 1 if the student is receiving school support, family_supporti = 1 if the student is receiving extra family support for education, extra_tutoringi = 1 if the student is receiving extra paid classes within math, and attended_nurseryi = 1 if the student attended nursery school before. Traveltime is a numerical variable, where 1 represents < 15 minutes travel time to school, 2 represents between 15 and 30 minutes travel time, 3 represents 30 minutes to 1 hr travel time, and 4 represents > 1 hr travel time.

For the parental education, since this value was not originally binary, we assessed that it would be most relevant to assess whether or not the education level was secondary or not since if the respondent's parents had reached the at least the same level of education as the respondent him/herself, then there would be some likelihood of greater parental support due to them having shared the same experiences. Regardless, as we will explain in the next section, switching the value to be higher than secondary education makes no impact.

```
## # A tibble: 12 x 7
##    term             estimate std.error statistic p.value conf.low conf.high
##    <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
##  1 (Intercept)        0.0115     1.36     -3.27  0.00106 0.000613     0.139
##  2 urban              1.67       0.667     0.768 0.443   0.491        7.04
##  3 famsmall           1.09       0.577     0.149 0.882   0.328        3.27
##  4 parents_together   0.336      0.733    -1.49  0.137   0.0861       1.67
##  5 mother_secondary   1.63       0.642     0.766 0.444   0.466        5.92
##  6 father_secondary   0.816      0.608    -0.333 0.739   0.248        2.75
##  7 school_support     1.70       0.735     0.719 0.472   0.334        6.51
##  8 family_support     0.772      0.577    -0.449 0.654   0.254        2.52
##  9 extra_tutoring     5.77       0.640     2.74  0.00616 1.78        22.7
## 10 attended_nursery   0.246      0.574    -2.44  0.0145  0.0786       0.773
## 11 internetaccess     1.14       0.828     0.153 0.878   0.262        7.92
## 12 traveltime         2.76       0.333     3.05  0.00228 1.43         5.38
```

Logistic Regression Analysis

Upon completing the logistic regression, we found three variables to be significant at a p-value of 0.05, which were whether or not a student had extra tutoring or not (B8, coefficient for extra_tutoringi), whether a student had attended nursery or not (B9, coefficient for attended_nurseryi), and the travel time to school for the respondent (B11, coefficient for traveltimei). All other estimates had much higher p-values, which were not close to being significant. Furthermore, we attempted to see how changing the level of education we identified as appropriate for our binary variables on parental education, but the p-value did not shift significantly (change from 0.58/0.60 to 0.66/0.80 for mother's education and father's education respectively). Thus, only for extra tutoring and nursery attendance do we reject the null hypothesis that their estimates are zero and we can state that there is relationship between these variables and workday binge drinking.

Moving forward with the extra_tutoringi variable, the estimate came out to be 4.285 with a p-value of 0.0122 and 95% CI between 1.45 and 14.73. Because the values are exponentiated, this indicates that the odds ratio for this value 4.285, which means that a student with extra paid math classes is 4.285 times as likely to be a binge drinker than someone who does not have extra paid math classes. This is quite interesting and contradicts our perception regarding this variable that educational support through tutoring might indicate less likelihood for binge drinking. Some speculation as to why tutoring may be correlated with binge drinking

is possibly due to a perceived need for tutoring for the respondent from the students' parents because of the drinking itself.

With regard to the attended_nurseryi variable, the estimate was 0.252 with a p-value of 0.0139 and 95% CI between 0.08 and 077. As this value is exponentiated, the odds ratio is 0.252 and thus a student who has attended nursery may be 0.252 times as likely to be a binge drinker. This is logical as nursery attendance may be indicative of a positive external factor, as it is possible that a family's prioritization of education and possible aversion to binge drinking for children.

For the traveltimei variable, the estimate was 2.758 with a p-value of 0.0023 and 95% CI between 1.43 and 5.38. As the estimate is exponentiated, the odds ratio for the variable is 2.758, indicating that a student who has a one unit increase in travel time to school, which is essentially a doubling of travel time based on the variable's parameters, the individual is 2.76 times as likely to be a binge drinker. This is interesting, as it shows that students who live farther away from school are far more likely to be binge drinkers.

