IDS690 - PROJECT STRATEGY OUTLINE

Oct 23, 2019

Team 2
Pair A: Yu Gu, Zifan Peng
Pair B: Jingyi Wu, Shota Takeshima

Repo:
https://github.com/MIDS-at-Duke/estimating-impact-of-opioid-prescription-regulations-team-2

Prescription Analysis (Pair A)

- ## Action Plan

  ➔ **Source Dataset:** The source dataset (all the .tsv files of Opioid Prescriptions Shipments from 2006 to 2012) are saved in dropbox. Each state has a .tsv file, namely Florida, Washington, Texas and several states for diff-in-diff comparison(have not decided which states to choose) . Then convert it to a .csv file and save it in the `20_intermediate_files` folder on our repo.

  ➔ **Initial Dataset:** Initiate basic datasets on county and state level for prescriptions shipments analysis, which includes basic subsetting and reshaping. Write code in .py file and save it in the `10_code` folder on our repo.

  ➔ **Merge Extra Dataset:** Use Population dataset at county level for normalization, which is the extra dataset. Save the source data of population in `00_source` on our repo. Then merge it with our initial dataset and conduct normalization.

  ➔ **Control Group Selection:** Develop the rules for selecting counties as control group. Write code in .py file and save it in the `10_code` folder on our repo.

  ➔ **Intermediate Dataset:** Subset the merged dataset and create binary variables for plotting and conducting regression analysis. Write code in .py file and save it in the `10_code` folder on our repo. Save the dataset as a .csv file and save it in the `20_intermediate_files` folder on our repo. Before moving to the next step, each person in pair A should review this intermediate dataset.

  ➔ **Pre-Post Analysis:** Plot the Pre-Post graphs for treatment counties using the intermediate dataset. Write code in .py file and save it in the `10_code` folder on our repo. Save the graphs (both for statistician and non-statistician) in the `30_results folder` on our repo.

  ➔ **Difference-in-Difference Analysis:** Plot graphs for both treatment and control groups using the intermediate dataset, and conduct analysis on their levels and trends. Use D-in-D regressions to find out more accurate estimates of the coefficients which can measure the effect of policy change. Write code in .py file and save it in the `10_code` folder on our repo. Save the graphs (both for statistician and non-statistician) in the `30_results folder` on our repo.

  ➔ **Code Review:** Use Pair programming method, which means one, the driver, writes code while the other, the observer, reviews each line of code as it is typed in. Two team members in pair A will switch roles frequently. Besides,

pair A will review pair B's code and graphs every time pair B finishes a part of their work.

➔ **Reports:** Write reports (using Latex?) based on the graphs and specific analysis. Save the latex file (.tex file) in `40_docs` folder on our repo.

➔ **Review Reports:** Review pair B's latex file. Then export the report to pdf as our final reports and save it in the `40_docs` folder on our repo.

● About Dataset (Source, Extra and Intermediate)

➔ **Source Data:**

Opioid Prescriptions Shipments from 2006 to 2012

➔ **Extra Population Data:**

Use an extra Poplupation dataset on country level to normalize the shipments.

➔ **Intermediate Dataset:**

In the intermediate dataset, variables should be like this:

Year: From 2006 to 2012 (be aware of the data shortage for Texas and Washington, could use monthly data for analysis.Use `TRANSACTION_DATE` from the source dataset.

State: Probably 2 states (if we only analyze Florida). One for policy-change states, Florida. Another for a non-policy-change state that has similar pre-trends with Florida. Use `BUYER_STATE` from the source dataset.

County: Analyze ALL the counties in Florida. For the non-policy-change state, we need to find several counties (maybe neighbors of FL) that has similar pre-trends with counties in Florida. Use `BUYER_COUNTY` in the source dataset.

Shipments_County: Total shipments for a specific county per year. Use `QUANTITY` and `UNIT` from the source dataset.

Shipments_Per_Cap: Normalize the shipments by dividing ***Shipments_County*** by ***Population of that county***.

Post: An indicator variable for whether we are in a period after implementation of the policy change.

Policy_State: An indicator variable for whether a given county is in a state that experienced a policy change.

The Intermediate Dataset may look like this：

Year	*State*	*County*	*Shipments_ County*	*Shipments_ State*	*Shipment_P er_Cap*	*Post*	*Policy_ State*
2006	FL	County1	500	12000	0.0417	0	1
2007	FL	County1	600	12000	0.0500	0	1
…	…	…	…	…	…	…	…
2011	FL	County1	500	12000	0.0417	1	1
2012	FL	County1	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.

- ## Sample Selection Rule

Our sample selection Rule is basically four-fold.

