Backwards Design Assignment Causal Inference

IDS 690 - Unifying Data Science

4/1/20

**1. Topic:**

*What is your project about? What problem are you seeking to solve, or in which domain do you think you can contribute meaningfully?*

Our project is about the spread of coronavirus (COVID-19) and how social distancing and self-quarantining practices by communities affect the spread of the virus. Social distancing will be measured by cell phone data using a metric called “distance travelled.” When we see a large dip (for example, if there is more than 60% of the rapid decrease in average daily distance travelled, we will have to look at the data first, which we do not have yet), we will judge this as the date where quarantine began and people started losing their jobs/working from home. We are seeking to answer to what degree does earlier social distancing and quarantining affect the spread of the disease. We will also be looking at unemployment numbers (data for the month of March comes out April 3rd, to somewhat validate our quarantine infection date) This analysis will provide insight into the types of measures that are most effective at reducing pandemic.

# 

# **2. Project Question**:

*What specific question are you seeking to answer with this project? For this project, this must be a* ***causal*** *question.*

What is the effect of the date of quarantine after the first case of COVID-19 within a state on the infection spread and mortality rate of COVID-19? We will be analyzing how the “leverage” of the lockdown/social distancing affects different states/counties.

# 

# **3. Ideal Experiment**

*If you were a god, what experiment would you run to answer your question? Define both your treatment variable, and your outcome of interest.*

In a defined state, randomly split half states/counties being quarantined and another half not being quarantined and view the infection spread and mortality rate within these subpopulations. Quarantining is the treatment variable while infection spread and mortality rate are the outcome variables.

# 

# 4. Pick a Study Context

*Where can you get data that (a) measures your outcome variable, and (b) includes variation in*

*your treatment variable?*

**Infection rate** - Will be calculated as infections/population in a given state/county. This data is gotten from a nytimes dataset: <https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv>

We will also compare between the NY Times dataset and a couple Kaggle datasets (<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>,

<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/discussion/129550>) to ensure reliability of our data.

**Mortality Rate** - This data will be calculated as deaths/(deaths + recoveries) or as deaths/(total confirmed infections) (We will look at the data to decide on this or maybe use both in our analysis). Data will be gotten from the NY Times dataset and extrapolated population from the 2018 World Bank population data.

**Variation in Treatment** (When a state quarantined) - This will be obtained from cell phone data, we will look for a significant drop (60%, need to look at data to actually define a percentage) in distance travelled and define this as the date where quarantine/ social distancing practices began kicking in.

# 5. Project Design

*Given the context you want to study (and data you can find), what design do you think would*

*be feasible?*

For this study, both a difference-in-difference and pre-post design are applicable. The diff-in-diff can be used to compare the rates of infection/death between states and counties that implemented social distancing/quarantine practices in relation to when the first. The pre-post can be used to look at individual states' infection and mortality rates with and without quarantine. Since epidemics follow an exponential (and eventually sigmoid) curve, we can use the data to plot the expected trend without an intervention. Note that since it takes about 2 weeks for the disease to show symptoms, we will want to see if measures such as the quarantine are effective after a 2 week period from implementation. We will primarily use the diff-in-diff design to compare the responses of different countries.

# 

# 6. Model Results

*One of the hardest parts of developing a good data science project is developing a question that*

*is actually answerable. Perhaps the best way to figure out if your question is answerable is to*

*see if you can imagine what an answer to your question would look like. Below, draw the graph,*

*regression table, etc. that you would consider to be an answer to your question. Then draw it*

*again, so you have a model result for if treatment has an effect, and a model result for if your*

*treatment does not have an effect. (If the answer to your question is continuous, not discrete*

*(like: what is the effect of health insurance on life expectancy), draw it for high values (high*

*inequality) and low values (low inequality)).*

**Result if your hypothesis is true Result if your hypothesis is false**

# 

Our hypothesis is that implementing a quarantine (based off a drop in distance travelled in cell phone data) early in the development of infections will be effective in reducing the cumulative infection rate. If true, we will see the cumulative rate decrease for countries that implement a quarantine while those without a quarantine continue to have rates that follow an exponential growth. If our hypothesis is false, we will see that the infection rates for countries with and without quarantines will follow the same trend.

Furthermore, we aim to investigate the impact of when the quarantine was implemented. For example, will earlier quarantines be significantly more effective than later ones? Does increased population density lead to a higher infection rate and mortality rate?

Note that there are other factors that can contribute to lower rates of infection, such as how early testing and rapid identification of sick individuals. We will have to address these in our project design.

# 7. Final Variables Required

*Now that you’ve specified what an answer to your question looks like, what data do you need to*

*generate that answer? For each variable, define both the variable you need and the population for which you need the variables to be defined.*

*You don’t have to be too specific (“I need variable 7 from the NHGIS 2019 census 1% sample release”) – just define it in the most general terms that are still specific enough to meet your needs (e.g. I need income data for a nationally representative sample of US citizens from both before and after 2012).*

For this project the final variables that we require are the following:

1. List of states and their respective populations
2. Number of COVID-19 infections for each day for each region
3. Number of COVID-19 deaths
4. Date of first confirmed case in a given state
   1. The x-axis will be the number of days since the first case, so we will subtract each date by the date of the first case
      1. We might also consider not just using the first infection but a cutoff number as a single member of a population can vary greatly based on the person. With a cutoff, using the law of large numbers, we can get a more generalizable/reliable model of how social distancing/quarantining affects the spread of the disease.
5. Date of intervention (quarantine)
   1. This will also be subtracted from the date of the first case
6. We are considering using population density as a variable in our model
7. Unemployment claims will also be looked at

# 8. Data Sources

*Finally, given the variables you need for your analysis, what actual data sources do you think*

*will have the data you need?*

*In specifying the datasets you need, if you list more than one also indicate how you think you*

*can relate these datasets (i.e. if you’re gonna merge them, what variables do you think those*

*datasets will provide that will allow you merge them? There’s no use saying “I’ll merge this*

*political survey with medical records of who has received bad care” if the political survey doesn’t*

*provide identifying information you can use to link survey respondents to medical records, even*

*if you have both the survey and medical records!)*

The data sources that are required for this analysis are the following:

1. Kaggle Dataset from WHO contains information about Country, confirmed cases, deaths, age, recoveries, symptoms, etc.
   1. <https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv>
   2. <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>
2. WHO Current Data Trends for additional information
   1. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>
   2. <https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd>
3. World Bank has information on population in countries and the percentage people working in a country
   1. <https://data.worldbank.org/indicator/SP.POP.TOTL>
4. The New York Times has cumulative counts of Coronavirus cases in the United States, at the state and county level, over time.
5. <https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv>
6. U.S. Unemployment Claims from the Bureau of Labor Statistics (still trying to find weekly data).
   1. <https://data.bls.gov/timeseries/LNS14000000>
7. Social distance data by state in the United States

We will merge the population dataset with the COVID-19 dataset on Country. We will also find the dates in which a quarantine was implemented (if it was) for each state in our primary list (TBD). This data can be obtained from reputable news sources or WHO. We have created a separate spreadsheet to contain the relevant dates that we can merge within the primary datasets.

# 