

Analyzing the relationship between student alcohol consumption and school performance

Fat GoldFish: Daniel Bolja, Damla Ozdemir, Elyse McFalls, Nathan Nguyen

2020-11-15

I. Introduction

There has been a recent study that looked at changes in the drinking habits of Americans, which finds that “Americans Are Drinking 14% More Often During Pandemic”, and “Instances of heavy drinking among women, which for women was defined as four or more drinks within a couple of hours, spiked by 41%”. Although these studies did not look at younger adults, we found this conclusion interesting, so we want to analyze data having something to do with alcohol consumption in a younger population. We want to investigate the relationship between alcohol consumption in younger age groups and its effects on academic performance. We referenced a study using various types of data to try and predict secondary school performance of Portuguese students, with weekend and workday alcohol consumption being two important variables to consider. We want to determine if alcohol consumption is a significant predictor of academic success, especially in grades, Weekend study time, absences, and the number of failed classes.

We are using the data from the study of Cortez and Silva (2008), which focused on the study performance of secondary students based on their alcohol consumption. The data contains 649 observations which are secondary students enrolled in Portuguese languages from two public schools in the Alenjetto region of Portugal during the 2005 - 2006 school year. Since at this time, the majority of Portuguese public schools’ information systems remained very poor, the authors built the database from two sources: paper-based school reports with few attributes from grades and number of absences, and questionnaires about several demographic, social/emotional, and school-related variables that were expected to affect student performance. Data with a lack of identification details were discarded. There are a total of 33 variables, where our main focus will be Workday (**Dalc**) and Weekend (**Walc**) alcohol consumption, first, second, and third-period grades (**G1**, **G2**, **G3**), Weekend study time (**studytime**), absences (**absences**) and the number of failed classes (**failure**).

Variables	Meaning
Dalc	workday alcohol consumption (from 1 to 5)
Walc	weekend alcohol consumption (from 1 to 5)
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 (n if $0 \leq n < 3$, else 4)
absences	number of school absences (from 0 to 93)
G1	first period grade (from 1 to 20)
G2	second period grade (from 1 to 20)
G3	final grade (from 1 to 20)

II. Methodology

Preliminary distribution: Firstly, we plot several variables to examine their distribution for our preliminary works. For the Weekend Alcohol Consumption and Workday Alcohol Consumption variables, we factor

the intervals of consumption, ranging from “very low” (1) to “very high” (5) to make the data easier to understand and work with. As we used factors for these variables, we visualized the data using a bar chart. Next, for the school grades, we visualize three different period grades by histograms (since grades are mostly continuous) to see if there were any noticeable changes with the passing time in school.

Our third type of variable was the Number of Absences, as we felt that students who drink more might have more absences due to drunkenness or hangovers. Our fourth variable indicated Weekend Study Time, which we included because we thought that students who consumed more alcohol might have less time to study between drunkenness, parties, and hangovers. Of course, we also included a variable for Number of Failed Classes, in the same vein as the variables for grades, as our project focuses on academic success. All in all, these variables picked out from a long list of variables used in the original research on Portuguese schools, because we saw them as being most relevant to academic success.

It is important to notice that we also did not separate the two schools, because we were not sure of the quality of education at each school (we were unfamiliar with the Portuguese education system and the data did not give a clear indication of the quality of education, only that they are both public schools). As shown in the third hypothesis test and appendix II, we will split the students based on variables’ values.

Exploratory Analysis: Once we had our preliminary data visualizations and were familiar with the dataset/general shapes of the variables, we began an exploratory analysis, focusing on each set of variables separately, looking at the relationship between academic performance and alcohol consumption from various perspectives. As our main focus was on the proportions of different levels of academic performance based on alcohol consumption, we used visualizations that included both Workday and Weekend alcohol consumption as variables, and our other variables of school grades, Weekend study times, absences, number of failed classes, in turn. Since both Workday and Weekend alcohol consumption are ordinal variables, we used the appropriate ggplot visualizations of boxplot and percent stacked bar charts. While these visualizations can be helpful to see the shapes or patterns in data, we also reduced our results from each variable relationship into summary statistics in tables.

First hypothesis: For our Hypothesis Testing, we had three separate hypotheses for three of our variables. Firstly, to make our testing more streamlined, we aggregated some of our data. For the testing on student grades vs. alcohol consumption, used only Workday alcohol consumption as the main indicator of consumption, and calculated on final grade (G3). We performed simulation-based hypothesis tests to evaluate the claim: “Alcohol consumption will significantly predict final grades after controlling for other variables. There will be a negative relationship between grades and alcohol consumption.” We also chose to do a Linear Model for this claim, but the validity of these results is not trustworthy—the equal variance, independence, and normality condition for using a linear model appear to be checked, but the linearity condition is not met by our data.

Second hypothesis: For our second hypothesis, which focuses on Weekend study times, we evaluated the claim that “Alcohol consumption will significantly predict Weekend study times after controlling for other variables. There will be a negative relationship between Weekend study times and alcohol consumption”. We used chi-square and simulation-based tests, as well as linear models to check for this relationship between our two variables. The linear model is not reliable so we concluded based on results of other tests.

Third hypothesis: In the third hypothesis, we turned our attention to absences as an indicator of academic performance. The claim: “Alcohol consumption will significantly predict absences after controlling for other variables. There will be a positive relationship between absences and alcohol consumption.” Again, we used a simulation-based approach to test the claim that there is a negative correlation. We also constructed a linear model to test further. However, for the absences variable, we thought that the students’ distance from school to home might play a part in absences, so to control that variable, we looked at the data of students considered “close” to their school. Again, we used simulation and linear regression.

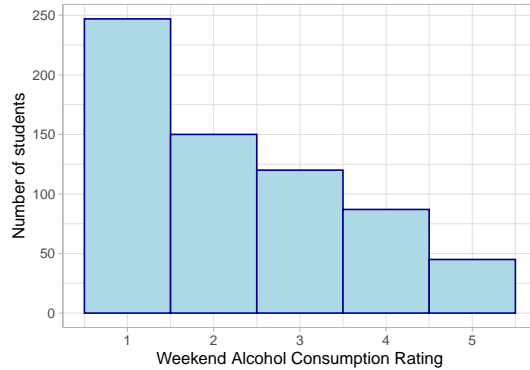
Final hypothesis: Finally, we considered the number of failed classes in another hypothesis. The claim: “Alcohol consumption will significantly predict the number of failed classes. There will be a positive relationship between the number of failed classes and alcohol consumption.” We followed the same procedure as in other hypotheses and created both chi-square and simulation-based distributions, and a linear model for this hypothesis test. The linear model is not reliable so we concluded based on results of other tests.

III. Results

1. Preliminary/ Distributions

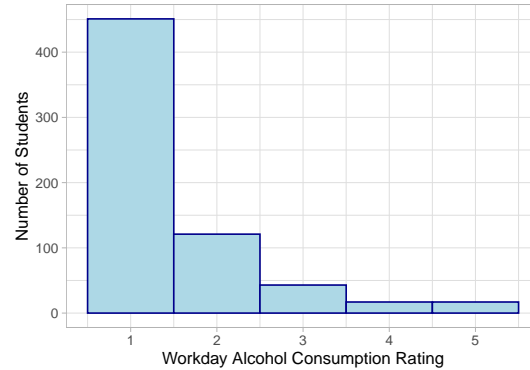
Distribution of Weekend Alcohol Consumption from Portuguese High Schoolers in 2005–2006

1 = very low, 5 = very high

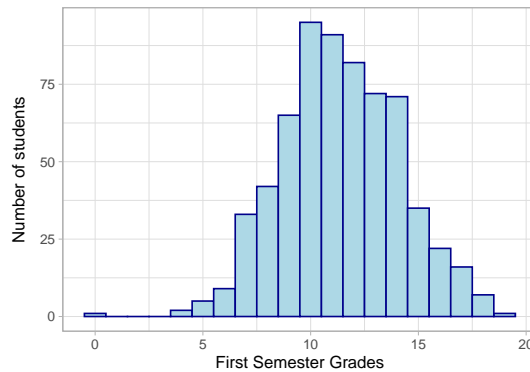


Distribution of Workday Alcohol Consumption from Portuguese High Schoolers in 2005–2006

