Project Update

CS396 Causal Inference

May 25, 2023

Instructions

This assignment is due on Thursday, May 25 at 11:59pm CDT. It will accepted up to 72 hours late, but with a one point penalty per day late. If your assignment is late but less than 24 hours late, you'll have one point deducted; between 24 and 48 hours late results in two points deducted; 48 to 72 hours late means three points deducted. If it's more than 72 hours late, you'll receive zero points.

Please upload your (group's) update to your GitHub repository as a PDF. You can use this assignment TeX to fill in your answers. Your project update should be at least two pages but no more than five.

As always, your work must be your own. It's fine to use published packages or prepossessing code, but don't claim credit for work that you didn't do. If you use information from other sources, you must cite those.

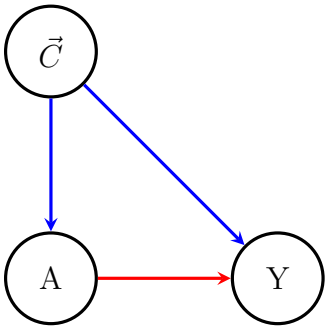
1 Group members

Shuo Han (SHV7753), Zeqiu Yu (ZYV9962)

2 Big changes

- For the reason of the complexity of the previous topics, which is hard to fulfill for just one quarter, we change the dataset and the topic of the project.
- We misunderstood the meaning of the treatment variable and change the explanation of its effect on the outcome variable.
- After some trials, we find there will be bias if we keep using the numeric outcome variables. Then we use the median as the threshold, with 1 denotes above the median and 0 denotes under the median.

3 Causal graph



Where $\vec{C}$ represent the vector of confounders, and we will use gender, race, parental level of education, and test preparation as confounders. And A is the treatment lunch type. And Y is the outcome, Test performance.

4 Counterfactual function

For your treatment A and outcome Y, write out $E[Y^a]$ as a function of the observed data. Our main purpose is to use different methods to compare the results.

$$E[Y^a] = \sum_{\vec{C}} P(Y|\vec{C}, A = a)P(\vec{C})$$

4.1 Assumptions

We are assuming SUTVA, Consistency, and Conditional exchangeability here. As for Consistency: we assume that the effect of students' lunch types on test performance is consistent for all students. It assumes that if a student had a different lunch type, their test performance would change accordingly, regardless of other factors. And as for SUTVA, we assume that there is no interference between students and each student can be seen as an independent individual. Finally, as for Conditional exchangeability, we assume that there is no difference between individuals who receive treatments when given the confounders.

5 Estimation and Interpretation

We plan to use risk difference to estimate the causal effect. We plan to implement naive ATE method, IPW and do-calculus backdoor estimator. We also try to implement some other estimators like learner methods, we are trying to figure it out and complement it later.

5.1 Point estimate and uncertainty

1. Naive ATE:
Because it is essentially an simple DAG model, we can calculate it in a naive way by calculating the difference in outcome on two groups: controlled and the intervened ones. Use the notation for control and treated group:

$$CG := \{i \; s.t. \; a_i = 0\}$$

$$TG := \{i \; s.t. \; a_i = 1\}$$

We will have:

$$ATE = \frac{1}{|TG|} \sum_{i \in TG} Y_i - \frac{1}{|CG|} \sum_{i \in CG} Y_i = -0.22$$

Then we use bootstrap method (random seed = 396, n = 1000, ci = 95), we have the confidence interval: [-0.29, -0.16].

2. Backdoor:
Then we use the backdoor estimator to calculate the causal effect. With the do-Calculus, we can use association to estimate the causal effect:
ATE = E[$Y^{a=1}$] - E[$Y^{a=0}$] = E[$Y|do(A = 1)$] -E[$Y|do(A = 0)$] = E[E[Y |$\vec{C}$, A = 1]] - E[E[Y |$\vec{C}$, A = 0]] = -0.25. In the same way, after using bootstrap method (random seed = 396, n = 1000, ci = 95), we Can get the confidence interval: [-0.31, -0.20].

3. IPW:

Then we use the Inverse Propensity Weighting (IPW) to estimate the causal effect.

