DOES DRINKING ALCOHOL AFFECT CHILDREN'S GRADES?

Adir Toledano and Daniel Yehezkel

IEM Faculty, Technion – Israel Institute of Technology

Abstract

The following project is part of the overall work done as part of the causal inference course in the spring 2022 semester. In the project we would like to apply the variety of methods learned in the course and answer the following question: Does drinking alcohol among high school students affect their grades in studies.

CONTENTS

1 INTRODUCTION OF THE PROBLEM

Alcohol is widely used by young people, the 2019 National Drug Strategy Household Survey found 2.8 percent of 14–17-year olds drink weekly (while for the 18–24 age group, the figure is 27.9 percent). While drinking alcohol may lead to physical health risks, such as: hangovers, headaches, nausea and vomiting, shakiness, exposure to sexually transmissible infections (STIs) and possible pregnancy it can also lead to behavioral risks such as fighting or brawling, drowning, drug overdose, self-harm and suicide.

In addition, alcohol consumption, often a normative part of the college experience (Chen Kandel, 1995), is associated with a multitude of negative consequences, including problems with academics, interpersonal relations, and the legal system THOMAS R. SYRE, JACQUELINE A. PESA and DAVID COCKLEY (1999)(e.g., Syre, Pesa, Cockley, 1999). These negative consequences appear to be a particularly problematic aspect of college student drinking (Wechsler, Davenport, Dowdall, Moeykens, Castillo, 1994)K Chen and D B Kandel (2011). However, little research has examined students' own perceptions of the consequences of their drinking (cf. Nystrom, 1992, Sadava Pak, 1993)Henry Wechsler (1994).

As the previous paragraphs show that drinking alcohol might lead to a variety of negative impacts; we think that strengthening the knowledge about young student educational impact can strengthen the understanding that resources must be dedicated to solving the drinking problem among youth. Such solution, for a widespread problem may be expensive, therefore countries and large organizations are in no rush to implement solutions.

We are concerned that we will be able to show a direct connection between drinking and alcohol and educational performance. Therefore, in our opinion when we show the connection we can encourage organizations to find and implement a solution to the drinking problem among youth.

We will try to estimate the following effect:
Does Drinking alcohol (T) affect children's grades (Y)?

2 THE DATA

2.1 Introduction

The data were obtained in a survey of students'math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students. The dataset is composed of two tables, one for each course. The math table has data of 395 students and the Portuguese language has data of 649 students. The data consists of many features: demographic, personal, educational, familial, behavioral,etc. There are two columns that describe the alcohol consumption of the students, which we can aggregate in order to represent drinking alcohol. The next explanations refer to the raw data. We will later explain the transformations we did over the data, and the definition of consuming alcohol (T=1).

2.2 Links

Kaggle dataset - Student Alcohol Consumption (2016)
link to our code

2.3 Data-set's Features

- school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: 'U' - urban or 'R' - rural)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- traveltime - home to school travel time (numeric: 1 - lower than 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - greater than 1 hour)
- studytime - weekly study time (numeric: 1 - lower than 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - greater than 10 hours)
- failures - number of past class failures (numeric: n if $1 <= n < 3$, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)

- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

2.4 Data Cleaning

While initially analyzing the data, we've noticed that a few outliers exist w.r.t the student age. Few students are over the age of 21. Additionally to the fact that these students are relatively older, they are also legally eligible to drink alcohol, which might differ their way of consuming alcohol and might biased the data ,as *drinking alcohol* conceal also *way of drinking alcohol* (forums, volumes, etc). Therefore, we've decided to drop those students.

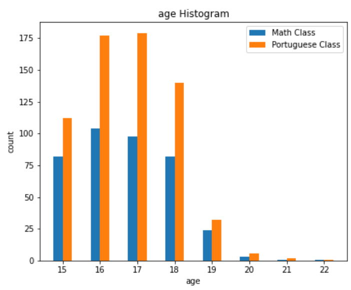
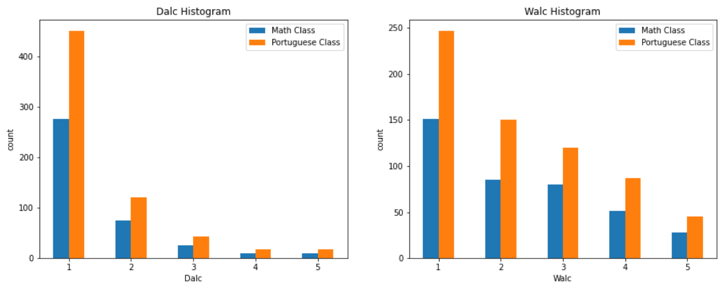


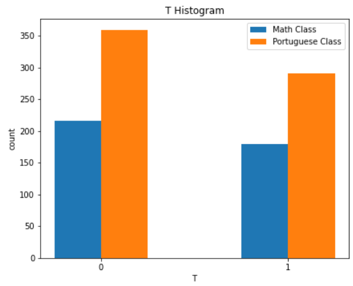
Figure 1. Age histogram before cleaning, few records are over 21 year-old.

