17.835: Problem Set 3

Professor: In Song Kim
TA: Jesse Clark, Nicole Wilson, Sean Liyu

Due: Wednesday, October 7th 11:00 AM EST
(Late submissions will not be accepted)

Please post all questions to the Piazza discussion forum, which you can access through the link on the class website or at:

https://piazza.com/mit/fall2020/17835

For this and future problem sets, you will need to submit two files onto the Canvas (under `Assignment`): 1) your write-up with your answers to all questions (including codes and results for computational questions and any related graphics) as a PDF file, and 2) your code as an `R` file (so that we can verify it runs without errors). Both files should be identified with your last name (e.g., `clark.R` and `clark.pdf`). Please ensure that all of these are completed *before* class on the day of the due date – late submissions will be automatically flagged to us by Canvas. For problems that require calculations, please show all the steps of your work. For problems involving `R`, please ensure that your code is thoroughly commented

This week's problems focus on different forms of inference. As we progress through the problems, you will learn about using OLS for estimating causal effects in experiments and about regression discontinuity designs, one type of regression methods that allow us to draw causal inferences when we do not have experimental data. For all the plots in this problem set, you may use the base R graphics or the `ggplot2()` library in R (more information on `ggplot` can be found here `http://ggplot2.tidyverse.org/`)

Problem 1: Retrospective Voting and Recency Bias

In a previous lecture we learned that under certain assumptions, the OLS coefficient $\hat{\beta}$ can be interpreted as a causal effect.[1] However, the treatment effects on different groups of people may vary. For example, a flexible early voting law may be more likely to encourage young voters who attended college outside their states to vote than senior citizens. Much experimental research in political science explores the treatment effects across different groups of subjects. One notable case is the research on people's retrospective voting behavior.

Retrospective voting refers to the case in which citizens make their voting decisions based on how well the incumbent politicians performed. For instance, if the economy is doing well, the incumbent president is likely to be re-elected. However, political scientists, campaign strategists, and commentators are debating about whether ordinary citizens are competent in evaluating the performance of incumbent politicians. First, do average voters have recency bias? That is, do they make voting decisions based on election-year economy instead of the overall economic performance during incumbent's tenure? Second, are voters' perceptions susceptible to political manipulations? To address these two questions, Huber, Hill, and Lenz (2012) designed a survey experiment and the problem below is based on their experimental research.[2]

[1] See the OLS lecture handout for specific assumptions needed to interpret OLS coefficient as a causal effect.

[2] Huber, Hill, and Lenz's article can be found here: `https://doi.org/10.1017/S0003055412000391`. You don't have to read the paper to complete this problem, however.

To emulate the incumbent politicians, researchers set up two types of allocators that make regular payments to respondents via Amazon's MTurk platform. Type-A adopts a flat rate, paying respondents \$2 a day for 60 days. Type-B paid respondents \$1 a day for the first 30 days and \$3 a day for the remaining 30 days. The total amount of money each respondent received would be exactly the same, regardless of which type of allocator they had. After 60 days, respodents were asked whether they would keep (re-elect) the current allocator (incumbent) or switch to a new one. Researchers also told all respondents that if they chose to keep, the re-elected incumbent would repeat the same payment pattern (e.g., the re-elected Type-B incumbent would still pay \$1 in the first period and \$3 in the second period), and that if they chose to switch, a new allocator would be chosen in a totally random fashion and therefore there is no guarantee that the new allocator would perform better/worse than the previous one.

To simulate the political manipulation, respondents were randomly chosen to receive a message immediately prior to the decision to keep/switch the allocator: "Looking back over the past 60 days, how satisfied were you with your allocator?" with five closed-end responses ranging from "very satisfied" to "very unsatisfied." The message emulated the political rhetoric that might influence voters' assessment of the incumbent politicians in the campaign season. For instance, Ronald Reagan famously asked, "Are you better off now than you were four years ago?" when he was campaigning against the incumbent President Jimmy Carter in the 1980 election.

To summarize, respondents were randomly assigned to four groups: (1) Group one received the Type-A (flat rate) allocator and did not receive the prime/message. (2) Group two got the Type-B (high payment in the late period) without receiving any prime. (3) Group three received the Type-A allocator and the campaign message. (4) Group four got the Type-B and the campaign prime. The randomization was properly executed and the assignment of allocators was independent from the assignment of the campaign prime. This design allows researchers to better understand "recent" performance becuase it allows them to directly manipulate the time in which the "politician"– here played by the allocator– performs. Conversely, if we were to use a different type of estimation, such as a naive difference estimator that we previously learned about, we would be open to confounding due to the fact that politicians may select when to perform based on their electoral chances.