Thus, in terms of external and familial factors impacting whether or not a student respondent is a school/workday binge drinker, we assess that there are only three significant variables, which are whether or not the student has extra paid math classes, whether the student has attended nursery school or not, and the travel time of the student to school.

```
## # A tibble: 4 x 7
##   term             estimate std.error statistic p.value conf.low conf.high
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)        0.0115     1.36      -3.27 0.00106 0.000613     0.139
## 2 extra_tutoring     5.77       0.640      2.74 0.00616 1.78        22.7
## 3 attended_nursery   0.246      0.574     -2.44 0.0145  0.0786       0.773
## 4 traveltime         2.76       0.333      3.05 0.00228 1.43         5.38
```

Fit for Logistic Regression

To assess how well our logistic model fits, we used the data set to create a training and testing data set. We used seed 0 and split the data 80% training and 20% test. For our small model, we considered only the three significant variables, extra tutoring, nursery attendance, and travel time. For the big model, we used the original model with all the original predictors, and for the small model, we used all the significant predictors that we had rejected the null hypothesis for.

Unfortunately, we could not fit the charts on this final report, but we were able to find information about the area under the curve for each model. For the large model, the area under the curve is 0.974, indicating very good fit. For the small model, the area under the curve is 0.686, which is not is great, but still better than just the estimate, which would be 0.50, or a random fit. Thus, although the large model contains many predictors that are not statistically significant, the large model is a very effective prediction model. This could be a result of the data regarding those predictors not being sufficiently large to provide greater insight into their statistical significance. Thus, since the small model is still a decent fit at 0.686, and the predictors are within our acceptable p-values, we will continue with our analysis that the variables of having paid math classes, attended nursery, and travel time to school as the statistically significant predictors for alcohol binge drinking.

Multiple Linear Regression - Factors Influencing Workday Alcohol Consumption

How does student life impact daily alcohol consumption?

A multiple linear regression was run to understand how student life impacts alcohol consumption. We use Dalc (daily alcohol consumption) as response variables and failures(number of past class failures) , goout (going out with friends) , studytime (weekly study time) , freetime (free time after school), famrel (quality

of family relationships), activities (extra-curricular activities), and G3 (final grade in class), romantic (in romantic relationship), health (health status), and absences (number of absences) as explanatory variables. Through this linear regression model, our goal is to answer our second question by understanding how student internal factors explain one's own drinking level.

It is additionally important to note how these variables are read in the dataset, where workday alcohol consumption (Dalc) is a self-reported scale from 1-5, where 1 is "very low consumption" and 5 is "very high consumption." Number of classes failed (failures) is an integer value of the number of previous classes failed if the number is 1, 2, or 3, and if it's above 3, the value is always 4. Going out with friends (goout) is also a 1-5 self-reported scale of how often the respondent goes out where 5 is very high and 1 is very low. Weekly study time (studytime) is scaled for 1 as < 2 hours, 2 is 2-5 hours, 3 is between 5 to 10 hours, and 4 is greater than 10 hours. Free time after school (freetime) is also a 1-5 self-reported scale where 5 is very high and 1 is very low, and family relationships is similarly a 1-5 scale where 5 excellent and 1 is very bad. Activities is a categorical variable of yes or no, where yes indicates the respondent partakes in extra-curricular activities. The final grade (G3) is a numerical grade ranging from 0 - 20. Being in a romantic relationship (romantic) is a categorical variable of yes or no. Health is a subjective variable, with 5 being very good and 1 being very bad, and absences is an integer amount from 0 to 93.

