Determine whether the romantic relationship is has causal effect on the youths' alcohol consumption

Yi Qin

December 9, 2020

I.Abstract

This study base on the dataset about Portugal secondary students to find out the influential factors on the youths' alcohol consumption level. There is no causal association between the youths' romantic relationship status and the youths' alcohol consumption. However, the quality of family relationship, the frequency of going out with friends, gender and self-rated health status are strong related to the students' alcohol consumption. This report would cover the used concepts and dataset in the study, the introduction of the model, the analysis and conclusion of the results.

Keywords: Casual Inference, propensity score, Romantic relationship, alcohol consumption

II. Introduction

Alcohol consumption usually linked to violence, car accidents, commitment to crimes, alcohol addiction, and damage to physical health. As people are paying more attention to mental health, alcohol abuse is a major concern in many countries. The causes of alcohol abuse are usually complicated. By previous research, early-drinking and unstable emotional status would increase the risk of lifetime alcohol addiction (NIAAA, 2006). The United Stations defines that the youths are people aged from 15 to 24 years old. Youth is an important stage for people to build up their self-concept. In this stage, adolescents (10-18 years old) would gradually transit to young adulthood (Wikipedia contributors, 2020). The curiosity and the excitement to the adult's world drive young adults or even adolescents to have their first try on alcoholic beverages at this time period. World-widely, approximately 26.5 percent of youth aged 15 to 19 years old are current alcohol drinkers. Specifically, compared to the total population heavy episodic drinking (HED), which is an indicator of regular alcohol consumption, is higher among the 15-19 years age group (World Health Organization, 2018).

Meanwhile, according to the research, youth is also closely related to psychosocial and developmental challenges, including dealing with their first romantic relationship and unstable emotions (Price et al., 2016, p. 9). Most of the youths would have a romantic relationship at their age of 17 years old. (Romantic Relationships in Adolescence, 2020). However, due to lack of experiences and self-confidence, mental health problems associated with romantic relationship concerns are one of the most common issues among the youths (Price et al.,2016,p,9).

The youth may face more emotional ups and downs in their romantic relationships, and unstable emotions would increase the chance of risk drink. It seems there may exist a link between the youths' romance and early alcohol consumption, but how about in the reality? This question will be answered in the rest of my study.

Through reading this report, psychologists and parents would learn about the factors related to young adults' alcohol consumption level. Thus, they can have a better understanding of the youths who are likely to have an alcohol-related mental illness.

In the rest of the report, the dataset from Kaggle will be used to inspect the casual association between the youth's romantic relationship status and alcohol consumption by the propensity sore method which was

introduced by Rosenbaum and Rubin in 1983. In the Methodology section, the description of datasets and model that was used in propensity score matching analysis is provided. Results and discussion about this study are displayed in the Result and Discussion section respectively.

III. Model

1.Data

(Figure 1: Overall description of the table)

Table 1: Data summary

Name	d1
Number of rows	649
Number of columns	33
Column type frequency:	
factor	17
numeric	16
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
school	0	1	FALSE	2	GP: 423, MS: 226
sex	0	1	FALSE	2	F: 383, M: 266
address	0	1	FALSE	2	U: 452, R: 197
famsize	0	1	FALSE	2	GT3: 457, LE3: 192
Pstatus	0	1	FALSE	2	T: 569, A: 80
Mjob	0	1	FALSE	5	oth: 258, ser: 136, at_: 135, tea: 72
Fjob	0	1	FALSE	5	oth: 367, ser: 181, at_: 42, tea: 36
reason	0	1	FALSE	4	cou: 285, hom: 149, rep: 143, oth: 72
guardian	0	1	FALSE	3	mot: 455, fat: 153, oth: 41
schoolsup	0	1	FALSE	2	no: 581, yes: 68
famsup	0	1	FALSE	2	yes: 398, no: 251
paid	0	1	FALSE	2	no: 610, yes: 39
activities	0	1	FALSE	2	no: 334, yes: 315
nursery	0	1	FALSE	2	yes: 521, no: 128
higher	0	1	FALSE	2	yes: 580, no: 69
internet	0	1	FALSE	2	yes: 498, no: 151
romantic	0	1	FALSE	2	no: 410, yes: 239