2.5 Defining the Treatment

The raw data contains two attributes that refer to drinking alcohol: Dalc and Walc, the former represent consuming alcohol on work-days and the latter represent the same on weekend-days. First, we want to exploit both attributes in order to determine whether a student drinks alcohol. However, it's also important to distinguish between the two, as drinking in workdays is mostly rare and so in our data. Therefore we've decided to build our treatment via a linear interpolation between the two: 0.7 weight for Walc and 0.3 for Dalc. Lastly, we define a drinking student if the linear interpolation is greater or equal to two.



(a) Work-days & Weekend-days alcohol consuming histograms.

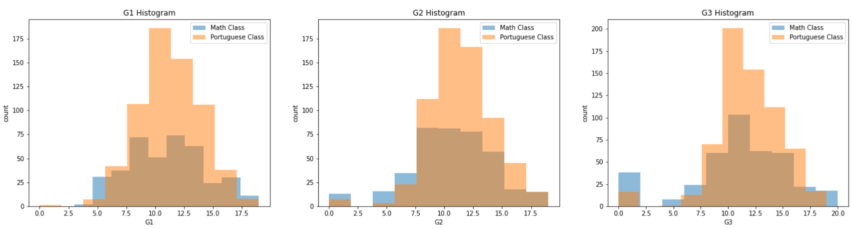


(b) Treatment (T) histogram after linear interpolation.

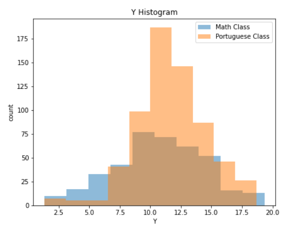
Figure 2. Histogram of alcohol consuming, drinking on work-days is relatively rare.

2.6 Defining the Outcome

The raw data contains three attributes that refer to grading student: G1, G2 and G3. First period grade, second period grade and final grade relatively. The final grade is not simply the average of the first grades, probably due to additional exams/tasks. We do not want to evaluate student educational ability based only on final grade, therefore we took their average as an outcome (Y).



(a) Raw Grades attributes histograms.



(b) Outcome (Y) histogram after averaging.

Figure 3. Linear Regression between classes grades results

2.7 Additional Manipulations

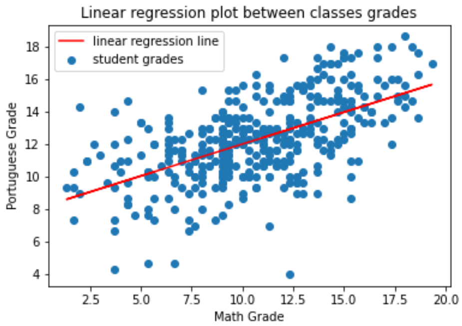
Our data is composed of two components, Portuguese language course attendants and math course attendants, some takes both courses. In particular, 57% of the students take both classes. This might lead to the independent violation. The naive way to handle this is to drop all relevant students, however it holds two main probelms:

1. As these student takes both classes, their records are 73% of the total records. Dropping them will leave us with too little data-set.

2. There might be a difference between the two populations (single class vs dual class). For example: student who takes two classes are put more effort in their education.

Therefore, we've decided to sample one record for each student in this group. We've avoided deterministic dropping in order to insert another bias to the data. In addition, we refrained from referring to the treatment variable, in order to maintain the randomness (if any) of the treatment delivery method.

If a student succeeds in Math but fails in Portuguese (or vice versa), dropping one of the classes might lead to bias. In order to validate correlation between the two, we've performed a linear regression on the dual classes students.