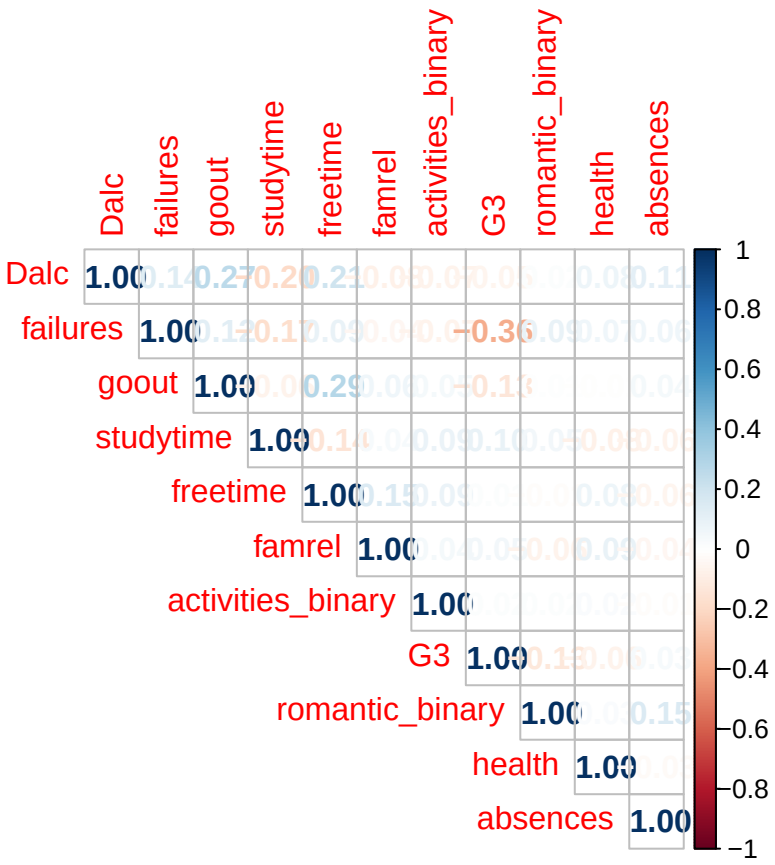
We chose to use the final grade as opposed to a mix of the first and second period grades since we want to assess how one's overall academic performance is related to workday alcohol consumption, and we are not looking for specific granularity with regard to a specific period's assessment's relationship with workday alcohol consumption. Additionally, we converted all of the categorical variables to numerically binary variables where 1 is yes and 0 is no.

The linear regression used is the following one: Dalc= B0 + B1(failures) + B2(goout) + B3(studytime) + B4(freetime) + B5(famrel) + B6(activities_binary) + B7(G3) + B8(romantic) + B9(health) + B10(absences)

In addition to building a multiple linear regression model, we also created a correlation matrix to visualize the correlation coefficients between workday alcohol consumption and what we found to be the significant variables from the linear regression model. Through this matrix, we plan to visualize how well the explanatory variables are correlated with daily alcohol consumption, but more importantly, assess how the variables may be dependent on one another as well.

Linear regression analysis - Factors Influencing Workday Alcohol Consumption:

```
## # A tibble: 12 x 5
##    term              estimate std.error statistic   p.value
##    <chr>                <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)        1.01      0.351      2.88   0.00425
##  2 failures           0.0695    0.0636     1.09   0.275
##  3 goout              0.175     0.0400     4.37   0.0000158
##  4 studytime         -0.139     0.0523    -2.65   0.00836
##  5 freetime           0.129     0.0453     2.85   0.00454
##  6 famrel            -0.106     0.0478    -2.21   0.0280
##  7 activities_binary -0.125     0.0851    -1.47   0.144
##  8 higheryes          0.00423   0.204      0.0207 0.984
##  9 G3                 0.00260   0.0100     0.259  0.796
## 10 romantic_binary   -0.00647   0.0917    -0.0706 0.944
## 11 health             0.0449    0.0306     1.47   0.143
## 12 absences           0.0106    0.00537    1.97   0.0491
```



In this regression output, only the estimates for goout, studytime, freetime, famrel, and absences were significant at a p-value of 0.05. Thus, for these variables, we reject the null hypothesis that their estimates are 0, and can state that there exists a relationship between these variables and workday alcohol consumption.

From our linear model output, it appears that the more frequently a student goes out with friends, the more the student drinks. A one unit increase in a self-reported, subjective amount of going out with friends (such as high to very high) increases the value of workday alcohol consumption by 0.176. In this case, we can say that going out with friends impacts student daily alcohol consumption. Additionally, a one unit increase in studytime decreases alcohol consumption by 0.146, indicating that those who dedicate more time to studying or have more time for studying have lower alcohol consumption. Furthermore, the higher the quality of family relationships, the less alcohol consumption is. A one unit increase in the quality of family relationships decreases daily alcohol consumption by 0.1. The model's outputs also portray how a one unit increase in freetime also increases alcohol consumption by 0.13. Lastly, daily alcohol consumption is also significantly related to absences, where a single increase in the number can be traced to a 0.01 increase in alcohol consumption.

With regard to dependence among the variables as seen in the correlation matrix, the highest correlation coefficients we see are with failures being negatively correlated with average grade with a correlation coefficient of -0.38, and going out time is positively correlated with free time with a correlation coefficient of 0.29. Although it is reasonable to assume that these variables are somewhat dependent and correlated because low grade average is likely related to failing class and people who have more free time are likely to go out more, these correlations coefficients are moderate to low. Therefore, we will not use them to rule out any significant explanatory variables.

