Investigation on Whether Alcohol Consumption Affects Student's Academic Performance

Zichen Gong - 1005682469

December 17, 2021

Abstract

Although some countries have developed advanced modern education system, Portugal still has its education at the tail of Europe [6]. The dataset used for this report comes from a survey generated between year 2005 to 2006 by the University of Minho[3]. The survey was distributed to students of two secondary schools in Alentejo region in Portugal. After basic data cleaning, I intend to investigate whether high or low level of alcohol consumption will affect students' academic performance at school. First, I set up a propensity model using logistic multiple linear regression, with alcohol consumption level as the treatment variable being predicted, and a set of variables that might have potential influence on the treatment as indicators. AIC values are used to select the best subset of variables, balancing between goodness of fit and complexity. The final version of propensity model is used to predict the treatment variable, so that extra influence from these indicators is excluded in the final model. To manually assign one treatment (high alcohol consumption) and one control (low alcohol consumption) group helps set up a meaningful causal inference in the final model. The final model suggests that students who drink a lot of alcohol are more likely to fail courses in school due to the significance analysis via p values. Since this dataset might be outdated and was only collected from two secondary schools in one region of Portugal, whether or not this model is applicable to a wider range might require further researches.

Keywords: Education, Causal Inference, Propensity Score Matching, Logistic Multiple Linear Regression, Model Selection.

Introduction

The Lisbon Strategy's guideline argued that some European countries were still far from reaching the minimum requirement of education implementation [6]. This guideline also revealed that education should be put into economic and social priority. Unfortunately, Portugal is one of these countries. Its schooling system falls behind the modern society requirements, while maintaining a typically pre-modern style.

Not until the second half of the 20th century Portugal had its first universally schooled generation entered the society [6]. High illiteracy still existed. The direct visualization of the influence was Portugal fell behind those most developed countries, with its GNP did not even reach the half of the average recorded in OECD. Although the Portuguese government has realized this protraction, the universalization of secondary education is still an almost unattainable goal for not only Portugal, but also many other European countries. In Portugal, early school-leaving rate and high school failing rate are still high, particularly for young people from non-ideal social backgrounds, despite the Portuguese government's increasing budget on education. **This explains why it is necessary to investigate potential factors that might contribute to students' academic performance**.

For particularly the core course Portuguese language, the failing rate kept high regardless of the overall improved education level in Portugal over the last few decades [3]. The dataset used for this report comes from a study by the University of Minho, in which 649 students taking the Portuguese language course in secondary school took part in a survey. The dataset contains identity information such as sex, family background and study situation of students.

This analysis aims to investigate the causal relationship between students' drinking behavior and whether that will affect their academic performance. In order to achieve that, propensity score matching is incorporated for causal inference. I will first set up a logistic multiple linear regression model to eliminate influence from other potential factors other than the level of alcohol intake. The response variable for this propensity will be the treatment variable, alcohol level, and the set of potential influential variables that might affect the treatment variable are selected as indicators in this propensity model. AIC value will be applied to this propensity model for model selection, to make sure the model is balanced between goodness of fit and complexity. After the subset of independent variables are decided, the dataset will be split into treatment group and control group based on propensity score for each observation. Finally, the causal relationship will be demonstrated using one simple linear regression model, with alcohol level as the sole predictor and the number of failed courses as the response variable. After we manually set up a randomly controlled experiment through propensity score matching, we will be able to determine whether alcohol drinking will affect academic performance, and in which way it will affect academic performance. A detailed explanation about methods used in this report will be introduced in the "Method" section.

In this report, we assume that **whether a student is drinking much or little alcohol is going to affect his/her academic performance**. **For set of potential factors, parents' education background, study time, internet access, health condition, family relationship, intention for proceeding higher degree, and the number of absences, that might affect our treatment "alcohol level", their influence to the treatment will be eliminated in the final model**. Our **null hypothesis is students' drinking behavior will affect their school performance, while the alternative hypothesis is students' drinking behavior will not affect their school performance**.

In **Data** section, I will describe the data collection process, as well as a summary of our dataset. There will be description of the data cleaning process, and a set of important variables summary through tables and plots. in **Methods** section, I will introduce methodologies used for analysis in this report, including linear regression models, propensity score matching, Akaike's Information Criterion (AIC), and p values. In **Results** section, I will showcase my propensity score model and my final causal relationship simple linear regression model between alcohol level and the number of failed courses. Model selection process and model coefficients will be put into tables. In **Conclusion** section, I will wrap up this report through a recap of hypothesis, methods and results. I will discuss the weakness of my model and the next steps. The **Bibliography** section will list out sources and R packages used for this report. Finally, the **Appendix** section will include ethics statement, a glimpse of my dataset, one supplementary table and some rigorous mathematics calculation for my methods.

Data

Data Collection Process

The dataset collects information from students during the 2005-2006 school year from two public schools of the Alentejo region in Portugal [3]. Information was collected in two ways. One came from paper sheets provided by the school, which provided information for the last four columns of variables "Absences", "First Period Grades", "Second Period Grades" and "Final grades". These information came from paper reports, since the majority of Portuguese public school information systems were still poor at that time. The remaining variables were collected through a designed survey. Paper sheets were distributed in class to 788 students. Before the survey was released, the questionnaire was first tested within a small group of 15 students. 111 answers were removed due to a lack of identification information, with 649 records left for the Portuguese class.