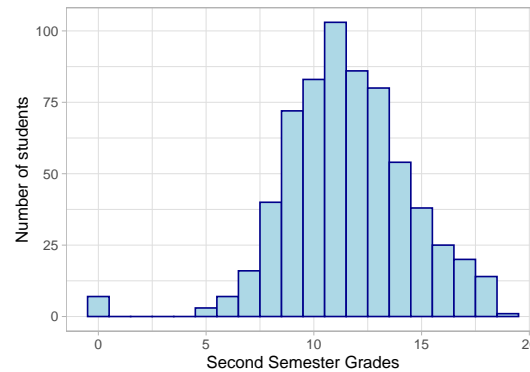
1 = very low, 5 = very high



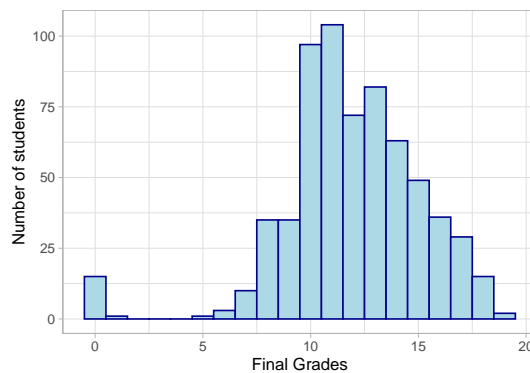
Distribution of First Semester Grades from Portuguese High Schoolers in 2005–2006



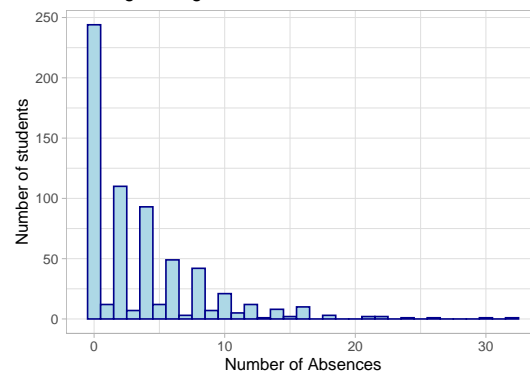
Distribution of Second Semester Grades from Portuguese High Schoolers in 2005–2006



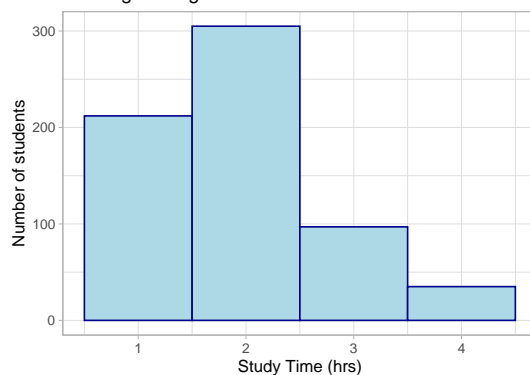
Distribution of Final Grades from Portuguese High Schoolers in 2005–2006



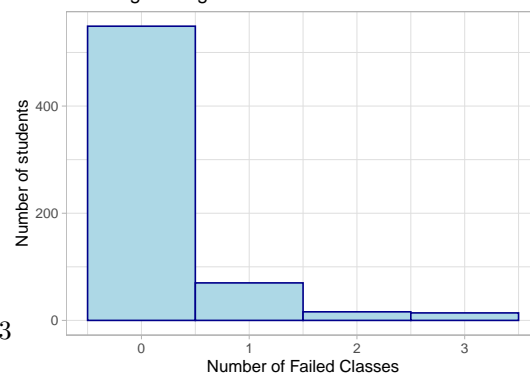
Distribution of Absences from Portuguese High Schoolers in 2005–2006



Distribution of Study Time (hrs) from Portuguese High Schoolers in 2005–2006



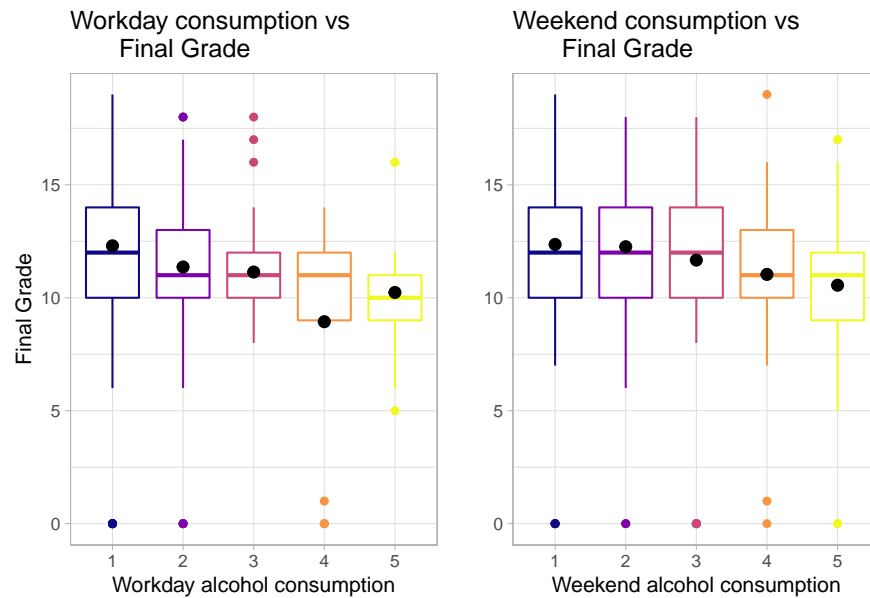
Distribution of Failed Classes from Portuguese High Schoolers in 2005–2006



The variables with the most normal distribution in our data come from the first, second, and final grades. On the other hand, the weekend alcohol consumption, Workday alcohol consumption, absences, and failed classes variables are highly skewed to the right. This indicates that the vast majority of individuals in our dataset rank very low on these dimensions (i.e. low weekend alcohol consumption, very few absences, etc.). The study time distribution is also slightly skewed to the left, but it's not as dramatic as the other distributions.

2. Exploratory Analysis

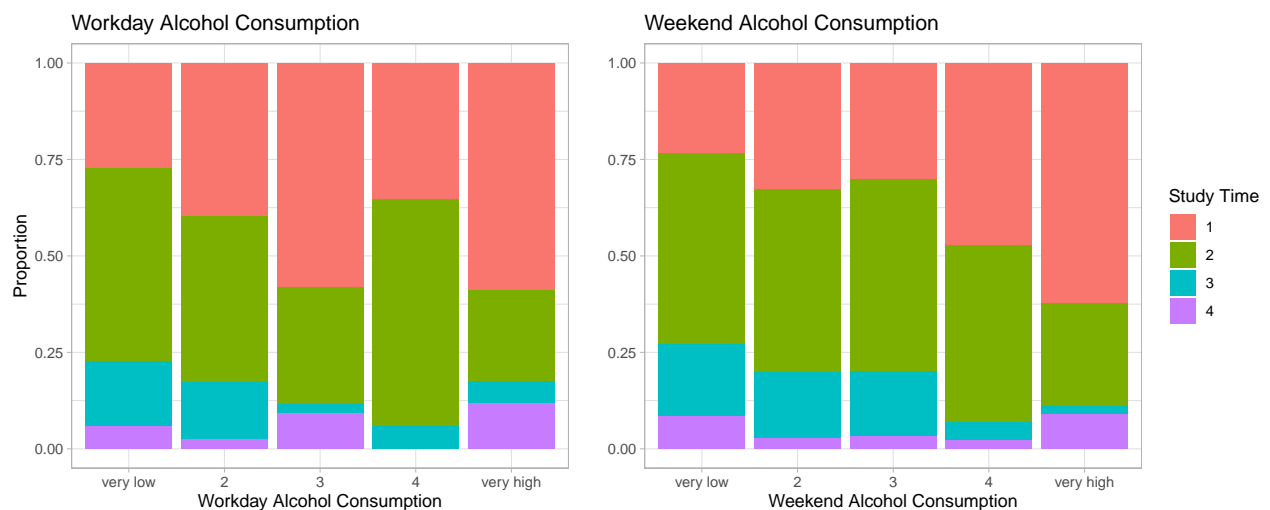
2.1 Grades



Final grade: The median of final grade decreases when the level of alcoholic consumption increases and the average mean taken between each group also has a decreasing trend when considered by both Workday and Weekend consumption. The sample correlation of Workday and Weekend consumption with final grade (-0.205, -0.177) are negative.

