Analyzing the Effects on Young People's Health Index: No Casual Effect Between Drinking Alcohol and Health Index

Jiayue Li

12/6/2020

Code and data supporting this analysis is available at: https://github.com/Carriejiayue/Young-People-s-Health-Index

Abstract

Young people's health problems always draw public attention; significantly, young people's parents are concerned about keeping them healthy and what factors could affect young people's health. Obtained a dataset from Kaggle about Student Alcohol Consumption to investigate what will affect young people's health and whether there is a causal effect between the young people's health index and drinking alcohol. Used the propensity score to match treated (drinking alcohol) and non-treated (no drinking alcohol) groups, then built a linear regression model to predict young people's health index based on some relevant variables in the dataset. In the results, there is no causal effect between the young people's health index and drinking alcohol; and sex, quality of family relationships and young people's final grade will impact their health.

Keywords

Propensity Score, Linear Regression, Causal Inference, Alcohol Consumption, Health Index

Introduction

The health problems of young people who are 10-24 years old, according to WHO (Adolescent health), are always arousing broad public concern. More than 2.6 million young people die each year (#YouthStats: Health - Office of the Secretary-General's Envoy on Youth), this number is incredible and breathtaking, since each young people is a hope of a family and even a country. Thus, young people's health problems draw public concern. The health problems consist of injuries, mental health, alcohol and drugs, physical activities (Adolescents: health risks and solutions). Particularly, young people have more significant risks of alcohol-related harm than adults because young people's brains are still developing, drinking more alcohol may damage their brains, which will lead to problems in later life (Department of Health & Human Services, 2020).

Due to the importance of young people's health problems, this report will analyze what factors could impact young people's health and construct a causal link between drinking alcohol and health index. One popular way to construct causal inference is using propensity score matching (Propensity score matching 2020). In the observational study, we cannot compare the treatment and control groups directly because they are different at the baseline. However, the propensity score that has balancing property plays an essential role in making them comparable. Similar propensity score represents treated, and untreated groups have similar/identical property. Balancing property means if we control the propensity score, then we turn an observational study into a randomized experiment (n.d.). This report uses the propensity score to observe if there is a causal

link between young people drink alcohol and their health index.
In the following section, the dataset "Student Alcohol Consumption" from Kaggle (Learning, 2016) will be used to investigate how propensity score matching could make a causal inference between drinking and health index and what factors could affect their health. In the Methodology section, the dataset will be introduced, and the model will be built to perform the propensity score analysis. In the Results section, comprehensive and detailed results will be provided based on the linear regression model and some plots. Finally, a summary, conclusion and limitations will present in the Discussion section.

Methodology

Data

The Student Alcohol Consumption data is from Kaggle (Learning, 2016), but it is original from the UCI Machine Learning Repository (P. Cortez and A, 2008). It was used to predict secondary school student performance and investigate what factors can affect Portugal's student achievement. This data was collected during 2005-2006 from two public schools from Portugal's Alentejo region. The database was built from two sources: school reports and questionnaires. School reports contain the variables three-period grades and the number of school absences. They conducted the questionnaires to collect personal questions, such as the mother's / father's education, family income, alcohol consumption, number of past class failures, etc. The questionnaire was reviewed by schools and tested on 15 students to get feedback. The final version of the questionnaire included 37 questions, and 788 students answered it. Afterwards, 111 answers and some variables were removed due to lack of identification details and discriminative values. Finally, the data was separated into two datasets the Mathematics (395 observations) and the Portuguese language (649 observations).
The target population of this data is that students are between 15-24 years old in Portugal. The frame population is lists students of two public schools in the Alentejo region of Portugal. The sample is 677 students who answered the questionnaire without missing identification details and discriminative values. This study's key features are they used a data mining model to extract high-level information from the raw data, and it also provided some automated tools to help the education field. They used lots of decision trees as well, each tree indicates one random feature, and it is easy to understand by people. In terms of its strengths, they collected information from many aspects that are related to student performance. Variables are all meaningful, interesting and clean. On the contrary, the data size is too small, which will reduce the accuracy of the results. Mostly, they want to see what factors can affect student achievement in Portugal. The data that is only from two schools is inadequate.
This report will use the Portuguese language dataset because it has more observations than the Mathematics dataset. This report analyzes the young people's health, so pick health as the outcome variable from the dataset. Health represents students' current health status from 1 (very bad) to 5 (very good). According to Statistics Canada, health status contains people's general health, mental health, and life stress (Health status 2016). Parental relationships are essential to young people (2016); it will affect young people's psychology and social behaviours. I will choose the variable "famrel" (quality of family relationships) because this might impact young people's mental health. This variable is from 1 (very bad) to 5 (excellent). Next, I will pick the variable "Walc" (weekend alcohol consumption) that is from 1 (very low) to 5 (very high), as we know drinking more alcohol will impact health. There is another variable weekday alcohol consumption in the dataset. It is too similar to the variable weekend alcohol consumption, and young people have more free time during the weekend, they tend to have higher alcohol consumptions than on weekdays. Thus, the variable weekday alcohol consumption will be more meaningful and useful in this study. Additionally, I create a new dummy variable drinking based on "Walc". Assuming that young people do not drink alcohol, if "Walc" is 1, then drinking will be 0; otherwise, it will be 1, which means young people drink alcohol. Drinking will be treated in the following step to do the propensity score matching. Meanwhile, I pick the variable "studytime" (weekly study time) from 1 to 4; 1 means weekly study time is less than 2 hours; 2 means weekly study time is between 2-5 hours; 3 means weekly study time is between 5-10 hours; 4 means weekly study time is greater than 10 hours. Also, select the variable "G3" (final grade) from 0 to 20. In Portugal, the marking scale is only 20 points, 0 is the lowest score, and 20 is the highest score. Since young people's stress may come from