One foreseeable drawback of this collection process is the information is purely descriptive, which needs additional model construction and propensity score matching to set up treatment and control groups by myself. Another problem is this dataset is not conclusive enough. It is hard to make conclusions for the overall students population, not even the European students population, since the information only came from two schools in one region of Portugal. Bias might occur since records collected by this survey might indicate students from similar background. Finally, since this survey was conducted between year 2005 and 2006, the information is not up-to-date and this time lag might cause significant difference. Back to the times when most Portuguese schools were still using paper sheets for information storage, nowadays both the information system can be much more advanced. At the same time, we are not sure to which extend Portuguese education developed during this period.

Data Summary

This dataset was originally combined from two different sources of collection [3]. For the part from paper records provided by the two public schools, it gives information about the number of absences, the first period grade, the second period grade, and the final period grade for each individual student. For the part from paper survey distributed in class, besides identification questions such as sex, age and address, a set of more closed questions with predefined answers were also provided. For instance, demographic questions like parents' education level and family relationship, social questions like alcohol consumption and romance relationship, and school related questions like past failures and extra paid after-school classes. Some sensitive questions, for example, the question related to family income, received only a few responses, so they were removed from this dataset by researchers.

Data Cleaning

First, for two variables "Medu" and "Fedu", they represent mother's education and 'father's education, respectively. For each variable, the level of education is represented using numbers. 0 means receiving no education. 1 means primary education, which is 4th grade. 2 means 5 to 9th grade. 3 means secondary education. 4 means higher than secondary education. Since I want the overall average education level of both mother and father, I create a new variable "avg_edu" by calculating the average of mother's education and father's education. If the average is higher than 2, I will consider it as "High"; otherwise I will consider it as "Low". Whether parents' average education level is high or low is stored into the variable "parents_edu".

Similarly, we have two variables "Dalc" and "Walc", each representing workday alcohol consumption and weekend alcohol consumption, respectively. Each variable contains values from 1 to 5, with 1 representing very low level of consumption and 5 representing very high level. If someone is drinking a lot of alcohol during workdays but none during weekends, or vice versa, he/she is still drinking fairly much alcohol. As a result, I create a new variable "alc" representing the average of workday alcohol consumption and weekend alcohol consumption. If the average is higher than 2.5, it will be considered as high level alcohol consumption;

otherwise it is low level alcohol consumption. Whether students drink high or low level of alcohol is stored into the variable "alc_level".

After that, I select only variables that I think might be useful for my analysis: parents' average education (parents_edu), students' weekly study time (studytime), number of past failed classes (failures), internet access at home (internet), current health condition (health), average level of alcohol consumption during workdays and weekends (alc_level), family relationship quality (famrel), intention for higher education (higher), and the number of school absences (absences), in which alcohol consumption level is the treatment variable and the number of passed failures is the response variable.

The next step is to make variable values meaningful. For the weekly study time variable, study time is separated into four different levels and represented through numbers from 1 to 4. I transferred these values into their real meanings. 1 is changed into Less than 2 hours. 2 is changed into 2 to 5 hours. 3 is changed into 5 to 10 hours. 4 is changed into More than 10 hours.

For the family relationship variable, relationship is rated by numbers from 1 to 5, with 1 representing very bad and 5 representing excellent. I changed 1 into its meaning Very bad. 2 is changed into Bad. 3 is changed into Medium. 4 is changed into Good. 5 is changed into Excellent.

For the variable meaning current health conditions, each level is health condition is rated with five levels, with 1 representing Very bad and 5 representing Very good. I turned 1 into Very bad. 2 is turned into Bad. 3 is turned into Medium. 4 is turned into Good. 5 is turned into Very good.

Finally, in order to make my upcoming model construction more convenient, I turned the data type of my categorical variables, parents education level, weekly study time, home internet access, health condition, alcohol consumption level, family relationship and intention for higher education from character into factor.

Important Variables

Parents_edu is a variable created by myself. It means the average level of academic backround of father and mother. Father and mother education background are separated into four levels, which are 1 representing primary education (4th grade), 2 representing 5th to 9th grade, 3 representing secondary education and 4 representing higher education. The average level of parents' education is considered as high if the average value is bigger than 2, and is considered as low if the average is smaller than 2.

Studytime represents the weekly study time for each observation. Study time is separated into four categories. Ranking from low to high is less than 2 hours, 2 to 5 hours, 5 to 10 hours, and more than 10 hours.

Failures is a numerical variable, representing the number of past failed classes for each observation. **This is the response variable for this analysis.**

Internet indicates whether students have internet access at home. If so, it is a yes. If no access at home, it is a no.

Health stands for the current health condition for students. The health condition is rated with five levels, from very bad to very good.

Alc_level is a variable created by myself. It represents the average level of alcohol consumption for each observation. As in the dataset, alcohol consumption is split into two time slots, which are weekday consumption and weekend consumption. Consumption level is rated with five levels, from 1 representing very low to 5 representing very high. After calculating the average alcohol consumption level of these two time slots, if the average is higher than 2.5, it is considered as a high level of alcohol consumption. If the average is low than 2.5, it is considered as a low level of alcohol consumption. **This is the treatment variable for this analysis.**

Famrel indicates family relationship quality. The quality is rated with five levels from very bad to very good.

Higher records whether students intend to take higher education or not. Each is recorded as yes or no.

Absences is a numerical variable provided by the school. The number of school absences for each observation is recorded in this variable. The range is between 0 and 93.

