SDS 384: Causal Inference Methodology
Homework 1*

February 18, 2020

Professor Zigler

Spencer Woody

*Code available at `github.com/spencerwoody/sds384causal`

Work with Preston Biro, Jingjing Fan, and Qiahui Lin.

Problem 1

The canvas site provides an excerpts from the published HEI Research Report 148, Impact of Improved Air Quality During the 1996 Summer Olympic Games in Atlanta on Multiple Cardiovascular and Respiratory Outcomes *by Peel et al.. The first four pages are excerpts from an overall summary and critique of the report, which provide general background. The next five pages are excerpts from the more detailed research report. In both cases, the text appearing in red boxes should be sufficient to answer the following questions.*

(a) *For both the analysis of ozone concentrations and the analysis of ED visits, what are the potential outcomes defining the effects of interest?*

This study analyzes the impact of a short-term, temporary intervention designed to reduce car traffic in Atlanta during the 1996 Summer Olympic Games on (i) daily ozone concentrations in the city measured at a specific site, and (ii) the number of visits to emergency departments related to cardiovascular and respiratory cases for specified cohorts of interest aggregated across twelve hospitals in the city.

Let Y_i denote the ozone concentration in Atlanta on a single day i, Z_i denote the number of ED visits on day i, and $W_i \in \{0, 1\}$ be an indicator for whether the intervention is applied to day i. Then the causal effects of interest compare the two sets of potential outcomes:

$$(Y_i(W_i = 0), Y_i(W_i = 1)) \quad \text{and} \quad (Z_i(W_i = 0), Z_i(W_i = 1)).$$

That is, the causal effects concern the differences between ozone concentration when the intervention is applied versus when it is not applied, and likewise for ED visits.

(b) *For both the analysis of ozone concentrations and the analysis of ED visits, provide a possible violation of the "No Multiple Versions of Treatment" (or "consistency") part of SUTVA.*

There are several possible violations of the consistency portion of the stable unit treatment value assumption (SUTVA), mainly stemming from the city's adherence to the intervention. For instance, one provision of the intervention was to encourage businesses to provide telecommuting work options and alternative work hours for employees, and to advocate the use of vacation time. However, the degree to which businesses uphold these policies could vary throughout the length of the Olympic Games. There could be some days when many businesses do not allow for telecommuting, for instance due to need to produce quarterly results. This would result in multiple versions of the treatment, in that the extent of the treatment would change.

(c) *For both the analysis of ozone concentrations and the analysis of ED visits, describe the assignment mechanism and provide one reason why it may not be* unconfounded.

The choice of when to deploy the intervention was deterministically chosen by the timing of the 1996 Olympics. Any event which affects the measured outcomes and takes place during the Olympics could confound the treatment assignment. For instance, there is likely an increase in plane traffic arriving at the Atlanta airport carrying spectators and athletes, and these planes could also increase the level of

ozone in the air. Also, there are expected to be many more people in the city for the Olympics compared to normal, and this would increase the amount of car traffic.

Problem 2

Describe an example of an assignment mechanism that is ignorable but not unconfounded and support your argument with statement(s) about the assignment mechanism.

The ignorability assumption (Rubin, 1978) may be expressed as

$$\mathbf{W} \perp\!\!\!\perp \mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{X} \tag{1}$$

The assumption of unconfoundedness (Rosenbaum and Rubin, 1983) can be expressed as:

$$\mathbf{W} \perp\!\!\!\perp (\mathbf{Y}(0), \mathbf{Y}(1)) \mid \mathbf{X}. \tag{2}$$

We may create an assignment mechanism that is ignorable but not unconfounded using a sequential treatment assignment that conditions on observed outcomes. For example, consider sampling treatment by taking draws of balls from an urn with replacement, and we put more balls in the urn depending on the outcome of Y_i^{obs}. To give a specific example, first define $N_s^{[0]} = N_d^{[0]} = 0$. The outcome Y is either 1, for survival, or 0, denoting death. Let treatment assignments for units $i = 1, \ldots, N$ be determined by iterating through the following steps:

(i) Calculate probability to treatment for unit i

$$p^{[i]} = \frac{1 + N_s^{[i-1]}}{2 + N_s^{[i-1]} + N_d^{[i-1]}}.$$

(ii) Assign treatment via

$$W_i \sim \text{Bernoulli}(p^{[i]}).$$

(iii) Observe $Y_i^{\text{obs}} = Y_i(W_i) \in \{0, 1\}$

(iv) If unit i is assigned to treatment, then update the weights for assignment to treatment and control depending on the outcome of $Y_i(W_i = 1)$. If unit i is assigned treatment and dies, give more weight to assignment to control for the next unit; if unit i is assigned treatment and survives, give more weight to treatment for the next unit. Specifically, the weights are updated by

$$N_s^{[i]} = \begin{cases} N_s^{[i-1]} + 1 & \text{if } W_i = 1 \text{ and } Y_i = 1 \\ N_s^{[i-1]} & \text{otherwise} \end{cases} \quad \text{and} \quad N_d^{[i]} = \begin{cases} N_d^{[i-1]} + 1 & \text{if } W_i = 1 \text{ and } Y_i = 0 \\ N_d^{[i-1]} & \text{otherwise.} \end{cases}$$

(v) If $i < N$, return to step (i).

