Project Proposal

CS396 Causal Inference

May 23, 2023

1 Group members

Please list your group members.
Zeqiu Yu
Shuo Han

2 Problem Statement

The problem that we want to study is the effects of lunch types (free/reduced and standard) on how the students perform the exam. Our rationale behind this question is as follows: This variable is a direct correlation to the income of parents and whether the students take enough food after class in the US. First ,the requirement for applying is that your parents income is below a certain level. The income of the parents also affects the development of the students. Also, the free or reduced lunch might not be as good as the standard one, and this variable also shows the students' eating style outside of campus (Since students who can afford lunch are more likely to eat better at home dinner and during weekends while the students who have free lunch might not eat enough after school), which may affect the physical conditions of the students, like brain development. If we can find out the effect does exist, we can give some concrete evidence to relevant policy makers to make some changes like give such group of students more benefice such as free dinner when there is enough funding. Then we might encourage more students who suffer from poverty to have better performance at school.

3 Causal Questions or Hypotheses

Describe at least one causal question or hypothesis you would like to investigate.

(a) What are the treatment(s) and outcome(s)? Why?

(b) Frame your question as a contrast of counterfactual random variables.

(c) If you haven't been able to decide which causal question(s) you want to ask, please list multiple possible treatments and outcomes and possible arguments for or against each.

Questions:

(a) **Causal Questions:** The treatment A in our project is the students' lunch types, namely free/reduced and standard. The outcome Y is the performance of their test score. And to better analysis the effect, we will do on-hot transformation on the categorical data: As for treatment 'lunch': $standard \rightarrow 0, reduced/free \rightarrow 1$. As for the outcome: there are three outcomes indicating score performance, and we plan to take all of them into consideration as students' total performance.

(b) **Causal Estimator:** We are interested in the Average Treatment Effect $E[Y^{a=1}] - E[Y^{a=0}]$. Where $E[Y^{a=1}]$ represent the expected value of total test performance when students are under treatment (reduced/free), and $E[Y^{a=0}]$ is the expected value of total test performance when students are under control(standard). And if we can prove that this Average Treatment Effect is significantly smaller than 0, we can tell that students who suffer form eating qualities are also likely to suffer from study.

(c) **Comment:** Overall, this causal question is indeed important investigating the impact of lunch types (Also the students' family income) on test performance has broad implications for education-based decision making, and academic achievement. It can help shape policies and practices that promote student well-being and academic success and overall-well-being.

4 Dataset(s)

We plan to use **Students Performance in Exams** data set from Kaggle. This is a top voted data set in kaggle visualization section.

(a) **Background and Contents:** Our idea is impressed by the nutritional impact on cognitive function, and we want to use this casual inference problem to explore how deficiencies in nutrients (Students who take free lunch are more likely to suffer from such problem even if they eat enough at lunch. Because most meals they eat are at their home.) may impact the test performances.

(b) **limitations:** First, there are three outcomes regrading the test performance. And we need to find out a way to comprehensively consider these three outcomes scores and these scores are of great scales, which might cause bias. Also, the behavior data is not enough here, when given more behavior variable such as students life style or their study style and help us analysis the real effect of the treatment.

(c) **Format:**This data set is a tabular data set, it has 8 columns. With 65% as control and 35% as treatment. After loading the dataset we see in Figure 1:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
5	female	group B	associate's degree	standard	none	71	83	78
6	female	group B	some college	standard	completed	88	95	92
7	male	group B	some college	free/reduced	none	40	43	39
8	male	group D	high school	free/reduced	completed	64	64	67
9	female	group B	high school	free/reduced	none	38	60	50

Figure 1: DataFrame Preview

(d) Variables:

i. `gender:` This variable describes the students' gender.

ii. `race` This variable describes which race the students belongs to. It has 5 categories.

iii. `parental level of education:` This variable describes the students parents' educational level, this variable can be considered as a confounder. Because parents' educational level can highly influnce the students life style and grade.

iv. `lunch:` This variable describes what kind of lunch the student is having. And it has two outcomes the standard one and the free or reduced one. It is also our treatment.

v. `test preparation course` This variable describes whether the student complete the preparation or not.

vi. `math reading writing scores` These three variable describe three test scores that the students have. And they are our casual problem's outcomes.

(e) **Unmeasured Confounds:** There are many unmeasured Confounds in this dataset. Factors such as students IQ, life style, or attitudes of individual students may affect their behavior during testing.

5 Expectations and Concerns

Our objective is to determine the Average Treatment Effect of lunch types (free/reduced and standard) on how the students perform the exam. Since we have limited data, we might started with Naive ATE and Backdoor adjustment to estimate ATE. We might also try to use IPW later and we will compare our results on these methods.

I will also try to perform the analysis on some existing python package like dowhy.

Also, our data consists mostly of categorical variables, we might chose to utilize the Random Forest Classifier or logistic regression to make predictions when applying the above methods.

6 References

- Spscientist. (2020). Students Performance in Exams. Retrieved from https://www.kaggle.com/datasets/performance-in-exams