Homework Assignment 2

Econ 31380 Causal Machine Learning
Max H. Farrell

Due November 8. Submit your answers on Canvas typed using LaTeX or Markdown.

1 Multiple Testing and Heterogeneous Treatment Effects

We have i.i.d. data from a randomized experiment with a binary treatment $T \in \{0,1\}$, a continuous outcome of interest Y, a set of binary covariates $\mathbf{X} = (X_1, X_2, \ldots, X_d)' \in \{0,1\}^d$ and a set of continuous covariates $W \in \mathbb{R}^l$. Both $\mathbf{X}$ and W are pre-treatment and all our usual assumptions are met.

(a) Define the univariate conditional average treatment effects with respect to each X_j as $\tau_j(x) = \mathbb{E}[Y(1) - Y(0) \mid X_j = x]$ for $j \in \{1, \ldots, d\}$ and $x \in \{0,1\}$. Use a linear regression to propose an estimator for $\tau_j(x)$ and establish its asymptotic distribution. Provide all necessary regularity conditions. Construct your estimation so that the estimators for $\tau_j(x)$ and $\tau_k(x')$ are based on independent data if $j \neq k$ or $x \neq x'$.

(b) Construct a 5% level test of $H_0 : \tau_1(1) = 0$ versus $H_1 : \tau_1(1) \neq 0$ based on the t-statistic from the asymptotic distribution above. Call the test statistic t_{11}. Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject. Prove that your test is consistent.

(c) Construct a 5% level test of $H_0 : \tau_1(0) = 0$ versus $H_1 : \tau_1(0) \neq 0$ based on the t-statistic from the asymptotic distribution above. Call the test statistic t_{10}. Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject. Prove that your test is consistent.

(d) By your own argument from part **(a)**, the tests **(b)** and **(c)** are independent, and you just established that they are 5% level tests. Find the probability that at least one of the tests gives a false positive. What does this tell you about how often you will make mistakes?

(e) I would like to test the null hypothesis that the treatment is not effective in both groups,

$$H_0 : \tau_1(1) = 0 \text{ AND } \tau_1(0) = 0,$$

against the alternative that the treatment is effective for at least one group,

$$H_0 : \tau_1(1) \neq 0 \text{ OR } \tau_1(0) \neq 0.$$

I will reject the null if $|t_{11}| > c$ OR $|t_{10}| > c$. For what c is a 5% level test?

(f) The previous part refers to testing two restrictions. What happens to the value of c as the number of restrictions grows, but the level stays fixed at 5%? Give an expression for c (as a function of d) such that we can test the null that $\tau_j(x) = 0$ for all $j \in \{1, \ldots, d\}$ and $x \in \{0,1\}$.

Notice that we have not even begun to explore subgroup effects in earnest, because the above does not consider any interactions.

2 Propensity Score Weighting & ATT Estimation

Assume that the random variables $(Y_1, Y_0, T, \mathbf{X}')' \in \mathbb{R} \times \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ obey $\{Y_1, Y_0\} \perp\!\!\!\perp T \mid \mathbf{X}$. The researcher observes $(Y, T, \mathbf{X}')'$, where $Y = Y_1 T + Y_0(1-T)$. Define the propensity score $p(\mathbf{x}) = \mathbb{P}[T = 1 \mid \mathbf{X} = \mathbf{x}]$ and assume it is bounded inside $(0,1)$. Define $\mu_t = \mathbb{E}[Y(t) \mid T = 1]$ and $\mu_t(\mathbf{x}) = \mathbb{E}[Y(t) \mid \mathbf{X} = \mathbf{x}]$. The average treatment effect on the treated (ATT) is $\tau = \mu_1 - \mu_0$.

(a) A function $f(\mathbf{X})$ is called a *balancing score* if $\mathbf{X} \perp\!\!\!\perp T \mid f(\mathbf{X})$. Prove that the propensity score is a balancing score.