The treatment assignment vector $\mathbf{W}$ therefore depends on the observed values in $\mathbf{Y}^{\text{obs}}$, but it is still independent of the missing outcomes $\mathbf{Y}^{\text{mis}}$; therefore condition (1) holds, but condition (2) does not.

Statistics	All (614)	No High School Degree (387)	High School Degree (227)
T^{rank}	0.282	0.420	0.791
$T^{\text{rank-gain}}$	< 0.000	0.007	0.005
T^{dif}	0.336	0.499	0.888

Table 1: Fisher exact tests for the sharp null hypothesis $H_0 : Y_i(1) - Y_i(0) = 0$ for the Lalonde dataset.

Problem 3

For this exercise, you will follow Chapter 11 of the Imbens and Rubin textbook, but use a different data set to implement some of the techniques we have learned about in class. While the textbook describes analysis of a social program called the Saturation Work Initiative Model (SWIM) program, in this exercise you will analyze the lalonde data available in the `R` *package* `MatchIt` *with the command* `data(lalonde)`*. Use the* `help(lalonde)` *command to get a basic description of the data. Use the variable called* `treat` *to denote the randomized treatment assignment, and use* `re78` *as the outcome of interest.*

(a) *Produce a table analogous to Table 11.2 of the textbook. Use the same test statistics as in Section 11.3, and conduct the tests for the entire data set (corresponding to the All column in Table 11.2) and stratified by the* `nodegree` *variable (corresponding to the two rightmost columns of Table 11.2).*

See Table 1 for p-values corresponding to these tests. The null distributions and observed test statistics are shown in Figure 1.

(b) *Unlike in Section 11.3 of the textbook, provide a Fisher interval for the analysis of the entire data set (i.e., not stratified by* `nodegree`*).*

Note that we can determine confidence intervals from inverting a series of hypothesis tests $H_0^C : Y_i(1) - Y_i(0) = C$. That is, the confidence interval is the set of values of C for which we fail to reject H_0^C. Here are three (approximate) 95% confidence intervals resulting from the tests given in Part (a).

$$\begin{aligned} T^{\text{Rank}} &: \ (-1050, 0) \\ T^{\text{Rank-Gain}} &: \ (1000, 3250) \\ T^{\text{dif}} &: \ (-1877.55, 693.88) \end{aligned}$$

The interval coming from the rank test includes zero because the presence of many 0's in Y^{obs} leads to the test being very sensitive to small perturbations of C.

(c) *Produce a table analogous to Table 11.3 of the textbook, but with estimates for the same three categories as in part (a) (All and both levels of* `nodegree`*).*

See Table 2.

(d) *Fit a regression model with interactions as in Section 11.5 of the textbook, using the following variables as covariates:* `age`*,* `educ`*,* `black`*,* `hispan`*,* `married`*,* `nodegree`*. Report the*