(a) Scattering of students grades who takes both classes, and the linear regression line

Metric	Value
Coefficient	8.1
R^2	0.34
R^2_{Adj}	0.88

(b) Results of the linear regression. High $R^2_{Adj} \rightarrow$ high amount of variance of the dependent variable represented by independent variable

Figure 4. Linear Regression between classes grades results

T		0		1			
Att.	Value	%	#	%	#	χ^2	P-value
school	GP	66.85	242.0	63.85	189.0	0.52	0.47
	MS	33.15	120.0	36.15	107.0		
sex	F	71.82	260.0	43.58	129.0	52.58	<0.05
	M	28.18	102.0	56.42	167.0		
address	R	28.73	104.0	32.09	95.0	0.72	0.396
	U	71.27	258.0	67.91	201.0		
famsize	GT3	74.03	268.0	64.53	191.0	6.53	<0.05
	LE3	25.97	94.0	35.47	105.0		
Pstatus	A	14.64	53.0	10.14	30.0	2.6	0.107
	T	85.36	309.0	89.86	266.0		
Mjob	at_home	22.1	80.0	20.27	60.0	2.44	0.655
	health	7.18	26.0	7.43	22.0		
	other	40.61	147.0	37.16	110.0		
	services	19.61	71.0	24.32	72.0		
	teacher	10.5	38.0	10.81	32.0		
Fjob	at_home	7.73	28.0	6.08	18.0	10.12	<0.05
	health	4.97	18.0	2.03	6.0		
	other	56.08	203.0	56.42	167.0		
	services	24.59	89.0	31.76	94.0		
	teacher	6.63	24.0	3.72	11.0		
reason	course	44.48	161.0	42.23	125.0	3.92	0.271
	home	21.82	79.0	24.32	72.0		
	other	9.39	34.0	13.18	39.0		
	reputation	24.31	88.0	20.27	60.0		
guardian	father	21.55	78.0	24.66	73.0	1.78	0.411
	mother	70.99	257.0	66.22	196.0		
	other	7.46	27.0	9.12	27.0		
schoolsup	no	88.12	319.0	90.88	269.0	1.03	0.311
	yes	11.88	43.0	9.12	27.0		
famsup	no	36.74	133.0	42.23	125.0	1.83	0.176
	yes	63.26	229.0	57.77	171.0		
paid	no	82.6	299.0	78.72	233.0	1.34	0.246
	yes	17.4	63.0	21.28	63.0		
activities	no	53.87	195.0	50.0	148.0	0.83	0.363
	yes	46.13	167.0	50.0	148.0		
nursery	no	18.23	66.0	23.31	69.0	2.27	0.132
	yes	81.77	296.0	76.69	227.0		
higher	no	8.01	29.0	13.85	41.0	5.24	<0.05
	yes	91.99	333.0	86.15	255.0		
internet	no	25.69	93.0	21.62	64.0	1.27	0.26
	yes	74.31	269.0	78.38	232.0		
romantic	no	61.88	224.0	62.84	186.0	0.03	0.864
	yes	38.12	138.0	37.16	110.0		

Table 1. Balancing table, for four attributes (out-of seventeen), we reject H_0, i.e. only 25% of our attributes are not balanced well.

3 CHALLENGES

3.1 Challenges

Despite the challenges we've discussed so far, like the independent violation we've treated, there are still few challenges concealed in the data. One of the assumptions is that we can not validate that the treatment given to all students is the same, i.e that the "Drinking" phenomenon is expressed in the same or even similar way between different students. Do all student consumes the same volumes of alcohol? Do all consume the same type of alcohol? Do all consume it on the same type of occasions? Do they rate their rating in the same approach? Do they feel equally comfortable answering how much alcohol they consume? We can partially deal with his challenge by validating that there is sufficient amount of common support in the probability to get the treatment.

Additional con-founder that we might miss is the student socioeconomic state. It is possible that students from lower socioeconomic state have less potential to succeed in school because they need to put effort in other things (work for example) or their emotional state onerous them, or even alternatively students who come from a higher socioeconomic status have more professional support for example private tutors and additional practice. Socioeconomic is not directly collected in the data, but it can be measured indirectly through other attributes: internet, address, family size, parents status, parents educations & jobs.

3.2 Causal effect can still be estimated

The data we have presents many challenges as we discussed above. However, It's a big data w.r.t number of features and covariates that can help us to measure the causal effect. We might miss additional con-founders. But we believe that the most significant features are in the data, thus ignorability can be applied in a way. In addition, the way of getting the treatment is not given to us but because we look on students of the same age group, we believe that alcohol tends to be consumed similarly among teens, unlike adults where the difference in their drinking habits is greater.