Important Variable Summaries

Table 1: Table with summary important categorical variables

Variable	Category	Proportion
Parents Education Level	High	50.5
	Low	49.5
Study Time	2 to 5 hours	47.0
	5 to 10 hours	14.9
	Less than 2 hours	32.7
	More than 10 hours	5.4
Internet	no	23.3
	yes	76.7
Health	Bad	12.0
	Good	16.6
	Medium	19.1
	Very bad	13.9
	Very good	38.4
Alcohol Level	High	18.5
	Low	81.5
Family Relationship	Bad	4.5
	Excellent	27.7
	Good	48.8
	Medium	15.6
	Very bad	3.4
Higher Education Intention	no	10.6
	yes	89.4
Total cases		649.0

Table 2: Table with summary important numerical variables

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Absences	0.000	0.000	2.000	3.659	6.000	32.000
Failures	0.000	0.000	0.000	0.2219	0.000	3.000

Table 1 lists a set of categorical variables. The first column is a list of variables. The second column are factors of these variables, and the third column is the proportion of each factor among 649 observations.

For the variable describing parents' average education level, 50.5% of all students' parents have received relatively higher education, while the other 49.5% have comparatively lower education. The proportion is very close. We might argue that the overall education background for the last generation is fair, but not very strong.

If we take a look at study time, we find out that the two biggest proportion come from the 2 to 5 hours range and the less than 2 hours range, which are 47.0% and 32.7%, respectively. This two categories take over 79.7% of the survey samples, while only 5.4% of these students state that they study more than 10 hours per week. Almost 80 percent of these students study an average of less than 5 hours per week, which means a significantly larger population's study time is on the lesser side. This can be a potential explanation for why high school failing rate is so high.

For internet access at home, we can see that the majority of students have access to internet, which is 76.7% of them. The rest 23.3% otherwise indicate that they do not have internet access.

The largest proportion health status among five categories is very good, which is 38.4%. The other four categories have comparable proportion, which are 13.9% for very bad, 12% for bad, 19.1% for medium, and 16.6 for good.

It is unusual that secondary school students have already been addicted to alcohol, so we see that 81.5% of students have on average low alcohol consumption. This might also because some of them only have a relatively higher alcohol consumption on weekdays or on weekends, and this is averaged out during the data cleaning process. However, there are still 18.5% of students have high alcohol consumption on both weekdays and weekends.

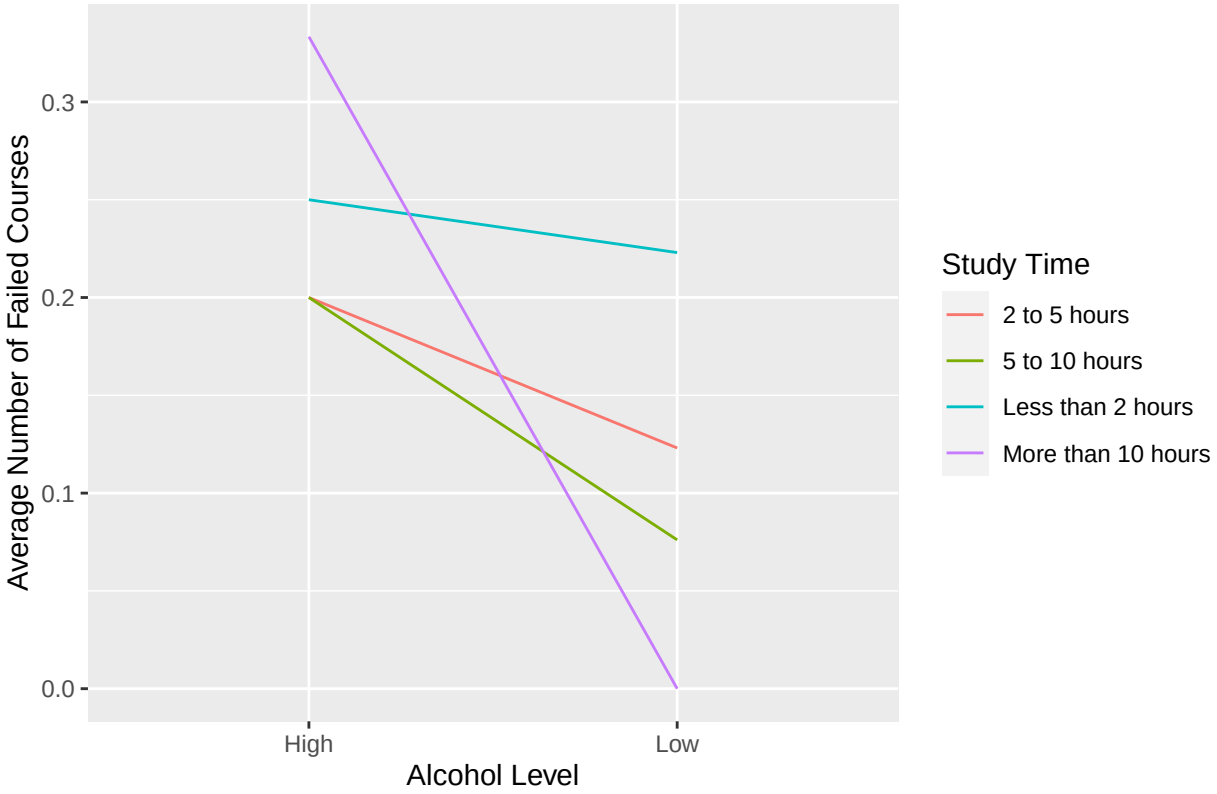
We can see that the majority of students' family relationship is on the better side. 27.7% of them believe that they have excellent family relationship. 48.8% of them believe that their family relationship is good, and 15.6% of them have medium family relationship. Only 4.5% of them think their family relationship is bad and 3.4% of them think they suffer from very bad family relationship.

