Final report

- Group Topic: Politics and Policy
- Team members: Ming DeMers, Danielle Donovan, Julian Kanu, Daisy Meng, Kate Sloan, Jemima Yoon

Causal Question

Describe your causal question in a way that someone who has not taken this class would understand. Why are you interested in this question? How could answering this question allow for better decision making? Include any necessary background or context. Cite outside sources you use. How might the distance that a constituent lives from their legislator affect the likelihood of the constituent getting a response to a request sent to their legislator.

This question is interesting because legislators oftentimes have to make decisions in situations that are difficult to observe, e.g. behind closed doors and in high security locations. In the context of this question, it is difficult to observe a given legislator in deciding what kind of requests from their constituents they will and will not respond to. The location of the constituent's residence can be a factor that influences the legislator's decision to respond to their constituents. Moreover, we as constituents do not always pay attention to the types of issues that legislators are presented with (and have to make a decision on). This question has significant real-life implications; for example, if legislators are found to be more likely to respond to the constituents who live within the same district, this may encourage higher voter turnout among the local constituents who feel like their voices are being heard.

Describe your causal question in the language of causal inference we've learned in this course: What is the treatment? What is the outcome? What are the potential outcomes? Write these out in words and in the math notation we have used in class.

$$A_i = \begin{cases} 0 & \text{If location far} \\ 1 & \text{If location is close} \end{cases}$$

$Y^{a=1}_{constituent}$

$Y^{a=0}_{constituent}$

Treatment: Proximity of constituent

Outcome: Usefulness of the legislator's response to the constituent's letter.

Draw a DAG representing your causal question that includes at least three relevant variables besides treatment and outcome that are included in your dataset. You may include more than three variables. You may include variables that are not in your dataset, but at least 3 of your variables (excluding treatment and outcome) must be included in your dataset. If you use letters to denote variables, make sure they are clearly defined Answer

Explain your DAG: tell us in words what is meant by each edge in your DAG. A->Y This is a direct causal path for location and the usefulness of the response. A<-M->Y This shows how the median household income which is M impacts the location in which you are living which also influences the response that you get. R->P->Y: The amount of rural housing(R) will influence the legislative party as rural areas tend to be redder, and it will also affect the response as there can be more or fewer incentives based on your party. A<-P->Y: This shows how the legislative party(P) of an area impacts the location as we are assuming that red areas tend to be more rural, which also impacts the response since some politicians may prefer to represent certain areas or demographics.

Discuss your DAG. How realistic is it? Are there variables or edges you excluded from your DAG that someone else might argue should be included? Playing devil's advocate, how would you critique the reliability of your DAG? This DAG is realistic because the variables included are fairly broad, and their effects on multiple other variables are demonstrated. Although our study focuses on location and response, median household income and the legislative party can impact both of these variables, as explained above. We excluded race, however, which we initially considered a confounder. However, our study chooses to focus on other factors that may also be affected by race.. Race may affect location and response outside of median household income because of other systemic bias or deliberate manipulation we couldn't formalize in our dag:systemic segregation, redlining, and gerrymandering, for example. Another variable that one may argue should be included is age; some politicians make it a goal to listen to their younger constituents, while some politicians' voter base may be older, which will impact the helpfulness of the response.

Method and Identification

What method are you using to estimate a causal effect? What causal effect are you estimating (ATE vs LATE vs ATT)? What assumptions are required to identify the causal effect via your chosen method? We use nearest neighbor model as it favors a fast, but accurate result over medium and large sample sizes. This approach matches treated units (constituents close to the legislator) to control units (those farther away) based on their similarity in observed covariates. The causal effect is estimated by average treatment effect on treated (ATT). This measures the group as a whole, and finding if they are better or worse off given a treatment, on average.

Some of the assumptions that are made to identify the causal effect via ATT are the idea of conditional exchangeability. And we ultimately chose ATT because we want to find the average treatment effect of the treated, due to the design of the experiment we would only want to see the outcome to those who actually specified their location(treatment) and not those who do not match the the far or near criteria.

Explain what conditional exchangeability means in the context of your causal question. Is it important? Why or why not? How do sufficient adjustment sets relate to conditional exchangeability? M(Median income) and P(political Party) are confounding variables and need to be blocked. By blocking the backdoor paths A<-M->Y and A<-P->Y, we satisfy the conditional exchangeability assumption. This is important because the location of constituents is not randomized; we are working with observational data. As we have explained, location is affected by political party and income, unless we condition on these variables, we will not satisfy any type of exchangeability. If we do not satisfy the exchangeability assumption, we cannot tell if the response rate is caused by income or party rather than location. When we satisfy conditional exchangeability, we can be more confident on the causal effect of location.