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	16.74	1.22	15	16	17	18	22	
Medu	0	1	2.51	1.13	0	2	2	4	4	
Fedu	0	1	2.31	1.10	0	1	2	3	4	
traveltime	0	1	1.57	0.75	1	1	1	2	4	
studytime	0	1	1.93	0.83	1	1	2	2	4	
failures	0	1	0.22	0.59	0	0	0	0	3	
famrel	0	1	3.93	0.96	1	4	4	5	5	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
freetime	0	1	3.18	1.05	1	3	3	4	5	
goout	0	1	3.18	1.18	1	2	3	4	5	
Dalc	0	1	1.50	0.92	1	1	1	2	5	
Walc	0	1	2.28	1.28	1	1	2	3	5	
health	0	1	3.54	1.45	1	2	4	5	5	
absences	0	1	3.66	4.64	0	0	2	6	32	
G1	0	1	11.40	2.75	0	10	11	13	19	
G2	0	1	11.57	2.91	0	10	11	13	19	
G3	0	1	11.91	3.23	0	10	12	14	19	

1.1 Source of Data

The data package Student Alcohol Consumption (2016) provided by the University of Minho is downloaded from a crowd-sourced platform Kaggle. The data package contains two datasets of the students from two public secondary schools in the Alen- Tejo region of Portugal from the school year 2005-2006. One dataset is collected from the students in math class, while another one is collected from the students in a Portuguese language class. Through reading the guide provided by the University of Minho (Porto, 2008), I gained some insights about this dataset.

1.2 Data Collection

Researchers collected data by using paper sheets. Two main sources of the database are school reports and questionnaires. Researchers access the data about three-period grades and the number of school absences for every student from school reports. They also designed a questionnaire with 37 closed questions (refers to questions with predefined options). Researchers collect students' personal information about students' socioeconomic status, health conditions, and school life through filled questionaries.

• The target population of this study is secondary school students in Portugal, and • the sampling frames are the lists of students in both public secondary schools. Furthermore, • the sampled population is the secondary students who fill the questionnaire without missing identification information.

There are 778 students in Portugal language class who were asked to fill the questionnaires, but researchers decided to discard 111 students' answers because of the lack of identification information.

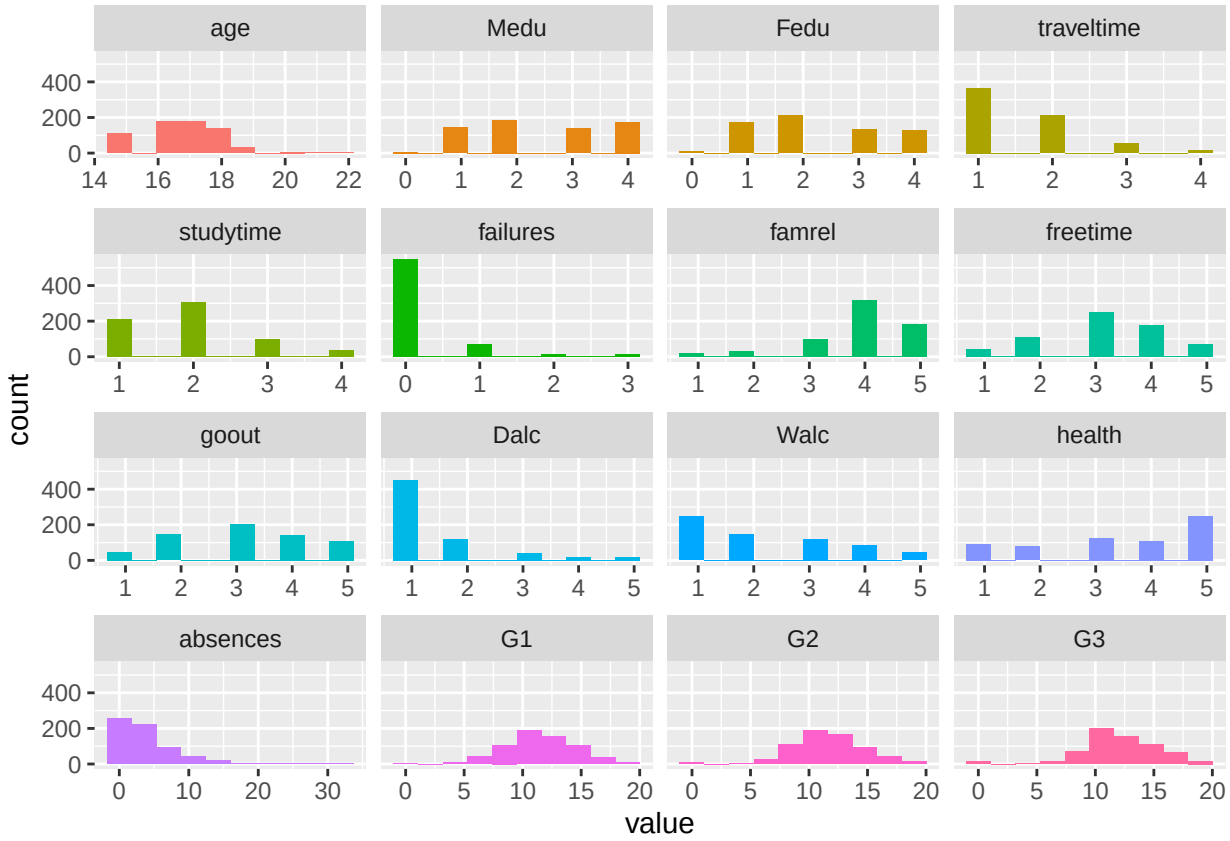
In this study, I will only use Portugal language class datasets, because it contains more observations, and I can have a larger sample. The Portugal language class dataset contains 33 variables for 649 observations, including 17 qualitative variables and 16 numeric variables [Figure 1].

