**IDS690 - PROJECT STRATEGY OUTLINE**

**Oct 19, 2019**

**Due Date: Oct 22, 2019**

Pair A: Yu Gu, Zifan Peng

Pair B: Jingyi Wu, Shota

**Please look at where I highlighted and please feel free to give comments and modifications!!!**

**Prescription Analysis (Pair A)**

1. **Data Cleaning**

* **Initial Dataset:** Opioid Prescriptions Shipments from 2006-2012
* **Variables in the Intermediate Dataset:** In the intermediate dataset for this analysis, variables should be like this:

***Year:***From 2006 to 2012 (be aware of the data shortage for Texas and Washington, could instead use monthly data if we have time.

***State:***Probably 2 states (if we only analyze Florida). One for policy-change states,Florida. Another one for a non-policy-change state that has similar pre-trends with Florida.

***County:*** Probably choose the top 3 counties with the most shipments (or randomly choose 3 counties?) in Florida (in case of losing generality and we don’t want to analyze ALL the counties in Florida). Then in that non-policy-change state, we have to find 3 counties that has similar pre-trends corresponding to the 3 counties in Florida. (Or we could randomly choose counties that may not have similar pre-trend, since we can use the second regression which controls for change in trends. This will give us correct estimate regardless of different pre-trends.)

***Shipments\_County:*** Total shipments for a specific county per year.

***Shipments\_State:*** Total shipments for a specific state for the whole period from 2006 to 2012.

***Shipments\_Adjust:*** To account for different magnitude of shipments in the two states, we divide ***Shipments\_County*** by ***Shipments\_State*** to make an adjustment.

***Post:*** An indicator variable for whether we are in a period after implementation of the policy change. (have to create this variable by looking at the year and the state)

***Policy\_State:*** Anindicator variable for whether a given county is in a state that experienced a policy change. (have to create this variable by looking at the state)

* **Corresponding variables in the Opioid Prescriptions Shipments dataset：**

***Year:*** TRANSACTION\_DATE

***State:*** UseBUYER\_STATE.The sample non-policy-change states arenot decided yet, still have to look at the data and check the trends.

***County:*** UseREPORTER\_COUNTY and BUYER\_COUNTY. The sample policy-change and non-policy-change counties are both not decided yet. We still have to look at the data and choose the most representative policy-change counties. We also have to look at the trends to choose the sample non-policy-change counties.

***Shipments\_County*:** Use QUANTITY and UNIT to calculate the total of amount of shipments. Have to filter out the observations for all the counties we need. Then sum up the number of filtered observations by county and by year.

***Shipments\_State:*** Similar to above

* **An intermediate dataset may look like this：**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Year*** | ***State*** | ***County*** | ***Shipments\_County*** | ***Shipments\_State*** | ***Shipments\_Adjust*** | ***Post*** | ***Policy\_State*** |
| 2006 | FL | County1 | 500 | 12000 | 0.0417 | 0 | 1 |
| 2007 | FL | County1 | 600 | 12000 | 0.0500 | 0 | 1 |
| … | … | … | … | … | … | … | … |
| 2011 | FL | County2 | 500 | 12000 | 0.0417 | 1 | 1 |
| 2012 | FL | County3 | 400 | 12000 | 0.0333 | 1 | 1 |
| 2006 | CA | County2 | 200 | 8000 | 0.0250 | 0 | 0 |
| 2007 | CA | County2 | 250 | 8000 | 0.0313 | 0 | 0 |
| … | … | … | … | … | … | … | … |
| 2011 | CA | County2 | 260 | 8000 | 0.0325 | 1 | 0 |
| 2012 | CA | County2 | 280 | 8000 | 0.0350 | 1 | 0 |
| … | … | … | … | … | … | … | … |

Obviously, each single row is a county-year-level record.

1. **Pre-Post Plotting**

Use ***Year*** and ***Shipments\_Adjust*** to plot pre-post graphs. For each graph, there should be two lines for the policy-change county and the non-policy-change county respectively.

1. **Difference-in-Difference Regression**

If we can find similar trends for each pair of counties, then we could simply use the first regression since we already control for trends. But we could also do the second regression to verify our result and get a more accurate estimate for difference-in-difference coefficient.

