Project 7: Difference-in-Differences and Synthetic Control

Steven Herrera Tenorio

Introduction

For this project, you will explore the question of whether the Affordable Care Act increased health insurance coverage (or conversely, decreased the number of people who are uninsured). The ACA was passed in March 2010, but several of its provisions were phased in over a few years. The ACA instituted the "individual mandate" which required that all Americans must carry health insurance, or else suffer a tax penalty. There are four mechanisms for how the ACA aims to reduce the uninsured population:

- Require companies with more than 50 employees to provide health insurance.
- Build state-run healthcare markets ("exchanges") for individuals to purchase health insurance.
- Provide subsidies to middle income individuals and families who do not qualify for employer based coverage.
- Expand Medicaid to require that states grant eligibility to all citizens and legal residents earning up to 138% of the federal poverty line. The federal government would initially pay 100% of the costs of this expansion, and over a period of 5 years the burden would shift so the federal government would pay 90% and the states would pay 10%.

In 2012, the Supreme Court heard the landmark case NFIB v. Sebelius, which principally challenged the constitutionality of the law under the theory that Congress could not institute an individual mandate. The Supreme Court ultimately upheld the individual mandate under Congress's taxation power, but struck down the requirement that states must expand Medicaid as impermissible subordination of the states to the federal government. Subsequently, several states refused to expand Medicaid when the program began on January 1, 2014. This refusal created the "Medicaid coverage gap" where there are indivudals who earn too much to qualify for Medicaid under the old standards, but too little to qualify for the ACA subsidies targeted at middle-income individuals.

States that refused to expand Medicaid principally cited the cost as the primary factor. Critics pointed out however, that the decision not to expand primarily broke down along partisan lines. In the years since the initial expansion, several states have opted into the program, either because of a change in the governing party, or because voters directly approved expansion via a ballot initiative.

You will explore the question of whether Medicaid expansion reduced the uninsured population in the U.S. in the 7 years since it went into effect. To address this question, you will use difference-in-differences estimation, and synthetic control.

Data

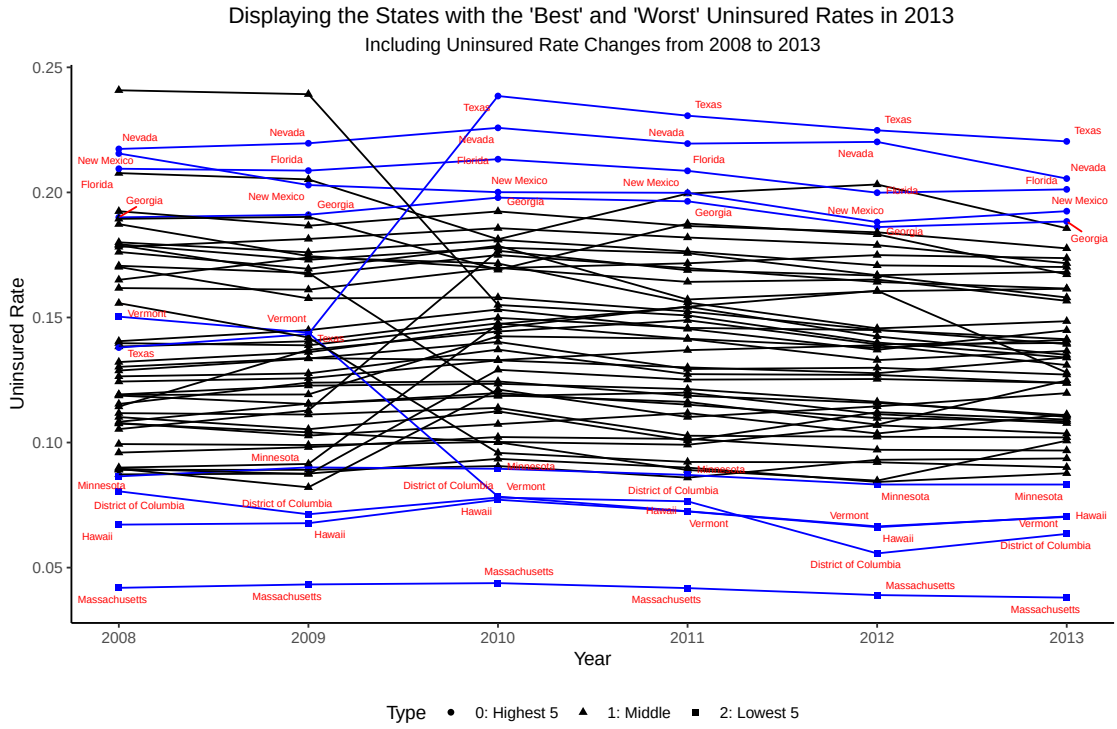
The dataset you will work with has been assembled from a few different sources about Medicaid. The key variables are:

- **State**: Full name of state
- **Medicaid Expansion Adoption**: Date that the state adopted the Medicaid expansion, if it did so.
- **Year**: Year of observation.
- **Uninsured rate**: State uninsured rate in that year.

Exploratory Data Analysis

Create plots and provide 1-2 sentence analyses to answer the following questions:

- Which states had the highest uninsured rates prior to 2014? The lowest?

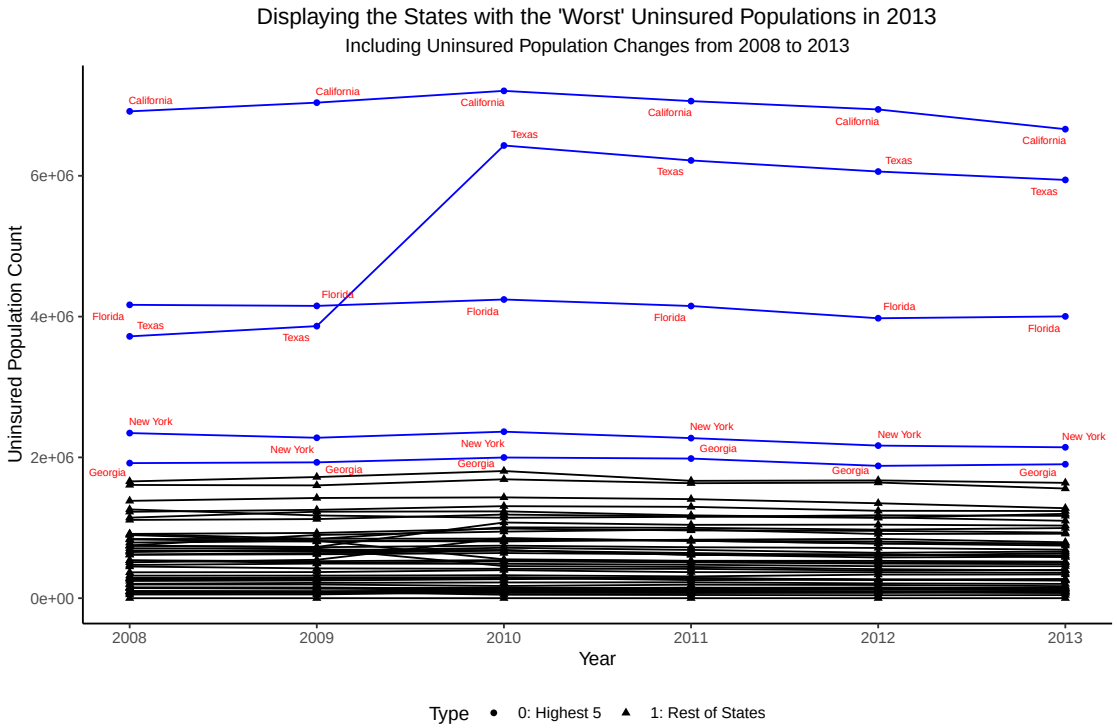


The states with the highest uninsured rates right before 2014–so in 2013–were Texas, Nevada, and Florida, while the states with the lowest uninsured rates were Massachusetts, Vermont, and Hawaii (and D.C., if you count the capital as a state). However, it is worth noting that in 2008, there was a state that had the highest uninsured rate, but then that decreased and the state joined the middle group of states. Below, we see that this state was Utah in 2008.