We are interested in two questions: (1) Whether respondents who received Type-B allocators would be more likely to keep the allocator than those who received the Type-A. (2) Whether showing respondents campaign messages would affect their decision. Let's analyze the data from Huber et al. (2012). You can download and import the dataset `retro_vote.csv` from the course website. The dataset contains the following variables:

- `user_id`: Researchers assigned a unique ID number to each respondent
- `kept`: Whether the respondent decided to keep (re-elect) the allocator after 60 days of payment. TRUE = keep the allocator; FALSE = switch to another one
- `alloc_type`: The type of the allocator assigned to the respondent. "Flat" = Type-A (flat rate); "Late_High" = Type-B
- `message`: Whether the respondent got the campaign message. 1 = got message, 0 = did not get message.

1. First we prepare the data for analysis. Load the data and convert `kept`, `alloc_type`, and `message` to dichotomous variables. For example, for `kept`, you may want to recode TRUE as 1 and FALSE as 0.

2. Create a 2×2 table that presents the retention rate (i.e., the percentage of "keep") in each of the four groups. The row names and column names of the table should be properly labeled and a title should be added to the table. We will refer to the table in the later questions. The table should look like the example below:

	No Message	Message
Flat	.xxxx	.xxxx
Late High	.xxxx	.xxxx

3. We can also estimate the uncertainty of the retention rate in each group. Compute the standard errors of the retention rate estimates and construct the 95% confidence intervals for each group. Below is the formula for computing the standard error for retention rates:

$$SE_{retention} = \sqrt{\frac{\mathbb{V}[Y_i]}{n}}$$

Where $\mathbb{V}[Y_i]$ stands for the sample variance for `kept` and n represents the number of observations in each of four groups.

4. After obtaining the confidence intervals, we can visualize the results obtained in 1.2 and 1.3. Create a scatter plot and use points to represent the retention rates. The X-axis should indicate the name of each group defined by the types of treatments they received. The Y-axis should represent the sizes of the point estimates, i.e., the retention rates. Be sure to add a title to the plot and properly label the X- and Y-axes. Once you created the plot, you may notice that the confidence intervals of the retention rates in groups that receive the flat-rate allocator (both with and without receiving the messages) are fairly wide. Provide one reason for the wide estimated confidence intervals.

5. Use difference-in-means method to compute the average treatment effect (ATE) of receiving high payment in the second 30 days on `kept` among all respondents. Then, use OLS regression to estimate the ATE. That is, regressing `kept` on the type of allocator. Report the difference-in-means and $\hat{\beta}_{alloc\_type}$. What do you notice? Based on the results, what can you conclude about the recency bias of ordinary voters? Based on your substantive knowledge, what are the implications for this year's general election and the politics of the US in general?

6. Do the same exercise as we did in 1.5 to estimate the ATE for receiving a campaign message among all respondents. We want to compute the mean difference manually, and with OLS. You may notice that the ATE for the campaign priming was smaller than the ATE for receiving different types of allocators. Briefly explain why this is the case (*Hint*: You may want to take look at the 2×2 table you created in 1.2).

7. Compute the confidence intervals for $\hat{\beta}_{alloc\_type}$ and $\hat{\beta}_{message}$, i.e., the OLS coefficients obtained in 1.5 and 1.6. Visualize the point estimates and confidence intervals (using the OLS SE). Set `ylim` to (-0.3, 0.3). The X-axis should indicate the name of the treatment and the Y-axis should reflect the size of the average treatment effect. Add a dashed horizon line indicating 0. You may notice that the confidence interval of $\hat{\beta}_{message}$ tounched the horizontal line. What does this imply?

8. Now we are more interested in whether the effects of political rhetoric were different on people who received different types of allocators. This is often called the *conditional average treatment effects* or *heterogeneous treatment effects*. Use the difference-in-means method to compute the conditional treatment effects of political messages among those who received the Type-A allocators as well as those who received the Type-B allocators. Briefly discuss what you find. Also, we could modify the plot in 1.3 to reflect the within group difference. Plot the retention rate of each group and their confidence intervals. However, this time, the X-axis should only have the names of two groups: those who got the message and those who didn't. Then add two lines that connect the groups that got the message with the groups that didn't. Use different colors to represent different types of allocators each group received. Finally, add a legend that describes different types of allocators received. You may notice that the slopes of two lines are different. What do the slopes stand for? Your plot should look similar to the one below:

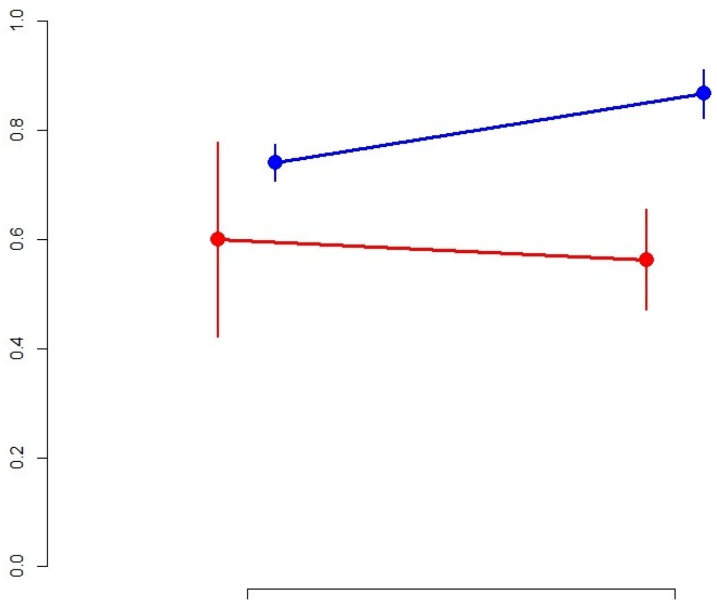

Figure 1:

9. ***Extra Credit***: You may also notice that the conditional average treatment effects are not equal to the OLS coefficients in previous questions. An OLS model with interaction term may unveil the conditional treatment effects of political rhetoric. Propose such a model and express the conditional average treatment effects in terms of OLS coefficients. An interaction term is gernally done by multiplying one term by another (using the " * " symbol) in the lm() command.

Problem 2: Incumbency Advantage or Disadvantage?

In this question, we will study the consequences of one election on future election outcomes. Specifically, we will examine whether there exists a causal effect of "being the current incumbent in a district" on the vote shares obtained in the following district election. For example, if a politician wins a seat in one election, is this politician therefore (i.e., causally) more likely to win the seat in the next election? Political scientists define this phenomenon as "incumbency advantage."

While the existence of an incumbency advantage among candidates for elected offices in the US is largely established, some authors have questioned whether this might be true in other parts of the world. In fact, some scholars have argued that outside the US and specially in the context of developing countries the opposite is true. That is, current incumbents are less likely to be elected and therefore there is an "incumbency disadvantage."

Brazil is one of the countries in which scholars argue that electoral results show the existence of "incumbency disadvantage." They claim that this phenomenon affects mayors in particular. Since the Brazilian party system is weak, mayors have uncertain career prospects which might lead them to engage in corrupt practices to "maximize" their gains while they are in office. Relatedly, in contexts where institutions whose role is to control politicians are weak, it is easier for mayors to divert public resources for their personal gains.[3]

1. Suppose that you compare the mean vote share of the incumbent party and that of non-incumbent party in the next election, and you find that the average vote share of the incumbent party is higher that that of non-incumbent party. Discuss why this finding may not be convincing evidence for the presence of incumbency advantage. Provide a real world example to support your argument.

2. **Regression discontinuity designs** can be used to estimte effects on "close elections." Specifically, we will focus on districts where the incumbent mayor won the previous election only by a close margin. The main idea behind this design is that the incumbency is determined almost as random in a close election. We will explore this question using data from the 1996 and 2000 mayoral elections in Brazil. In most municipalities, mayors are elected with a simple majority (one vote more than the second is enough to win). Unlike the US, there is generally more than one competitive party in each district. For this reason, instead of using 0.5 as our threshold, we will use relative vote margins (i.e., for all the loosing candidates, their vote margin will be their vote percentage minus the percentage of the winning candidate, while for the winning candidate it will be their vote percentage minus the percentage of the runner-up). In other words, our RDD cutoff rule will be zero, since candidates need to have a margin above zero to win. *Hint*: The RDD cutoff is where the forcing variables causes a "jump." For example, 50% of the vote is required to win in a two-way race; this would be our cutoff if that is what we are interested in. It is assumed that is is random if someone received 49.999% of the vote vs. if they received 50.00001% of the vote; this approximates an experiment.

   Below you can find the list of variables available on the dataset `brazil_mayor.csv` that we will use in this problem.