1.3 Key features of Data

The Portugal language class dataset is obtained in a survey of students in two Portugal secondary school. It contains school-related, demographic, and alcohol consumption-related information about students. This dataset contains no missing value and no strange value for all observations, so I spend too much time cleaning the data. Participators of the survey are the students who are at age between 15 and 22 years old. Therefore, this dataset is appropriate for me to study the influential factors of youth's alcohol consumption. In my study, I would assume these Portugal students can represent the youth around the world. Moreover, in this dataset, most of the variables are categorical. Thus, only limited data analysis can be applied to this dataset. Additionally, this dataset is only about Portugal students. Culture background of the youths is an important factor in their alcohol consumption but it is ignored in this dataset.

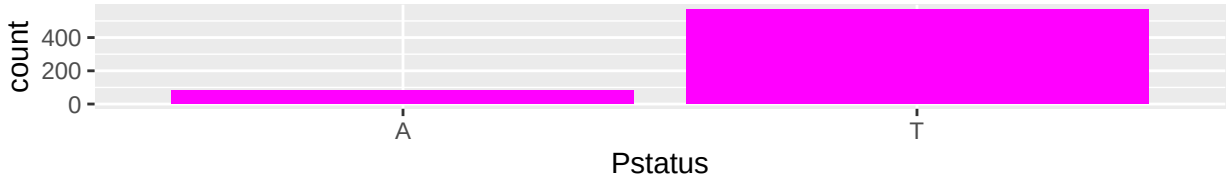
1.4 Variable selection

(Figure 2: Plot for all numeric variables)



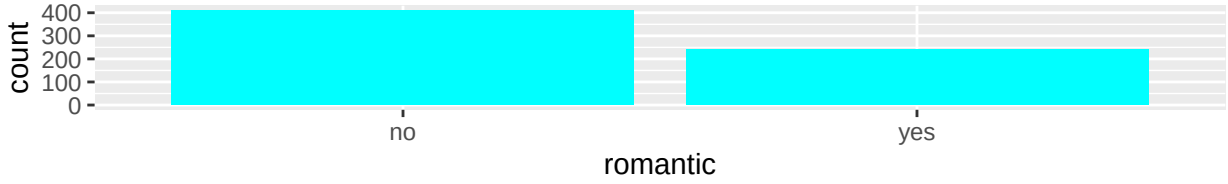
(Figure 3: Plot for some of qualitative variables)

1 bar chart for Pstatus



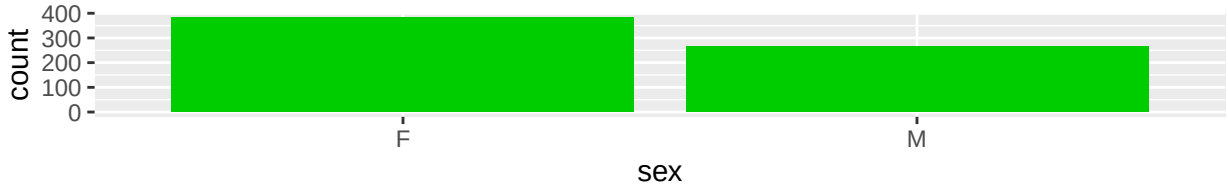
A: live apart, P: live together

2 bar chart for Romantic



no: not in romance, yes: in romance

3 bar chart for sex



F:Female, M:Male

By Explanatory Data Analysis, I select 5 predictor variables, including a treatment variable, and create a new variable "Alc_level" as my response variable.