➔ **First: Adjacent State**

Examine the adjacent states for each treatment state. Specifically, we want to get a list of adjacent states as the candidates for each treatment state.

➔ **Second: Less Policy Change**

As the control group, these states should experience policy changes as less as possible. Therefore, in the list of adjacent states, we will select the one that experience the least policy change around the time policy took place in the corresponding treatment state. In this way, we will select 3 control states for 3 treatment states respectively. We will then analyze counties in these states.

➔ **Third: Have Similar Trends**

Among all the control counties, we will choose those have the most similar trends with the treatment counties. Since we have many counties, we may want to group the treatment and control counties by their trends.

To measure their trends, we will create a regression model. We regress ***Shipments_PerCap_County*** on ***Year*** to get a linear line representing the trend for county i.

$$Shipment\_PerCap\_County_i = \alpha_i + \beta_i Year + \varepsilon_i$$

Where α_i represents the level and β_i represents the trend for county i. Now we can group the trend β by setting the threshold for each group.

We are more likely to select pairs if they are in the same trends group. For example, there may be 5 treatment counties and 3 control counties in group 1, and the same as group 2, group 3, etc.

➔ **Fourth: Have Similar Levels**

To get the pair in one specific trend group, we further look at their levels, α. For each treatment counties in one specific group, we will select the control county that has the most closed α. Note that it is possible that we end up pairing one control county to many treatment counties.

Mortality Analysis (Pair B)

- **Action Plan**

  ➔ **Source Dataset:** The source dataset (all the .txt files of Underlying Cause of Death from 2003 to 2015) are saved in dropbox. Integrate all the data together. Filter out all the drug overdose records. Then convert it to a .csv file and save it in the `20_intermediate_files` folder on our repo.

  ➔ **Initial Dataset:** Initiate basic datasets on county and state level for mortality analysis, which includes basic subsetting and reshaping. Write code in .py file and save it in the `10_code` folder on our repo.

  ➔ **Merge Extra Dataset:** We need population data at county level for normalization purpose, which is the extra data needs to find by ourselves. Each person in pair B are expected to look for this data. Save the source data of population in `00_source` on our repo. Then merge with our initial dataset and do normalization. Before moving to next step, each person in pair B should review on this merged dataset.

  ➔ **Control Selection:** Based on some examination on statistics, develop the rules for selecting sample counties as control group. Write code in .py file and save it in the `10_code` folder on our repo. Before moving to next step, each person in pair B should review on the code and validate the control group.

  ➔ **Intermediate Dataset:** After selecting our sample, we subset the merged dataset and create binary variables ready for plotting and regression analysis. Write code in .py file and save it in the `10_code` folder on our repo. Could also save the dataset as a .csv file and save it in the `20_intermediate_files` folder on our repo. Before moving to next step, each person in pair B should review this intermediate dataset.

  ➔ **Pre-Post Analysis:** Plot the Pre-Post graphs simply on treatment counties using the intermediate dataset. Write code in .py file and save it in the `10_code` folder on our repo. Also save the graphs ready for use in the final reports (both for statistician and non-statistician) in the `30_results folder` on our repo.

  ➔ **Difference-in-Difference Analysis:** Plot graphs for both treatment and control counties in a Pre-Post fashion using the intermediate dataset. Analyze both on the levels and trends. Together with graphs, use D-in-D regressions to get more accurate estimates of the coefficients measuring the effect of policy change. Write code in .py file and save it in the `10_code` folder on our

repo. Also save the graphs ready for use in the final reports (both for statistician and non-statistician) in the 30_results folder on our repo.

➔ **Code Review:** Review one another's .py code and check all the graphs to see if they make sense and are as expected. Have discussion and modification. Besides, pair B will review pair A's code and graphs every time pair A finishes a part of their work.

➔ **Reports:** Write reports (using Latex?) based on the graphs and specific analysis each person took responsible for before (Pre-Post or D-in-D). Save the latex file (.tex file) in 40_docs folder on our repo.

➔ **Review Reports:** Review one another's latex file to see if our analysis make sense and are consistent. Have discussion and modification. Then export to pdf as our final reports and save it in the 40_docs folder on our repo.

● About Dataset (Source, Extra and Intermediate)

➔ **Source Data:**

Vital Statistics Mortality Data from 2003 to 2015

➔ **Extra Population Data:**

To accounts for different magnitude of deaths in different states/counties, we want to normalize the deaths by population. There is no population data available, however, in the two source datasets the professor provides, so we have to find by ourselves. We basically want to get the population corresponds to all the counties in our mortality data. Since we have the state code and name to identify a county, we should look for population dataset that contains this two to match our records in case of merging. To make sure we get the correct population (since there may be typos in our dataset, things will be messier than expected), we may need extra information such as zip code to verify.