$\text{ATE} = \frac{1}{n}\sum_{i\in TG}\frac{y_i}{\hat{p}(A_i|\vec{C_i})} - \frac{1}{n}\sum_{i\in CG}\frac{y_i}{\hat{p}(A_i|\vec{C_i})}$

In the formula above, we look at each sample with treatment A_i, outcome Y_i and the confounder vector $\vec{C_i}$, and the $\hat{p}$ is the estimator of $p(A|\vec{C})$. We can estimate it with many methods: like linear regression models(as shown in our HW2), logistic regression models and others. We will present them later.

We are also currently trying some learner models and will complete the results later.

5.2 Interpretation

In naive ATE method, we get a ATE of -0.22 together with confidence interval [-0.29,0.16], and in do calculus backdoor, we get a ATE of -0.25 together with confidence interval [-0.31, -0.20]. And this value is significantly smaller than 0, and we can tell that students who have free/reduced lunch are more likely to perform badly during test. Recall from our proposal that this variable is a direct correlation to the income of parents and whether the students take enough food after class in the US 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. Now we find out such an effect do exist, we might consider giving some concrete evidence to relevant policy makers to make some changes like give such group of students more benefice after class such as free dinner or offer them nutritional supplement as an extra benefit.

Most importantly, all of our results are significant at a 95% confidence level so I think this is something that all high school should pay attention to.

6 Next steps for the project

Double Machine Learning What if the relationship between A and Y is confounded by a nonlinear relationship involving a high-dimensional confounder? Just training a linear regression $E[Y \mid A, C]$ won't work in general, but so-called double machine learning provides a way to get unbiased estimates of the causal effect. While implementing these methods is outside the scope of this class, you can use existing implementations from the DoWhy or similar packages.

6.1 One additional challenge

We would like to do the double machine learning analysis in this project. It is true that the relationship between A and Y can be confounded by a nonlinear relationship. And as we can see, our data consists of mostly categorical data. In this update working, we are seeing this problem as a regression problem and use one hot transformation so that allows us to make prediction based on linear regression. Later, we might consider using machine learning methods from SKlearn such as Random Forest, Decision Tree Classifier and KNeighbors Classifier and try to predict $E[Y \mid A, C]$ based on classification problem modeling. Also we are trying ti implement learner models like X-Learner and S-Learner.

6.2 Ask for feedback

Because our data set have been so well prepossessed before we download it. So we can't do many challenges such as missing value. I wonder if we can try to implement other causal inference method that we haven't covered during class as part of our final project? And also, We would like to see your feedback on our idea towards one additional challenge part, do you have other suggestions on how we can perform the double machine learning analysis on this dataset?

6.3 Extra credit

More details are in the note book: 6.3 Extra credit: We have a python notebook that go through all the analysis process on this dataset using dowhy. This package is powerful. I first do the data processing, change all the categorical variables into one-hot labeled. Then I use this package to form a graph model based on the data. This is easy to use, we only need to tell it what our outcome, treatment and confounders are, then it will form a model for us. It also has method identified_estimand, which can help us find some possible way to do the estimation. Then based on the suggestion from identified_estimand method, I used some backdoor method like "backdoor.linear_regression" and "backdoor.propensity_score_stratification" to find out the ATE = -0.2339. In the end, I used methods in this package to perform the refuting estimate process. I performed a "refute" analysis by Adding a random common cause and using a subset of the data. They all show that there is no strong evidence to support the claim that the new effect is significantly different from the estimated effect. And finally, we obtained similar results compared to our previous analysis.

7 Reference

- Amit Sharma, Emre Kiciman. DoWhy: An End-to-End Library for Causal Inference. 2020. https://arxiv.org/abs/2011.04216
- Goman, Daniel. 2023. "Causal-Inference-Final-Project." GitHub repository. Accessed May 21, 2023. https://github.com/DanielGoman/Causal-Inference-Final-Project
- "Students Performance in Exams." Kaggle dataset. Accessed May 25, 2023. https://www.kaggle.com/da performance-in-exams
- "DoWhy 0.9.1 Documentation." PyWhy. Accessed May 25, 2023. https://www.pywhy.org/dowhy/v0.9.1