Firstly, after viewing their distribution, I determine the binary variable "romantic" [Figure 3] to be my treatment variable. It indicates whether the student is in a romantic relationship. If the student has a value of "yes" in "romantic", that means the student is in a romantic relationship, otherwise, the student is single. Furthermore, the value of "romantic" is "yes". in the treated group are in a romantic relationship, while the control group students are not.

Secondly, for other predictor variables, I choose according to the internal and external factors that would influence students' alcohol consumption.

Observing the distribution of variables from [Figure 2 and 3], I select the internal factors to include students' sex, age. [1]), and physical health conditions. The numeric variable "health" is a self-reported health level. It can be used to represent the student's physical health status.

External factors include familial factors and environmental factors. In terms of the family-related factors, family relationship ("family") indicates how well the student gets along with his/her family members respectively. Environmental factors are mainly related to peer pressure and the friend cycle. The variable "goout" represents the frequency of the student going out with friends. I choose "goout", instead of the variable "activities"[2], although participating in an extra-curriculum activity or going out friends are both more likely to stay with peers outside of school.

Thirdly, the variable "Walc" represents student's weekend alcohol consumption, while "Doc" represents student's weekday alcohol consumption. In order to apply linear regression, I create a new variable "Alc_level"

[1]According to Portugal liquor laws, Legal drinking age is 18 years old. Globally average drinking age is around 18.6 (without considering countries without legal drinking age

[2]Extracurricular activity is an activity, performed by students and supervised by schools, but it is excluded in the common curriculum of school education (Wikipedia contributors, 2020b). Therefore, extra-curriculum activity is different from going out with friends.

according to the value of Walc. Alc_level indicates students' overall alcohol consumption level. Since students are still at a young age, I would be more cautious about students' alcohol consumption. If a student has value in "Walc" which greater than 2, he/ she would be regarded as the person who has "high-level alcohol consumption", otherwise the student has low alcohol consumption. Although "Dalc" describes students' alcohol consumption during 5 weekdays, the variable "Walc", weekend alcohol consumption, would be more likely to be students' real daily alcohol consumption without teachers' supervision. Therefore, to show overall consumption, "Alc_level" is my response variable.

More details about my selected variables are shown in Figure 4.

(Figure 4. Some important variables explanation)

Attributes	variable type	detail
sex	binary (female or male)	student's sex
age	numeric (from 15 to 22)	student's age
famrel	numeric(from 1-very bad to 5-excellent)	self-rated score for quality of family relationship
goout	numeric(from 1-rare to 5-very frequent)	self-rated score for frequency of going out with friends
health	numeric(from 1-very bad to 5-excellent)	self-rated score for current health status
romantic	binary(yes or no)	whether student is in a romantic relationship
Alc	binary (high or low)	overall alcohol consumption level
Walc	numeric(from 1-very low to 5-very high)	weekend alcohol consumption level

2.Model

Rstudio (R core team, 2013) with the function glm() was used to construct a propensity score model and the final model for the whole study. Both models are logistic regression models. The final model can be used to predict whether a student's alcohol consumption is low or high.

Propensity score matching was used which would lead to a serious reduction of the sample size and my dataset is relatively small, thus, all observations from the dataset are included in my sample.

2.1 Propensity score matching

In my study, propensity score method was used to see the effect of romantic relationship on students' alcohol consumption. The treatment is whether the student is in romantic relationship [3].], and the outcome is the alcohol consumption level "Alc_level". To estimate the propensity score, a binary logistic regression model was used. Generally it can be represented by the formula $log(\frac{p_1}{1-p_1}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$ where p_1 is the probability of the student is in romantic relationship.[4]. As lng as I get the score, one treated and one untreated observations with similar propensity scores will be matched.

In this case, my treated sample size (N=239) is smaller than the untreated sample size(N=410). Therefore, there are 239 matched pairs have a similar propensity score.