➔ **Intermediate Dataset:**

In the intermediate dataset for this analysis, variables should be like this:

Year: From 2003 to 2015. Use 'Year' from the source dataset.

State: All states from the source dataset. Use 'State' from the source dataset, which is actually the state code that can be an identifier. For our samples, there are three states as treatment group: Florida, Texas and Washington. There should be another three non-policy-change states as

control group. The sample control states are subject to the selecting rule, still have to look at data.

County: All counties from the source dataset. Use 'County' from the source dataset. But we need both *'*County' and 'State' to ensure a unique county. For our samples, all the counties in the treatment state (Florida, Texas and Washington) are treatment counties. There should be the same amount of non-policy-change counties to be control counties. The sample control counties are subject to the selecting rule, still have to look at data.

Deaths_County: Total deaths caused by overdose for a specific county and a specific year. Use 'Deaths' from the source dataset. Note that there are 3 specific reasons for drug overdose, but we just want sum them up and regard it as general drug overdose.

Population: Population in thousands corresponds to a specific county (or 10 thousands? we make the unit large because we don't want the per capita data below to have too many decimal points. We define it after looking at the data). Data comes from the merged dataset of extra population dataset and initial dataset.

Deaths_PerCap_County: = ***Deaths_County / Population***. Get deaths per capital by county, i.e. normalization.

Post: An indicator variable for whether we are in a period after implementation of the policy change. This variable depends on different samples and year, so we create this after sample selection. We assume an policy take effect in the next year, so if Florida change policy on Feb 2010, we set the 'Post' binary variable for Florida in 2010 to be False, and 2011 to be True.

Policy_State: An indicator variable for whether a given county is in a state that experienced a policy change. This variable depends on different samples, so we create this after sample selection.

The Intermediate Dataset may look like this：

Year	***State***	***County***	***Deaths_ County***	***Population***	***Deaths_PerCap_C ounty***	***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
…	…	…	…	…	…	…	…
2003	FL	County3	80	2000	0.0400	0	1
…	…	…	…	…	…	…	…

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

- **Sample Selection Rule**

  Our sample selection Rule is basically four-fold.

  ➔ **First: Adjacent State**

  We start by examining adjacent states for each treatment state. Specifically, we want to get a list of adjacent states as the candidates for each treatment state.

  ➔ **Second: Less Policy Change**

  Because we want to select control group, we need them to experience policy change as less as possible. Therefore, in the list of adjacent states, we select the one that experience the least policy change near the time policy took

place in the corresponding treatment state. In this way, we select 3 control states for 3 treatment states respectively. And also, the control states become the pools that contains all the corresponding control counties candidates.

➔ **Third: Have Similar Trends**

Among all the control counties candidates, we want those have the most similar trends with the treatment counties. But since we have so many counties, we may want to group the treatment and control counties by their trends.

To measure their trends, we create a regression model. We regress ***Deaths_PerCap_County*** on ***Year*** to get a linear line representing the trend for county i.

$$Deaths\_PerCap\_County_i = \alpha_i + \beta_i Year + \varepsilon_i$$

Where α_i represents the level and β_i represents the trend for county i. Now we can group the trend β by setting the threshold for each group.

We are more likely to select pairs if they are in the same trends group. As a result and for example, there may be 5 treatment counties and 3 control counties in group 1. So to get the pair, we move to the final step.

➔ **Fourth: Have Similar Levels**

To get the pair in one specific trend group, we further look at their levels, α. For each of treatment counties in one specific group, we select the control county that has the most closed α. Note that it is possible that we end up pairing one control county to many treatment counties.

Task Assignment After Both Pairs Finish Their Work

Each pair independently do their part firstly including the data wrangling, plotting and regression & statistics. Before each pair give review on one another pair's work, one person in each pair should give a brief presentation on how they play with the data and methods & findings on their analysis. Give some discussion and revision. Finally, integrate the Prescription and Mortality analysis for the final reports.

Combining the Prescription and Mortality Analysis?

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 want 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. Discussion needed for this part.

Questions Left

Extra dataset: FIPS Codes? Name dictionary for states and counties? **Population**?

In-class Q&A

➢ Have to analyze **all the counties** in FL, WX, TX.

➢ Could use **population data** to normalize both for shipments and mortality data.

➢ Coefficients in the regression actually catch **both trends and level**, which make them hard to interpret. So **graphs are fine.**