daily study, pick these two variables related to study. Too much stress will cause mental problems such as depression and eating disorder (McCullough). Finally, pick two general variables sex and age. Overall, the outcome variable is health and explanatory variables are sex, age, "famrel", drinking, "studytime" and "G3".

Table 1: Data summary

Name	dataset
Number of rows	649
Number of columns	8
Column type frequency:	
factor	4
numeric	4
Group variables	None

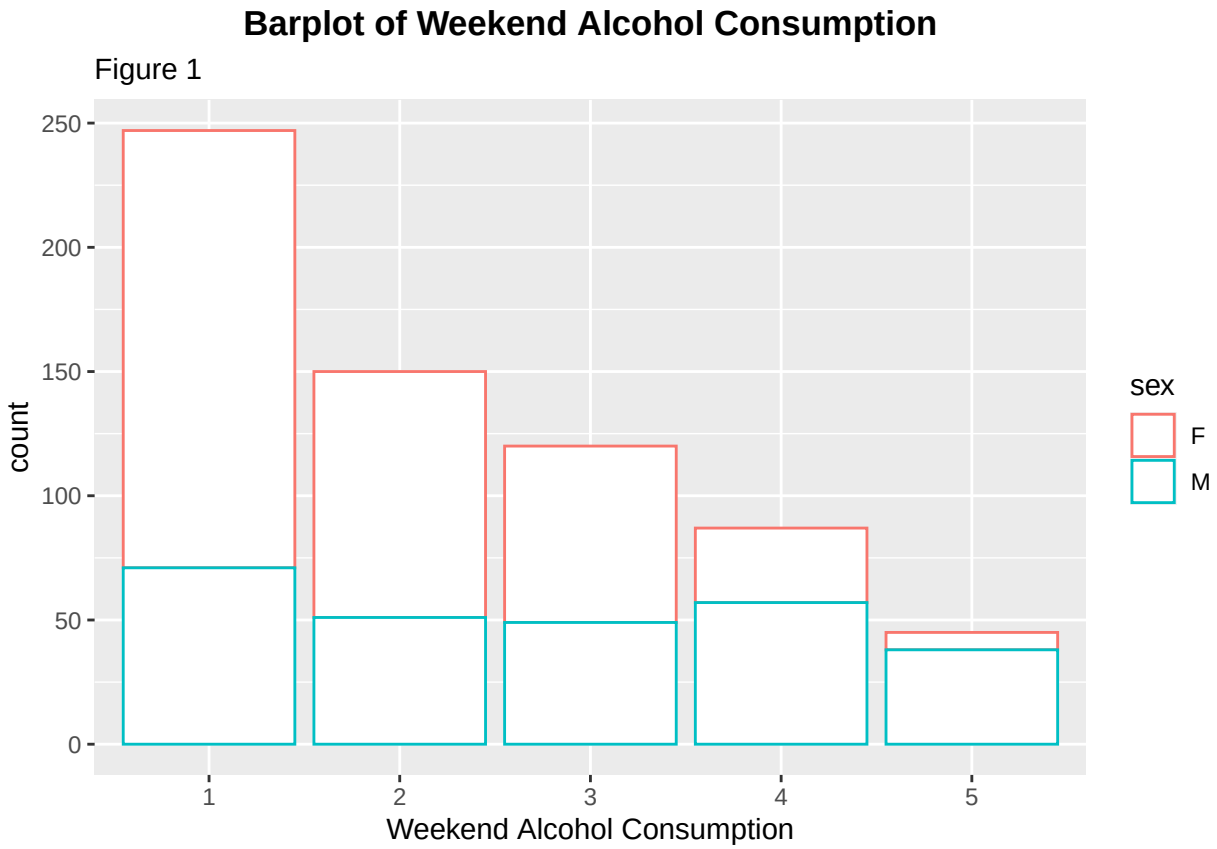
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	F: 383, M: 266
famrel	0	1	FALSE	5	4: 317, 5: 180, 3: 101, 2: 29
studytime	0	1	FALSE	4	2: 305, 1: 212, 3: 97, 4: 35
Walc	0	1	FALSE	5	1: 247, 2: 150, 3: 120, 4: 87

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	
health	0	1	3.54	1.45	1	2	4	5	5	
G3	0	1	11.91	3.23	0	10	12	14	19	
drinking	0	1	0.38	0.49	0	0	0	1	1	

There are 649 observations with eight variables in the dataset. The variables "famrel", "studytime" and "Walc" contain 5, 4 and 5 levels, respectively. Thus, "famrel", "studytime", "Walc" and sex are the categorical variables; and drinking is a dummy variable. Other variables are numerical. No missing values are in the dataset. In the dataset, most young people have the quality of family relationships ("famrel") at level 4, weekly study time ("studytime") is at level 2 and weekend alcohol consumption ("Walc") is at level 1.



