PDS Midterm Project Strategy Outline

Erika Fox, Shining Yang, Preet Khowaja, Minjung Lee

I. Introduction

Our goal for this midterm project is to explore the effect of opioid drug prescription regulations on the volume of opioids prescribed and drug overdose deaths. The policy changes we are specifically interested in include a 2010 Florida change, a 2007 Texas change, and a 2012 Washington change. We will take two approaches in order to tackle this causal question, including a pre-post comparison and a difference-in-difference approach.

For each approach, we will be looking at four comparisons. The effect of each state's respective policy changes on opioid overdose deaths, and the effect of only Florida's policy change on opioid shipments, due to the nature of our data.

Pre-post analysis:

Each of our four comparisons requires a plot with 'Years from Policy Change' on the x-axis. For the y-axis for opioid shipments, we can use opioids per capita, and for the y-axis for opioid overdose deaths, we can use the number of deaths per capita.

Difference-in-difference:

The plots we will make for this part of the project will be similar to the plots for the pre-post analysis approach, but here, we will put a threshold at the policy year. Also, for comparison, states without policy changes will be added as control states. After the policy year, the treatment state is supposed to have a different pattern with the control state.

II. Data

Working backward from the goal, which is to have opioid consumption per capita and deaths per capita by overdose, we want a final dataset that has county-wise observations for these two variables for each year from 2006-2012.

We are planning on using three separate datasets in order to conduct the above analyses. The first gives us opioid consumption, the second gives us the number of deaths from opioid drug overdose and the third gives us United States county populations. Merging these datasets will give us the necessary data to create the opioid consumption per capita and deaths per capita variables.

Opioid Dataset

The first dataset we are going to use will consist of county data for Florida, Washington, Texas and other control states. The variables of interest from this data are *transaction_date, buyer_county, measure* (what form the drug is in), *dos_str* (strength of dose), *drug_code* and *MME*. Morphine milligram equivalent, or *MME*, is our response variable which will tell us the quantity of morphine equivalent drug that was in each transaction.

This dataset is messy and large and lots of work will be needed to make it usable. Every row in this dataset is a county, year observation for a transaction of opioid split by the form of the drug, name of the drug supplier, the kind of drug it was (oxycodone and hydrocodone) and the dosage. For this, we will have to add up the opioid consumption units across the different rows in order to aggregate for county and year. *MME* ensures that we have a standardized unit for these aggregations and that we can add the doses of the two types of drugs without any significant loss of information. We will use *MME per capita* in our final analysis.

Deaths by overdose data

This dataset is not as messy as the one above. It contains information on *county*, *county_code*, *year*, *the number of deaths* each year and categories of what the fatalities were caused by. After filtering out the deaths that are not caused by overdosing on drugs, we can use the remaining data in our analysis. Since the data in this dataset is sparse, we will have to carefully choose our control states and ensure that they appear here.

Population dataset

In order to have final variables that are per capita we needed a dataset that gave us populations in the counties over 2006-2012. We used the Census Bureau website to find this dataset. The dataset consists of one file for each decade, and the number of populations in each county is aggregated by year in each file. However, the data here is column-wise. This means that each row has a population observation for each year from 2000-2020.

Merging

Once we are done with the data cleaning, we will merge the datasets above using FIPS codes and year, so that the finished dataset has row-wise observations for *county*, *year*, *opioid per capita* and *deaths per capita*. These are the four variables of interest that will allow us to do our analysis. Based on the counties we select, we will also need to add a variable for whether the given county is a control or treatment one.

III. Role

Since there are two datasets that need to be cleaned up, two analysis methods, and two reports, the team is divided into two and each is responsible for different combinations of main and backup. This role can be changed as it progresses.

Erika: Cleaning opioid shipment data; merging; Pre-Post analysis for Florida and plotting; final report for Nick

Preet: Cleaning opioid shipment data; merging; Pre-Post analysis for Texas, Washington and plotting; final report for Nick

Shining: Cleaning overdose death data; merging; Diff-in-Diff analysis for Florida and plotting; final report for a policy maker

Minjung: Cleaning overdose death data; merging; Diff-in-Diff analysis for Texas, Washington and plotting; final report for a policy maker

IV. Workflow

All work will be managed and shared through github, following the folder structure below.

00_raw_data: raw data, will not be changed
10_modified_data: cleaned or merged datasets (the opioid shipment, overdose death, population, etc)
20_modeling: code files for cleaning, merging, pre-post and difference-in-difference analysis
30_visualization: tables, figures, mas, etc
40_reports: proposal and final reports