Assuming your DAG is true, list out all non-causal paths between treatment and outcome and list one sufficient adjustment set to identify the causal effect of the treatment on the outcome. If a sufficient adjustment set does not exist, add additional variables to your DAG so that one does exist. All non-causal paths in the DAG above include:

A<-M->Y, median household income impacts the location in which you are living which also influences the response that you get.

R->P->Y: Amount of rural housing will influence the legislative party will also affect the response as there can be more or fewer incentives based on your party.

A<-P->Y: This shows how the legislative party of an area impacts the location as we are assuming that red areas tend to be more rural, which also impacts the response.

One sufficient adjustment set is {P, M}. By adjusting for M and P, you will see the isolated causal effect of A to Y since you would be blocking the back door paths.

Discuss the plausibility of conditional exchangeability in your setting. If your sufficient adjustment set contains variables that are not in your dataset, discuss the implications. The sufficient adjustment set does not have any variables that are not found in the data set, as the dataset and its observations are extensive. Because the data accounts for multiple confounders, conditional exchangeability is plausible. Additionally, suppose we adjust the sufficient adjustment set identified previously {P, M} and assume that no other factors affect the effect of A to Y. In that case, it can be assumed that the direct causal path from A to Y would be legitimate. By controlling for political party and median household income, we can assume that the most significant confounding variables have been accounted for. Many smaller factors can be mostly accounted for within these variables.

As mentioned previously, it is possible that there are variables not in the data set that could be confounders on their own, such as race (included in data set) or age of the constituent (not included in the data set). If these confounders are relevant, it could impact the plausibility of conditional exchangeability. However, we have reasoned that these variables are not relevant to our research since there is no effort to identify or separate the age groups of the constituents, and race would not necessarily affect distance.

Discuss any other identification assumptions for your method here, such as positivity and consistency. What do they mean in the context of your causal question and are they plausible? Positivity may be violated in cases if only rural areas or low-income areas align one party (like Republicans) and live very far away from the location. However, positivity would not be violated if there are variations in party alignment P or variations in M (median income) and the location is close. We assume that the positivity would not be violated as income and political alignment is not a determining factor of whether or not one lives further away from the politician. Thus the treatments exist.

While we can assume that a certain type of people may choose to live more rural or urban, there is not a clear cut-off. There is some correlation between income and political party that separates people into different neighborhoods based on both of these factors. However, there is no reason to believe that neighborhoods are dispersed in such a way that one group (high or low income/Republican or Democrat) would be mostly either in the treated or untreated group. We are also assuming that there is consistency in our method since the distance is already defined in the data set. The distance is precise and divides everyone into two exclusive categories. Y=Yˆa and A=a. It is also consistent because there is clarity about interference. This is true because people's messages and responses should be independent from one another.

However, one thing that might cause interference is that in some cases, someone may be more likely to receive a response if their neighbor does. If an issue is affectinga specific area and many people complain about it, a resident in that group may be more likely to receive a response than someone who makes a unique complaint.

Discussion: Analysis and Results

Give some context for your dataset. Who is included in your dataset? How was the data collected? When was the data collected? Make sure to cite the dataset. The dataset consists of demographic information regarding the legislators, the districts they represent (including voter turnout and urban/rural housing), whether or not the legislators responded to the constituents' requests, and whether or not the legislators were helpful in their response. The original data on the legislators and their districts involved in the experiment was collected through public state legislative websites for all 50 states within the USA. This information includes the legislators' race, party affiliation, and gender, as well as the racial diversity of the districts and the extent of urbanization within each district. The researcher that collected this data did not note when he collected this data, but given that his findings were published in 2013, the data was most likely collected about 2-3 years prior (2010-2012) (cleaned dataset: https://github.com/Ming-DeMers/3900-final/blob/main/data/poli1.Rdata) (original Data set: https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/AMMI3S/QCV5DG&version=2.1&toolType=PREVIEW)

Discuss any choices you made regarding data cleaning and processing: Did your data have missing values or outliers? How did you handle them? Were there any variables you dichotomized (i.e. made binary), or variables that you changed the format (e.g. yes/no to 1/0)? The original dataset was already very well-organized, so there was not much to clean. While there was one variable with missing values (percentage of those in a legislator's district who voted Democratic), we were not interested in the variable and so the missing values did not pose an issue.