According to Figure 1, when Weekend Alcohol Consumption is at level 1, females' number is higher than males. When Weekend Alcohol Consumption is at level 5, the number of females is hugely less than males. Meanwhile, this bar plot shows that when the level of Weekend Alcohol Consumption increases, the number of females is decreasing, and the number of males is roughly remaining the same.

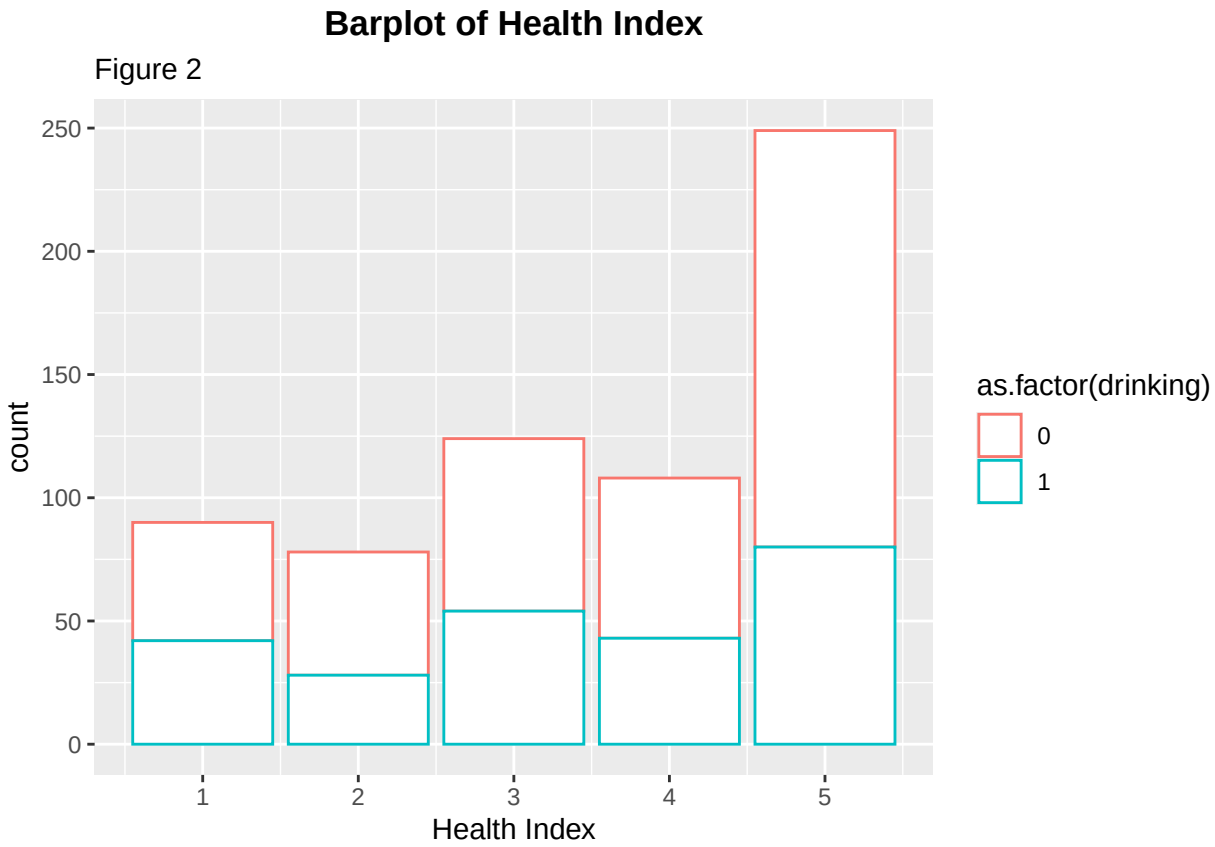


Figure 2 shows that the red parts represent no drinking alcohol, and the blue part represents drinking alcohol. When the Health Index is 5, the number of no drinking alcohol is hugely higher than drinking alcohol. Most young people have health index 5 in this dataset; thus, they are very healthy.

```
##                     Stratified by drinking
##                      0             1             p      test
##   n                    402           247
##   sex = M (%)          195 (48.5)     71 (28.7)  <0.001
##   age (mean (SD))    16.83 (1.18)  16.61 (1.27)   0.029
##   famrel (%)                                      0.132
##      1                  15 ( 3.7)      7 ( 2.8)
##      2                  20 ( 5.0)      9 ( 3.6)
##      3                  71 (17.7)     30 (12.1)
##      4                 196 (48.8)    121 (49.0)
##      5                 100 (24.9)     80 (32.4)
##   studytime (%)                                  <0.001
##      1                 154 (38.3)     58 (23.5)
##      2                 183 (45.5)    122 (49.4)
##      3                  51 (12.7)     46 (18.6)
##      4                  14 ( 3.5)     21 ( 8.5)
##   health (mean (SD))  3.64 (1.43)   3.37 (1.46)   0.020
##   G3 (mean (SD))     11.63 (3.31)  12.36 (3.06)   0.005
```

This table is providing baseline characteristics of the dataset, and it is separated by treatment group drinking. There are two strata in this table, drinking 0 means no drinking alcohol, and 1 indicates drinking alcohol. The number of young people who do not drink alcohol is greater than the number of young people who

drink alcohol. This table shows the mean and standard deviation of numerical variables, the number and percentage of each level in every categorical variable in the sample dataset. The mean age is around 17 years old. The mean health of no drinking alcohol is slightly higher than drinking alcohol. Also, young people who drink alcohol have a higher mean of G3(final grade). About half of young people have the 4th level of "famrel" (quality of family relationships), and "studytime" (weekly study time) is at level 2.