89.4% of students think they will keep proceeding higher degrees, which is almost all of them. Only 10.6 of them think they might leave school after graduate.

Table 2 lists two numerical variables. For the number of absences provided by school records, at least 25% of students have no absences. At least 50% of students have absences fewer than 2 times, might due to occasional sickness or accidents. At least 75% of students have absences fewer than 6 times, which is still low frequency. The average absences is 3.659, which is higher than the mean value due to outliers. We see that the student with the highest absences frequency have 32 absences, which is approximate one third of 93.

Perhaps because students entering secondary schools are more willing to study than those who have not, we see that at least 75% of students have never failed any past classes. The maximum is 3 times, and the mean failing time is 0.2219.

Graph 1: Alcohol Level and Study Time Interaction on Failed Courses



```
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

We should not make easy conclusion that students who spend the most time studying will have better academic performance, but is very likely that students who are more familiar with course materials generally do not fail a lot, and they need certain time after class to review and practice. Although effort is not always equal to success, if we take a look at the interaction plot with alcohol consumption on the x-axis, failures on the y-axis, and colors representing study time, we see that no matter how many hours students spend on study, high level alcohol consumption always leads to high failing rate.

However, before any further data manipulation and model construction, what is presented in this graph is more of descriptive than indicating any causal relationship. This makes our investigation necessary. We intend to construct a causal study trying to interpret the causal relationship between alcohol level and academic performance.

All analysis for this report was programmed using `R version 4.0.4`. Functions used for data cleaning are from library `tidyverse` [13]. Functions used for creating tables are from library `expss` [7].

Methods

For this analysis, the final model is set up using both normal **Simple Linear Regression** and **Logistic Multiple Linear Regression**. I use **Propensity Score Matching** to create a cleaned dataset for better causal relationship determination. **Akaike's Information Criterion (AIC)** is applied for model selection. **P value** will be used to briefly determine the statistical significance.

All rigorous mathematics calculation and proofs for this section are listed in the appendix.

Propensity Score Matching

Instead of saying Propensity Score Matching as one type of analysis, it is more appropriate to conclude it as one cleaning technique. Usually what we have are observational datasets, in which variables contain only descriptive contents. We did not set up randomly selected experiments with clear control and treatment groups. However, what we do is trying to produce causal relationship instead of observational conclusions. Using Propensity Score Matching will automatically separate this dataset into one treatment group and one control group. These two groups are matched as pairs and influence brought by variables other than the treatment is violated.

To implement this method, first we need to decide our treatment variable. In this analysis, my treatment variable is student's drinking behavior, whether they are drinking high or low levels of alcohol. Two assumptions should be fulfilled. First, we need to make sure that it is possible to predict this treatment variable with other indicators. Second, Our dataset is large enough. Ideally we should have at least 100 observations for both treatment and control groups.

Second, I will use a logistic regression model to calculated the probability of being treated. In this step, we will get a propensity score from our selected variables. Based on these variable, we use the regression model to calculate the probability that one student will drink a lot of alcohol.

The third step is the matching. In this step, one observation in the treatment group and another observation from the control group are matched into pairs when they have similar propensity score. This means that two students with similar probability of being treated, but one drink much alcohol while the other rarely drink alcohol, are matched into one pair.

The forth step is to reduce our dataset. Observations that do not have a matched pair will be removed from this dataset.

The last step is to analyze using our matched dataset. After previous operations, potential influence brought by other variables has largely been removed. We are able to use one simple linear regression model to demonstrate the causal relationship between students' drinking behavior and their academic performance.

Linear Regression Model

What we can easily conclude is correlation between our independent variable and response variable, which is explained with Simple Linear Regression. As we add more independent variables to our model, we have more controls, so Multiple Linear Regression has stronger explanations on causal relationship. The reason why we are using linear relationship is that is easy for interpretation.

In a **Simple Linear Regression**, we assume that the response variable can be determined by merely the sole independent variable we have, and the true relationship is described through $E(Y|X = x) = \beta_0 + \beta_1 x$. With specific values of X $(x_1, x_2, ..., x_n)$, mean value of Y is related to each value of X. β_0 and β_1 are unknown parameters representing the intercept and the slope, respectively. If we collect real-life data, the mean value of Y is not going to always sit on X values. As a result, we have $y = E(Y|X = x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$, in which the error term ϵ_i are fluctuations of y_i values. $E(\epsilon_i|X) = 0$. The relationship between Y and X are linear in parameters β_0 and β_1. After propensity score matching, a simple linear regression, with the function `lm()` from basic `R`, is able to explain the causal relationship between the sole independent variable and the response variable.

In a multiple linear regression with p predictors, relationship between conditional mean of Y and X values is set up as $E(Y|X_1 = x_1, ..., X_p = x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$. Again, if the dataset comes from real-life investigations, we will have random fluctuations:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon_i$$

, such that $E(\epsilon_i|X) = 0$. The response variable Y_i is predicted based on p predictors and each relationship is linear in parameters $\beta_0, \beta_1, ..., \beta_p$. When we are trying to interpret the relationship between one independent variable and the response variable, we hold all other $x_i's$ as constant. They work as control variables in this interpretation.

Since our model output for our propensity score matching is a categorical variable, with the function `glm()` from basic `R`, we will be using **Logistic Multiple Linear Regression** for this matching. The response variable for this analysis is whether the students' drinking high or low level of alcohol. Let p denote the probability that one student is drinking a lot, the response will be denoted as

$$\hat{y_i} = log(\frac{p}{1-p})$$

. This model will not directly output the value of response, wince the response is categorical. Instead, the response is transformed into numerical values through this fraction. For each parameter coefficients $(\beta_i's)$, each coefficient indicates when holding all other variables constant, one unit increase in this variable will result in β_i unit(s) change in the response variable. For the intercept coefficient β_0, when all variables have values equal to zero, the response variable Failure will be equal to β_0. For categorical independent variables, coefficients on these variables represent the effect of each factor within this categorical variable. For a categorical variable with 2 factors, the final model will demonstrate only one factor, since the other factor is the base factor. All other factors demonstrate their effects with respect to this base factor.

Under this assumption, our propensity model will be in this format:

$$alc\_level = \hat{\beta}_0 + \hat{\beta}_1 parents\_edu + \hat{\beta}_2 studytime + \hat{\beta}_3 internet + \hat{\beta}_4 health + \hat{\beta}_5 famrel + \hat{\beta}_6 higher + \hat{\beta}_7 absences$$

with *alc_level* representing alcohol consumption, *parents_edu* representing parents' academic background, *studytime* representing weekly study time, *internet* representing internet access, *health* representing health condition, $famrel$ representing family relationship, *higher* representing intention of proceeding higher degrees, and *absences* representing number of school absences.

Our final model will be in this format:

$$failures = \hat{\beta}_0 + \hat{\beta}_1 alc\_level$$

with $failures$ representing the number of past failed courses.