(Figure 5. Summary table for the estimated coefficient of predictor variables)

```
## Warning in build_tabular(ht): Multiple horizontal border widths in a single row;
## using the maximum.
```

[3]My propensity score model was $log(\frac{p_1}{1-p_1}) = \beta_0 + \beta_1 age + \beta_2 sex_{male} + \beta_3 goout + \beta_4 health + \beta_5 famrel$

[4]treatment variable is a binary variable, "romantic"

	(1)
(Intercept)	-3.347 *
	(1.608)
sexM	1.318 ***
	(0.224)
famrel	-0.320 **
	(0.117)
goout	0.686 ***
	(0.099)
health	0.162 *
	(0.076)
romanticyes	-0.029
	(0.212)
age	0.046
	(0.088)

*** p < 0.001; ** p < 0.01; * p < 0.05.

2.2 Choice of the final model

Base on the procedure of propensity score matching, the final model can be generally represented by the formula $log(\frac{p_2}{1-p_2}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$ where p_2 is the probability of the student has high alcohol consumption level and x_js are explanatory variables for predicting log-odds $log(\frac{p_2}{1-p_2})$.

After applying the propensity score method, I constructed the final model with the response variable "Alc_level". Since my response variable "Alc_level" is binary, a binary logistic regression model was used again. From Figure 5, I realize that the variables "sex", "famrel", "health" and "goout" are significant, because they all have small p-values which are less than 0.05[5]. thus we have evidence to reject the null hypothesis: the estimated regression coefficients are zero. Thus, there exists a relationship between these four variables and alcohol consumption level. However, the treatment variable "romantic" has an insignificant p-value, which means there is no causal association between romantic relationship status and alcohol consumption level.

More specifically, the final model is $log(\frac{p}{1-p}) = \beta_0 + \beta_1 sex_{male} + \beta_2 famrel + \beta_3 goout + \beta_4 health$where $log(p/1-p)$ represents the log-odds of the proportion of the student has a high alcohol consumption level. Moreover,

β_0 is the intercept of the model, which cannot be interpreted under the background of this dataset[6]]. Additionally,

β_1 represents how much the log-odds of the proportion we expect to change for a male student compared with a female student, holding all other variables fixed.

[5]set significance level as 0.05

[6]In this dataset, famrel, goout, health are rated from 1 to 5, they cannot be 0, thus the intercept cannot be interpreted

$\beta_2$represents how much the log-odds of the proportion we expect to change when the self-rated score for quality of family relationship increases by 1 unit, keeping all other variables fixed.

$\beta_3$represents how much the log-odds of the proportion we expect to change when the self-rated score for the frequency of going out increases by 1, keeping all other variables fixed.

$\beta_4$represents how much the log-odds of the proportion we expect to change when the self-rated score for the current health status of going out increases by 1, keeping all other variables fixed.

2.3 Model specifics

I want to point out that numeric variables "famrel","goout", and "health" will be treated as quantitative variables in my study because they are scores for self-evaluation of students' conditions. I assume between any consecutive integer score are space equally. For example, Walc can be valued from 1-very low to 5-very high, for a change from 1 to 2 has the same length as the change from 3 to 4, et. Cetera (Grace-Martin, 2020b). If we treat these variables as qualitative variables, it would make the model too long and complicated. Therefore, only "sex" will be treated as a qualitative variable in my study. "sex" would be a dummy variable, when "sex" has value 1, it represents the student is a male, otherwise the student is female. Furthermore, instead of using age group, "age" is more reasonable, because the age range is relatively narrow, it is unnecessary to separate into different age groups.