4 METHODS

We will focus on propensity-based methods, when we will examine the different methods by two criteria according to which we will decide with the help of What method will we estimate the causal effect. This is of course before we have seen the ATE calculated by any method. When we use the following covariates to build the classifiers: school, studytime, failures, schoolsup, famsup, activities, higher, internet, health, absences.
Our criteria are

4.1 Quality of the classifier model we use for calculating the propensity score

First, we would like the classifier to be able to evaluate well the probability to get the treatment for each unit. Thus, we can look at two metrics: Calibration graph and Brier Score.

- **Brier score is define as:** $\frac{1}{N}\sum_{i=1}^{N}\left(f(x_i)-t_i\right)^2$ where $f(x_i)$ is the probability to get the treatment for the i^{th} unit, given by the classifier; t_i is the true treatment of the i^{th} unit.
  In this phase we will also the the common support assumption still holds, and important to note

that if it seems that one of the methods has more common support it doesn't state that it is better, as the truth common support might be smaller.

- **Calibration curves** compare how well the probabilistic predictions of a binary classifier are calibrated. When the intercept is close to 0, a slope close to 1 indicates that good calibration is also maintained across the range of individuals or subgroups, whereas a slope greater than or less than 1 indicates that there are individuals or subgroups in whom calibration is poor.

4.2 Achieving a better balance

We would like to choose the model that achieves a better balance. Since we have many variables, we believe that we will not be able to achieve a good balance in all of them, so we would like to improve mainly the: study time, school support, family support and absences. Those characteristics that we believe may be con-founders.

4.3 Calculating the Propensity Score

We will calculate the propensity score, using three methods:

- Logistic Regression
- Gradient Boosted Trees with depth = 1
- Gradient Boosted Trees with depth = 3, enable catch interactions between multiple variables

4.4 Estimating the ATE

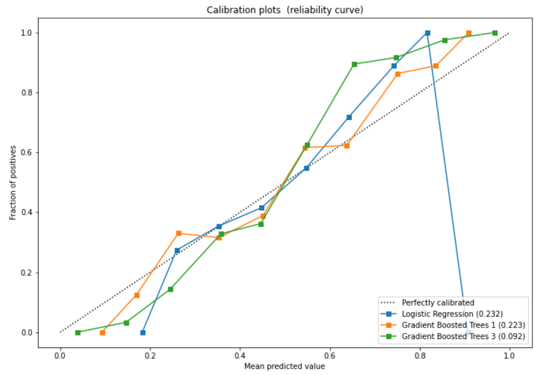
- IPW: $\frac{1}{N}\sum_{i=1}^{N}\frac{t_i \cdot y_i}{\widehat{e(x_i)}} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-t_i)\cdot y_i}{1-\widehat{e(x_i)}}$, where $\widehat{e(x_i)}$ given by the propensity score method.
- 1-NN Matching, when the matching is done on the basis of the propensity score.

Overall, we got 6 different methods, we will select the best one by the metrics we've mentioned earlier and than estimate the ATE using it. Than we can estimate the confidence interval by using bootstrap method. We will also calculate the ATE for the other methods, including the non-fixing one. As we mentioned in class, if ATEs and CIs are very different this is a bad sign for the quality of our estimation.

5 RESULTS

5.1 Quality of Classifiers

The following graphs present the comparison we did for the three models using calibration and their Brier scores. First we can see that LR has the worst brier score and not the best calibration, therefore we will concentrate on GBT1 and GBT3. It can be observed that GBT1 didn't get the best Brier score, the GBT3 did; However, it has the best correlation w.r.t calibration, as its plot is closest the the identity line (perfect classifier). We prefer to use GBT1, which receives a fairly good brier score, best calibration. In addition, it is also important to validate the existence of the common support assumption. Based on all model there is some common support. By the logistic regression and GBT1 classifiers it seems like there is a sufficient amount of common support, by the GBT3 classifier the commin support is low. However we've already rated the GBT1 as the strongest w.r.t the evaluations metrics, therefore we believe that it is fair to say the common support assumption holds.