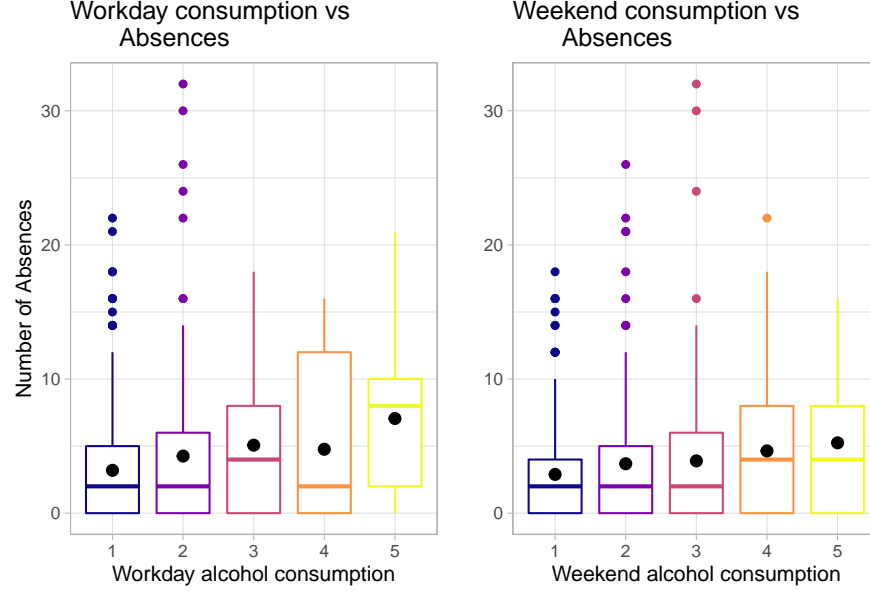
2.2. Weekly study times

The Relationship between amount of Study Time and Alcohol Consumption
From Portuguese High Schoolers during 2005–2006



The proportion of students spending less Weekly study time is generally higher when alcohol consumption (Workday and Weekend) is higher. However, this trend isn't clearly seen with workday alcohol consumption. Nevertheless, the "very high" alcohol consumption group has a higher proportion of high Weekend study time than some other groups'. The correlation between workday alcohol consumption and weekly study time is -0.138, while the correlation between weekend alcohol consumption and weekly study time is -0.215. We can see that both sample correlations are negative.

2.3. Absences



Generally, the absence is positively associated with alcohol consumption. Groups with higher alcohol consumption level will have a higher mean and median of absences, especially when considered by weekend. However, outliers concentrate mostly in the group with lower alcohol consumption, with some absences up to 32 days.

Table 2: Average number of absences considered by workday and weekly consumption

Workday alcohol consumption	1	2	3	4	5
Average absences	3.195	4.256	5.070	4.765	7.059

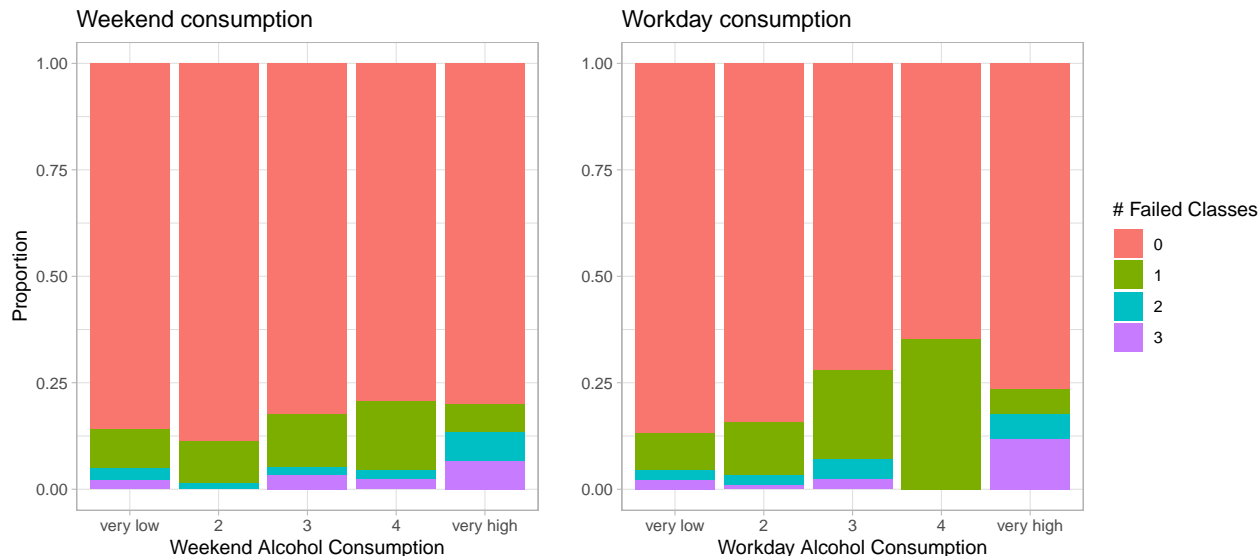
Weekend alcohol consumption	1	2	3	4	5
Average absences	2.891	3.867	3.9	4.644	5.244

The sample correlation between workday alcohol consumption and weekly study time is 0.173, while the correlation between weekend alcohol consumption and weekly study time is 0.156. We can see that both sample correlations are positive.

2.4. Number of failed classes

The Relationship between amount of Failed Classes and Alcohol Consumption

From Portuguese High Schoolers during 2005–2006



A large proportion of students did not fail any classes. However, in the very high alcohol consumption group (both Weekend and Workday), the proportion of students failing at least 2 classes is much larger compared to other groups. Sample correlations of alcohol consumption vs failures are positive (0.106, 0).

3. Hypothesis test

3.1. Final Grades

First hypothesis: (Alternative) *Alcohol consumption will significantly predict final grades after controlling for other variables. There will be a negative relationship between grades and alcohol consumption.*

Null: *Alcohol consumption will not significantly predict final grades after controlling for other variables. There will be no relationship between grades and alcohol consumption.*

Linear model: *Weekend Alcohol Consumption (Walc)*

term	estimate	std.error	statistic	p.value
(Intercept)	12.3603239	0.2027433	60.9653881	0.0000000
walc_fct_num2	-0.1003239	0.3298345	-0.3041643	0.7611010
walc_fct_num3	-0.6936572	0.3545593	-1.9563925	0.0508507
walc_fct_num4	-1.3258411	0.3972465	-3.3375776	0.0008938
walc_fct_num5	-1.8047683	0.5164540	-3.4945381	0.0005075

We obtain the model

$$\widehat{G3} = 12.36 - 0.100 \times \text{Walc2} - 0.694 \times \text{Walc3} - 1.326 \times \text{Walc4} - 1.805 \times \text{Walc5}.$$

With weekend alcohol consumption as the only predictor, we expect a student's final grade to decrease by 0.100 - 1.805 points with an increase in weekend alcohol consumption. The coefficients continually decrease as the difference in the factors increase. The p-values of these relationships range from insignificant to significant at an alpha level of 0.05. The significant relationships are between very low and somewhat high ($\sim 0 < 0.05$) and very low and very high (< 0.05). These relationships provide sufficient evidence to reject the null. Overall, this result can be used to support a negative relationship between final grades and alcohol consumption.

Linear model: *Workday Alcohol Consumption (Dalc)*

term	estimate	std.error	statistic	p.value
(Intercept)	12.2993348	0.1488567	82.625321	0.0000000
dalc_fct_num2	-0.9356984	0.3236486	-2.891093	0.0039684
dalc_fct_num3	-1.1597999	0.5045425	-2.298716	0.0218404
dalc_fct_num4	-3.3581583	0.7810286	-4.299661	0.0000198
dalc_fct_num5	-2.0640407	0.7810286	-2.642721	0.0084242

We obtain the model

$$\widehat{G3} = 12.30 - 0.936 \times \text{Dalc2} - 1.160 \times \text{Dalc3} - 3.358 \times \text{Dalc4} - 2.06 \times \text{Dalc5}.$$

With workday alcohol consumption as the only predictor, we expect a student's final grade to approximately decrease by 0.936 - 3.358 points with an increase in daily alcohol consumption. All of the relationships are significant at an alpha of 0.05. Thus, there is sufficient evidence to reject the null. Like the previous result, this finding can support a negative relationship between the two variables. Surprisingly, however, the relationship between very low and somewhat high consumption has a lower p-value and coefficient than the relationship between very low and very high consumption. This is divergent from the trend seen in the previous analysis where the coefficients and p-values continually decreased as the difference in factors increased. These findings are also different from the previous analysis in that all of the relationship observed are significant and the range of coefficients is higher. This indicates that workday alcohol consumption may have a greater impact on final grades than weekend alcohol consumption.

For both models, we obtained small r-squared and adjusted r-squared values, thus suggesting that alcohol consumption accounts for small variability in final grade, even though the relationship is significant (Appendix III.1).

Simulation-based testing: The sample correlation of final grade vs workday and weekend alcohol consumption is -0.205 and -0.177, respectively. When performing simulation-based hypothesis testing to assert whether there is a negative correlation between alcohol consumption and final grade, the p-values for both Workday and Weekend consumption hypothesis tests are extremely small that can't be detected among the null distributions. So we conclude that there is a negative correlation between alcohol consumption and student's final grade.

3.2. Weekly study times

Second hypothesis: (Alternative) *Alcohol consumption will significantly predict weekly study times after controlling for other variables. There will be a negative relationship between weekly study times and alcohol consumption.*

Null: Alcohol consumption will not significantly predict weekly study times after controlling for other variables. There will be no relationship between weekly study times and alcohol consumption

Chi-square test: We use a Chi-Square distribution to test whether there is a relationship between alcohol consumption and Weekly study times.