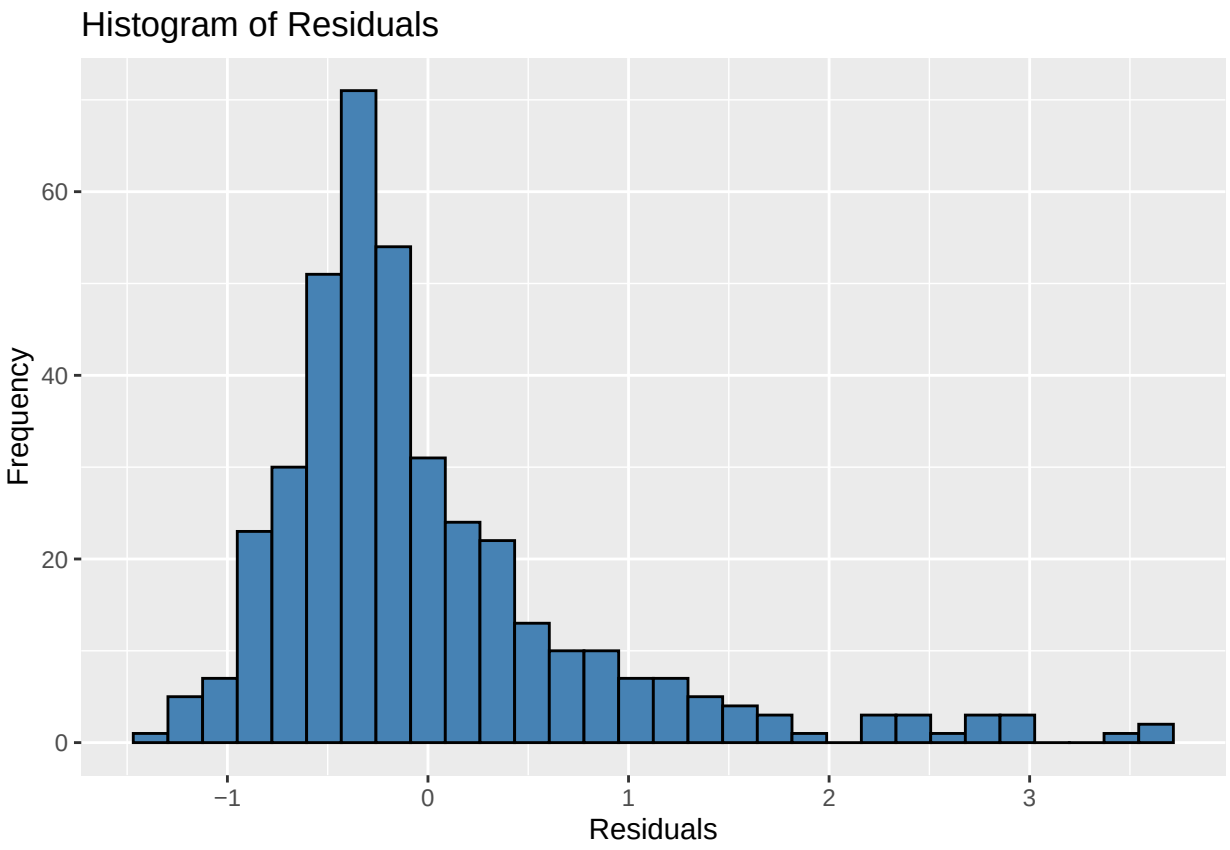
Furthermore, it is important to note that the y-intercept is 1.28, indicating that the base level of drinking is very low amongst the respondents.