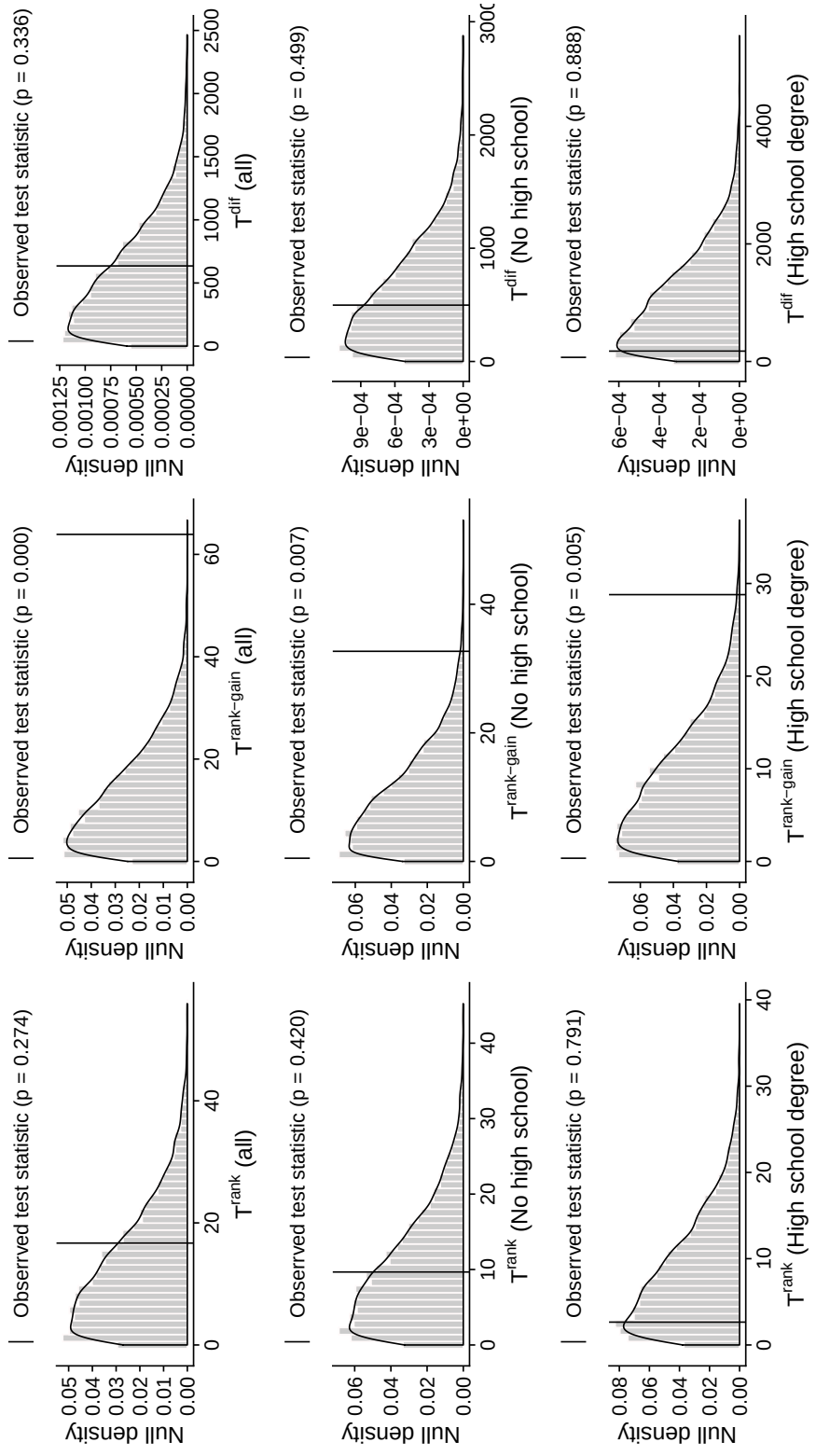

Figure 1: Null distributions, observed test statistics, and *p*-values corresponding to comparisons made in Table 1.

	All (614)	No High School Degree (387)	High School Degree (227)
Est	–635.03	–497.615	–176.36
$\widehat{\text{(s.e.)}}$	677.20	774.94	1317.78

Table 2: Estimates for average treatment effects on earnings from the Lalonde dataset based on Neyman's repeated sampling approach.

Covariates	Est	$\widehat{\text{s.e.}}$	Est	$\widehat{\text{s.e.}}$
`intercept`	6984.17	360.71	6404.82	391.99
`treatment`	–635.03	657.14	863.07	1040.78
`age`			48.26	36.44
`educ`			526.20	188.29
`black`			–1672.81	923.90
`hispan`			723.87	1059.16
`married`			2369.01	786.56
`nodegree`			213.62	1064.92
Interactions with treatment				
`treat` × `age`			34.04	86.79
`treat` × `educ`			102.78	411.75
`treat` × `black`			598.43	2049.87
`treat` × `hispan`			–263.67	3020.45
`treat` × `married`			–1028.01	1621.14
`treat` × `nodegree`			–591.56	1947.89

Table 3: Regression analyses for the Lalonde dataset, with and without covariates and treatment interactions.

population average treatment effect from this model and compare it with the effect estimates from a regression model that adjusts only for the treatment indicator.

See Table 3. Note that the regression without covariates returns a negative point estimate for the treatment effect, while the regression with covariates returns a positive point estimate for the treatment effect. However, neither estimate is statistically significant at the $\alpha = 0.1$ level.

