Project Proposal

CS396 Causal Inference

April 25, 2023

1 Group members

Amir Dhami, Dennis Wu, Jose Cordova

2 Problem Statement

We want to understand to what extent do job assistance programs result in greater future earning power for non-degreed individuals. This is significant for individuals who cannot attend college but aspire to pursue a career in an alternative manner (i.e. apprenticeship).

3 Causal Questions or Hypotheses

(a) Our treatment A is participation in a job assistance program and our outcome Y is real earnings greater than or equal to the average for all non-degreed individuals.

(b) We are interested in the causal risk ratio $E[Y^{a=1}]/E[Y^{a=0}]$, or the expected rate of GEP (greater earning power) had everyone participated in the job assistance program divided by the expected rate of GEP had no one participated in the job assistance program.

(c) We might also be interested in examining the same treatment but for degreed individuals, which might provide a new perspective on the value of internships. However, this might introduce covariates such as the type of degree obtained, the major that was chosen, and the quality of the institution.

4 Dataset(s)

What dataset(s) do you plan to use?

We use the National Supported Work Demonstration (NSW) job-training program experiment dataset.

(a) **Background**: This dataset is a labor market experiment in which participants were randomized between treatment (on-the-job training lasting between nine months and a year) and control groups.

(b) **Limitations**: The dataset was published in 1986, so some of the results might be outdated. Some of the important features related to job training might be missing, e.g. prior work experience, household income, etc.

(c) **Data Format**: The data contains 445 observations and 11 variables. It can be described as a 445x11 data.

(d) **DataFrame**:

	data_id	treat	age	educ	black	hisp	marr	nodegree	re74	re75	
0	Dehejia-Wahba Sample	1	37	11	1	0	1	1	0.0	0.0	9
1	Dehejia-Wahba Sample	1	22	9	0	1	0	1	0.0	0.0	3
2	Dehejia-Wahba Sample	1	30	12	1	0	0	0	0.0	0.0	24
3	Dehejia-Wahba Sample	1	27	11	1	0	0	1	0.0	0.0	7
4	Dehejia-Wahba Sample	1	33	8	1	0	0	1	0.0	0.0	
5	Dehejia-Wahba Sample	1	22	9	1	0	0	1	0.0	0.0	4
6	Dehejia-Wahba Sample	1	23	12	1	0	0	0	0.0	0.0	
7	Dehejia-Wahba Sample	1	32	11	1	0	0	1	0.0	0.0	8
8	Dehejia-Wahba Sample	1	22	16	1	0	0	0	0.0	0.0	2
9	Dehejia-Wahba Sample	1	33	12	0	0	1	0	0.0	0.0	12

(e) For at least six variables (columns) in your dataset:

i. Describe that variable: what is it measuring? Is it a discrete or continuous variable? Does it have any missing values? What is its mean and standard deviation?

treat: A binary variable represents whether a sample has participanted in the job-training program. It doesn't have any missing values. 185 were given the treatment and the remaining 260 were not.

age: A continous int variable represents the age of each sample during the program starts. It doesn't have any missing values. It has a mean of 23.37 and a standard deviation of 7.10.

black: Binary variable, denotes whether a sample's race is black.t doesn't have any missing values. There were 371 samples with black race.

hips: Binary variable, denotes whether a sample's ethnicity is Hispanic. It doesn't have any missing values. There were 39 samples with hispanic ethnicity.

marr: Binary variable, denotes whether a sample is married. It doesn't have any missing values. There were 75 samples married.

nodegree: Binary variable, denotes whether a sample holds a college degree. It doesn't have any missing values. There were 348 samples with no degree.

ii. What are the possible causal relationships between this variable and the other variables (in part e)?

There are possible causal relationships between age,black and hisp with nodegree. It is possible that the race carries causal mechanismns (e.g., discrimination) that lead to a lower probability of obtaining college degree.

There could also be a causal relationship between black/hisp and marr. It is possible that cultural diferences between these groups can lead to a different probability of getting/staying married.

Another causal relationship is the relation between age and marr. It is likely that the probability of being married increases with age.

(f) What is at least one variable that your dataset doesn't contain but might be a causal factor? How might such a variable complicate your causal question(s)?

The geographical location of the sample. Salary ranges change dramatically between different cities and even between neighborhoods in the same city. Additionally, the effect of this

variable could be confounded with the effect of other variable sin the dataset e.g., black, hisp or age.

The type of industry that each sample was working in. Different jobs can have a huge difference on salary. E.g. the person who works on software engineering has a higher chance to obtain a high salary comparing to the national average.

5 Expectations and Concerns

Throughout this project,

5.1 Expectations

- We expect to utilize casual inference methods that we learn from the lecture (IPW, backdoor, frontdoor).
- We expect to learn how to identify whether a variable is confounded or not.
- We expect to learn how to estimation the expectation value even when the sample size is relatively small.
- We can utilize matching to calculate the expectation value by differentiating samples into different groups.

5.2 Concerns

- One concern is that we are unsure whether this dataset is still representative nowadays since it was collected in the 80s. Some of our findings might be surprising.
- Another thing is the type of job will significantly affect the salary, which is missing in this dataset. Thus, we do not know if the estimation can still be accurate without it.

6 References

- Rajeev H. Dehejia and Sadek Wahba. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." Review of Economics and Statistics 84.1 (2002): 151-161.
- LaLonde, Robert J. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." American Economic Review 76.4 (1986): 604-620.