(b) Prove that $\{Y_1, Y_0\} \perp\!\!\!\perp T \mid \mathbf{X}$ implies $\{Y(1), Y(0)\} \perp\!\!\!\perp T \mid p(\mathbf{X})$.

(c) Prove that

$$\mu_1 = \mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}\left[\frac{TY}{\mathbb{E}[T]}\right]$$

and use this to propose a consistent estimator of μ_1. Notice that estimation of the μ_1 half of the ATT does not require estimation of $p(\mathbf{X})$ or $\mu_0(\mathbf{X})$. Explain why.

(d) Prove that

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\mathbb{E}\left[\frac{(1-T)p(\mathbf{X})Y}{(1-p(\mathbf{X}))}\right].$$

Assume you have access to $\hat{p}(\mathbf{x})$ that is a uniformly[1] consistent estimator of $p(\mathbf{x})$. Propose a consistent estimator of μ_0.

(e) Prove that

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\mathbb{E}\left[T\mu_0(\mathbf{X}) + \frac{(1-T)p(\mathbf{X})(Y - \mu_0(\mathbf{X}))}{(1-p(\mathbf{X}))}\right].$$

Further, prove that this moment condition is "doubly robust", meaning that it still holds even if one of $p(\mathbf{X})$ or $\mu_0(\mathbf{X})$ is not correctly specified (or cannot be estimated consistently). Replace $p(\mathbf{X})$ in the above by some other function $\tilde{p}(\mathbf{X})$ and show that the equality still holds. Do the same for $\mu_0(\mathbf{X})$.

3 Application – Pricing Experiment

We have data from a pricing experiment from an online recruiting service. The unit of observation is a customer of this service, which is a firm looking to hire (applicants use the service for free). The firms are charged a fixed price for access to the online recruiting system and its tools. Currently, the price is \$99. But they are concerned this price is too low, so they ran an experiment. Arriving customers were randomly assigned a `price` of either \$99 or \$249. We observe the decision to either `buy` the service or not and we observe the `customerSize` for each firm. The data is in the file `priceExperiment.csv`.

(a) Run a regression of the binary outcome `buy` on the `price`. Is this regression causal? What do the intercept and slope in this regression represent? Use potential outcomes.

(b) Create a dummy variable indicating the different prices. Regress `buy` on this variable. Is this regression causal? What do the intercept and slope in this regression represent? Use potential outcomes.

[1] Why does uniformity matter here? What can happen if $\hat{p}(\mathbf{x})$ is consistent in (e.g.) L_2 but not uniformly?

(c) Create a variable that measures revenue. Regress this outcome on the dummy variable you just created. Is this regression causal? What do the intercept and slope in this regression represent? Compare explicitly to the previous question. Use potential outcomes.

(d) At the 95% level, are the effects in parts **(b)** and **(c)** statistically significant? Justify your choice of standard errors.

(e) Should the firm stick with $99 or switch to $249? Justify your answer using the results from what you've done so far.

The data includes a variable `customerSize` that gives the size of the customer firm (remember, the customers of this business are themselves firms). The sizes are ranked 0, 1, 2, for small, medium, and large firms.

(f) Using a *single* regression (e.g., one `lm()` command), estimate the revenue effect for each firm size individually. That is, obtain estimates of the CATEs $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, for $x = 0, 1, 2$. Verify your answer manually using a difference in means for each group. Does this pattern make sense to you?

(g) Using these results, decide on the optimal pricing strategy to maximize revenue when the service can charge different prices to different customers based on their size. We are imagining that when a firm goes to the service, they first fill out several questions, including their firm size, and then are shown a price based on these answers. (This is known as *third-degree* price discrimination.)