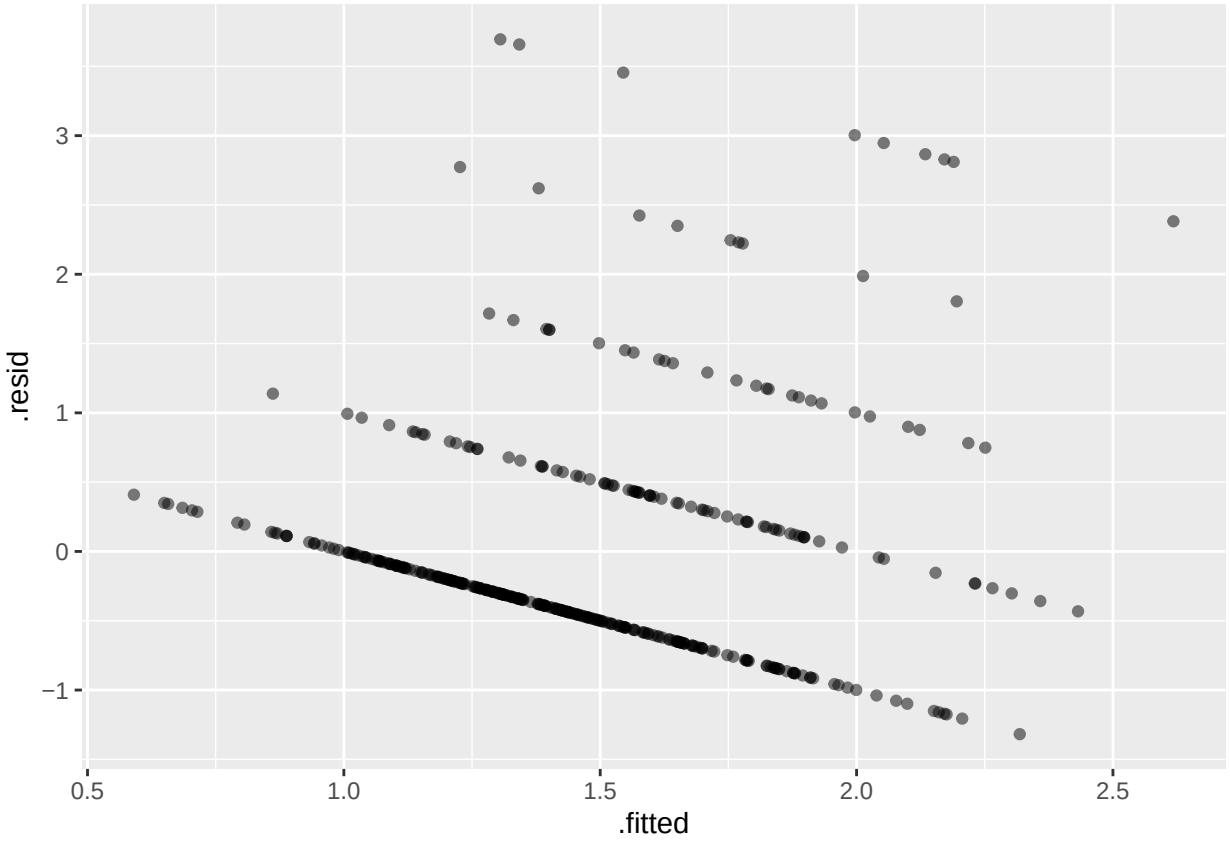
It is extremely important to proceed with caution in interpreting the analysis of these variables and

relationships because most of them are heavily subjective and not very-well quantified. Especially with regard to the estimates, it is hard to assess what a 0.1 or 0.01 increase in workday alcohol consumption truly indicates since workday alcohol consumption was originally a self-reported, subjective value.

Thus, our multiple linear regression model takes the following form: Dalc= 1.28 + 0.176(goout) + 0.146(studytime) + 0.1(famrel) + 0.13(freetime)

```
## Warning: Use of `model_regression_follow$.resid` is discouraged. Use `.resid`
## instead.
```





The R adjusted squared here is approximately 0.13, which implies that 13% of the variation in the response variable is explained by our linear model, which is relatively low, indicating how our model may not be a good predictor for variation in workday alcohol consumption.

For our RMSE and R^2 analysis, to assess the performance of our model, we split the student data into 75% training and 25% testing data. We then evaluated the RMSE and R^2 of our model for the training and testing data, and found the RMSE and R^2 for the training data to be 0.788 and 0.150 respectively and the RMSE and R^2 value for the testing data to be 0.937 and 0.153 respectively. The RMSE is lower for the training data by a significant amount, which makes sense since the model is built for hte training data set. Interestingly, the R^2 essentially remained the same (increased very slightly) from the training set to the test set, indicating the model does not degrade so much with data beyond the training data set, so that the model is a fairly competent fit.

The histogram of the residual is approximately symmetric which satisfies the condition of normality for linear regression. The mean of the residuals (-1.38079e-16) which is enough close to 0 to strengthen our linear regression model.

However, though the conditions of normality was satisfied, looking at the residual plot, the conditions of independence was not satisfied due to an identified pattern of the residuals. Residuals are linearly distributed which shows some flaws of our linear regression model.Moreover, the conditions of homoscedasticity was not satisfied because residual points are not constant. In some regions of the plot, residual points are more compacted than in other.