Model Selection

Although we have added many variable into our propensity model, we are not sure whether all of them are influential to our treatment variable. In this case, we will apply **AIC** value for model selection [11]. AIC will help us choose the best subset of variables for our propensity model.

The AIC value measures the goodness of fit of the my model and balances the model with a penalty term, ensuring that this model is not over complicated. Goodness and complexity are under consideration at the same time.

For measuring the goodness of fit, we want a large log-likelihood value or a small negative log-likelihood.

The Maximum Likelihood Principle is to choose the parameter(s) of interest in a way that the data is most likely, given a dataset $x_1, x_2, ..., x_n$. $x_1, x_2, ..., x_n$ are realizations of a random sample $X_1, X_2, ..., X_n$ from a distribution characterized by a parameter θ. Consider $X_i's$ are in a discrete form and are independent of each other, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

In this context, $p(x|\theta)$ does not represent a notation for conditional probability. Instead, it is a probability mass function, which can also be denoted as $p_\theta(x)$. If we consider the continuous form, the probability density function can be expressed as $f_\theta(x)$ or $f(x|\theta)$. The maximum likelihood estimate of the unknown parameter(s) θ can be expressed in the realization $t = h(x_1, x_2, ..., x_n)$ that maximizes the likelihood function $L(\theta)$. Its corresponding random variable is the maximum likelihood estimator of θ:

$$T = h(X_1, X_2, ..., X_n)$$

This can also be denoted as $\hat{\theta}$.

For continuous random variable, $\epsilon > 0$ stands for a random fixed and small number. We have $\hat{\theta}$ that maximizes the probability:

$$P(x_1 - \epsilon \leq X_1 \leq x_1 + \epsilon, ..., x_n - \epsilon \leq X_n \leq x_n + \epsilon) \approx f(x_1|\theta) \cdot ... \cdot f(x_n|\theta) \cdot (2\epsilon)^n$$

Since the last term $(2\epsilon)^n$ does not affect $\hat{\theta}$, we have the likelihood function for a dataset of continuous random variables:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Log-likelihood function $l(\theta)$ is easier to use in most situation:

$$l(\theta) = \sum_{i=1}^{n} log(f(x_i|\theta)) \text{ if the } x_i^{'}s \text{ are independent}$$

The reason is $y = log(x)$ is a monotonically increasing function. The value that maximizes the log-likelihood function will also maximizes the likelihood function.

Besides goodness of fit, we add a penalty term balancing this model based on on the number of parameters being estimated in the goodness step.

The AIC takes the form

$$AIC = -2[log(L(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p,\ \hat{\sigma}^2|Y)) - (p+2)]$$

Our preferred model is the one with the smallest AIC. A better fitted model is the one with a larger log-likelihood value. If the model is over fit, it takes away some of the log-likelihood through its penalty term. This is trade-off that balances between goodness of fit and complexity. We want a relatively complicated model only when this model has substantially powerful explanation.

After our model is selected with several predictor indicated by AIC, we will calculate **p values** for each coefficient. Operation of p value is based on the Student's T-test:

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{var[\hat{\beta}_i]}}$$

In this formula, $\hat{\beta}_i$ represents the estimated coefficient and β_i represents the β_i in the null hypothesis. In hypothesis testing, the null hypothesis and alternative hypothesis are

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

In the null hypothesis, we have $\beta_i = 0$, indicating that the coefficient on this variable is zero. This means in the presence of this variable, no relationship occurs between this indicator and the response variable. The second step is to calculate the probability of observing a value that is at least as extreme as our test statistic under the null hypothesis, which is the p value: p-value $= P(|t_{df=n-p-1}| > T)$, in which n represents the number of observations and p represents the number of predictors. The 1 represents the intercept term. In this case, $t_{df=n-p-1}$ is a T distribution with a degree of freedom of $n - p - 1$. If the p-value is small, it is unlikely that we will observe a value as or more extreme than the test statistic under the null hypothesis, so our test statistic is not due to change.

Usually we use a significance level of $\alpha = 0.05$ and compare it with the p value. If the p value is smaller than 0.05, we will conclude that there is a linear relationship between this indicator and the response variable, and therefore the null hypothesis is rejected.

Results

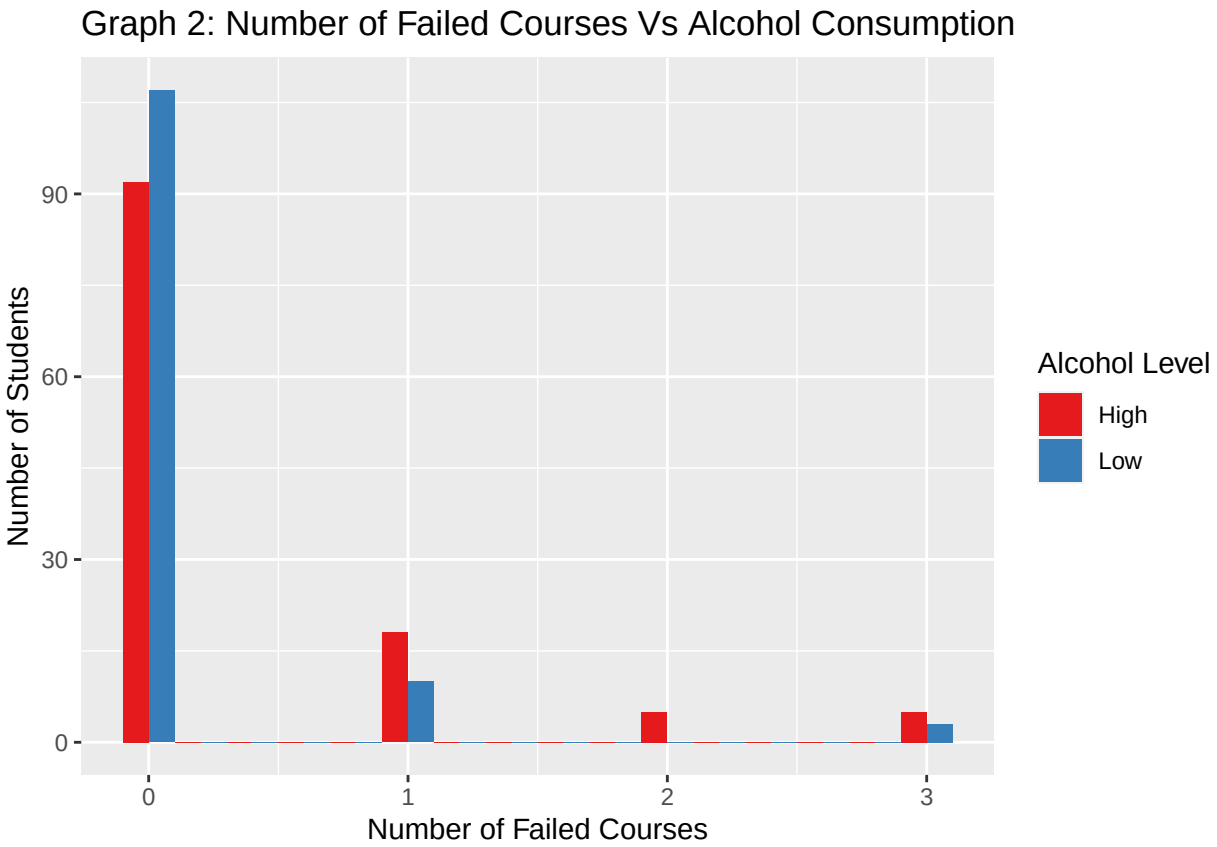
We put variables that might affect students' drinking behavior, which are parents education, study time, internet access, health condition, family relationship, intention for higher education, and absences times, into a propensity logistic multiple linear regression model as indicators, with the treatment variable alcohol level as the response variable. After model selection with AIC values, only study time and absences left in this model as indicators. After propensity scores for each observation is predicted via this second propensity model, observations with similar propensity score but different treatment are paired up and form the treatment and control groups, which are students with high alcohol consumption and low alcohol consumption, respectively. Observations without pairs are removed. After this cleaning process, I create a model with alcohol level, study time and absences as indicators and failures as the response. This time I find out that alcohol consumption is statistically significant. In my final model, two variables used to predict alcohol consumption, study time and absences are removed. The causal relationship between alcohol consumption and failures is constructed using a simple linear regression model.

Table 3: Table with summary AIC values for subset models

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Parents Education Level	+	+	-	-	-	-
Study Time	+	+	+	+	+	+
Internet	+	+	+	-	-	-
Health	+	+	+	+	+	-
Family Relationship	+	-	-	-	-	-
Higher Education Intention	+	+	+	+	-	-
Absences	+	+	+	+	+	+
AIC Value	598.24	592.01	590.01	588.57	587.9	587.86

In Table 3, it lists out subsets of variables for our propensity model. As indicated by the table, the first model is when we put all selected indicators into our propensity model. Following model 2 to model 6 each removes one variable from the previous model based on AIC selection. The removed variable is indicated by a minus sign. After this model selection, the model with the lowest AIC value is the one with only study time and absences. This model balances the best between complexity and goodness of fit. The next step we should do is to use our final propensity model to predict the propensity score for alcohol consumption. Observations with similar propensity score, but different value of alcohol consumption will be paired up. Observations that do not have pairs will be removed.

After previous cleaning operations, since the treatment variable alcohol consumption is predicted by study time and absences, these two variables should no longer be influential to our response variable - the number of past failed classes. However, we still want to make sure whether this is the case. The table with estimates of coefficients for the model with three indicators, study time, absences and alcohol consumption, is listed in the appendix. As introduced in the method section, with a significance level of 0.05, our alcohol consumption with a p value of 0.0195 is proved to be statistically significant, while other variables have fairly high p values, except for the variable absences. The absences variable with a p value of 0.0318 is still statistically significant, but we can see that this value is larger than the one for the alcohol consumption.



Before we move to our final model, we can have a glimpse of our data after propensity score matching. According to Graph 2, we see that students with low level of alcohol consumption are more unlikely to fail a course. For students who have failed courses, no matter for 1, 2, or 3 failed courses, students who consume more alcohol are more likely to fail. Although this dataset was originally an observational one, after we manually clean it and split it into treatment and control group. This graph illustrates that the number of observations for the control and for the treatment group are the same, each represented via the color blue and red, respectively. This new dataset after propensity score matching has some explanation power on causal relationship instead of purely observational, compared to what has been demonstrated by Graph 1.

Table 4: Final Model Coefficient Esitmates

	Estimate	p Value
(Intercept)	0.3583333	0.0000000
alc_levelLow	-0.2000000	0.0185237

