Project Update: Job Training

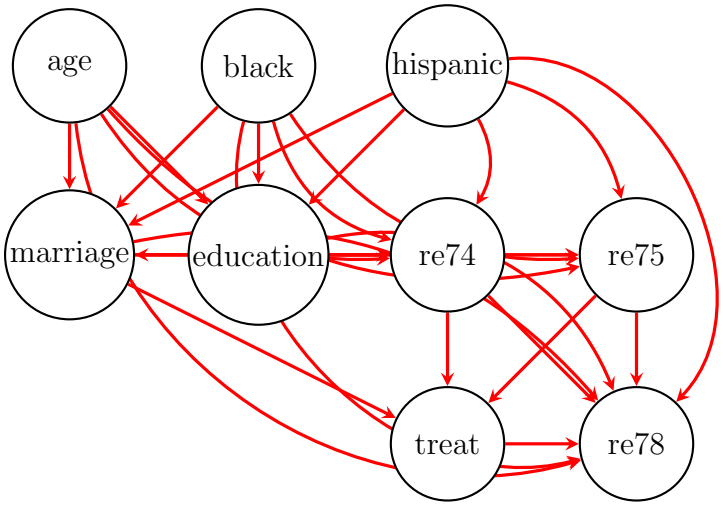
Jacob John, Sai Ganesh Nellore, Michael Hartmann, Shravan Srinivasan

May 20, 2024

1 Big changes

We've changed our project entirely from the impact of exercise on depression to the effects of job training on employment and revenue. When working through depression datasets, we realized that we either didn't have enough data or confounders. The Dehejia-Wahba subset (1999) [1] of the LaLonde dataset (1986) [3] was an excellent choice for this course as it comprised an observational and an experimental dataset with all its confounders and no missing values.

2 Causal graph



- **treat**: 1 indicidates job training was provided, 0 indicates that the individual had to fend for themselves.
- **age**: Report age of the individual at the beginning of the study.
- **educ**: Total years of formal education.
- **black**: Binary indicator (1 = individual is Black, 0 = otherwise).
- **hispanic**: Binary indicator (1 = individual is Hispanic, 0 = otherwise).
- **married**: Binary indicator (1 = individual is married, 0 = otherwise).
- **re74**: Real earnings of the individual in 1974, before the treatment. This allows for a more extended view of their economic status.
- **re75**: Real earnings of the individual in 1975, before the treatment.
- **re78**: Real earnings of the individual in 1978, after the treatment.

Note that all earnings are adjusted for 1982 US dollars thereby circumventing the need to adjust for inflation.

3 Counterfactual function

Assume Y represents our outcome variable, i.e., `re78`. Our set of confounders is $Z = \{$`education, re74, re75, black, hispanic, married`$\}$. A is our intervention variable which represents `treat` or job training. Our derivation for the counterfactual function looks below:

$$P(Y^a) = \sum_Z P(Y^a \mid A = a, Z)P(A = a, Z) \tag{1}$$

$$= \sum_Z P(Y^a \mid A = a, Z)P(A = a \mid Z)P(Z) \tag{2}$$

$$= \sum_Z P(Y \mid A = a, Z)P(A = a \mid Z)P(Z) \tag{3}$$

$$= \sum_Z \underbrace{P(Y \mid A = a, Z)}_{Y \sim A+Z} \underbrace{P(A = a \mid Z)}_{\text{propensity}} P(Z) \tag{4}$$

1. Marginalizing over Z.
2. Chain rule: $P(A = a, Z) = P(A = a \mid Z)P(Z)$
3. Consistency: $P(Y^a \mid A = a, Z) = P(Y \mid A = a, Z)$
4. Assume: $P(Z) = \frac{1}{N}$

3.1 Assumptions

In our causal graph, we can't d-separate our outcome from our treatment by conditioning on our confounders. One way to see this is by seeing that `education` is a collider. When unobserved, it blocks the path from `age, black`, however, it unblocks the path from the outcome to the treatment. Now, if we condition on it, it opens up a path. This is why we can't use the backdoor criteria. We can't apply frontdoor as we have no mediators.

For our derivation, we assumed conditional consistency in Step 3. We can't use conditional exchangeability as we're assuming confounders such as `age`, `black` and `hispanic` to have a dependency on `re74`, the outcome and the treatment (and so forth).