We considered making the distance binary and setting up a threshold to what constitutes far or near. However, this dataset was constructed where the treatment is already considered "faraway city" or "nearby city," and a column in the dataset indicating whether or not the legislator received the treatment.Thus, we didn't think the actual quantified distances we have in the dataset are too important. Ultimately we did not dichotomize or change the format of any variables.

Discuss the impact of any choices you made regarding your dataset, such as choices you made in data cleaning or processing. Initially, we removed the urban/rural housing variables because we did not think it would be included in the DAG. However, redecided when designing the DAG. We reincluded the urban/housing variables. This is because, depending on the area, living more urban or rural may impact one's political affiliation which may also in turn, impact the politician's response rate.

Explain how you estimated a causal effect.

- If you used matching, explain and discuss your choices. What formula did you use and why? What matching strategy did you use and why? Are there any advantages or drawbacks to the strategy you chose? How many units did your matching drop? How was the covariate balance in your matched sample? Discuss the implications of any choices you made and the quality of your matching.
- If you didn't use matching, explain any choices you made related to the method you used and discuss their implications. Think about advantages or drawbacks to any choices you made, possible bias-variance trade-offs, and assessing how well your method did.

We estimated the causal effect of proximity on legislator responsiveness using propensity score matching. The treatment variable, treat_out, indicating proximity, was modeled as a function of covariates to ensure matching based on all variables likely to influence both treatment assignment (proximity) and the outcome (responsiveness), as shown by our DAG, thus reducing confounding.

We used nearest-neighbor matching with replacement. This method pairs treated units with the closest control units ensuring immediate matches with minimal differences.Allowing replacement improved match quality, especially when control units were limited, but also meant that some control units were reused multiple times, potentially reducing representativeness.

Matching dropped 1,393 control units, reflecting a potential positivity violation. After matching, covariate balance improved substantially between treated and control groups, suggesting the matching process was effective in creating comparable groups.

The advantages include its ability to reduce bias from observed confounders and to make causal estimates more credible. However, its reliance on the previously mentioned assumptions on confounders and exchangeability means that any unmeasured confounders could still bias the estimate. Additionally, dropping unmatched units could limit the generalizability of the findings to the matched sample. Overall, the method provided a credible estimate of the ATT, but the quality of the matching and potential for unobserved confounding highlight areas where the analysis could be improved.

Report your causal effect estimate and interpret it in the context of your causal question. Our estimated causal effect, the ATT, is -0.12. This means proximity, on average, decreased the likelihood of a legislator responding to a request by approximately 12 points. That is, for our constituents that are closer

to their legislator, they are less likely to receive a response to a request that than those who live further away, by around 0.12%

This result suggests that proximity might negatively influence responsiveness, possibly due to factors like higher demand from nearby constituents, strategic political behavior, or other unobserved confounders associated with being "close." One possible mathematical explanation would be that, given a steady density of inhabitants at any given distance from the legislator, there will be more total inhabitants in the circumference of a larger radius than a smaller one.

Overall, the causal effect estimate suggests that simply being geographically closer to a legislator does not guarantee greater responsiveness and, in this case, appears to reduce it.

Discuss the limitations of your analysis: what are the limitations of your dataset? Is there other data you would have wanted to have to bolster your analysis? Playing devil's advocate, how would you critique the reliability of you causal estimate? Our analysis is ultimately limited by the nature of the dataset and the assumptions required for the methodology. Foremost, the dataset is synthetic, meaning it is constructed to simulate real-world scenarios, yet may not fully capture the complexities of legislative behavior and constituent-legislator dynamics.

The binary treatment variable (treat_out), which indicates proximity, oversimplifies distance into a treated or control category. A continuous measure of proximity could've better capture the nuances of how distance affects responsiveness. However, we would have binarized any linear measurement as otherwise would've been outside the scope of this project.

To bolster the analysis, we would have liked to include data on the nature and content of constituent requests (e.g., their urgency or importance), as this likely influences whether a legislator responds. What are the actual contents of the requests? Legislators likely have an agenda they focus on for each term, and thus targe certain requests more than others, despite the distance of the residence of requester. Legislative factors such as staff capacity, workload, and district size could also provide a clearer understanding of constraints on responsiveness.