Now our final model indicates a causal relationship between the treatment variable, alcohol consumption, and the response variable, past failed courses. As indicated by the p value, 0.0185 suggests that this indicator is statistically significant. Compared to the base factor, which is high level of alcohol consumption, low level of alcohol consumption will cause the number of failed courses to drop by 0.2. This result fits our previous graph visualizations, and it fits with common sense as well. After drinking, study efficiency drops, or students cannot study at all if they are drunk. At the same time, they need to spend time drinking, so they have less time for study. If they get drunk, they might miss classes as well, so that they cannot keep track with school progress. As time goes on, their chance of failing is going to rise.

All analysis for this report was programmed using `R version 4.0.4`. The function used for AIC model selection is from library `MASS` and library `car`. Functions used for propensity score matching are from library

`arm`.

Conclusions

Based on our research results, we want to investigate factors that might affect students' academic performance. Before any further analysis, we assume that whether or not drinking a lot of alcohol might affect students' school work. In order to build up causal relationship between students' drinking behavior and their academic performance, we use propensity score matching to manually set up the treatment group and the control group. The propensity model for predicting the treatment variable alcohol consumption is set up using logistic multiple linear regression model. The final causal model is presented with a simple linear regression model with the number of past failed courses as the response variable regressed on the sole predictor alcohol consumption.

After model selection, students' weekly study time and their school absences are kept in the propensity model to predict the treatment variable, the alcohol consumption level. After propensity score matching, alcohol drinking is proved to be statistically significant. As indicated by our final model, a low level of alcohol consumption cause the number of failed courses to drop. In other words, if students are addicted to alcohol, they are more likely to fail in school.

This suggests that in education, schools should avoid materials that might promote alcohol drinking behavior. Any alcohol exposure to children should be under caution. At the same time, schools might also consider education about side effects of alcohol drinking. If situation gets non-ideal, for example, students are generally interested alcohol, and this has greatly affect their health condition and academic performance, alcohol prohibition is necessary. In many countries, like Canada, these countries have strict law against teenager drinking. In Canada, people who are under 19 years old are not legally permitted to drink.

Weaknesses

Since the dataset was collected only from one region in Portugal, it is likely that the final model is not conclusive enough. Although we seem to figure out one factor that might affect students' studying behavior, we cannot guarantee that this model is applicable in other situation. For example, for countries that prohibit teenager drinking, as secondary students are generally illegal to drink, this factor is no longer going to affect their academic performance. After these students grow up and enter some colleges or universities, since this survey was based on only secondary school students, we might be cautious if we want to apply this model to maturer students. Although university students might drink, alcohol consumption might not be that influential to their study.

For the model itself, since the propensity model is constructed using linear regression model, we should be aware of assumptions like constant variance of residuals. If the relationship is not even close to a linear regression, we might consider transformations like Boxcox transformations to make sure that our model does not violate assumptions.

In addition, we have limited knowledge about the data collection process, since I am not the person who designed the survey and collected data points by myself. If we have more prior information, a Bayesian approach with prior and posterior distribution is going to set up models that have stronger explanation power.

Next Steps

Since we already have a clear data cleaning and analysis procedure, we might want to collect a new dataset. The dataset used for this analysis is from year 2005 to 2006 [3], which is very outdated. For this time, we might consider a wider range for our survey distribution instead of merely two secondary schools from one region of Portugal. If we are interested in just education of Portugal, we might consider collecting

data from more schools in every Portuguese city. The data collection scale can be even larger if we want a understanding of wider range students.

If possible, for example, researchers have enough time and raise enough fund, they can set up an experiment with treatment and control groups during the data collection process. Although we use propensity score matching to manually simulate this process, minor unavoidable error during the propensity score predicting and matching step might still occur. A dataset of randomly controlled experiment generally has stronger causal inference power than observational studies.

Discussion

Based on the open dataset collected through a survey distributed to students of two secondary schools in Alentejo region in Portugal [3], I decide to set the alcohol consumption of students as the treatment variable and investigate whether this will cause any influence on students academic performance. Study time and absences times are two indicators left after model selection to predict the treatment. After the treatment and control groups are manually matched, the cleaned dataset suggests the causation relationship that if students' drinking little alcohol, they failed courses in school will drop by 0.2. This suggests that alcohol have negative influence on students school performance. Schools might be aware of alcohol alcohol drinking behavior and decide whether they should take steps to prevent this negative influence.

Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/.

2. Andrew Gelman and Yu-Sung Su (2021). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.12-2. https://CRAN.R-project.org/package=arm

3. Cortez, Paulo & Silva, Alice. (2008). *Using data mining to predict secondary school student performance* [Data set]. EUROSIS. https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performance

4. David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.9. https://CRAN.R-project.org/package=broom

5. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

6. Dias, M. (2014). Education, Development and Social Inclusion in Portugal: Policies, Processes and Results. *Social and Behavioral Sciences, 116* (1877-0428), 1864-1868. https://doi.org/10.1016/j.sbspro.2014.01.485

7. Gregory Demin (2020). expss: Tables, Labels and Some Useful Functions from Spreadsheets and 'SPSS' Statistics. R package version 0.10.7. https://CRAN.R-project.org/package=expss

8. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html.

9. John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

10. Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Y. Kim, Ben Baumer, Chester Ismay, Nick Paterno and Christopher Barr (2021). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. R package version 2.2.0. https://CRAN.R-project.org/package=openintro

11. Rossi, R., Murari, A., Gaudio, P., & Gelfusa, M. (2020) *Upgrading Model Selection Criteria with Goodness of Fit Tests for Practical Applications.* entropy. Retrieved November 2, 2021, from https://www.mdpi.com/1099-4300/22/4/447/pdf.

12. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

13. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

14. Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.35.

Appendix

A1: Ethics Statement

First, since this report has listed out clear data cleaning and method using for analysis, and the dataset is accessed as open data, it is possible if readers want to reproduce the same analysis. Readers can access this dataset and follow the cleaning and analyzing process if they intend to reproduce the investigation process.

Second, all advanced methods included in this report has been properly cited in the Bibliography section. Background information about this dataset and background knowledge paper are also properly cited. The dataset used for this report is also cited.

Third, assumptions for the main method propensity score matching are fulfilled. First, it is possible to predict our treatment variable. We use logistic multiple linear regression model and AIC model selection to create the propensity model. Second, our dataset is large enough. Before data cleaning, we have 649 observations. After the matching process, we have 240 observations left. The numbers of observations left for treatment group and control group are both greater than 100.

A2: Materials

Supplimentary Data

Here is a glimpse of my dataset after the cleaning process.

