Opioids: Backwards Desgin

1) Define Your Problem

The problem is to evaluate the effectiveness of policy interventions to limit opioid abuse. The purpose of this project is to measure the impact of policy changes on opioid prescribing and drug overdose mortality in multiple U.S. states.

2) Define the Question You Wish to Answer

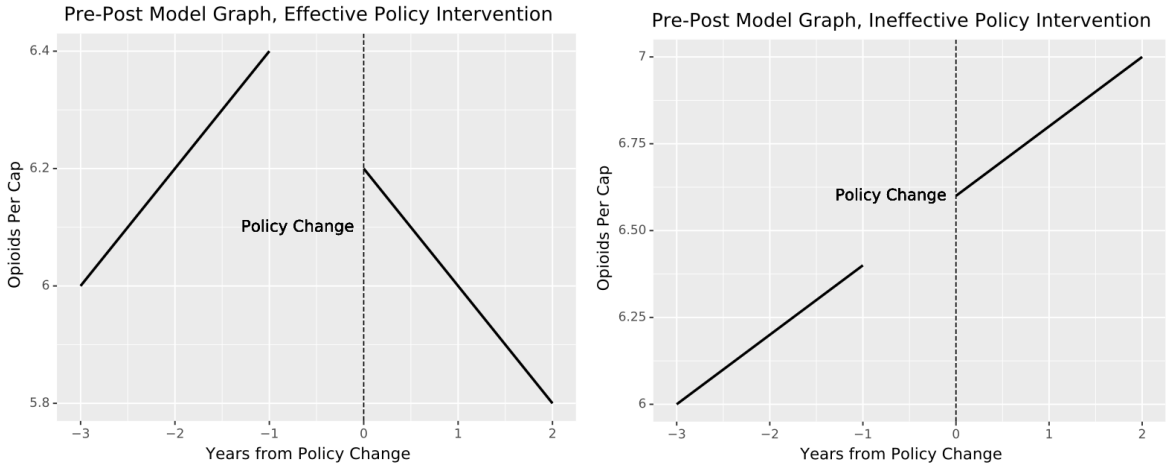
The question we aim to answer is: "What is the impact of opioid prescription regulations on

(1) the volume of opioids prescribed and (2) drug overdose deaths in multiple U.S. states?"

3) Write Down What An Answer Would Look Like

a) pre-post plots draft:

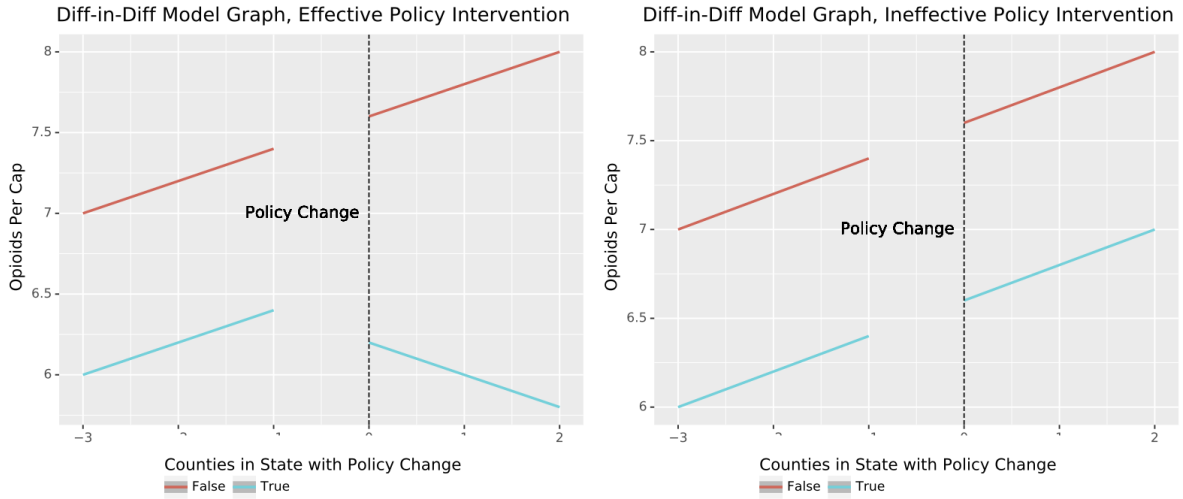
Figure 2: Pre-Post Model Figures



For the Florida and Washington cases, we draw two plots to analyze the effect of its policy change on BOTH opioid shipments and overdose deaths.
For Texas, we only need to draw and analyze the effect of its policy change on overdose deaths.

b) difference in difference plots drafts:

Figure 3: Difference-in-Difference Example Plots



For the Florida and Washington cases, we draw two plots to analyze the effect of its policy change on BOTH opioid shipments and overdose deaths compared with other states. For Texas, we only need to draw and analyze the effect of its policy change overdose deaths compared with other states.

4) What Data Do You Need?

∗ What **variables** will I need in this data?
For opioid volume analysis: year, county, population, volume of opioid,
For drug overdose deaths analysis: year, county, population, drug overdose deaths

∗ What **sample** (what years, what counties, etc.) needs to be covered in this data?
For pre-post: years before policy and years after policy. counties in the state
For difference-in-difference: years before policy and years after policy. counties in the state, as well as counties in states that have similar trends in opioids usage before policy change.
Specific years for each state:

- Florida: years up to 2009 (inclusive), years since 2010 (inclusive)
- Texas: years up to 2006 (inclusive), years since 2007 (inclusive)
    - Note: if we do the bonus analysis on opioid shipment, we will be using months: months of 2006 (12 observations), and the months afterwards.
- Washington: years up to 2011 (inclusive), years since 2012 (inclusive)

∗ What should a **single row** of this data look like (i.e. what's a unit of observation?)
A unit of observation should be county-year (one observation per county per year).
For opioid volume analysis: A single row of data should have year, county, population, and volume of opioid.
For drug overdose deaths analysis: A single row of data should have year, county, population, and amount of overdose deaths.