   - `state`: State code
   - `incumbent`: Whether the candidate was in office during the election, the treatment indicator (0, 1)
   - `vote_margin_pct`: Vote margin with respect to winner/runner-up in 1996 (in percentage points)
   - `margin_2000`: Vote margin with respect to winner/runner-up in 2000 (in percentage points)
   - `vote_share_2000`: Vote percentage received in 2000

[3] For a good discussion of "incumbency disadvantage" in developing countries you can consult this paper by Marko Klasnja and Rocio Titiunik (who has also made significant contributions to the development of RDD methods): `https://doi-org.libproxy.mit.edu/10.1017/S0003055416000575`.

- `votes`: Total votes received in 2000
- `cand_2000`: Whether the candidate who ran in 1996 also ran in 2000 (0, 1)
- `win_2000`: Whether the candidate won the 2000 election (0, 1)
- `party`: Candidate's party
- `cand_code`: Unique candidate identifier
- `municode`: Name of the municipality
- `codes`: Unique state-municipality name

(a) Since often incumbents do not run again (either because they have no more terms allowed, or because they run for higher offices, or for other reasons), you will first need to load your data and remove all the cases of candidates who did not run in the 2000.

(b) Trim your data so that we can only focus on districts where candidates won or lost the previous election by a close margin. Specifically, trim your data such that the final data will only include the observations in which a candidate won or lost the election by 2 percentage points. Is this is a reasonable threshold? Why or why not?

(c) Divide your data into two parts, such that one of them includes only the treatment group and the other only the control group.

(d) Run the regression specification given below for both subsets and store your results (i.e., intercepts and slopes) for the treatment group as `mod.tr` and for the control group as `mod.cl`. In other words, you will run two regressions, one for each group separately. Report the summary from these models.

$$\texttt{margin\_2000} = \alpha + \beta \cdot \texttt{vote\_margin\_pct} + \epsilon$$

3. Now we will graphically examine the presence of incumbency advantage. Please draw a plot with the following components.

(a) Draw a scatterplot of all data, where you will plot the current vote margins against the previous vote margins, i.e., `vote_margin_pct` on the x-axis and `margin_2000` on the y-axis. Please label each axis properly.

(b) Set the `cex` option in the `plot` command to 0.5 to adjust the size of the dots. Set `pch` to 16.

(c) Overlay two lines to this plot using the regression coefficients that you estimated from above. The first line will be for the control group, meaning that it will calculate the predicted vote margin as a function of the vote margin from the previous election below the cutoff point 0, using the `mod.cl` coefficients . The second line will be for the treatment group, which means that it will calculate the predicted vote margin for the values of previous vote margin above the cutoff point 0, using the `mod.tr` coefficients. You may find `curve` command useful for drawing the lines.

(d) Set the color of the first line to blue and the color of the second line to red, and set the line width to 3 (using `lwd=3`).

(e) Include a legend to show that the blue line corresponds to "Incumbent" and the red line corresponds to "Barely lost."

4. Do the two lines meet at the center or does the predicted value suddenly jump at the cutoff point 0, where the x-axis switches from the control group to the treatment group? According to this finding, do you think incumbency provide an advantage for the next elections? What is the difference in the fitted values of the two lines at zero?

5. **Extra Credit:** In the previous step we used two different regression lines and we were able to evaluate graphically whether there was an effect. But we can also fit a single model using the entire dataset (instead of one model for those who are above/below the cutoff) and include a variable that is the treatment indicator (`incumbent` in this case). In other words, instead of splitting the dataset in two parts based on `incumbent` and fitting two models, we will use `incumbent` as a variable in the regression for the following model:

$$\texttt{margin\_2000} = \alpha + \tau \cdot \texttt{incumbent} + \beta_1 \texttt{vote\_margin\_pct} + \beta_2 \cdot \texttt{incumbent} \cdot \texttt{vote\_margin\_pct} + \epsilon$$

   Since `incumbent` is a dummy variable, numerically this is the same as fitting two different models (note that we are letting the slopes vary by interacting `incumbent` with `vote_margin_pct`). The advantage is that now we can quantify exactly the Local Average Treatment Effect (LATE) by looking at $\hat{\tau}$, that is, the coefficient on `incumbent` is our estimate the magnitude of the ATE at the cutoff (our causal quantity of interest).

   What is the estimated LATE of incumbency for Brazilian mayors at the cutoff? Is the result statistically significant? How does this result to the difference in fitted values you calculated above?