Overall, we can conclude that the linear regression model we used is competent at predicting student workday alcohol consumption due to our performance analyses. It is important to note that our linear model is not perfect due the lack of satisfaction of some assumption for linearity and low R^2 adjusted.

Linear Regression - Alcohol Consumption Impact on Academic Performance:

With regard to our second question, we also wanted to further analyze the relationship between alcohol consumption and academic performance. Thus, we conducted two linear regression models. For both models, workday alcohol consumption is the explanatory variable while the response variable for the first model is the final grade and the response variable for the second model is the number of class failures. The variables are the same as described and discussed from the previous multiple linear regression model.

Analysis of Alcohol Consumption's Impact on Academics

G3 = B0 + B1(Dalc)

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   10.8       0.448     24.2  7.13e-80
## 2 Dalc          -0.281     0.259     -1.09 2.78e- 1
```

The linear regression here was insignificant as the p-value for the estimate, B1, for workday alcohol consumption of 0.28, which is not significant at our preferred p-value of 0.05. Thus, we do not reject the null hypothesis that B1 = 0.

failures = B0 + B1(Dalc)

```
model_dalc_failures<-linear_reg()%>%
  set_engine("lm")%>%
  fit(failures~Dalc, data=student)
print(tidy(model_dalc_failures))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    0.166    0.0721      2.30 0.0218
## 2 Dalc           0.114    0.0417      2.72 0.00677
```

However, the linear regression for the number of failures as the response variable does produce a significant estimate for Dalc, with a p-value of 0.007, far below our preferred p-value of 0.05. Thus, we can reject the null hypothesis and state that there is evidence to support a relationship between workday alcohol consumption and the number of course failures with the following model:

failures = 0.11(Dalc) + 0.17

With this regression model, the result shows us that a one unit increase in daily alcohol consumption increases the number past failures by 0.11. Although we can state that daily alcohol consumption negatively impacts one's academic performance by contributing to the number of failed classes, the impact is relatively small, as someone who increases from moderate to very high drinking would only increase their number of past class failures by 0.22.

Conclusion

The logistic regression was able to answer our first research question regarding the importance of certain family and external factors in impacting whether a student was a workday alcohol binge drinker. From the logistic regression model, we discovered that amongst many family and external factors, including student's residence, family size, parental separation, mother and father education levels, school support, family support, additional paid math classes, and attendance of nursery school, only attendance of nursery school and

additional paid math classes appeared to be significant. Moreover, if a student attended nursery school, he or she was 0.252 times as likely to be a binge drinker while if a student's travel time effectively doubled, then he or she would be 2.758 times as likely to be a binge drinker. Most significantly, if someone had additional paid math classes, he or she was 4.285 times as likely to be a binge drinker. Thus, there are only a few external factors involved in binge drinking, and how the variables are connected is to be further investigated.

Through the regression analysis, we have found some significant factors that contribute to workday alcohol consumption. We found that an increase in freetime, going out with friends, and number of school absences can lead to an increase in weekday alcohol consumption. Moreover, we have seen that having good quality of family relationships and spending more time studying decreases weekday alcohol consumption. Furthermore, we found that workday alcohol consumption does not have a significant relationship with academic performance in terms of grades, but it does play a significant (in terms of p-value) role in the number of past class failures, although its actual impact is relatively small on that front. Thus, there are a few internal factors in one's personal life that contribute to workday alcohol consumption, and alcohol consumption is not very negatively impactful on academic performance.

All of these results, however, need to be interpreted with caution as this data represents only a small subset of all secondary school students, as it was taken only taken in two secondary schools in a region of Portugal, and only addressed to math students in those two schools. For future analysis to better understand the relationships among family and external factors, alcohol consumption, and academic and personal life, further data needs to be collected from a wider populace with more random sampling.

To make further data analysis more compelling, there needs to be less subjectivity and greater objectivity in the variables, especially workday alcohol consumption. In this dataset, it is a self-reported value of very low to very high alcohol consumption, but this provides no bearing as to how many actual drinks someone is consuming, and weakens our current analyses of the relationships between workday alcohol consumption/workday alcohol binge drinking with other factors.

Citations:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

"Underage Drinking." National Institute on Alcohol Abuse and Alcoholism, U.S. Department of Health and Human Services, https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/underage-drinking.