2.4 Validation of Model

2.4.1 Check Assumptions

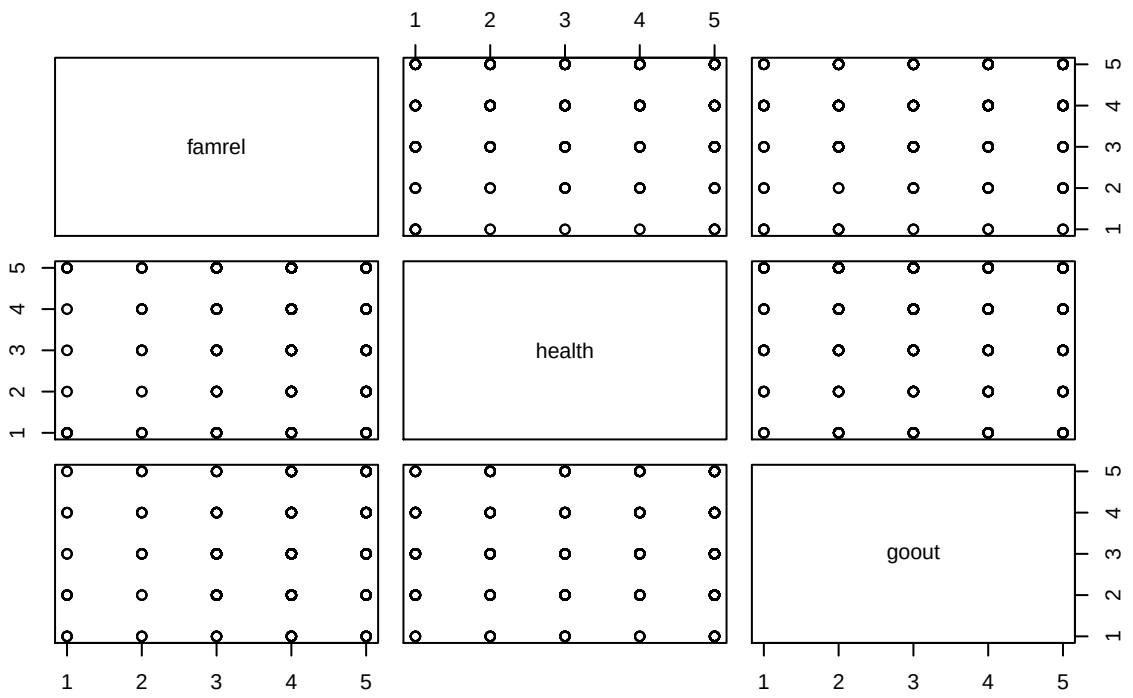
According to the guidance in Assumptions of Logistic Regression (2020), there are some assumptions I need to check for my model.

First of all, my response variable is binary which is appropriate to do the binary logistic regression model.

Secondly, all observations are independent of each other, because every student from the two Portugal Secondary schools only corresponds to one observation in the dataset. Thirdly, from Figure 6 and Figure 7, we can see the correlation coefficient is small between any two predictor variables and there is no obvious correlation between any two predictor variables. This satisfies the assumption that only a little or no collinearity among predictor variables.

Finally, Logistic requires a large sample size, a recommended minimum sample size for my study is about N=400. My dataset has a sample size of N=649, matched dataset in propensity score matching is N=410. This assumption is also satisfied. Therefore, my model satisfies basic assumptions for the logistic regression.

(Figure 6. Scatterplot matrix for numeric predictor variables)



(Figure 7. Tables for correlation coefficient between numeric predictors)

```
corrx
```

```
##        famrel health  goout
## famrel   1.00  0.110  0.090
## health   0.11  1.000 -0.016
## goout    0.09 -0.016  1.000
```

2.4.2 Validate Model Pragmatically

It is reasonable that the final model is related to the students' self-reported health status, gender, the quality of the family relationship,s and the frequency of going out with friends. First of all, alcohol consumption is usually related to people's health both physically and psychologically. Secondly, males and females are biologically different and they may treat alcohol differently. Thirdly, the relationship with family members would heavily influence the youths' emotions, and family members' drinking behaviors also have a significant impact on the students. Forth, the frequency of going out with friends would be related to the peers' drinking behaviors impact on the students, especially for the youth, peers' impact even more powerful than their family members. Furthermore, it is not difficult to collect data about these four aspects of real life, so the model is usable.

2.5 Alternative Model

Alternative model is $log(\frac{p}{1-p}) = 0.090 + 0.250sex_{male} - 0.060famrel + 0.136goout$. The model is selected by BIC in backward elimination and it has the lowest value of BIC. BIC strongly penalized free parameters in the model, thus the alternative model contains fewer parameters which may lead to underfitting. Also, the regression coefficient is biased due to the properties of BIC selection.

IV.Result

According to Figure 5, we know that there is no causal association between a romantic relationship and the youths' alcohol consumption from the Portugal language class dataset. However, I found out other 4 factors that are related to students' alcohol

The final model is determined, which is $log(\frac{p}{1-p}) = -3,347 + 1.318sex_{male} - 0.320famrel + 0.686goout + 0.162health$. The coefficients are interpreted as below:

$\beta_1 = 1.318$: Keeping all other variables unchanged, the log-odds of the proportion we expect to increase by 1.318 for a male student compared with a female student. Exponentiating log odds and get the Odds ratio equals 3.736 which means the odds for males are about 273.6% higher than females. (female is the reference group, male=0)

$\beta_2 = -0.320$:Keeping all other variables unchanged, the log-odds of the proportion we expect to decrease by 0.340 when the self-rated score for quality of family relationship increase by 1 unit. The odds ratio equals 0.726. That means for each additional score for quality of the family relationship, we expect to see the probability of the students have high alcohol assumption decrease by 27.3%, when quality of family relationship increase by 1 unit.