Critiquing the causal estimate, the analysis relies on several key assumptions, positivity and consistency, and ultimately conditional exchangeability. If unmeasured factors, like the complexity of constituent issues or district-specific political dynamics, influence both treatment and outcome, the estimate may be biased. While propensity score matching improved balance, residual imbalances in covariates and the exclusion of 1,393 unmatched control units suggest potential violations of the positivity assumption. This could restrict the generalizability of the findings to the matched sample rather than the broader population.

Moreover, the lower r-squared value in the regression model could indicate that only a small proportion of the variance in responsiveness is explained by the covariates included, suggesting other unobserved factors may play a significant role. While the analysis provides insights into how proximity might influence responsiveness, we can, at best, say our conclusions apply to this synthetic dataset, and may apply to a real-world example. But more data collection and research would be necessary.

Code:

```
# This is where your code goes for the data cleaning/processing and analysis.
# Make your code clean and easy-to-follow. Add short comments to explain what you are doing.
# If a classmate who wasn't as familiar with R were to read through this section,
# would they be able to follow along?

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(MatchIt)
```

Data Cleaning

```
#|label: dataset-clean

# Cleaning data
View(x) # x was automatically named while saving the original dataset
poli <- x # renamed the dataset
poli <- poli[ -c(23,24:25) ] # removed a few columns I deemed unnecessary
poli <- poli[-c(28)]
poli[-c(31)]
poli <- cbind(poli, x['ruralhousing']) # readded this column and the following from the original datase

poli <- cbind(poli, x['ruralhousing'])
poli <- cbind(poli, x['urbanhouses'])
poli <- cbind(poli, x['totalhouses'])

save(poli, file = "data/poli1.Rdata") # saved the file to my computer
```

Matching

We attempt to match.

```
#|label: matching

load("data/poli1.Rdata")
```

```r
# create matches based off the treatment (treat_out) and outcome
# (code_some_response), along with possible covariates leg_party and medianhhincome
# matching is done using the nearest neighbor method (method = "nearest"`)
# with a logistic regression model for estimating propensity scores (distance = "logit")
# the causal effect measurement is by ATT (estimand) = "ATT").
# `ratio = 1` specifies 1-to-1 matching, and `replace = T` allows controls to be reused
match <- matchit(treat_out ~ code_some_response_given + leg_party + medianhhincom,
                 data = poli,
                 method = "nearest",
                 distance = "logit",
                 estimand = "ATT",
                 ratio = 1,
                 replace = T)

# extract the matched dataset resulting from the propensity score matching
match.data(match) -> matched_df

# fit a linear regression model on the matched dataset
# this model estimates the treatment effect (causal effect) of treat_out
lm_mod <- lm(code_some_response_given ~ treat_out + leg_party + medianhhincom,
             data = matched_df)

# display a summary of the regression model, including coefficient estimates,
# standard errors, and statistical significance
summary(lm_mod)
```

```
##
## Call:
## lm(formula = code_some_response_given ~ treat_out + leg_party +
##     medianhhincom, data = matched_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6310 -0.3358 -0.2717  0.5990  0.7812
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.297806   0.025338  11.753  < 2e-16 ***
## treat_out     -0.115187   0.015168  -7.594 3.80e-14 ***
## leg_partyI    -0.269057   0.329060  -0.818   0.4136
## leg_partyP     0.381073   0.268757   1.418   0.1563
## leg_partyR     0.032668   0.014578   2.241   0.0251 *
## leg_partyR+D  -0.326947   0.465443  -0.702   0.4824
## medianhhincom  0.020698   0.005071   4.081 4.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4651 on 4194 degrees of freedom
## Multiple R-squared:  0.01991,    Adjusted R-squared:  0.0185
## F-statistic:  14.2 on 6 and 4194 DF,  p-value: 4.469e-16
```

```r
# extract the estimated treatment effect (ATT) from the regression model,
# and print the actual causal estimate
att_est <- coef(lm_mod)["treat_out"]
print(att_est)
```

```
##  treat_out
## -0.1151866
```

```r
# show how many units matched and unmatched
summary(match)$nn # number of matched
```

```
##                 Control Treated
## All (ESS)     2814.0000    2779
## All           2814.0000    2779
## Matched (ESS)  940.7773    2779
## Matched       1422.0000    2779
## Unmatched     1392.0000       0
## Discarded        0.0000       0
```

```r
# print the causal estimate (ATT) rounded to 3 places

print(round(att_est, 3))
```

```
## treat_out
##    -0.115
```