Table 6: Level of alcohol consumption by weekly study times

Weekend Alcohol Consumption	1	2	3	4
very low	58	122	46	21
2	49	71	26	4
3	36	60	20	4
4	41	40	4	2
very high	28	12	1	4

Workday Alcohol Consumption	1	2	3	4
very low	123	226	76	26
2	48	52	18	3
3	25	13	1	4
4	6	10	1	0
very high	10	4	1	2

The chi-square statistic of 37.371 in a chi-square distribution with degree of freedom 12 give a really small p-value 1.945×10^{-4} . Similarly the chi-square test for Weekend alcohol consumption and Weekly study time also has a small p-value 4.273×10^{-7} , thus giving enough evidence to show that there is a relationship between alcohol consumption and Weekly study times.

Simulation-based testing: There is a greater negative correlation between Weekend alcohol consumption with study times (-0.215), compared to workday alcohol consumption. When performing hypothesis test to test whether alcohol consumption would negatively correlate with Weekly study time, null distributions with 10^4 generated replications show no possible correlation as extreme as our observed statistic, leading to our conclusion that there is a significant negative trend between alcohol consumption status and Weekly study time of the student, and it is more clearly shown when considered consumption level by weekend.

Linear model: When estimating study time by alcohol consumption, we obtain the model

$$\text{Study time} = 2.246 - 0.003\text{Dalc} - 0.134\text{Walc} - 0.001\text{Dalc} \times \text{Walc}.$$

This model suggest that we have a negative relationship between alcohol consumption and Weekly study time based on the data we observed. However, when checking the conditions of this model (Appendix III.2), we find the conditions very insufficiently met, so it is not reliable to make conclusion from this linear model.

3.3. Absences

Third hypothesis: (Alternative) *Alcohol consumption will significantly predict absences after controlling for other variables. There will be a positive relationship between absences and alcohol consumption.*

Null hypothesis: Alcohol consumption will not significantly predict absences after controlling for other variables. There will be no relationship between absences and alcohol consumption

Generally, Workday alcohol or Weekend alcohol consumption positively associates with the average absences of students. This positive trend is clearer when considering Weekend alcohol consumption.

Linear model: When constructing a linear model to predict numbers of absences with predictors workday and Weekend alcohol consumption, we get positive slopes of both Workday (1.433) and Weekend (0.594) consumption (see appendix III.2 for conditions checked). P-values for both coefficients are small.

term	estimate	std.error	statistic	p.value
(Intercept)	1.0530912	0.8169954	1.288981	0.1978669
Dalc	1.4330122	0.6282146	2.281087	0.0228685
Walc	0.5938054	0.2794038	2.125259	0.0339445
Dalc:Walc	-0.2166327	0.1539181	-1.407455	0.1597744

Simulation-based testing: When we perform hypothesis testing to test whether there is a positive relationship between alcohol consumption and absences, we didn't observe any values as extreme as our observed statistics in the null distributions (p-values of 0 for Weekend and Workday consumption tests). Hence, we have enough evidence to conclude that alcohol consumption is positively correlated to the number of absences in students.

However, the high number of absences may correspond to whether the student lives far from the school or

not. Now, we define a student living “closed” to their school if the reason they choose the school is that it is closed to their home (**reason** = “home”). We will perform an additional test to consider only these “closed-to-school” students.



There is still a positive correlation between alcohol consumption and absences, after controlling students “closed” to their school. When we perform **simulation-based hypothesis testing** to test whether there is a positive relationship between alcohol consumption and absences, we get small p-values (0.0171 and 3×10^{-4}) for both hypotheses tests. Hence, we have enough evidence to conclude that alcohol consumption is positively correlated to the number of absences, regardless of the geographical condition.

One thing to notice is that the same **linear model** when controlling the variable has positive slopes for alcohol consumption (0.328 and 1.193 for workday and weekday, respectively). But p-values for both coefficients is high (0.857 and 0.077), so it is not reliable, but rather it suggests there are generally a positive association between absences and alcohol consumption in our data.

3.4. Number of failed classes

Fourth hypothesis: (Alternative) *Alcohol consumption will significantly predict the number of failed classes after controlling for other variables. There will be a positive relationship between the number of failed classes and alcohol consumption.*

Null: *Alcohol consumption will not significantly predict the number of failed classes after controlling for other variables. There will be no relationship between the number of failed classes and alcohol consumption.*

Chi-square test: We use a Chi-Square distribution to test whether there is a relationship between alcohol consumption and Weekend study times.

Table 9: Level of alcohol consumption by the number of failed classes

Weekend Alcohol Consumption	0	1	2	3
very low	212	23	7	5
2	133	15	2	0
3	99	15	2	4
4	69	14	2	2
very high	36	3	3	3

Workday Alcohol Consumption	0	1	2	3
very low	392	39	10	10
2	102	15	3	1
3	31	9	2	1
4	11	6	0	0
very high	13	1	1	2

The chi-square statistic of 28.981 in a chi-square distribution with degree of freedom 12 give a really small p-value 0.003966. However, the chi-square test for Weekend alcohol consumption and the number of failure classes has a large p-value (0.134). So we have enough evidence to conclude that there is a relationship between Workday alcohol consumption with the number of failed classes, but for Weekend alcohol consumption, we can't.

Simulation-based testing: When performing a hypothesis test on whether alcohol consumption would positively correlate with the number of failed classes, we obtain two p-values of 0.0064 and 0.0196 for Workday and Weekend alcohol consumption, respectively. At the level $\alpha = 0.05$, these p-value are really small, so we have enough evidence to show that alcohol consumption is positively correlated with the number of failed classes.

Linear model: We perform linear model testing for this hypothesis and get a model $\text{Study time} = 0.0538 + 0.1\text{Dalc} + 0.0286\text{Walc} - 0.0114\text{Dalc} \times \text{Walc}$. However, when checking the conditions of this model (Appendix III.4), we find the conditions very insufficiently met, so it is not reliable to make conclusion from this linear model.

IV. Discussion

Through our detailed statistical analysis of the link between alcohol consumption and academic performance, it is safe to conclude that increased alcohol consumption negatively with academic performance. This can be seen through our various visualizations, including Workday and Weekend alcohol consumption and grades, the relationship between study time and alcohol consumption, and through the hypotheses, tests conducted where a strong statistical connection between alcohol consumption and academic performance was observed. Some ways we can improve our analysis is examining the other variables besides alcohol consumption that might correlate with academic performance to determine the level of impact that drinking has on grades compared to other factors. This would be able to provide a more comprehensive view of the magnitude of the effect that drinking has on students.

As touched upon briefly in the introduction, the reliability of the data collected can be put into question. Taking data from only two schools is adequate on a small scale, but not enough to make a large-scale generalization of students in the Portuguese public school system. Also, we don't know the accuracy of the students' answers, as high school-aged children might be reluctant to admit their alcohol use or poor study habits, regardless of the anonymity of the study. One other ramification is that we lack some data variables' backgrounds to have some more in-depth conclusions. The original study also did not give any contexts about the failing threshold of the grading system in those schools. Therefore, we can only conclude significant negative correlations between grades and alcohol consumption, but we can't make further claims about whether students tend to have failed grades or not, given their alcohol consumption. Also, many of the variables are collected on the 1 to 5 scale. Therefore the information can be misleading, which prompts false conclusions.

There are some things our group would do differently in conducting our statistical analysis. As stated earlier, we would analyze more about the relationship of the rest of the variables and see if we could get a better comprehensive conclusion based on more conditions (family status, internet, love status,...). We had many variables in our data set (around 33) and we only used nearly 10 in our analysis. Another thing we would do differently is using a different statistical test, namely the ANOVA test, because the linear regression model is not as effective with our data as other tests could be.

Bibliography

1. Cortez P and Silva A. 2008. Using data mining to pre-dict secondary school student performance.
2. Student Alcohol Consumption dataset: <https://www.kaggle.com/uciml/student-alcohol-consumption>
3. Chodkiewicz J, Talarowska M, Miniszewska J, Nawrocka N, and Bilinski P. 2020. Alcohol Consumption Reported during the COVID-19 Pandemic: The Initial Stage. *Int. J. Environ. Res. Public Health* 17:4677. 10.3390/ijerph17134677
4. <https://www.npr.org/2020/10/05/920437811/americans-are-drinking-14-more-often-during-pandemic-study-finds>

Appendix

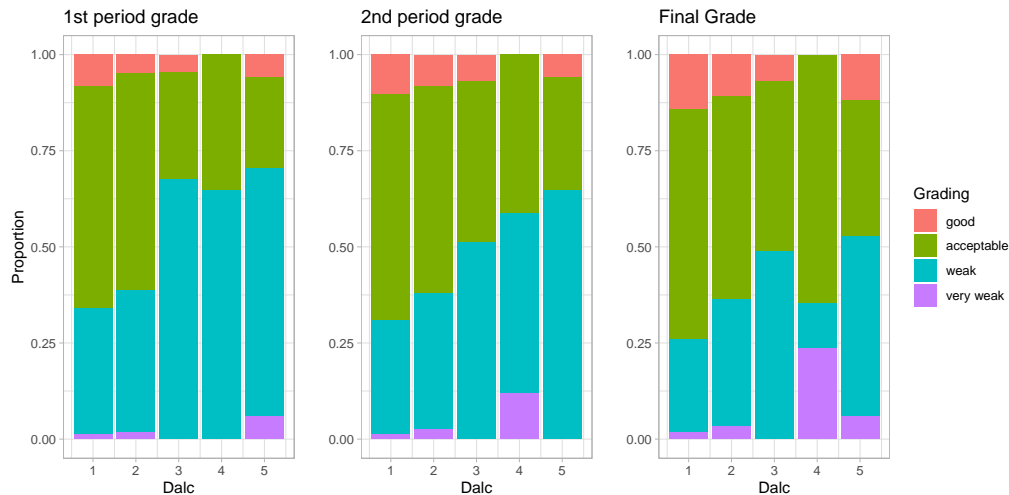
These are additional analyses for further conclusions between alcohol consumption and student study performance. More variables will be controlled when analyzing the hypotheses. We also present include checking of model's conditions and visualizations of null distributions used in the hypothesis testing. Analysis of models' plausibility is presented here.