5) Where Can You Get That Data?

We want to have two final analysis datasets: one for opioid usage in counties over the years (opioid dataset) and another for drug-related deaths in different counties (deaths dataset).

To get this dataset, what do I need to do? What do I need my input datasets to look like to get to this final analysis dataset?

For opioid volume analysis dataset:

1. A dataset with county, year, opioid usage
2. A dataset with county, year, population

For drug overdose deaths analysis dataset: A

1. A dataset with county, year, overdose deaths
2. A dataset with county, year, population

In both cases, we can merge the intermediate datasets on “county” and “year”

To get to these intermediate datasets, what source datasets do I need? What variables do they need?

For source datasets, we'll use the Opioid Prescriptions dataset provided by the Washington Post, the Vital Statistics Mortality dataset, and U.S. County Population Totals datasets.

Summary of source datasets

Vital Statistics Mortality Data Summary:

1. From 2003 to 2015, each year has its own txt file
2. The text files seem to follow tsv format, and the columns are “Notes, County, County Code, Year, Year Code, Drug/Alcohol Induced Cause, Drug/Alcohol Induced Cause Code, Deaths”
3. The unit of observation is county-year
4. The “County Code” refers to FIPS codes, a 2-digit State FIPS code and a 3-digit County FIPS Code. However, the dataset omits the first 0 for State FIPS code 1-9, resulting in 4-digit values.
5. The “Drug/Alcohol Induced Cause” value that we are interested in is “Drug poisonings (overdose) Unintentional (X40-X44)”
6. It has formatting issues, where some values are not properly separated by a tab.

U.S. County Population Data:

1. Links: https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-counties.html, https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html
2. County Population Totals, 2000-2010 and 2010-2019

3. Unit of observation: county
4. Columns: population totals in different years

Opioid:
The Washington Post's opioid data set represents a critical resource for analyzing the distribution of opioid pain pills in the United States from 2006 through 2019. This data set, derived from the Drug Enforcement Administration's Automation of Reports and Consolidated Orders System (ARCOS), was made public following a successful legal challenge by The Washington Post and HD Media. It encompasses detailed transaction records for oxycodone and hydrocodone pills, which constitute the bulk of opioid dosages during the specified period. This data set is pivotal in understanding the prescription opioid epidemic, which has been linked to over 210,000 overdose fatalities within the 14-year scope of the records.
The Washington Post has provided this extensive data in a .tsv (around 90GB) format. For ease of access and analysis, summary data is also available, highlighting the key distributors, manufacturers, and pharmacies within various locales.

Steps to create the intermediate datasets from them:

For the mortality dataset, we will follow these steps:
1. Concatenate the txt files to form one dataset/dataframe.
2. Drop unrelated variables that are not relevant to our analysis.
3. Filter the data based on the cause/code column, so we only have overdose rows.
4. Drop the cause and code columns, now that we only have overdose rows.

For the population dataset, we will follow these steps:
1. Drop unrelated variables that are not relevant to your analysis.
2. Reshape (wide-to-long) so that we have county-year as unit of observation.
3. Concatenate so that we have 1 dataset/dataframe from 2000-2019.

For the opioid dataset, we will start by downloading the "arcos_all_washpost" dataset and then proceed with the following steps:
1. Drop unrelated variables that are not needed for our analysis.
2. Create a new “year” column based on “TRANSACTION_DATE”.
3. Calculate the sum of Morphine Milligram Equivalents (MME) for each county and year, to create a new “MME_sum” column.

By processing these source datasets as described, we will create the intermediate datasets needed for final analysis.

6) Tasks Assignment

Team members: Zeying Huang(ZH), Yueting Luo(YL), Junyu Liu(JL)
Code Review Sequence: (ZH -> YL -> JL -> ZH)

Data Preprocessing: (Download, Merge)

1. Create intermediate datasets for population and overdose - JL
2. Create intermediate datasets for opioid - ZH, YL

- Opioid Data: Extract county, year, and calculate per capita values for 2 base states and 3 control states for each. (can do Texas for bonus)
- Death Data: Extract county, year, and death counts for 3 base states and control states

Visualization: (Pre-Post, Difference, State Verification)
Pre-Post Analysis:
- Create 5 plots for opioid data (3 base states).
- Create 4 plots for death data (2 base states) and 1 plot for the state with a death bonus.

Difference Analysis:
- Create 5 plots for opioid data (3 base states + 3 other states).
- Create 4 plots for death data (2 base states + 3 other states) and 1 plot for the state with a death bonus.
- Zeying Huang: Opioid Data (Difference Analysis)
- Yueting Luo: Opioid Data (Pre-Post Analysis)
- Junyu Liu: Overdose Deaths Data

Analysis: (Report)
- Present the project's motivation.(ZH)
- Explain the rationale behind the research design.(YL)
- Detail how different datasets have been related to one another.(ZH, YL, JL)
- Provide summary statistics for the data.(ZH, YL,JL)
- Perform the analysis (presented for a non-statistician).(ZH, YL,JL)
- Interpret the analysis results, highlighting strengths and weaknesses without using statistical jargon.(JL)