(e) *The second paragraph of Section 11.6 describes an analysis assuming a normal distribution for the two potential outcomes with the correlation between $Y_i(0)$ and $Y_i(1)$ fixed to be 1.0 and unknown mean and variance parameters. Perform the described model-based inference for the model with no covariates using the priors specified in the text and report your estimate of the treatment effect.*

The joint distribution of the potential outcomes is given by

$$\begin{bmatrix} Y_i(0) \\ Y_i(1) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_c \\ \mu_t \end{bmatrix}, \begin{bmatrix} \sigma_c^2 & \rho\sigma_c\sigma_t \\ \rho\sigma_c\sigma_t & \sigma_t^2 \end{bmatrix} \right)$$

The correlation between the potential outcomes, unidentifiable from the data, is fixed to be $\rho = 1$. The

unknown parameters to estimate are μ_c, μ_t, σ_c^2, and σ_t^2. We also need to impute the missing potential outcomes to estimate the treatment effect. We assign the independent priors

$$\mu_c \sim \mathcal{N}(0, 100^2), \quad \mu_t \sim \mathcal{N}(0, 100^2)$$
$$1/\sigma_c^2 \sim \mathcal{G}(1/2, 0.005), \quad 1/\sigma_t^2 \sim \mathcal{G}(1/2, 0.005)$$

where $\mathcal{G}(a, b)$ denotes the gamma distribution with shape parameter a and rate parameter b.

The full conditionals for μ_c, μ_t, σ_c^2, and σ_t^2 are easy enough to find due normal-inverse gamma conjugacy. Note that the Y_i^{obs} are observed independently. The full conditionals for the Y_i^{mis} for $W_i = 1$ can be found via

$$(Y_i^{\text{mis}} = Y_i(0) \mid Y_i^{\text{obs}} = Y_i(1), \ldots) \sim \mathcal{N}(\mu_c', \nu_c')$$
$$\mu_c' = \mu_c + \rho\sigma_c\sigma_t/\sigma_t^2 \cdot (Y_i^{\text{obs}} - \mu_t), \quad \nu_c' = (1 - \rho^2)\sigma_c^2$$

and for $W_i = 0$,

$$(Y_i^{\text{mis}} = Y_i(1) \mid Y_i^{\text{obs}} = Y_i(0), \ldots) \sim \mathcal{N}(\mu_t', \nu_t')$$
$$\mu_t' = \mu_t + \rho\sigma_c\sigma_t/\sigma_c^2 \cdot (Y_i^{\text{obs}} - \mu_c), \quad \nu_t' = (1 - \rho^2)\sigma_t^2$$

This sampling scheme can be easily vectorized by

$$(Y_i^{\text{mis}} \mid Y^{\text{obs}}, W_i, \ldots) \sim \mathcal{N}(W_i \cdot \mu_c' + (1 - W_i) \cdot \mu_t', W_i \cdot \nu_c' + (1 - W_i) \cdot \nu_t').$$

Also, note that the conditional variance of the missing potential outcome given the observed potential outcome is 0 because $\rho = 1$.

Figure 2 shows the posterior distribution for the finite sample average treatment effect τ, which has a posterior mean of -77.3 and a 95% posterior credible interval $(-876.0, 767.7)$

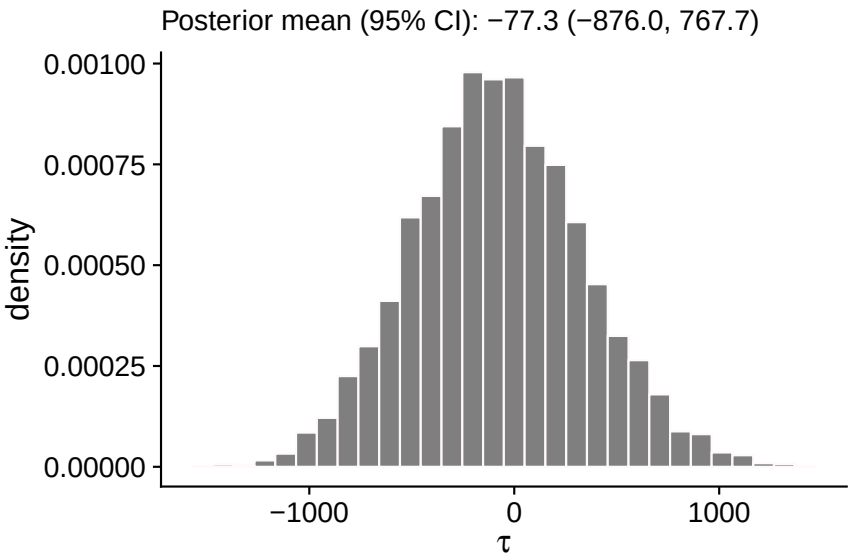

Figure 2: Posterior distribution for average treatment effect τ

References

Jennifer L Peel, Mitchell Klein, W Dana Flanders, James A Mulholland, Paige E Tolbert, HEI Health Review Committee, et al. *Impact of improved air quality during the 1996 Summer Olympic Games in Atlanta on multiple cardiovascular and respiratory outcomes*. Health Effects Institute Boston, MA, 2010.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, 6(1): 34–58, 01 1978. doi: 10.1214/aos/1176344064. URL https://doi.org/10.1214/aos/1176344064.