I. Overall grade

In our hypothesis testing of alcohol consumption and grades, we only consider the final grade as we think it is the most important grade. However, we may need a general view of grades in the whole process as well. So in this additional test, we consider a new variable which is the average grades of all three-period grades G1, G2, G3. In our exploratory analysis, we have mentioned the test statistics of student's proportions with three period score below 10 for each Workday alcohol consumption groups. For **average grade**, the proportion of students with mean grades lower than 10 increase when workday alcoholic consumption increase (from 0.266 to 0.529).

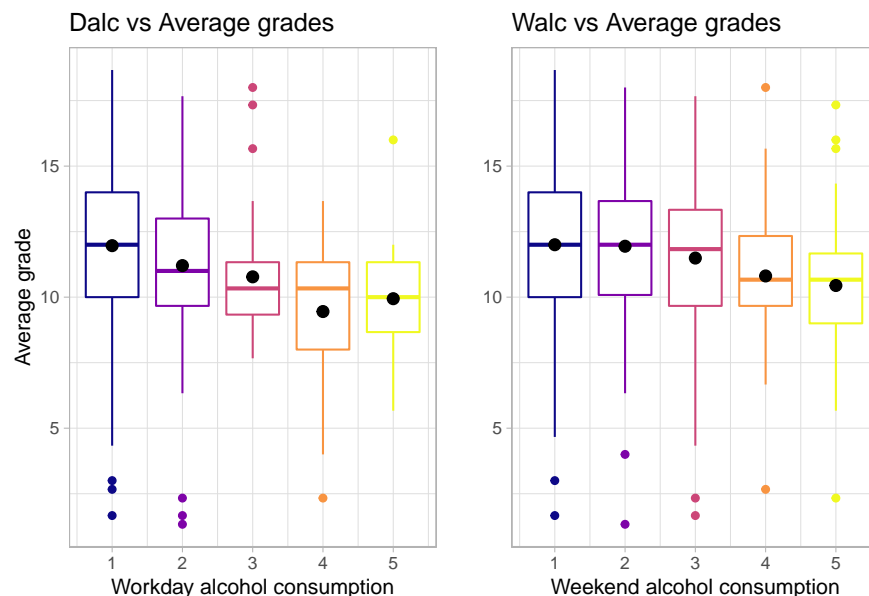
Table 11: Proportion of weak students by workday alcohol consumption (average grade)

Workday alcohol consumption	1	2	3	4	5
Proportion of weak students	0.266	0.347	0.465	0.471	0.529



For all grades, we split the students into 4 groups by their grades (good: grade ≥ 16 , acceptable: $11 \leq \text{grade} \leq 15$, weak: $6 \leq \text{grade} \leq 10$, and very weak: grade ≤ 5). From the first graph, higher workday alcohol consumption associates with a higher proportion of weak and very weak students.

Hypothesis test: Now we will perform a hypothesis test to check if alcohol consumption negatively affects the average grades. Our significant level will be 0.05.

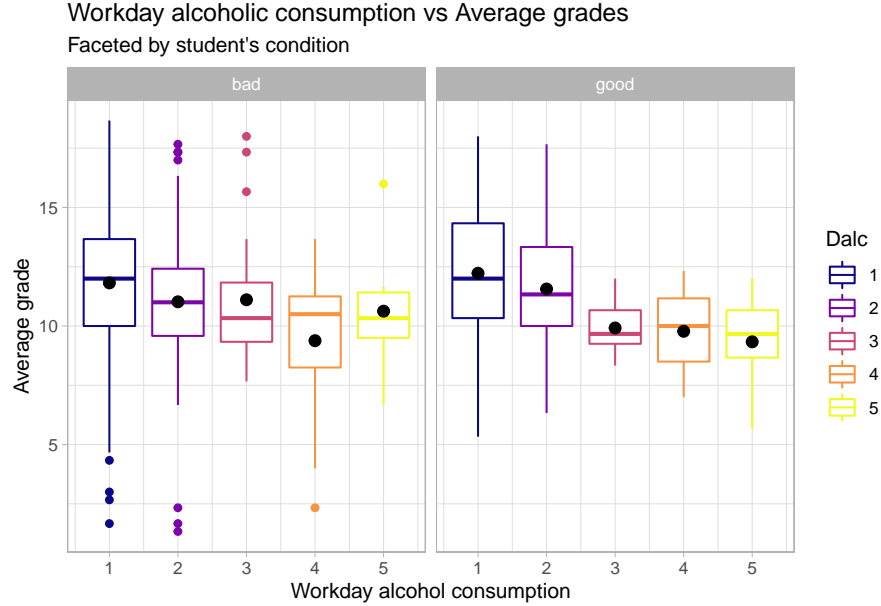


When considered by average grades, we can see that higher alcohol consumption associates with lower mean and median average grades among each consumption group. The declining grade trend is more clear when considered by the weekend.

When performing simulation-based hypothesis testing to assert whether there is a negative correlation between alcohol consumption and average grade, the p-values for both Workday and Weekend consumption hypothesis tests are again extremely small that can't be detected among the null distributions. So we conclude that there is a negative correlation between Workday alcohol consumption and student's average grade.

II. Good students vs bad students

This is a new attempt to control more variables to give a more general conclusion about alcohol consumption and study performance. We think that students with hard conditions (identified by the variables), may find it harder to perform well in their education. Therefore, we split the students into two groups. Based on their living status, we define them as having *good* condition or *bad* condition. The *good* conditioned student have their parent living together (`Pstatus = T`), have family educational support (`famsup = "yes"`), have internet access (`internet = "yes"`) and quality of family relationship is good (`famrel > 3`). The *bad* conditioned student lack at least one of these properties.



When dividing students into two groups of good and bad conditions, we see that good students have a clear decreasing average grade when Workday alcohol consumption increases. Generally, in both conditions, lower grades are associated with higher Workday alcohol consumption.

1. Conditions vs Grades

We consider the hypothesis that good-conditioned students may have higher mean average score compared to bad-conditioned students.

statistic	t_df	p_value	alternative
1.429365	508.1341	0.076757	greater

When controlling the condition of the student, we can see that the negative trends in grades of both groups, and in the group with good condition, the grades are generally higher, but not too much higher. CLT-based testing gives a t-distribution of 508.134 degree of freedom, and the p-value of 0.0768, when considering the hypothesis that the average grade of good-conditioned students is higher than bad-conditioned students'. We don't have sufficient evidence to show that this is true, since the p-value is high.

2. Conditions vs Grades and Alcohol consumption

Now we will consider the conditions and grades with more predictors from alcohol consumption.

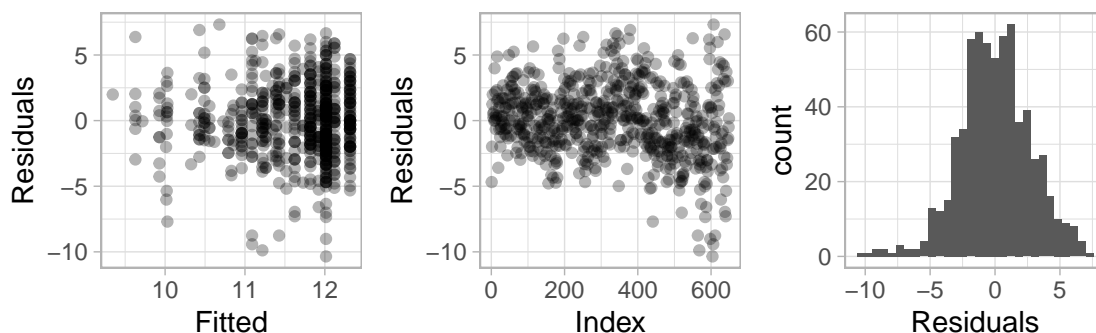
term	estimate	std.error	statistic	p.value
(Intercept)	12.9446747	0.5073846	25.5125485	0.0000000
conditiongood	0.3010688	0.2287734	1.3160131	0.1886379
Dalc	-0.7341612	0.3820844	-1.9214635	0.0551143
Walc	-0.2625168	0.1697365	-1.5466139	0.1224476
Dalc:Walc	0.0666654	0.0936039	0.7122082	0.4765938

We obtain the model

$$\text{Average Grade} = 12.945 + 0.301\text{conditiongood} - 0.734\text{Dalc} - 0.263\text{Walc} + 0.067\text{Dalc} \times \text{Walc}$$

which suggest that we expect from the data, student with better condition, holding other constant, will have in average 0.301 more point than bad-conditioned students. Holding other constant, we can expect that higher alcohol consumption will result in lower average score (with coefficients -0.734 and -0.263 for Dalc and Walc respectively). It is suggested from the data that students with better condition and lower alcohol consumption will perform better than students with bad condition and higher alcohol consumption.