Model

This section uses logistic regression to calculate the treatment group's propensity score drinking alcohol and match them based on the nearest propensity score. Afterwards, a new matched dataset is created; each treated observation is paired with a non-treated observation with a similar propensity score. Next, build a linear regression to predict young people's health index using the variables sex, age, "famrel" (quality of family relationships), drinking, "studytime" (weekly study time), and "G3" (final grade) in the new matched dataset. Meanwhile, we can investigate whether there is a causal effect between health and drinking alcohol and what factors can influence young people's health. All of the procedures will be done in RStudio.
Among these variables, sex is the categorical variable because it contains two categories M(male) and F(female). The variables "famrel" (quality of family relationships) and "studytime" (weekly study time) are also categorical because they have level effects, they contain different levels, and each level represents various meanings. Thus, different levels will have different effects on young people's health index. Drinking is a dummy variable; 0 indicates no drinking alcohol, 1 indicates drinking alcohol. Sex and "G3" (final grade) are numerical variables. Additionally, use the variable health to be the outcome variable, then build a linear regression to predict the young people's health index. In this dataset, using age is better than age-groups because the range of age is only between 15-22, age-groups are redundant, and they will reduce the model's accuracy.

```
propensity_score <- glm(drinking ~ age + sex + famrel + studytime + G3,family = binomial,
                        data = dataset)

dataset <- augment(propensity_score,
          data = dataset,
          type.predict = "response") %>%
  dplyr::select(-.resid, -.std.resid, -.hat, -.sigma, -.cooksd)

dataset <- dataset %>%
  arrange(.fitted, drinking)

dataset$treated <-
  if_else(dataset$drinking == 0, 0, 1)

dataset$treated <-
  as.integer(dataset$treated)

matches <- arm::matching(z = dataset$treated,
                         score = dataset$.fitted)

dataset <- cbind(dataset, matches)

dataset_matched <- dataset %>%
  filter(match.ind != 0) %>%
  dplyr::select(-match.ind, -pairs, -treated)
```

The R code above demonstrates how to calculate the treatment group's propensity score and match the treated and non-treated groups. First of all, calculate the propensity score of treatment group drinking

according to all variables except the outcome variable health. I construct a logistic regression model to explain whether a person was treated as a function of variables. It demonstrates the probability of observations being treated or not based on some independent variables. Next, use the forecast to the dataset to create matches. Then, for each treated (drinking alcohol) young person, find an untreated (no drinking alcohol) young person with a similar propensity score to match them. I am using a matching function from the arm package to accomplish this step. Afterwards, I get a new dataset_matched with 494 observations, 246 treated, and 246 non treated. Finally, construct a linear regression to examine the factors of being treated on the average health index. The linear regression model is:

$$health\ index = \beta_0 + \beta_1 x_{age} + \beta_2 Male + \beta_3 x_{famrel_2} + \beta_4 x_{famrel_3} + \beta_5 x_{famrel_4} + \beta_6 x_{famrel_5} +$$

