Backwards Design Assignment
Causal Inference

Nick Eubank

April 5, 2022

For the next step of your group data science project, you must document a plan using this backwards design template.

In completing this assignment, you should follow the template presented below. Note that this is an exercise that is fundamentally about what you do *before* you start working with data. You **should** look into what data is available, and you need to report specific datasets that will enable you to run the analyses you propose, but completing this assignment does not require the presentation of any data analysis.

Deadlines

- **Due**: Tuesday, March 15nd, Start of Class via Gradescope

I hope you're able to have fun with this exercise. It is rare in school that we get to invest in answering precisely the questions we find really exciting, and I hope you will see this as an opportunity to invest in learning about (and potentially helping address) a problem you care about personally. Moreover, as we've discussed before, this is a potential portfolio piece – one unique to your team – you can show future potential employers.

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?

The *Taliban Movement in Pakistan* or TTP is an Islamic anti-state armed organization based along the Afghan–Pakistani border. In 2007, the organization started a terrorist campaign against the Pakistan armed forces and the state. In response, the Pakistani government launched an aggressive military operation, coming out in victory for the first time in Swat Valley.

The Pakistani government had numerous attempts of peace deals with sub-groups belonging to the TTP, both official and unofficial. Most of these peace deals, however, lasted only a few months.

Terrorist attacks continue to strike Pakistan, mostly in its largest cities.

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 legacy the TTP terrorist attacks have left on Pakistani communities? More specifically, what effect has the TTP terrorist incursions that started in 2007 had on women's access to education and labor?

Our teams hypothesis is that women's rights (access to education and labor) are weakened by proximity of the extremist groups.

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.

The ideal experiment would measure women's enrollment in school and women's participation in labor in two parallel versions of Pakistan: One where there is no TTP terrorist campaign, and one where they do carry out their terrorist incursion starting in 2007. Then, we would compare women's access to education and labor across districts in the two parallel versions of Pakistan to see if, for districts that are closer to where attacks would happen, Pakistani women are enrolling less in schools or participating less in labor.

4 Pick a Study Context

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

(a) To obtain data that measures the outcome variable, we will use district level enrollment in K-12 schools (or equivalent) by sex. We are also interested in the participation in labor by sex. If the education enrollment variable cannot be easily obtained, we will only focus on the labor participation by sex. (b) The treatment variable will be the sex. Women will be considered the treatment and men will be considered control group. We would also control for proximity to the largest attacks (we have information about the top 20) if geographical location data of each district is available.

5 Project Design

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

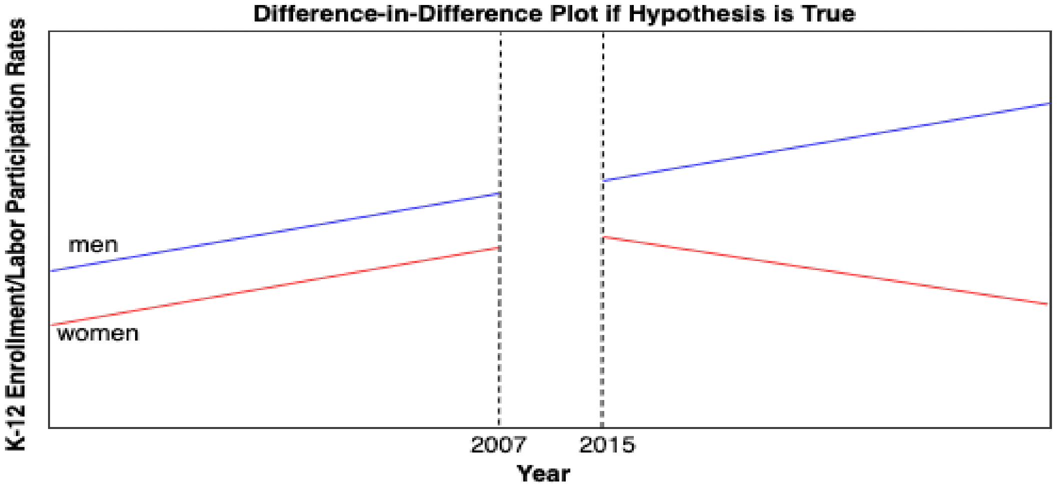
We will use the difference-in-difference framework to compare our outcome variable (enrollment rates in K-12 schools, and/or participation rates in labor, at the district level) before and after the 2007-2015 period of major Pakistani-Taliban attacks for men (control) versus women (treatment), possibly as a function of distance to the closest mayor attacks.

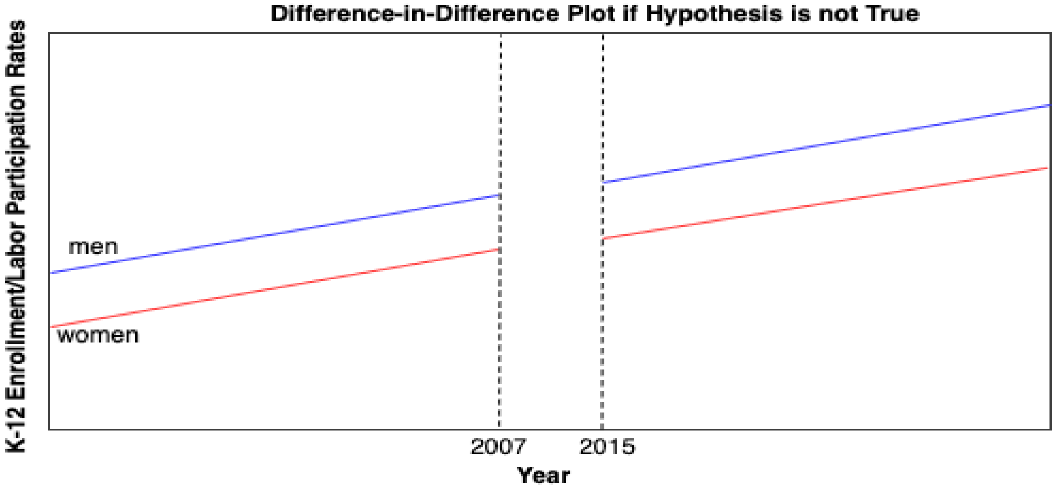
The outcome variable rates will be calculated based on the total population of each: total school enrolment counts per district by sex / district population for the age group; OR total number of workers per district by sex / district population in the working age.

There are 159 districts in Pakistan.

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)).





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).

To calculate the difference in difference effect, we would need a table with the following columns:

- Year, preferably from 1999 to 2020 (or some period before 2007 and after 2014 or 2015)
- District name, for all districts in Pakistan.
- District total population in the schooling age group OR working age (count)
- District total enrollment in schools OR participation in workforce (count)
- Sex
- Distance to the closest of the 20 major Taliban attacks from 2007 to 2015*

With these columns we can calculate the outcome variables as rates based on the adequate demographic population in the district, as well as the pre-post variable for the terrorist "intervention".

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!)*

To obtain the outcome variables we need (education and labor data by district, by sex, by year) seems to be available in the Pakistani Bureau of Statistics website. We may be able to obtain the enrollment in education and labor counts from the Pakistan Social And Living Standards Measurement survey (we have not selected which question in the survey answers this yet) and merge with Census data to obtain total population per district by sex and age.

It is possible the geospacial data for Pakistani districts is in the IPUMS dataset. We would use these to obtain the distance to the top 20 Taliban attacks. The list of the major Taliban attacks in Pakistan (mostly bombings) is available with specific addresses and death/wounded totals in a Stanford website. We would only need to find spacial data for these 20 addresses manually.