Now we test the conditions for our linear models. The R^2 and adjusted R^2 for this model is 0.0494 and 0.0435.



We can see that the model fits the linearity, equal variance, normality, and independence conditions.

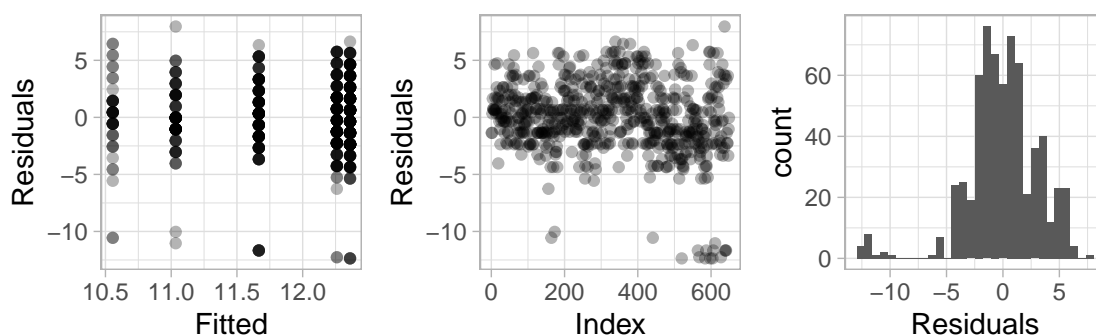
III. Assessing linear models fit

We have used some linear models throughout four main hypothesis tests. We will provide more context for these models' conditions and also analyze some more models. Also, we have seen that in Appendix II.2, when considered by condition, the models work really well. We will test by student's condition to see whether the conditions of the model are more sufficient.

1. Linear model in the first hypothesis test

Weekend alcohol consumption vs Final Grade:

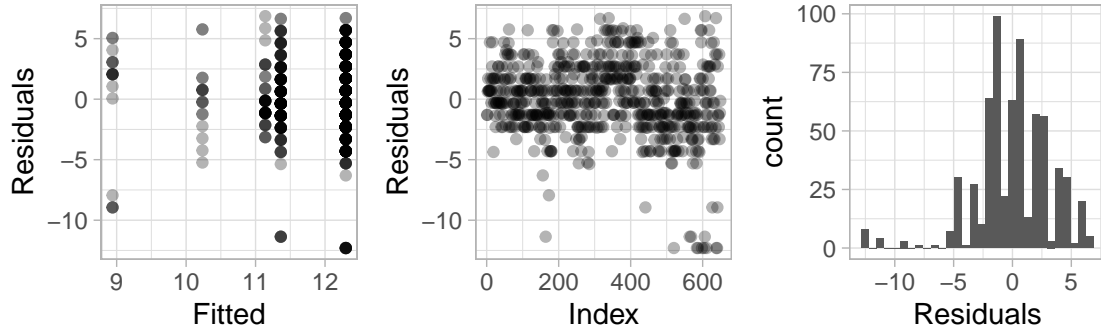
The r-squared value for this relationship is 0.033 and the adjusted r-squared value is 0.027. Therefore, although some of these relationships are significant, weekend alcohol consumption alone doesn't account for much of the variability seen in final grades.



The equal variance, independence, and normality condition for using a linear model appear to be checked. The linearity condition, on the other hand, is not met. Therefore, the validity of these results is a little rocky.

Workday alcohol consumption vs Final Grade:

The r-squared value for this relationship is 0.048 and the adjusted r-squared value is 0.043. Therefore, daily alcohol consumption still doesn't account for much of the variability seen in final grades, even though the relationships are significant.



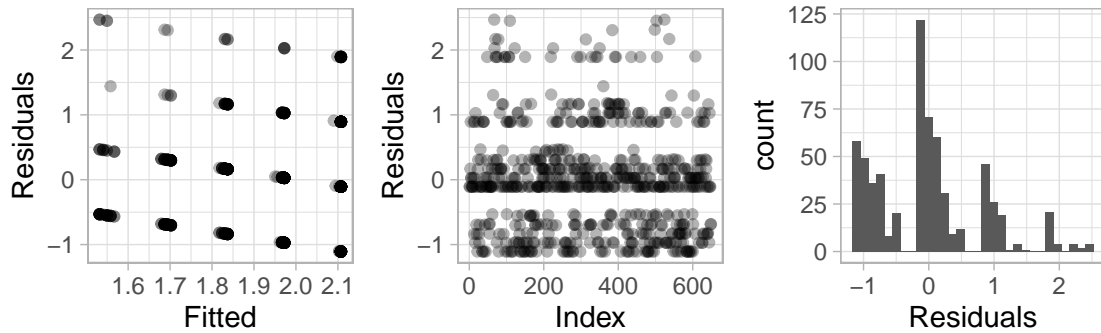
The linear model conditions for this model matched the conditions of the previous model. Therefore, the validity of this test is also rocky considering the failure to meet the linearity requirements.

2. Insufficient model in the second hypothesis testing

We've presented this model in the hypothesis 2:

$$\text{Study time} = 2.246 - 0.003\text{Dalc} - 0.134\text{Walc} - 0.001\text{Dalc} \times \text{Walc}.$$

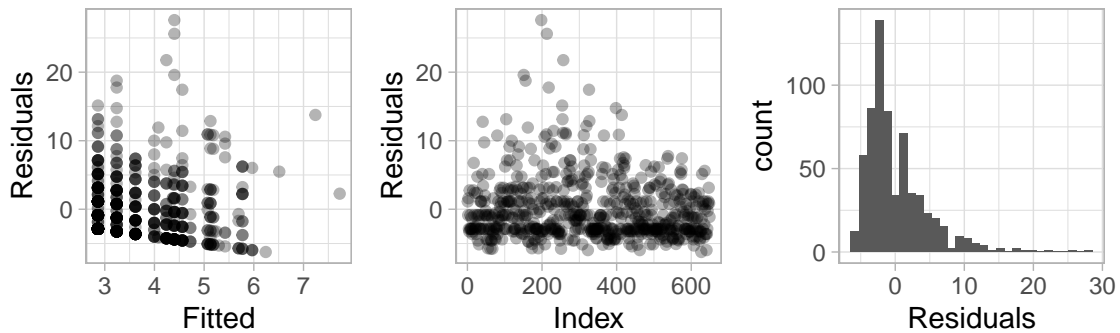
However, we mentioned that the conditions of this model is not sufficiently met. This is our testing for the model.



The linearity, independence, normality conditions are not really sufficient in this model. Also, equal variance condition is also questionable.

3. Linear model in third hypothesis test

We said that in the third hypothesis, we obtain a linear model predicting numbers of absences with predictors Workday and Weekend alcohol consumption, and get positive slopes of both Workday (1.433) and Weekend (0.594) consumption.



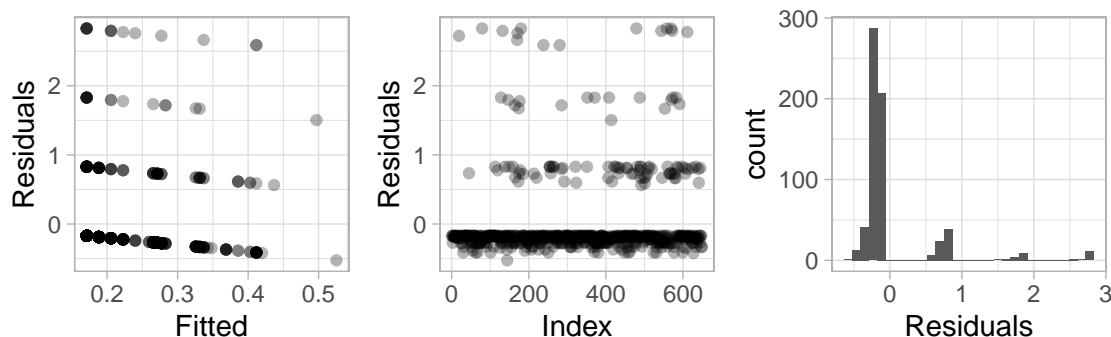
The conditions for this model is more sufficiently met than the previous one. Linearity, independence, and normality are sufficiently met. However, the equal variance condition is questionable.