4 Estimation and Interpretation

Due to our complicated confounding, `DoWhy` fails to find any estimators for a backdoor estimate. However, there are two ways we are going to calculate this estimate:

- Based on our point estimand above, we can regress over our treatment and confounders to calculate an estimate of our Y. We then calculate propensity using our confounders.
- In the original paper, they created matched stratum based on propensity and took an average causal risk difference of each strata. This is what we're currently using for our causal estimates, we're iterating through a list of number of stratas and choosing an appropriate number based on our observable estimate.

For propensity, one way is to regress the treatment on the confounders. An alternative approach is to use k-means or other clustering methods for matching and then comparing within stratas. Our forthcoming goal is to create stratas that are interpretable so one can say that the effect is

more apparent for a certain race, pre-treatment revenue, etc. Furthermore, we also found a list of methods to try via Sizemore's Blogpost [4]. Some of these cover Nearest-Neighbor Propensity Score Matching using propensity scores calculated through logistic regression, random forest, and XGBoost.

4.1 Point estimate

Albeit, the same, there are multiple ways to calculate our point estimate or outcome. We are planning to look at the following:

- $E[re78^{treat=1}] - E[re78^{treat=0}]$: This is the counterfactual treatment effect for the post-treatment revenue`re78`. This is the standard approach followed by Dehejia and LaLonde.
- $E[\frac{re78}{1000}^{treat=1}] - E[\frac{re78}{1000}^{treat=0}]$: Inspired by [2],this is the same difference as above, except we divide revenues and represent them in the 1000s. This way models like linear regression should be able to converge faster.
- $E[(re78-re74)^{treat=1}]-E[(re78-re74)^{treat=0}]$: This is the counterfactual treatment effect for the difference in revenue pre and post-treatment. This tells us how much someone's revenue increase due to the treatment since participants tend to have different revenue pre-treatment and we need to account for it.
- $E[employ^{treat=1}] - E[employ^{treat=0}]$: This represents the difference between the rate of employment between the treatment groups. Rather than looking at revenue, we assume anyone with a salary is considered employed and see the proportions before and after treatment.

4.2 Interpretation

Our causal risk estimate would allow policy makers to understand the scope of job training for the disadvantaged. The original paper described this cohort as "ex-drug addicts, ex-criminal offenders, and high school dropouts". Since we're working with a male-only subset of this data, we're limiting the applicability of our subset to the male disadvantaged population of the late 70s.

The organization behind the program estimated a cost of $6,800-$9,100 per participant for training. An interesting way to think about the benefit of job training would be to see whether the expected increment in annual salary would justify the investment in the program. Albeit, the salary increment for those with higher pre-treatment salaries would be diminishing.

Our employment based estimate, although simple, can guide policy makers to understand the impact of job training on jail recidivism. An extension of this study could be used to help understand how job training could help employment and hence reduce jail recidivism (employment would then be a mediator). A similar study can also be found by Rodolfa et al. [5]

5 Synthetic data

We use the observational dataset to obtain an estimate of the population statistics and use this to sample a random ethnicity (black, hispanic or white). We sample "treat" with $p = 0.25$ and increase p by 0.25 if "nodegree" is 1 and (independently) by 0.25 for if the person has no degree.

We also assume that impact of education is limited to to "nodegree" (encoded by "at least 12 years of education" in the dataset) rather than the number of years spent in school. Because the people under consideration are already out of school, we assume for simplicity that this variable does not depend on age (i.e. people never go back to school), but only on ethnicity and sample with

the corresponding conditional probabilities $P[nodegree \mid ethnicity]$. We then sample (`age, re74, re75, re78`) from a multivariate normal distribution with means and covariances obtained from the observational data (seperate for each group based on combination of race, marriage, degree status). We exclude the datapoint if age is less than 16 and clip the revenue variables at 0. We assume that marriage is dependent only on age and is described by a saturating exponential which we fit "by hand" to the observational data.

```
def marriage_prob(x):
    return 0.9 * (1 - np.exp(-0.25 * (x - 16)))
```

Listing 1: Marriage using exponential saturation

We created the exponential data by modeling age on marriage as show in Figure 1 using observational and experimental data.