(h) Can the recruiting service improve its revenue? By how much? *(Careful computing the revenue from your strategy. When using the data, think about which observations were exposed to which price, and how many of each type of firm you have.)*

(i) The law of iterated expectations says that the ATE obeys

$$\begin{aligned}\tau = \mathbb{E}[\tau(X)] &= \mathbb{E}\big[Y(1) - Y(0) \mid X = 0\big]\mathbb{P}[X = 0] \\ &\quad + \mathbb{E}\big[Y(1) - Y(0) \mid X = 1\big]\mathbb{P}[X = 1] \\ &\quad + \mathbb{E}\big[Y(1) - Y(0) \mid X = 2\big]\mathbb{P}[X = 2] \\ &= \tau(0)\mathbb{P}[X = 0] + \tau(1)\mathbb{P}[X = 1] + \tau(2)\mathbb{P}[X = 2].\end{aligned}$$

Use this and your single regression to compute plug-in estimator of the ATE:

$$\hat{\tau} = \hat{\tau}(0)\hat{\mathbb{P}}[X = 0] + \hat{\tau}(1)\hat{\mathbb{P}}[X = 1] + \hat{\tau}(2)\hat{\mathbb{P}}[X = 2].$$

Why does this value not match what you found in part **(c)**? Explain rigorously and propose a different way of aggregating data from each value of X so that you obtain exactly the value in part **(c)**.

4 Application – NSW

The National Supported Work (NSW) Demonstration was a randomized experiment done in the 1970s to study the impact of job training on earnings. We will use the data to study subgroup effects and to illustrate selection on observables.

4.1 Randomized Experiment

The data from the experiment is in the the file `nsw_RCT.csv`.We observe the following variables

- `income.after` = earnings after training, the outcome,
- `treat` = 1 if you had job training, 0 if not,
- `age`, `education` = demographics measured in years (continuous),
- `black`, `hispanic`, `married`, `hsdegree` = binary demographic variables,
- `income.before1`, `income.before2` = two years of data on earnings prior to the study.

We will use this data to study the effect of job training on average and for subgroups.

(a) Estimate the ATT using the difference in means. Is job training (on average) beneficial?

(b) Plot/display the distribution of earnings for the treatment and control groups. What does this tell you about the effect of job training? How does this inform how you would use the ATT estimate for policy making? What type of uncertainty is shown here?

(c) We wish to assess the statistical uncertainty around the ATT estimate. Do this by (i) obtaining an influence function representation for the difference in means estimate, (ii) using this result to prove that the estimator is asymptotically normal and characterize the asymptotic variance, and (iii) propose consistent standard errors. Compute a 90% confidence interval.

(d) We want to maximize welfare using targeting rules. To make the resulting policy easy to implement and transparent, it must be a threshold policy based on a single covariate, so search for rules of the form $d(\mathbf{x}) = \mathbf{1}\{x_j > c\}$ or $d(\mathbf{x}) = \mathbf{1}\{x_j < c\}$ for a specific covariate x_j and some cutoff value c. What is the welfare maximizing rule?

(e) What role does the size of the group flagged by $d(\mathbf{x})$ play in your conclusions?

(f) Explain how you would statistically test the effectiveness of your targeting rule on improving welfare.

(g) Examine the demographic characteristics of who your program targets compared to who is not targeted. What do you find and is this pattern concerning? *(A real study should compare the targeted demographics to the relevant/eligible population, e.g. the whole city.)*

4.2 Observational Data

Suppose there was no experiment. We have data from the same 185 men that received job training but we do not have access to the NSW control sample. For a comparison sample we found 2490 men from the Panel Study of Income Dynamics (PSID) that did not have training. The data is in `nsw_PSID.csv`.

(h) Before beginning the analysis, summarize the main concern when it comes to using observational data for the analysis. Why might the PSID *comparison* group not be a good *control* group?

(i) Using the PSID control sample as though it were the control group for a randomized trial, estimate the average treatment effect. Explain what you find and why you found it.

(j) Does the PSID sample appear to be a good control group for this purpose? That is, using the covariates, does the treatment appear to be randomly assigned?

(k) Using the above to guide you, build a linear regression that attempts to control for any sources of nonrandomization. Does your regression-based treatment effect estimate recover the experimental benchmark treatment effect estimate? Discuss the uncertainty of your regression-based estimate and how this relates to the experimental benchmark.