4. Insufficient model in the fourth hypothesis testing

We've presented this model in the fourth analysis:

$$\text{Study time} = 0.0538 + 0.1\text{Dalc} + 0.0286\text{Walc} - 0.0114\text{Dalc} \times \text{Walc}.$$

However, we mentioned that the conditions of this model is not sufficiently met. These are our testing for the models.



The linearity, independence, normality conditions are not really sufficient in this model. Also, equal variance condition is also questionable.

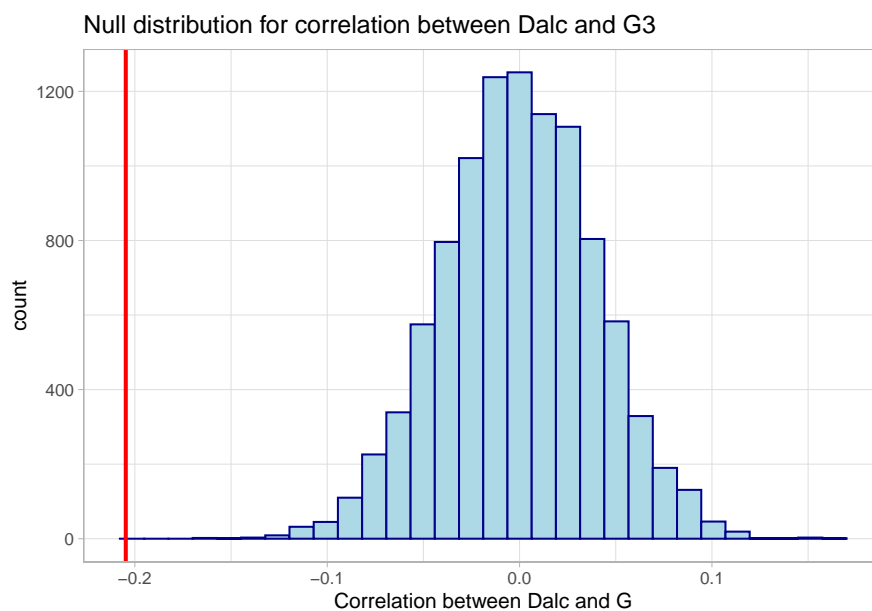
IV. Simulation-based testing information

We have shown that we used various simulation-based tests when testing the relationship between alcohol consumption and study performance. This part will explicitly display the null distribution and our calculated statistics.

1. First hypothesis

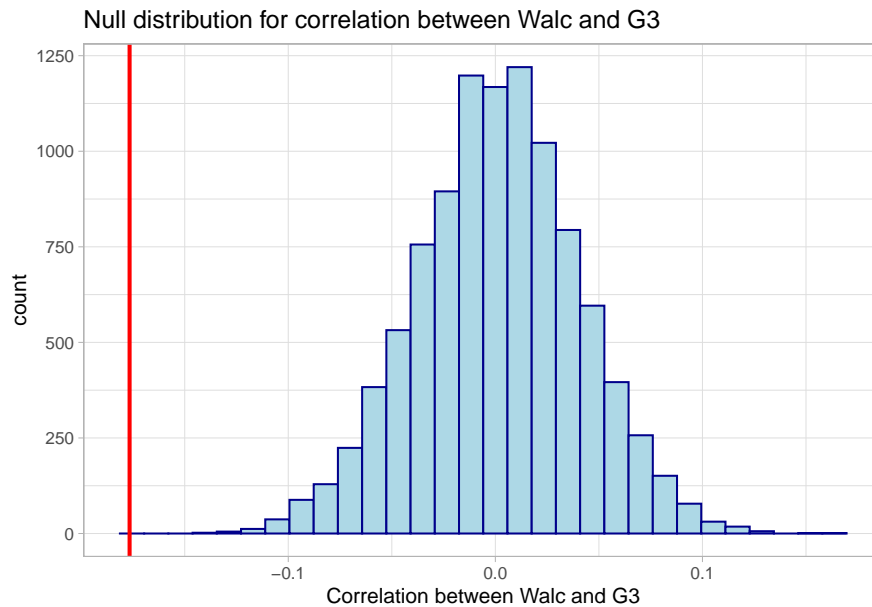
Correlation between workday alcohol consumption and final grade

- **Test statistic:** -0.205.
- **p-value:** 0.



Correlation between weekend alcohol consumption and final grade

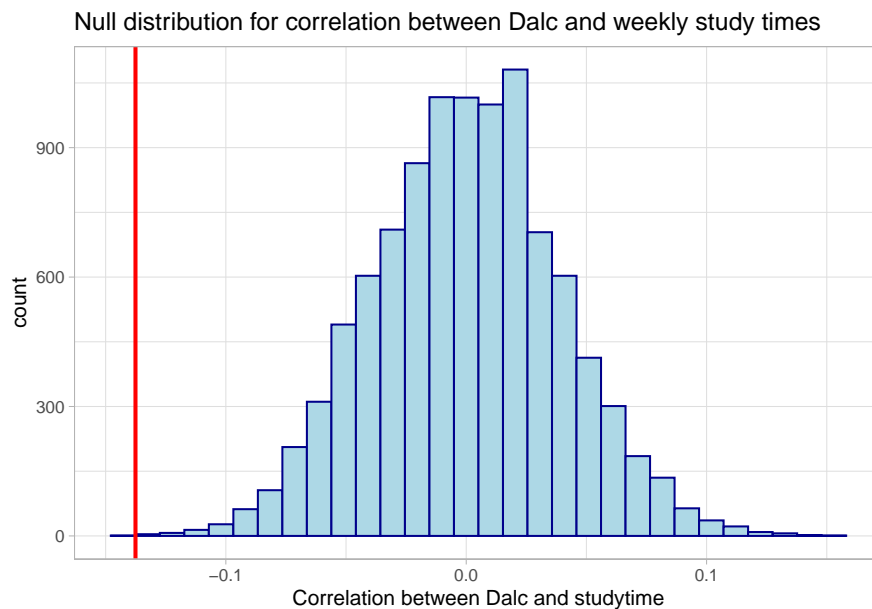
- **Test statistic:** -0.177.
- **p-value:** 0.



2. Second hypothesis

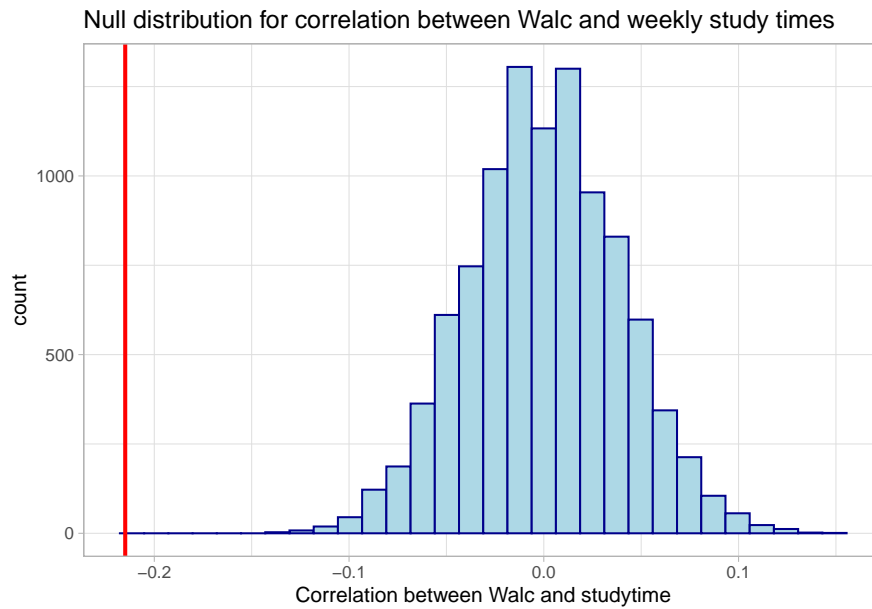
Correlation between workday alcohol consumption and weekly study time:

- **Test statistic:** -0.138.
- **p-value:** 1e-04.