```
## Rows: 240
## Columns: 11
## $ parents_edu <fct> High, Low, High, High, High, High, High, Low, High, High, ~
## $ studytime   <fct> Less than 2 hours, Less than 2 hours, Less than 2 hours, 2~
## $ failures    <dbl> 0, 0, 0, 0, 0, 1, 0, 3, 1, 1, 0, 0, 3, 1, 1, 0, 1, 0, 0, 0~
## $ internet    <fct> yes, yes, yes, yes, yes, yes, yes, no, yes, yes, yes, yes,~
## $ health      <fct> Very good, Very good, Medium, Very bad, Medium, Very good,~
## $ alc_level   <fct> High, Low, High, Low, High, High, Low, High, Low, Low, Hig~
## $ famrel      <fct> Good, Excellent, Good, Excellent, Medium, Bad, Medium, Bad~
## $ higher      <fct> yes, yes, yes, yes, yes, no, no, yes, no, yes, yes, yes, n~
## $ absences    <dbl> 22, 22, 18, 32, 16, 16, 16, 14, 14, 14, 13, 12, 12, 12, 12~
## $ .fitted     <dbl> 0.4510545, 0.4510545, 0.5103028, 0.5239262, 0.5399226, 0.5~
## $ cnts        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Supplimentary Table