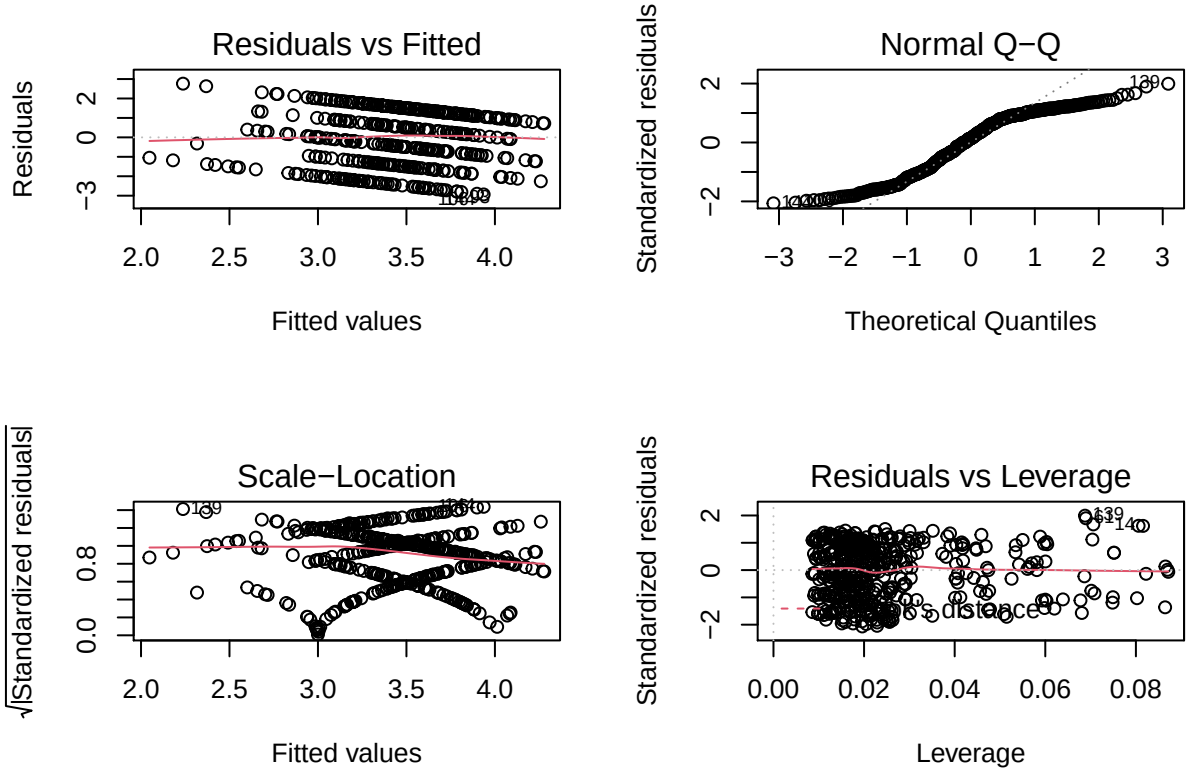
$$\beta_7 x_{study\ time_2} + \beta_8 x_{study\ time_3} + \beta_9 x_{study\ time_4} + \beta_{10} Drinking + \beta_{11} x_{G3}$$

```
##
## Call:
## lm(formula = health ~ age + sex + famrel + studytime + drinking +
##     G3, data = dataset_matched)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9365 -1.2739  0.1974  1.2934  2.7627
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.228856   1.045014   3.090  0.00212 **
## age          0.004751   0.056395   0.084  0.93290
## sexM         0.273362   0.149516   1.828  0.06812 .
## famrel2      0.500947   0.477617   1.049  0.29477
## famrel3      0.706076   0.396428   1.781  0.07553 .
## famrel4      1.046316   0.367938   2.844  0.00465 **
## famrel5      0.940255   0.373493   2.517  0.01214 *
## studytime2  -0.294732   0.162686  -1.812  0.07066 .
## studytime3   0.231818   0.206125   1.125  0.26130
## studytime4  -0.130460   0.281471  -0.463  0.64322
## drinking    -0.184780   0.129714  -1.425  0.15494
## G3          -0.051841   0.021758  -2.383  0.01758 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.437 on 482 degrees of freedom
## Multiple R-squared:  0.06026,    Adjusted R-squared:  0.03881
## F-statistic:  2.81 on 11 and 482 DF,  p-value: 0.001455
```

According to the summary table above, the R^2 of the model is 0.0603, which means only 6.03% variation in young people's health index can be explained by the model. However, the p_value of the whole model is 0.0015, which is less than 0.05; this supports that this model is significant. Additionally, the p_value of the variables famrel4, famrel5, and G3 are 0.0047, 0.0121 and 0.0176, respectively. They are all significant because p_values are less than 0.05. Furthermore, other variables' p_vlaues are all greater than 0.05, which means they are not useful to predict young people's health index. Therefore, the final reduced model should be:

$$\widehat{health\ index} = 3.2289 + 1.0463 x_{famrel_4} + 0.9403 x_{famrel_5} - 0.0518 x_{G3}$$

Diagnostic Plots (Figure 3)



To diagnose this model, we need to check whether it satisfies model linearity, normality and constant variance. From Figure 3, in the first Residuals vs Fitted plot, a horizontal red line shows, it indicates this model satisfies the linearity consumption. In the second Normal QQ plot, most points are fall in the middle line except two heavy tails show on both sides, so this model violates the normality consumption. However, normality is not super important when we determine the model's validity. If the sample size is large enough, we can assume that it is normally distributed due to the Central Limit Theorem (The Role of Probability). In the third Scale-Location plot, we can observe that a roughly horizontal line that represents constant variance. In the last Residuals vs Leverage plot, some points are outside the red dash line. If we need to check any influential points in the model, we should calculate the cook's distance; this plot is inadequate.
In terms of the model's advantage, this model has a strong linear relationship between the outcome variable health index and every explanatory variable, and it has constant variance. It is a valid model that can predict young people's health index in the future. On the other hand, the variables are not sufficient in this model. We may neglect some essential variables correlated with young people's health index, which might reduce the accuracy of the results. The alternative model can be using variable selection to select other variables are related to young people's health index in the data and build another linear regression model to compare with this model to find a model with a better prediction.

