POLC 4325 Assignment #3

Posted: October 22, 2024
Due: November 5, 2024 (11:59pm)

1 Regression Discontinuity Design (36%)

(Each has 6%.)

Revisit Grumbach, J.M. and Sahn, A., 2020. Race and representation in campaign finance. *American Political Science Review*, 114(1), pp.206-221.

1. What is the causal estimand you can identify with regression discontinuity in this study? Relatedly, what is the population of interest under regression discontinutiy in this study?

2. Explain the regression discontinuity design. Use the words **treatment assignment**, **deterministic**, **running variable**, and **cutoff**.

3. Discuss the plausibility of the local randomization (or "as-if random") assumption discussed in this study. What is the assumption? Do you think this is justifiable in the given study? If so (if not), why?

4. Now consider the identification based on the continuity assumption. Suppose that the average potential outcome under the control can be represented by $\mathbb{E}[Y_i(0)|X_i, D_i = 0] = 1000 + 5X_i$, where X_i is the running variable and $D_i = 0$ (the treatment indicator takes 0) and thus disappears. Suppose that the causal estimand is 5. What is the average potential outcome under the treatment $\mathbb{E}[Y_i(1)|X_i, D_i = 1]$?

5. What is $\mathbb{E}[Y_i(1)|X_i = c, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = c, D_i = 0]$ (c is the cutoff point)? Provide a single value.

6. Suppose that you have three variables `y`, `x`, and `d` for the outcome, running variable, and treatment indicator in R. How would you estimate the treatment effect (i.e., write down a syntax)? Discuss why the syntax looks that way (Hint: think of the RD estimation as a difference-in-means with a single control variable).

2 Difference-in-Differences (36%)

(Each has 6%.)

Revisit Grumbach, J.M. and Sahn, A., 2020. Race and representation in campaign finance. *American Political Science Review*, 114(1), pp.206-221. Specifically, revisit the following discussion:

> "We also see in Figure 4 that candidate ethnorace is correlated with overall fundraising from individual donors. Asian American candidates receive the most, followed by Latino, white, and black candidates. However, we urge caution in drawing causal conclusions from this figure because the relationship is confounded by time, geography, and, more specula tively, fundraising from nonindividual sources. Campaigns grew more expensive in recent decades as increasing numbers of Asian and Latino candidates ran for office. Campaigns in the US South, which have greater numbers of black candidates, tended to be less expensive than those in other regions during this period" (214)

1. What is the causal estimand you can identify with the difference-in-differences design.

2. The essence of difference-in-differences is to include unit-specific and time-specific fixed effects. Suppose that you have four variables `y`, `x`, `district`, `year`. In R and using `lm()`, how would you implement the difference-in-differences model (hint: what are fixed effects?)?

3. Provide one new example of time-varying confounder that affects both the candidate race and candidate contributions. Note that the effect of such time-varying confounder must be identical to all districts.

4. Provide one new example of unit-varying confounder. Note that such unit-varying confounder must vary across districts, but remain the same over time.

5. Discuss the plausibility of the parallel trend (no unobserved confounder) assumption in Grumbach and Sahn (2020). Do you think if their assumption is justifiable? If so (if not), why?

6. Explain why our identification breaks in difference-in-differences when there exists an unobserved confounder that varies across years and districts?

3 Missing Data (28%, + bonus points 15%)

(Each has 7%)

1. Discuss the disadvantages of listwise deletion when addressing missing data. Mention under what assumption listwise deletion can be justified, while still being suboptimal.

2. Explain partial identification for missing data to those who have little to no statistical background.

3. Simulate a dataset of 500 observations with missing values based on the Missing at Random (MAR) assumption as follows. Here, assume that $\mathbb{E}[Y_i]$ is the quantity of interest.

   (a) First, consider a simple model $Y_i = \alpha + \beta X_i + \epsilon_i$, where $\epsilon_i \sim N(0, 5)$. Generate 500 outcomes by using where $\alpha = -10$, $\beta = 5$, $X_i \sim N(3, 2)$. When simulating the data, set random seed to (123) for reproducibility.

   (b) Now, create missing data by generating the following missing value indicator $M_i = \mathbf{1}(|X_i| > 4)$, where $\mathbf{1}(\cdot)$ is the indicator function that takes 1 if the condition inside holds and 0 otherwise. Create a new variable "y.star" that is missing if $M_i = 1$ and non-missing (i.e., same as "y") otherwise. Suppose that $Y_i \in [-35, 35]$ or $-35 \leq Y_i \leq 35$.

   (c) Create two histograms of `y` and `y.star`, respectively. Here, fix the range of x-axis to (-35, 35). Compute the estimated means $\widehat{\mathbb{E}}[Y_i]$ and add them to the histograms. Discuss whether the means look different and why. Note that if we only use `y.star` as is, we are making the MCAR assumption.

4. Estimate $\mathbb{E}[Y_i]$ via partial identification. (Hint: provide a range of all possible values of $\mathbb{E}[Y_i]$). Discuss how informative the bound is.

5. Estimate $\mathbb{E}[Y_i]$ via (simplified and illustrative) multiple imputation. Compare the result to the estimated mean based on complete data (e.g., are they close?).

   (**bonus points**, 15%, each part = 3%).

   (a) Hint 1: Use the non-missing data to run linear regression using `y.star` and `x`.

   (b) Hint 2: Draw α and β from normal distributions with their estimated means and variances

   (c) Hint 3: Compute $\mu_i = \alpha + \beta X_i$ (create a new variable called `mu.imp1`). Draw Y_i^{imp} from $N(\mu_i, \sigma = 5)$ and **impute** missing data with them (create a new variable called `y.imp1` that takes the imputed values if originally missing and `y` otherwise. Do not impute or overwrite non-missing values!).

(d) Hint 4: Estimate the expected value using `y.imp1`. Save the estimated mean $\widehat{\mathbb{E}}[Y_i^{\text{imp1}}]$.

(e) Hint 5: Repeat the above steps five times. Report the five estimates. Take the arithmetic mean of the five estimates $\widehat{\mathbb{E}}[Y_i] = \frac{1}{5}(\widehat{\mathbb{E}}[Y_i^{\text{imp1}}] + \widehat{\mathbb{E}}[Y_i^{\text{imp2}}] + \widehat{\mathbb{E}}[Y_i^{\text{imp3}}] + \widehat{\mathbb{E}}[Y_i^{\text{imp4}}] + \widehat{\mathbb{E}}[Y_i^{\text{imp5}}])$