Correlation between weekend alcohol consumption and weekly study time:

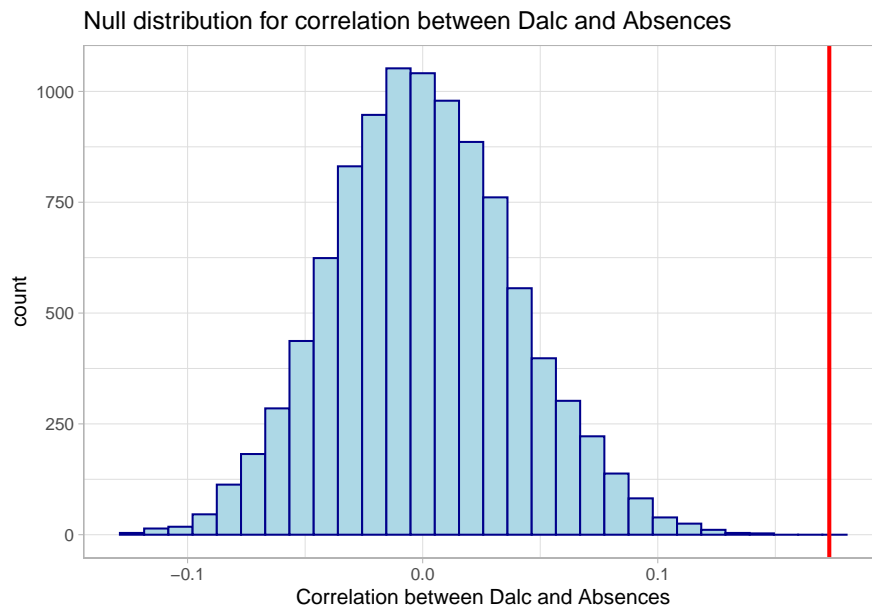
- **Test statistic:** -0.215.
- **p-value:** 0.



3. Third hypothesis test

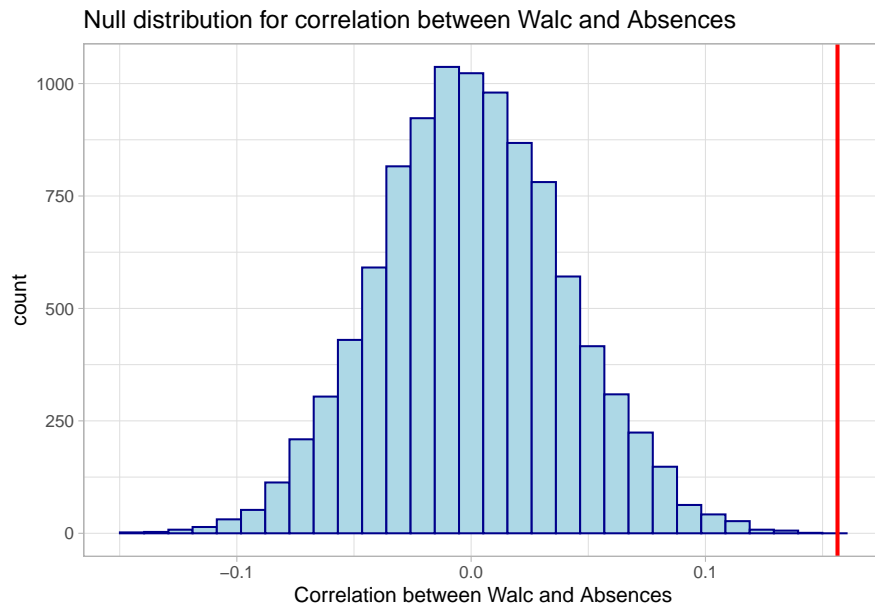
Correlation between workday study time and absences (before controlled)

- Test statistic: 0.173.
- p-value: 0.



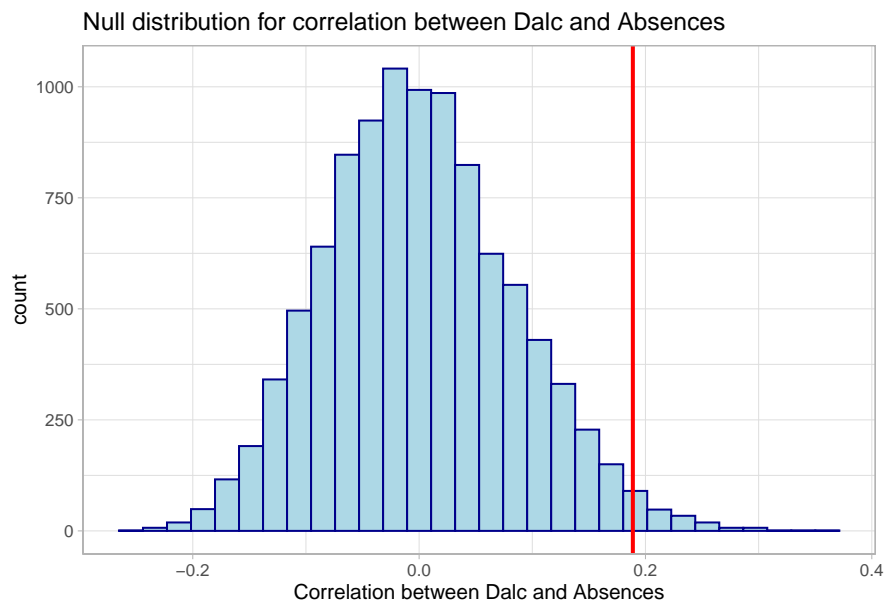
Correlation between weekend study time and absences (before controlled)

- Test statistic: 0.156.
- p-value: 0.

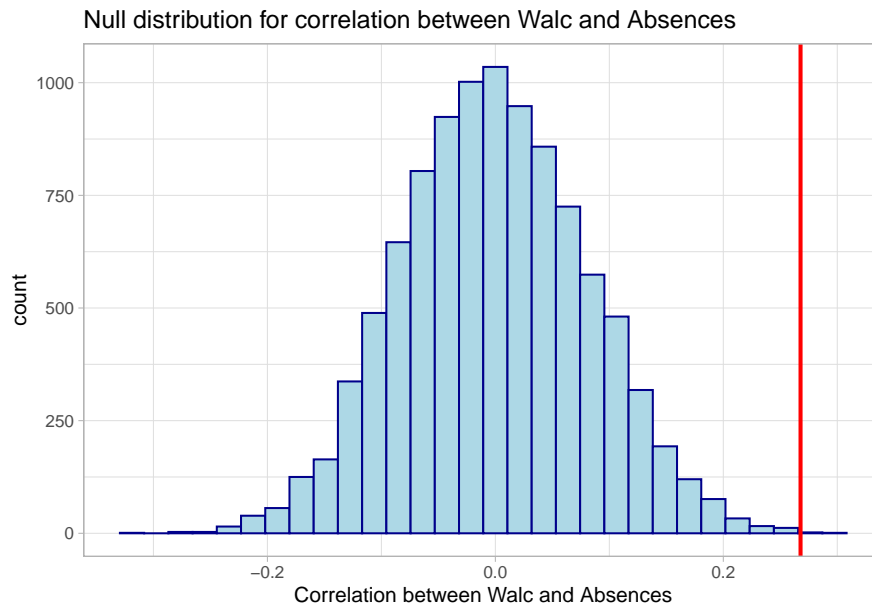


Correlation between workday study time and absences (after controlled)

- **Test statistic:** 0.189.
- **p-value:** 0.0171.



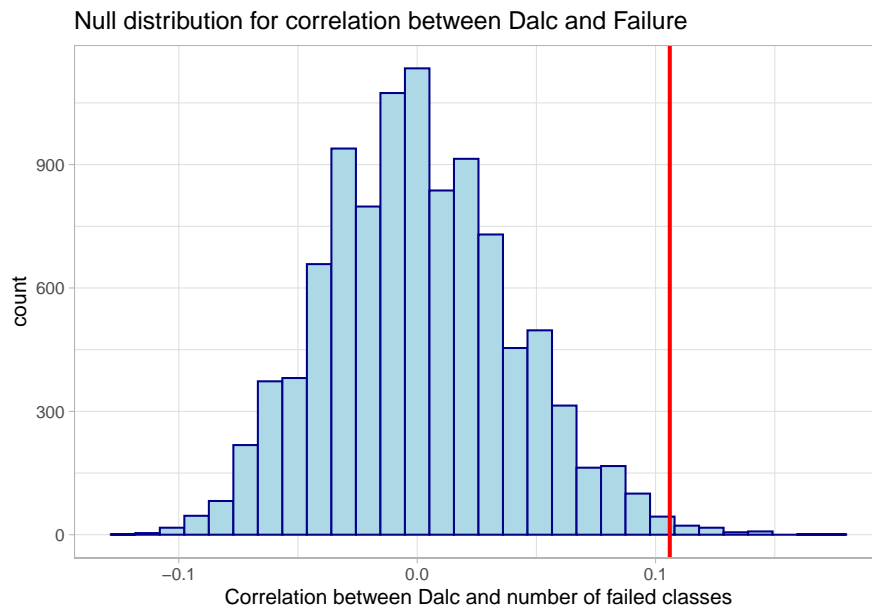
- **Test statistic:** 0.268.
- **p-value:** 3×10^{-4} .



4. Fourth hypothesis

Correlation between workday alcohol consumption and number of failed classes

- **Test statistic:** 0.106.
- **p-value:** 0.0064.



Correlation between weekend alcohol consumption and number of failed classes

- **Test statistic:** 0.082.
- **p-value:** 0.0196.