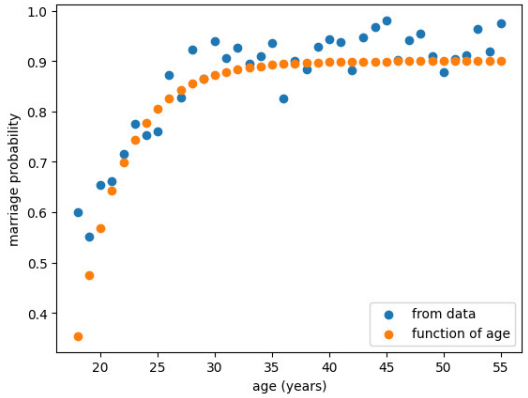


Figure 1: Probability of `marriage` against `age`

If "treat" is equal to one, we sample a treatment effect from a gaussian with mean 5000 and standard deviation 500 and add it to the re78 values of treated patients (Previously we tried to sample the ATE based on the experimental data, but because some combinations with one or two people in the experimental data had huge treatment effects, this skewed the estimate significantly and made it harder to see if the method was working correctly). In a way, we're exaggerating our ATE so it's easier to measure with our propensit-based strata method. This way, the confounders do influence treatment and the real income, but its increase upon treatment is very clear this way.

Although the real earnings in 1978 after the treatment are impacted by a number of confounders (and the likelihood of treatment depends on this as well), the average effect of setting "treat" to one is homogeneous and equal to an increase in income of 5000USD. We use the previously explained approach of computing stratified propensities and apply it to the synthetic dataset. Figure 2 shows the estimated effect as a function of the number of strata. The result is plausible and relatively close to the true value of 5000USD.

6 Next steps for the project

- Heterogeneous treatment effects for each subgroup: This would help us understand how job training affects a certain race, education, age group or those with lower starting salaries.
- Sensitivity Analysis: We want to understand how including additional covariates that represent potential unmeasured confounders impacts the propensity score. This would help assess how sensitive the results are to the inclusion of these unmeasured variables. For example, we could measure mental state or criminal history and understand its impact.

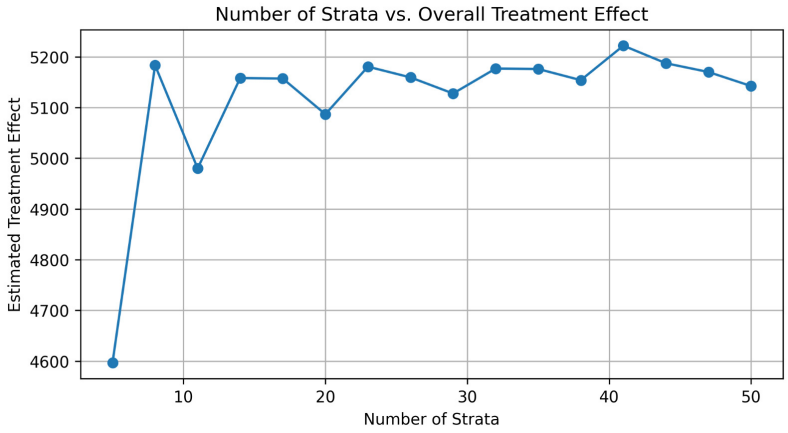


Figure 2: Number of Strata vs Average Treatment Effect

- Selection Bias: We need to adjust and understand how selection bias may impact our dataset "in the wild". Since the original study was a randomized control trial, a causal effect was readily available. However, for our observable dataset, we need to understand the impact of limiting our subset to the male only population and

7 Ask for feedback

We are unsure whether our causal risk estimate for the counterfactual was calculated correctly and we require feedback. We will discuss this over our weekly call. We also would like to understannd how HTEs are calculated through our upcoming classes.

References

[1] Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999. Accessed: 2024-05-20.

[2] Paul English. Causal inference on the lalonde dataset. `https://github.com/paul-english/causal-inference-notes/blob/master/Causal%20Inference%20on%20the%20Lalonde%20Dataset.ipynb`, 2016. Accessed: 2024-05-20.

[3] Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986. Accessed: 2024-05-20.

[4] Applied Predictive Modeling. Matching methods, 2019. Accessed: 2024-05-20.

[5] Kaila T. Rodolfa, Emily Salomon, Laura Haynes, Iliana H. Mendieta, Jeff Larson, and Rayid Ghani. Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 142–153, January 2020.