Results

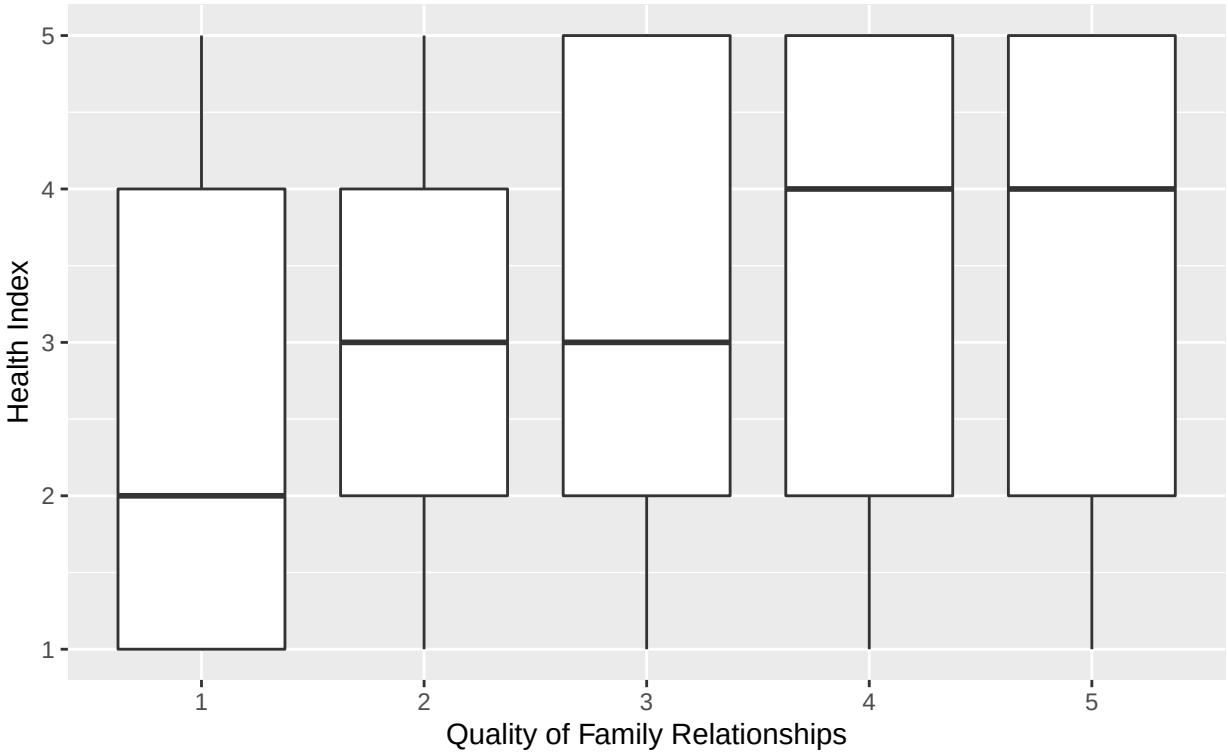
Table 4: Summary of the Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2289	1.0450	3.0898	0.0021
age	0.0048	0.0564	0.0842	0.9329
sexM	0.2734	0.1495	1.8283	0.0681

	Estimate	Std. Error	t value	Pr(>|t|)
famrel2	0.5009	0.4776	1.0488	0.2948
famrel3	0.7061	0.3964	1.7811	0.0755
famrel4	1.0463	0.3679	2.8437	0.0046
famrel5	0.9403	0.3735	2.5175	0.0121
studytime2	-0.2947	0.1627	-1.8117	0.0707
studytime3	0.2318	0.2061	1.1246	0.2613
studytime4	-0.1305	0.2815	-0.4635	0.6432
drinking	-0.1848	0.1297	-1.4245	0.1549
G3	-0.0518	0.0218	-2.3826	0.0176

In this summary model table, we get the final linear regression model $\hat{health\ index} = 3.2289 + 1.0463x_{famrel_4} + 0.9403x_{famrel_5} - 0.0518x_{G3}$ based on the p_value and coefficients. To interpret the intercept, if the young people are female who has level 1 in "famrel" (quality of family relationships) and G3 (final grade) is 0, then her mean health index is 3.2289. When other variables are same; if young female people have level 4 in "famrel" (quality of family relationships), then the mean health index difference between level 1 and level 4 is 1.0463. Also, keep other variables unchanged, if young female people have level 5 in "famrel" (quality of family relationships), then the mean health index difference between level 1 and level 5 is 0.9403. Moreover, if young female people who have level 1 in "famrel" (quality of family relationships), for each one unit increases in G3 (final grade), then the mean health index will decrease by 0.0518. We also notice that the treatment group drinking is not significant in this model; it is not useful to predict young people's health index. Hence, we cannot conclude any causal effect between young people's health index and drinking alcohol. Based on this result, we can only conclude that when young people have a high level of "famrel" (quality of family relationships), such as level 4 or 5, they tend to have a higher health index. Meanwhile, when young people's final grade is increasing, then their health index is decreasing. Besides, the p_value of the variables "sexM", "famreal3" and "studytime2" is pretty close to 0.05, and they could still have little impact on young people's health index. The variable "sexM" and "famreal3" have positive effects, but the variable "studytime2" has a negative effect.