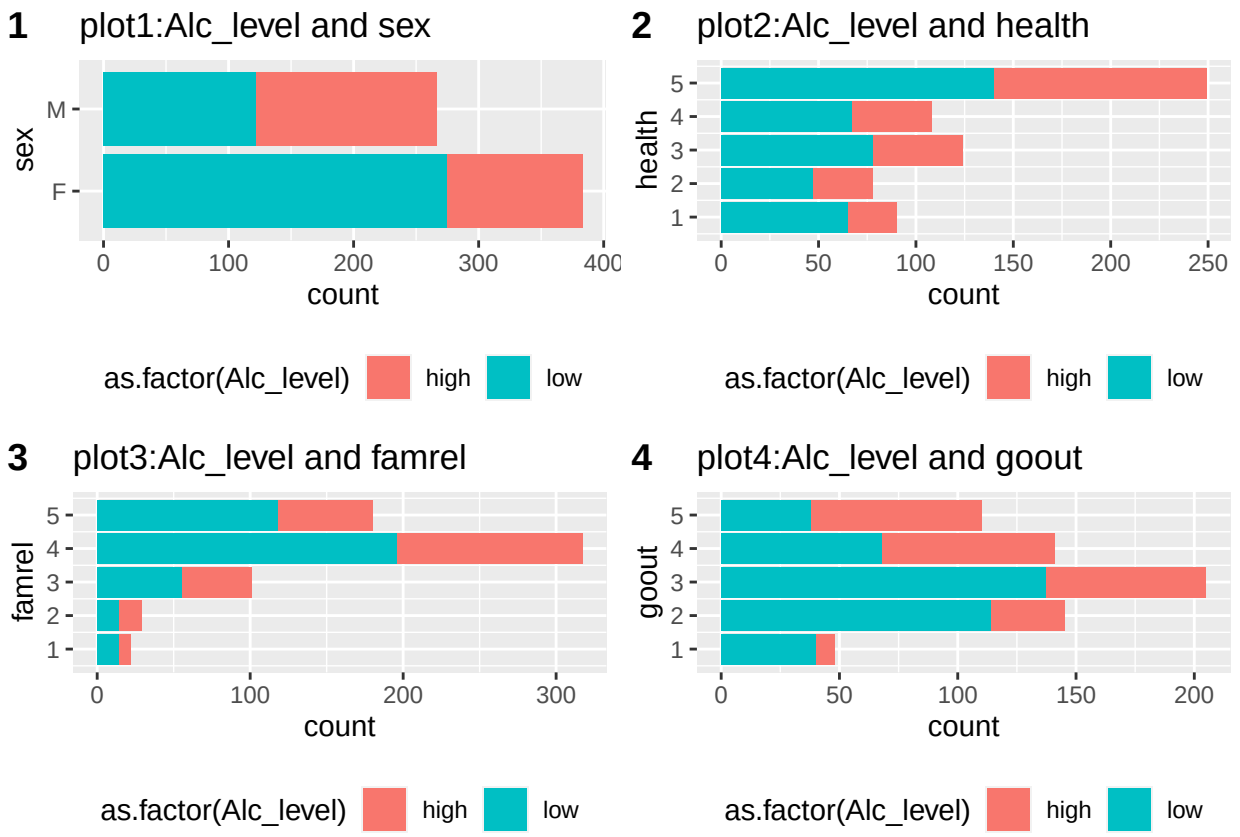
$\beta_3 = 0.686$:Keeping all other variables unchanged, the proportion we expect to increase when the self-rated score for the frequency of going out increases by 1 unit.The odds ratio equals 1.986 which means we expect to see 98.6% increase in the odds of having high-level alcohol consumption,when the frequency of going out with friends increases by 1 unit.

$\beta_4 = 0.162$:Keeping all other variables unchanged, the log-odds of the proportion we expect to increase by 0.136 when the self-rated score for the current health status of going out increase by 1.The odds ratio equals to 1.176 which means we expect to see 17.6% increase in the odds of having high-level alcohol consumption.