Table 5: Model Coefficient Esitmates

	Estimate	p Value
(Intercept)	0.2055211	0.0222496
studytime5 to 10 hours	-0.1302079	0.5489314
studytimeLess than 2 hours	0.1313638	0.1427480
studytimeMore than 10 hours	0.0253622	0.8994694
absences	0.0170968	0.0318065
alc_levelLow	-0.1968205	0.0195056

Supplimentary Math

The likelihood function of discrete $X_i's$ is given by:

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\ &= x_n|\theta) = p(x_1|\theta) \ \cdot \ p(x_2|\theta) \ \cdot \ ... \ \cdot \ p(x_n|\theta) \\ &= \prod_{i=1}^{n} p(x_i|\theta) \end{aligned}$$

For $\hat{\theta}$ that maximizes the probability:

$$\begin{aligned} &P(x_1 - \epsilon \leq X_1 \leq x_1 + \epsilon, ..., x_n - \epsilon \leq X_n \leq x_n + \epsilon) \\ &= P(x_1 - \epsilon \leq X_1 \leq x_1 + \epsilon) \ \cdot \ ... \ \cdot \ P(x_n - \epsilon \leq X_n \leq x_n + \epsilon) \\ &\approx f(x_1|\theta) \ \cdot \ ... \ \cdot \ f(x_n|\theta) \ \cdot \ (2\epsilon)^n \end{aligned}$$

The log-likelihood function:

$$\begin{aligned} l(\theta) &= log(L(\theta)) \\ &= log(f(x_1, x_2, ..., x_n|\theta)) \\ &= \sum_{i=1}^{n} log(f(x_i|\theta)) \text{ if the } x_i^{'}s \text{ are independent} \end{aligned}$$