Figure 4



I made a boxplot between the health index and the quality of family relationships to show whether or not better quality of family relationships will lead to a higher medium young people's health index. Figure 4 indicates that the medium of health index is highest when the quality of family relationships is at level 4 or 5, which has the medium health index is 4. The medium health index is lowest when the quality of family relationships is at level 1, and its medium health index is 2. When the quality of family relationships is at levels 2 and 3, the medium health index is 3. Therefore, we can see a pattern when the quality of family relationships improves; the medium health index is increased.

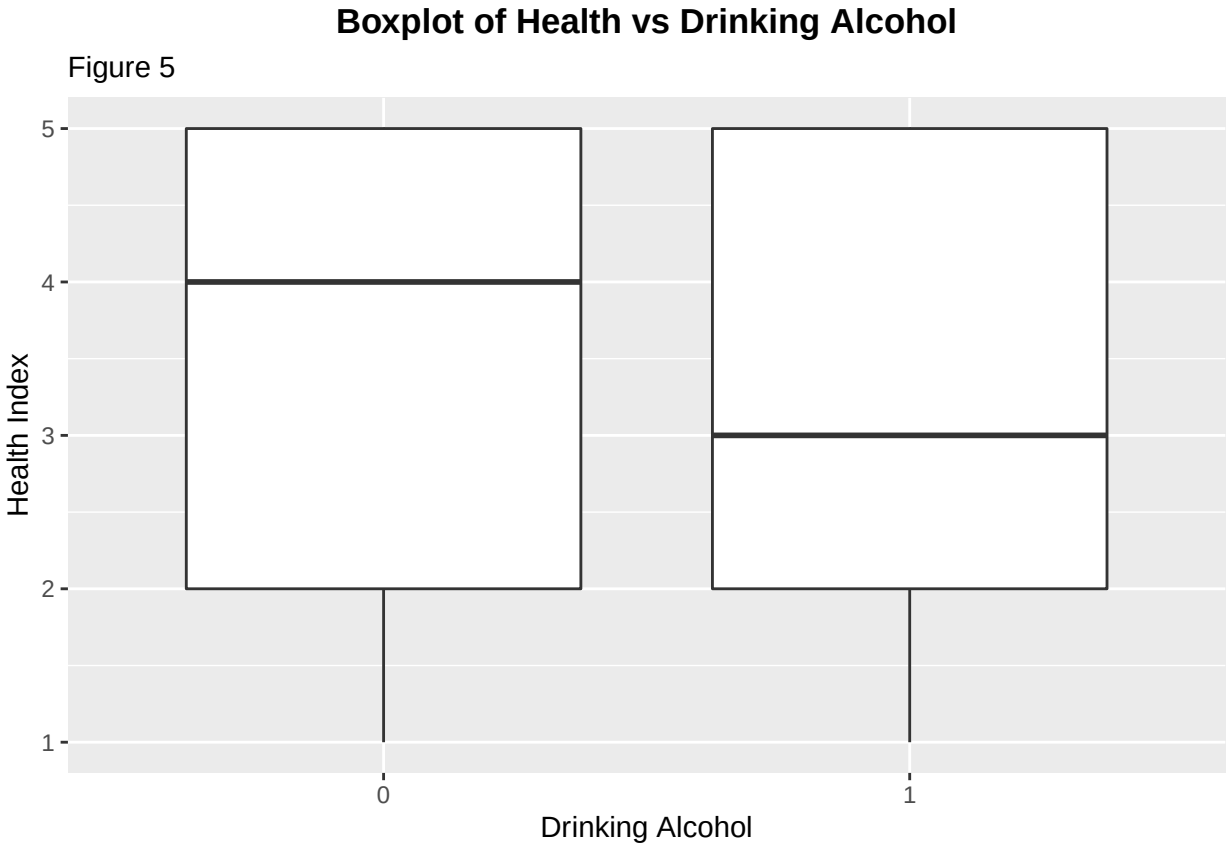


Figure 5

I produced a boxplot between the health index and the treatment group drinking alcohol to examine whether drinking alcohol will impact young people's health index. Even the treatment group is not significant in the linear regression model. Based on Figure 5, the median health index of no drinking alcohol is 4, the median health index of drinking alcohol is 3. Hence, we can get the result that drinking alcohol will decrease young people's health index.

Discussion

Summary

To summarize what was done earlier, it could be composed of five main parts. The first part is obtained data about "Student Alcohol Consumption" from Kaggle to investigate what will affect young people's health and whether there is a causal effect between drinking alcohol and young people's health index. The second part is selecting the variables relative to the objectives to create a new dataset for the following steps. The third part calculated the propensity score of treatment group drinking alcohol based on the variables in the dataset, next matching treated and non-treated groups with the closest propensity scores. A new dataset_matched was created; it dropped some observations without pairs. In the fourth part, building a linear regression model to predict young people's health index using dataset_matched, the final reduced model shows only the variable "famrel" (quality of family relationships) with level 4 and 5, and the variable "G3" (final grade) are significant. The interpretation of their coefficients shows in the results section. Thus, the treatment group drinking alcohol is not useful to predict young people's health index. In the last part, based on the model results, I made two boxplots in the results section; they are health between the quality of family relationships and drinking alcohol. Furthermore, I observed that the better the quality of family relationships, the higher of medium young people's health index. Additionally, drinking alcohol will decrease young people's health index.