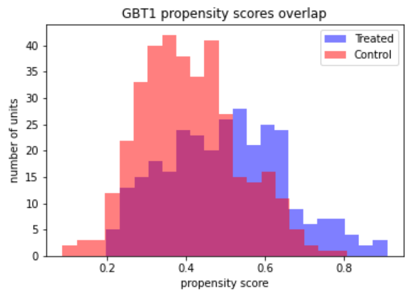
(a) Calibration plots of the three Propensity Score methods

Classifier	Brier score
Logistic Regression	0.232
GBT1	0.223
GBT3	0.092

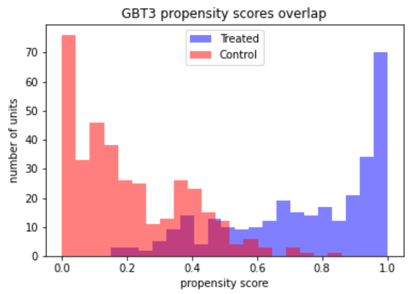
(b) Classifier's Brier scores

Figure 5. Three Propensity Score methods overlap graph and Brier scores

(a) Logistic Regression



(b) GBT, depth = 1



(c) GBT, depth = 3

Figure 6. Three Propensity Score methods overlap graph

5.2 Better Balancing

Since we are comparing 6 different methods, for convenience we will present a table comparing only the χ^2 statistic and the p-value of the different methods. A full comparison of the balance of the selected method to the original balance can be done by looking at the balance tables in the notebook. We calculated the indices for IPW methods using the propensity score scheme of the units in the same category, and normalized such that it sums to 1.

According to the balancing results it seems like the IPW method rules for all classifiers and in particular excellent for the GBT1 classifier we've selected earlier. The fact that GBT1 is the best in balancing empower our confidence for selecting it before. In addition, it is important to note that the Matching methods sometimes improved the original results which is an indication for a scent of reasonable performances.

To summary, by the evaluation metrics we've discussed, our selection is GBT1 IPW. However, it's still important to consider the other methods results. If a global statement appears, we can be more confident to receive its suggestion.

Method	Original		LR IPW		GBT1 IPW		GBT3 IPW		LR Matching		GBT1 Matching		GBT3 Matching	
Metric	χ^2	P-val.	χ^2	P-val.	χ^2	P-val.	χ^2	P-val.	χ^2	P-val.	χ^2	P-val.	χ^2	P-val.
school	0.52	0.47	0.00	1	0	1	0.00	1	–	< 0.05	–	< 0.05	–	< 0.05
sex	52.58	< 0.05	0.00	1	0	1	0.00	1	30.80	< 0.05	50.91	< 0.05	50.91	< 0.05
address	0.72	0.396	0.00	1	0	1	0.00	1	8.48	< 0.05	10.19	< 0.05	10.19	< 0.05
famsize	6.53	< 0.05	0.00	1	0	1	0.00	1	16.81	< 0.05	7.09	< 0.05	7.09	< 0.05
Pstatus	2.60	0.107	0.00	1	0	1	0.00	1	5.93	< 0.05	3.62	0.057	3.62	0.057
Mjob	2.44	0.655	0.01	1	0	1	0.02	1	4.12	0.389	7.51	0.111	7.51	0.111
Fjob	10.12	< 0.05	0.03	1	0	1	0.04	1	14.63	< 0.05	51.93	< 0.05	51.93	< 0.05
reason	3.92	0.271	0.01	1	0	1	0.00	1	24.00	< 0.05	5.72	0.126	5.72	0.126
guardian	1.78	0.411	0.01	0	0	0	0.02	0.989	1.50	0.473	8.39	< 0.05	8.39	< 0.05
schoolsup	1.03	0.311	0.00	1	0	1	0.00	1	0.10	0.748	0.62	0.429	0.62	0.429
famsup	1.83	0.176	0.00	1	0	1	0.00	1	1.28	0.258	4.08	< 0.05	4.08	< 0.05
paid	1.34	0.246	0.00	1	0	1	0.00	1	0.44	0.506	0.35	0.552	0.35	0.552
activities	0.83	0.363	0.00	1	0	1	0.00	1	0.84	0.36	1.90	0.168	1.90	0.168
nursery	2.27	0.132	0.00	1	0	1	0.00	1	1.35	0.246	1.20	0.273	1.20	0.273
higher	5.24	< 0.05	0.00	1	0	1	0.00	1	4.14	< 0.05	7.44	< 0.05	7.44	< 0.05
internet	1.27	0.26	0.00	1	0	1	0.00	1	0.44	0.508	1.03	0.31	1.03	0.31
romantic	0.03	0.864	0.00	1	0	1	0.00	1	1.08	0.299	3.88	< 0.05	3.88	< 0.05

