# Project Update

CS396 Causal Inference

May 26, 2023

## Instructions

This assignment is due on Friday, 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 preprocessing 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

Jose Cordova, Amir Dhami, Dennis Wu

## 2 Big changes

There have been no big changes to our project since our project proposal. We are continuing to have our treatment be participation in a job training program and our outcome be average real earnings for non-degreed individuals.

## 3 Causal graph

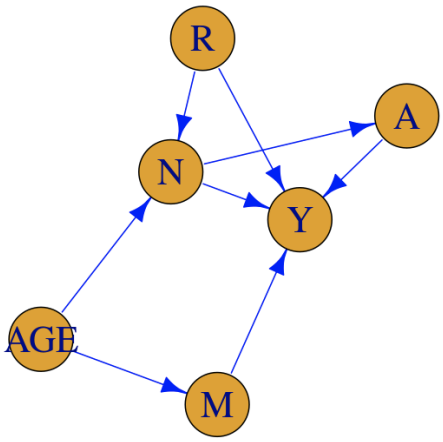


Figure 1: Causal Graph

A = Treatment: treat. Y = Outcome: [re74, re75, re76]. R = Race: [black, hispanic]. M = marriage. AGE = age.

# 4 Counterfactual function

$$\frac{\sum_{AGE,R,M} P(Y|AGE,R,M,N,A)P(N|AGE,R)P(M|AGE)P(R)P(AGE)}{\sum_{AGE,R,M,Y} P(Y|AGE,R,M,N,A)P(N|AGE,R)P(M|AGE)P(R)P(AGE)}$$

Figure 2: Function

## 4.1 Assumptions

The way we plot our causal graph is based on the assumption that the treatment assignment is not biased against people's demographics. While this assumption might not hold true in reality, we decide to have this assumption for the simplicity of our project scope.

# 5 Estimation and Interpretation

## 5.1 Dataset Stats

In order to have a better understanding of the result from our estimator, we include some important statistics of the NSW dataset 1. Overall, the earning has a very high variance across the whole dataset. We can also observe that there are some correlations between *nodegree* and their future earnings regardless of the job training program.

Table 1: Statistics of the NSW dataset

| | Average Income | Median income |
|---|---|---|
| Overall | $5300.76 \pm 6631.49$ | 3701.81 |
| $nodegree = 1$ | $4929.8413 \pm 6311.5156$ | 3557.14 |
| $nodegree = 0$ | $6631.4966 \pm 7557.5356$ | 4812.57 |

Table 2: Avg. Incomes of different subsets of NSW

| | $nodegree = 0$ | $nodegree = 1$ |
|---|---|---|
| $treat = 0$ | 4854.492 | 4495.415 |
| $treat = 1$ | 8046.518 | 5649.462 |

## 5.2 Point estimate and uncertainty

To answer the first question we have in our project proposal, we calculate the risk ratio as the following equation

$$E[Y^a|N = n] = \frac{E[Y^{a=0}|N = n]}{E[Y^{a=1}|N = n]}. \tag{1}$$

Intuitively, we use the backdoor adjustment due to the fact that there is no other variable between the treatment and outcome and we are not controlling for unobserved confounders. The risk ratio was showed in table 3. For bootstrapping, our result is shown in the table 4.

Table 3: Risk Ratios of different subsets w.r.t. nodegree

| | Overall | With degree | Without degree |
|---|---|---|---|
| Risk Ratio | 1.3874 | 1.6455 | 1.2673 |

Table 4: The confidence interval of our dataset

| | treat = 0 | treat = 1 |
|---|---|---|
| Confidence Interval | [3944.45, 5249.23] | [5272.80, 7468.70] |

### 5.3 Interpretation

From $E[Y^a]$, we can observe that people who did the job training program were expected to earn almost 1.4 times more than people who didn't. Generally, we can see that the job training program is beneficial for having higher expected income. On the other side, the job training program actually amplifies the positive effect on people who have at least high-school degrees. And for people who do not hold a high school degree, the effect of job training program is less but still positive.

Generally, having a highschool degree is beneficial to having a higher income. From pure stats, the job training program has a larger impact on earnings than having a highschool degree. However, having a highschool degree affects the effectiveness (the expected future earnings) of the job training program. Thus, one direction of improvement of this program can be providing oppotunities for subjects without a highschool degree to earn equivalent degree and to participate the job training program at the same time.

The result obviously cannot be generalized to the current job market due to the fact that highschool degrees are much more common now. However, we still can still learn the idea of the importance of education degrees on future earnings/job training program effects. Therefore, a reasonable interpretation/extension is to analogize highschool degrees in the 70s to college degrees nowadays.

## 6 Next steps for the project

### 6.1 One additional challenge

**Unobserved confounding** The National Institute of Justice mentions in the National Supported Work Demonstration Program Profile (https://crimesolutions.ojp.gov/ratedprograms/522pd) that criminals and substance abusers were part of the target population for the project. A criminal background or a failed drug test could come up during an employer's background check. This could influence such individuals' real earnings. The official findings report, which can be found at https://www.mdrc.org, presents an identical study that was instead conducted on juveniles instead of adults. Because a juvenile criminal record is expunged 2 years after the case is closed, there could be a control for failed background checks.

### 6.2 Ask for feedback

Apart from finding an existing dataset with information on the individuals' criminal records or using the juvenile dataset, are there any better ways of controlling for unobserved confounding?

### 6.3 Extra credit