**Mortality Analysis (Pair B)**

Actually, quite similar to Prescription Analysis except for different initial dataset and more sample counties, since we got plenty of data this time…

1. **Data Cleaning**

* **Initial Dataset:** Vital Statistics Mortality Data from 2003 to 2015
* **Variables in the Intermediate Dataset:** In the intermediate dataset for this analysis, variables should be like this (quite similar to the one in Prescription Analysis):

***Year:***From 2003 to 2015

***State:*** Probably 6 states. Three for policy-change states: Florida, Texas and Washington. Another three for non-policy-change states that has similar pre-trends with Florida, Texas and Washington respectively.

***County:*** Probably choose the top 3 counties with the most shipments (or randomly choose 3 counties?) in each policy-change state (in case of losing generality and we don’t want to analyze ALL the counties in the state). Then for each policy-change county, we have to find the corresponding non-policy change county that has similar pre-trends in the 3 non-policy-change states. (Or we could randomly choose counties that may not have similar pre-trends, since we can use the second regression which controls for change in trends. This will give us correct estimate regardless of the pre-trends.)

***Deaths\_County:*** Total deaths caused by overdose for a specific county per year.

***Deaths\_State:*** Total deaths caused by overdose for a specific state for the whole period from 2003 to 2015.

***Deaths\_Adjust:*** To account for different magnitude of deaths in each state, we divide ***Deaths\_County*** by ***Deaths\_STATE*** to make an adjustment.

***Post:*** An indicator variable for whether we are in a period after implementation of the policy change. (have to create this variable by looking at the year and the state)

***Policy\_State:*** Anindicator variable for whether a given county is in a state that experienced a policy change. (have to create this variable by looking at the state)

* **Corresponding variables in the Opioid Prescriptions Shipments dataset：**

***Year:*** Use ‘Year’

***State:*** Use ***‘***Note County’ and extract the state code.The sample non-policy-change states arenot decided yet, still have to look at the data and check the trends.

***County:*** Use ***‘***Note County’ and extract the county. The sample policy-change and non-policy-change counties are both not decided yet. We still have to look at the data and choose the most representative policy-change counties. We also have to look at the trends to choose the sample non-policy-change counties.

***Deaths\_County*:** Filter out the overdose using the ‘Drug/Alcohol Induced Cause’ column. Then filter out the counties we need. Finally sum up the ‘Deaths’ column by county and by year.

***Death\_State:*** Similar to above

* **An intermediate dataset may look like this：**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Year*** | ***State*** | ***County*** | ***Deaths\_County*** | ***Deaths\_State*** | ***Deaths\_Adjust*** | ***Post*** | ***Policy\_State*** |
| 2003 | FL | County1 | 50 | 1200 | 0.0417 | 0 | 1 |
| 2004 | FL | County1 | 60 | 1200 | 0.0500 | 0 | 1 |
| … | … | … | … | … | … | … | … |
| 2014 | FL | County1 | 50 | 1200 | 0.0417 | 1 | 1 |
| 2015 | FL | County1 | 40 | 1200 | 0.0333 | 1 | 1 |
| 2003 | CA | County2 | 20 | 800 | 0.0250 | 0 | 0 |
| 2004 | CA | County2 | 25 | 800 | 0.0313 | 0 | 0 |
| … | … | … | … | … | … | … | … |
| 2014 | CA | County2 | 26 | 800 | 0.0325 | 1 | 0 |
| 2015 | CA | County2 | 28 | 800 | 0.0350 | 1 | 0 |
| … | … | … | … | … | … | … | … |

Obviously, each single row is a county-year-level record.

**Combining the Two Analysis – Discussion needed for this part…**

We may want to merge the two intermediate datasets into an integrated dataset that both have shipments and overdose deaths data. We may do this by using merge on year, state and county. The reason we use the dataset is we may want to ask the question that, did increased prescriptions contribute to overdose deaths and did policy change make a difference on that contribution….

**About Task Assignment –Please modify this because I have NO idea at all…**

Each pair independently do their part firstly including the data wrangling, plotting and regression & statistics. Then each pair give review on other pair’s work and may discuss on some unexpected results. Finally, two pairs get together to discuss on the 2 reports.

**Question Need to Ask in Tuesday – Please come up with as many as possible!**

1. Extra dataset: FIPS Codes? Name dictionary for states and counties?
2. How do we choose our sample counties? How many should we analyze in each state?
3. Do we have to choose the treatment and control with same trends? How we define same trends? To what extent that we can say it’s a same trend?
4. Buyer and Reporter?
5. Can the trend of several counties represent the trend of the state? When analyze the trend of a state, why do we need to analyze counties instead of the state as a whole?