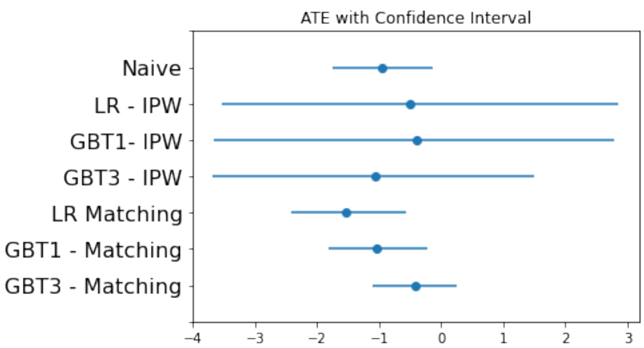
Table 2. Balancing table

6 The causal effect estimation

By the method we chose. we've gotten that: $\widehat{ATE} = -0.39$ which implies for the existence of the causal effect - drinking alcohol reflects with a reduction of 0.39 points to the student grade. That being said, the confidence interval with 95% confidence level is: [-3.49, 3.23], i.e. includes positive values, meaning it is hard to determine that the causal effect indeed exists.

Nevertheless, we can also observe the other methods results, by all, the $\widehat{ATE}$s are negative, and for all non-IPW of them, the whole intervals are lower than zero. We have already mentioned before that a global statement w.r.t all methods might brace or revoke an hypothesis.

We shall also remind that our data is not large, this might lead to the wide intervals we received for the IPW methods. We speculate that larger data would have shrink those intervals to the negative side and support the result we got. We will expand on it in the 'discussion' section.



(a) $\widehat{ATE}$ results graph

Method	$\widehat{ATE}$	LCB	UCB
Naive	-0.96	-1.75	-0.09
LR - IPW	-0.51	-3.87	2.95
GBT1 - IPW	-0.39	-3.49	3.23
GBT3 - IPW	-1.05	-3.41	1.50
LR - Matching	-1.53	-2.53	-0.5
GBT1 - Matching	-1.03	-1.89	-0.20
GBT3 - Matching	-0.41	-1.08	0.27

(b) $\widehat{ATE}$ results table

Figure 7. $\widehat{ATE}$ results

7 Possible weaknesses

There are several possible weaknesses in the methods we used and in the data available to us. First, as we mentioned at the beginning of the report: it is possible that there are additional variables that may be con-founders that we did not measure. In addition, the amount of our data is not particularly large. This could imply a lack of assumptions for the methods we proposed.

Another thing that could be a weakness is the fact that the method we chose wasn't the best in terms of **all** the aspects, brier score and calibration, as we saw in part 4 'Methods'. Of course, we would prefer the same method to be the best in every aspect, but unfortunately this is out of our control This is because there is a trade-off and we had to compromise on a slightly higher brier score in order to achieve the bets overall method.

Finally, although the causal methods we used are very accepted and widely used, they have significant weaknesses, known in advance (such as using a propensity score for matching), and it is possible that for this data it was necessary to use with more sophisticated causal methods.

8 Discussion

The initial step of our project was to find a suitable data for our mission and the topic of alcohol consuming among youth has a very special place in our hearts. Finding the right data-set wasn't an easy challenge at all: the multiple requirements and limitations cause us to reject many data-sets. After many efforts we were able to find data that would meet most of the requirements. It's important to remind: No Data is Perfect.

Throughout the project we have experienced the challenges of causal effect estimation. Working with authentic data from a questionnaire experiment was not easy. The results indeed imply the a causal effect exists and are align with our guess, back then, when we wrote the project proposal. However, they might be equivocal, and we think that larger data might solve this. Therefore, as future work, we would like to expand this work to a more comprehensive and large data and examine more methods.

In addition, we would like to emphasize that our work was very enjoyable. Along with the difficulties and debates, we saw the beauty of the world of causal inference. A different and interesting point of view. Thanks!

References

Henry Wechsler (1994). Health and behavioral consequences of binge drinking in college. *Retieved from: https://jamanetwork.com/journals/jama/article-abstract/383597*.

K Chen and D B Kandel (2011). The natural history of drug use from adolescence to the mid-thirties in a general population sample. *Retieved from: https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.85.1.41*.

Kaggle dataset - Student Alcohol Consumption (2016). https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?resource=download. *Retieved from: link to our kaggle dataset*.

THOMAS R. SYRE, JACQUELINE A. PESA and DAVID COCKLEY (1999). Alcohol problems on college campuses escalate in 1997-1998: Time for action. *Retieved from: https://go.gale.com/ps/i.do?id=GALE%7CA62894057sid=googleScholarv=2.1it=rlinkaccess*.