Problem Set 1

Due September 23rd, 6:30 pm before class

1. Specification searches

This is a reanalysis of the LaLonde observational data used by Hainmueller (2012). In `lalonderegression.R`, I run two linear regressions, one without covariates and the other with all covariates. The coefficients of the treatment have different signs.

In total, there are 10 covariates and therefore $2^{10} = 1024$ possible subsets of covariates. Run 1024 linear regressions with subsets of covariates, and report the regression coefficients of the treatment. How many are positively significant, how many are negatively significant, and how many are not significant? You can also report other interesting findings from these regressions.

2. More on racial discrimination

This is a reanalysis of the data based on the study of Bertrand and Mullainathan (2004). In class, I conduct an analysis based on the whole dataset, as shown in `resume.R`.

Conduct the same analysis for males and females. That is, conduct two subgroup analyses. What do you find?

3. Regression adjustment in the Fisher Randomization Test

This is a reanalysis of the LaLonde experimental data used in the lecture. In `FRTlalonde.R`, I conduct the Fisher randomization test using four test statistics. The Fisher randomization test can be more general with at least the following two additional strategies. Under the potential outcomes framework, *all potential outcomes and covariates are fixed numbers.*

First, we can use test statistics based on residuals from the linear regression. Run a linear regression of the outcomes on the covariates, and obtain the residuals (i.e., treat the residuals as the pseudo "outcomes"). Then define the four test statistics based on the residuals. Conduct the Fisher randomization test using these four new test statistics. Report the corresponding p-values.

Second, we can define the test statistic as the coefficient in the linear regression of the outcomes on the treatment and covariates. Conduct the Fisher randomization test using this test statistic. Report the corresponding p-value.

Why the five p-values from the above two strategies are exact p-values? Justify them.

4. Correlation and partial correlation

Consider a three-dimensional Normal random vector:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix} \right).$$

The correlation coefficient between X and Y is ρ_{XY}. The partial correlation coefficient between X and Y given Z is their correlation coefficient in the conditional distribution $(X, Y) \mid Z$, denoted by $\rho_{XY|Z}$.

Express $\rho_{XY|Z}$ using $(\rho_{XY}, \rho_{XZ}, \rho_{YZ})$. Then give an example with $\rho_{XY} > 0$ and $\rho_{XY|Z} < 0$.

5. Nonlinear causal estimands

With potential outcomes $\{(Y_i(1), Y_i(0)\}_{i=1}^n$ for n units under the treatment and control, we can define median treatment effect as

$$\delta_1 = \text{median}\{(Y_i(1)\}_{i=1}^n - \text{median}\{(Y_i(0)\}_{i=1}^n,$$

which is, in general, different from the median of the individual treatment effect

$$\delta_2 = \text{median}\{(Y_i(1) - Y_i(0)\}_{i=1}^n.$$

Given numerical examples which have $\delta_1 = \delta_2$, $\delta_1 > \delta_2$, and $\delta_1 < \delta_2$.

Which estimand makes more sense, δ_1 or δ_2? Why? Use examples to justify your conclusion.

6. A better bound of the variance formula

In class, I showed

$$\operatorname{var}(\widehat{\tau}) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n} \leq \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}.$$

Show the following bound

$$\operatorname{var}(\widehat{\tau}) \leq \frac{1}{n}\left(\sqrt{\frac{n_0}{n_1}}S_1 + \sqrt{\frac{n_1}{n_0}}S_0\right)^2.$$

When will the equality hold?

7. Vector version of Neyman (1923) – only for Stat 260 students

The classic result of Neyman (1923) is about a scalar outcome. It is common to have multiple outcomes in practice. Therefore, we can extend the potential outcomes to vectors. We consider the average causal effect on a vector outcome $\boldsymbol{V} \in \mathbb{R}^K$,

$$\tau_{\boldsymbol{V}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{V}_i(1) - \boldsymbol{V}_i(0)\right\},$$

where $\boldsymbol{V}_i(1)$ and $\boldsymbol{V}_i(0)$ are the potential outcomes of $\boldsymbol{V}$ for unit i. The Neyman-type estimator for $\tau_{\boldsymbol{V}}$ is the difference between the sample mean vectors of the observed outcomes under treatment and control:

$$\widehat{\tau}_{\boldsymbol{V}} = \bar{\boldsymbol{V}}_1 - \bar{\boldsymbol{V}}_0 = \frac{1}{n_1}\sum_{i=1}^{n} Z_i\boldsymbol{V}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)\boldsymbol{V}_i.$$

Consider a completely randomized experiment. Show that $\widehat{\tau}_{\boldsymbol{V}}$ is unbiased for $\tau_{\boldsymbol{V}}$. Find the covariance matrix of $\widehat{\tau}_{\boldsymbol{V}}$. Find a (possibly conservative) estimator for the variance.

REFERENCES

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles (with discussion). section 9 (translated). reprinted ed. *Statistical Science*, 5:465–472.