Conclusions

From previous sections, we conclude that young people's health can be affected by the quality of family relationships with level 4, 5 and final grade. And there is no causal effect between drinking alcohol and health index. According to Figure 1, most young female people have very low weekend alcohol consumptions and roughly the same amount of young male people at each level of weekend alcohol consumptions. In Figure 2, 250 young people have a health index of 5 in the data. When the health index is 5, the difference in number between young people who drink alcohol and do not drink alcohol is the largest. The number of no drinking alcohol is hugely higher than the number of drinking alcohol. According to Table 4, I summarized all variables' coefficients within the linear regression model; the variables "famrel4" and "famreal5" (quality of family relationships with level 4 and 5) and "G3" (final grade) are significant and useful to predict young people's heath index. Quality of family relationships with levels 4 and 5 positively impacts young people's health index, but the final grade negatively impacts on health index. The variables "famrel3", "sexM" and "studytime2" (weekly study time with level 2) have little effect on young people's health index. Thus, young people who have a better quality of family relationships tend to have a higher health index, and I also got a similar result in Figure 4. When young people are male, the mean health index is higher than females. If the weekly study time is at level 2, which means young people weekly study is between 2 to 5 hours, it negatively affects young people's health index. In the summary table, the treatment group drinking alcohol is not significant, so it is not useful to predict young people's health index. However, young people who drink alcohol have a lower health index than people who do not drink alcohol, and this result shows in Figure 3. In conclusion, about half of young people between 15-22 years old are very healthy. If parents wonder how to keep or improve young people's health, the most important thing is to provide a harmonious and happy family relationship. The results show that some factors relative to study have negative tendencies to young people's health; this is possible lots of study workloads bring them to many pressures, which may affect their mental health. Therefore, schools or parents should consider how to reduce their pressures or help them to find better and efficient study methods. Last but not least, although drinking alcohol does not have a causal effect with young people's health, but still drinking too much alcohol will impact young people's health. Thus, young people should avoid being addicted to alcohol.

Weakness

(1)The dataset size is not large enough, and this data was collected in 2005, it was 15 years ago, the current young people's health and situations are maybe different from this data. Thus, the accuracy of using this model to predict current young people's health index will be reduced. Additionally, young people's health can be affected by various aspects not only the factors in the dataset. (2) The variable drinking was created based on the assumption. If young people are at level 1 in the variable "Walc" (weekend alcohol consumption), I assumed they do not drink alcohol. It is inaccurate because we do not know what the actual weekend alcohol consumption with level 1 is, this might be improper to assume they do not drink alcohol. (3) In propensity score matching, I used the nearest neighbour matching method to match treated and non-treated groups. However, this method can be inaccurate in some cases, may need to consider using radius matching or calliper matching. Moreover, when dataset_matched was created, there were about 150 observations dropped. This is a massive amount of deduction in the dataset.

Next Steps

Due to the scarcity of the data, it brings many inconveniences to this study. We can try to find newer data that contains more variables and observations are relative to young people's health. Alternatively, we can conduct a questionnaire at the University of Toronto to get more updated data and, using the larger dataset goes through a similar method again. Then we will get a more accurate and useful model to predict young people's health index and investigate what factors could affect young people's health.

References

1. Adolescent health. (n.d.). Retrieved December 09, 2020, from https://www.who.int/southeastasia/health-topics/adolescent-health
2. #YouthStats: Health - Office of the Secretary-General's Envoy on Youth. (n.d.). Retrieved December 09, 2020, from https://www.un.org/youthenvoy/health/
3. Adolescents: Health risks and solutions. (n.d.). Retrieved December 09, 2020, from https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions
4. Department of Health & Human Services. (2020, October 13). Alcohol and teenagers. Retrieved December 09, 2020, from https://www.betterhealth.vic.gov.au/health/healthyliving/alcohol-teenagers
5. Propensity score matching. (2020, December 02). Retrieved December 09, 2020, from https://en.wikipedia.org/wiki/Propensity_score_matching
6. (n.d.). Retrieved December 09, 2020, from https://www.methodology.psu.edu/resources/propensity-scores/
7. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. http://www3.dsi.uminho.pt/pcortez/student.pdf
8. Learning, U. (2016, October 19). Student Alcohol Consumption. Retrieved December 15, 2020, from https://www.kaggle.com/uciml/student-alcohol-consumption
9. Health status. (2016, September 28). Retrieved December 15, 2020, from https://www150.statcan.gc.ca/n1/pub/82-229-x/2009001/status/int4-eng.htm
10. Why relationships are so important for children and young people. (2016, May 20). Retrieved December 15, 2020, from https://www.mentalhealth.org.uk/blog/why-relationships-are-so-important-children-and-young-people
11. The Role of Probability. (n.d.). Retrieved December 21, 2020, from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html
12. McCullough, D. (n.d.). The effect of stress on young people. Retrieved December 21, 2020, from https://www.cache.org.uk/news-media/the-effect-of-stress-on-young-people
13. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom
14. Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr
15. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
16. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
17. Kazuki Yoshida and Alexander Bartel (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone
18. Yoshida, K. (2020, July 25). Retrieved December 22, 2020, from https://cran.r-project.org/web/packages/tableone/vignettes/introduction.html
19. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.