V.Discussion

1. Further discussion about my study

(Figure 8. Plots for Alcohol consumption level vs predictor variables)



Above, we firstly select variables for the model by using explanatory data analysis. Secondly we done the propensity score analysis, and finding out that the romantic relationship has no significant impact on the students' alcohol consumption level. By propensity score analysis, it shows that males are have 273.6% higher in the probability of having high alcohol consumption level than females. Also, people increases their score of frequecy of going out with friends, they are 98.6% higher in the probability of having high alcohol consumption level. For more analysis, from Figure 8, we can see that a higher proportion of males have high alcohol consumption. And students with high self-evaluated health scores have a higher proportion of high alcohol consumption levels. This may because people who think themselves are physically healthy would ignore regulating their drinking pattern. This may be the start of risk drinking.

Additionally, when the quality of family relationships is better, the number of students is more likely to have low alcohol consumption. A good relationship with family members well can make students feel beloved, so students' emotions would be more stable. Students can also have the chance to express and release their stress by chatting with their other family members, instead of paralyzing their emotions by drinking alcohol. Also, As students go out with their friends more frequently, they are more likely to have a high alcohol consumption level.

The result gives me some insights into the youths' alcohol-related problems. Alcohol consumption is closely related to the environment, including our surrounding family members and friends. For parents or psychologists who want to help the youths with alcohol-related problems, rather than blaming the youths, it would be better to consider the issues from the aspect of the quality of the family relationship, the youths' friends' drinking pattern, and the youths' attitude toward their health, and help the youth to overcome the alcohol-related problems.

2. Weakness

There are several weaknesses in this study. First of all, my dataset is relatively small and it was collected 14 years ago. Therefore, the situation about the youths may be different from now, so the influential factors on the youths' alcohol consumption are different. Secondly, the dataset only includes Portugal youths'

information, the youths in other countries are not considered, the cultural background of the youths would be one of the biases. Thirdly, most of the variables are based on self-evaluated scores, every participator would have a different scale for the score which may impact the accuracy of the model. Fourth, this dataset does not have a well-defined variable for representing students' overall alcohol consumption, therefore I finally create Alc_level based on "Walc" which can reflect students' real alcohol consumption better.

3. Next step

For my next step, I would collect more new data about the youths all around the world to enlarge my dataset, and separate the new dataset into a training set and test set. Then I will do the cross-validation, train the model again, and test my model's binary classification ability by Receiver Operating Characteristic Curve. Also, I may add more relevant variables, especially a better new variable for reflecting the youths' overall alcohol consumption, to improve my model.

reference

1.Wikipedia contributors. (2020, December 11). Youth. Wikipedia. https://en.wikipedia.org/wiki/Youth

2.Price, M., Hides, L., Cockshaw, W., Staneva, A., & Stoyanov, S. (2016). Young Love: Romantic Concerns and Associated Mental Health Issues among Adolescent Help-Seekers. Behavioral Sciences, 6(2), 9. https://doi.org/10.3390/bs6020009

3.Romantic Relationships in Adolescence. (2020). Sexual Development - ACT for Youth. http://actforyouth.net/sexual_health/romantic.cfm

4.World Health Organization. (2018). Global status report on alcohol and health 2018. Licence: CC BY-NC-SA 3.0 IGO. https://www.paho.org/hq/index.php?option=com_docman&view=download&slug=who-s-global-status-report-on-alcohol-and-health-2018-1&Itemid=270&lang=en

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

6.BRITO, A. ; TEIXEIRA, J., eds. lit. – "Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008". [S.l. : EUROSIS, 2008]. ISBN 978-9077381-39-7. p. 5-12.

7.Grace-Martin, K. (2020, January 16). 3 Situations when it makes sense to Categorize a Continuous Predictor in a Regression Model. The Analysis Factor. https://www.theanalysisfactor.com/3-situations-when-it-makes-sense-to-categorize-a-continuous-predictor-in-a-regression-model/

8.Wikipedia contributors. (2020b, December 17). Extracurricular activity. Wikipedia. https://en.wikipedia.org/wiki/Extracurricular_activity

9.Juergens, J., & Parisi, T. (2020, November 30). Causes and Risk Factors of Alcoholism. Addiction Center. https://www.addictioncenter.com/alcohol/alcoholism-causes-risk-factors/

10.Young people, alcohol and influences. (2016b, March 16). JRF. https://www.jrf.org.uk/report/young-people-alcohol-and-influences

11.Wikipedia contributors. (2020c, December 21). Exploratory data analysis. Wikipedia. https://en.wikipedia.org/wiki/Exploratory_data_analysis

12.Assumptions of Logistic Regression. (2020, June 22). Statistics Solutions. https://www.statisticssolutions.com/assumptions-of-logistic-regression/

13. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.