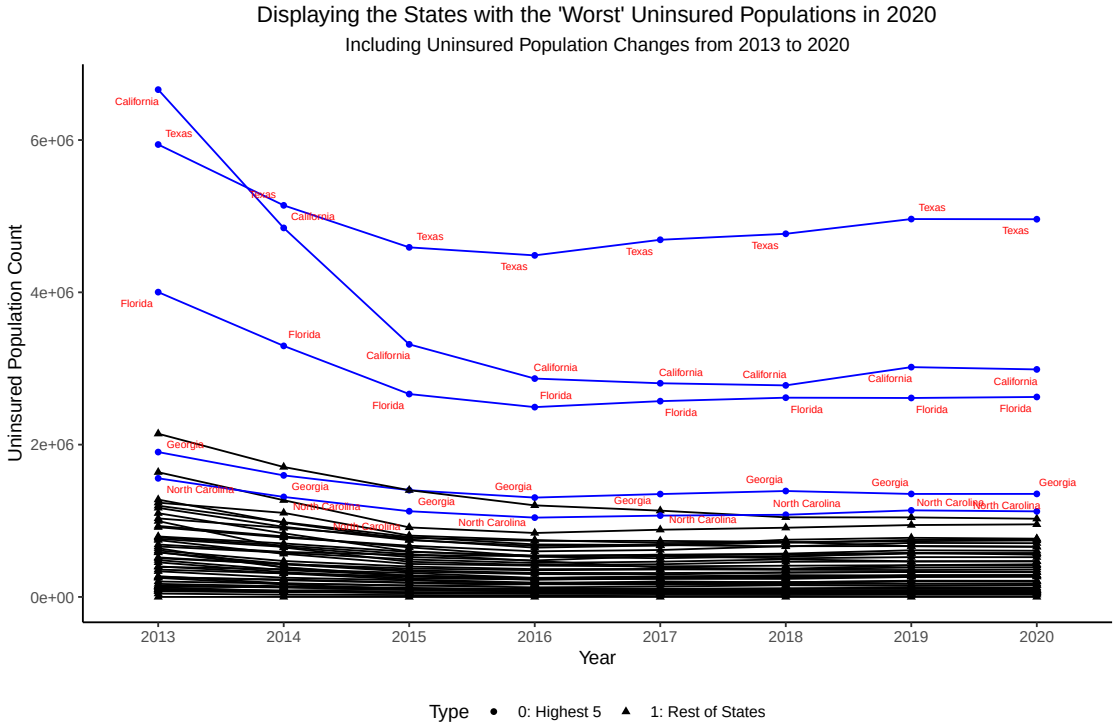
```
medicaid_expansion %>%
  filter(year == 2008) %>% arrange(desc(uninsured_rate)) %>%
  select(State) %>% head(1) %>% pull() %>%
  print()
```

```
## [1] "Utah"
```

- Which states were home to most uninsured Americans prior to 2014? How about in the last year in the data set? **Note**: 2010 state population is provided as a variable to answer this question. In an actual study you would likely use population estimates over time, but to simplify you can assume these numbers stay about the same.



California, Texas, Florida, New York, and Georgia were the states throughout the 2008-2013 period that had the largest number of uninsured Americans.



In the last year, Texas, California, Florida, Georgia, and North Carolina are the 5 states with the largest number of uninsured Americans. New York now leaves the top 5 "worst" and North Carolina joins.

Difference-in-Differences Estimation

Estimate Model

Do the following:

- Choose a state that adopted the Medicaid expansion on January 1, 2014 and a state that did not. **Hint**: Do not pick Massachusetts as it passed a universal healthcare law in 2006, and also avoid picking a state that adopted the Medicaid expansion between 2014 and 2015.
- Assess the parallel trends assumption for your choices using a plot. If you are not satisfied that the assumption has been met, pick another state and try again (but detail the states you tried).

```
unique(medicaid_expansion[which(medicaid_expansion$Date_Adopted=="2014-01-01"),]$State)
```

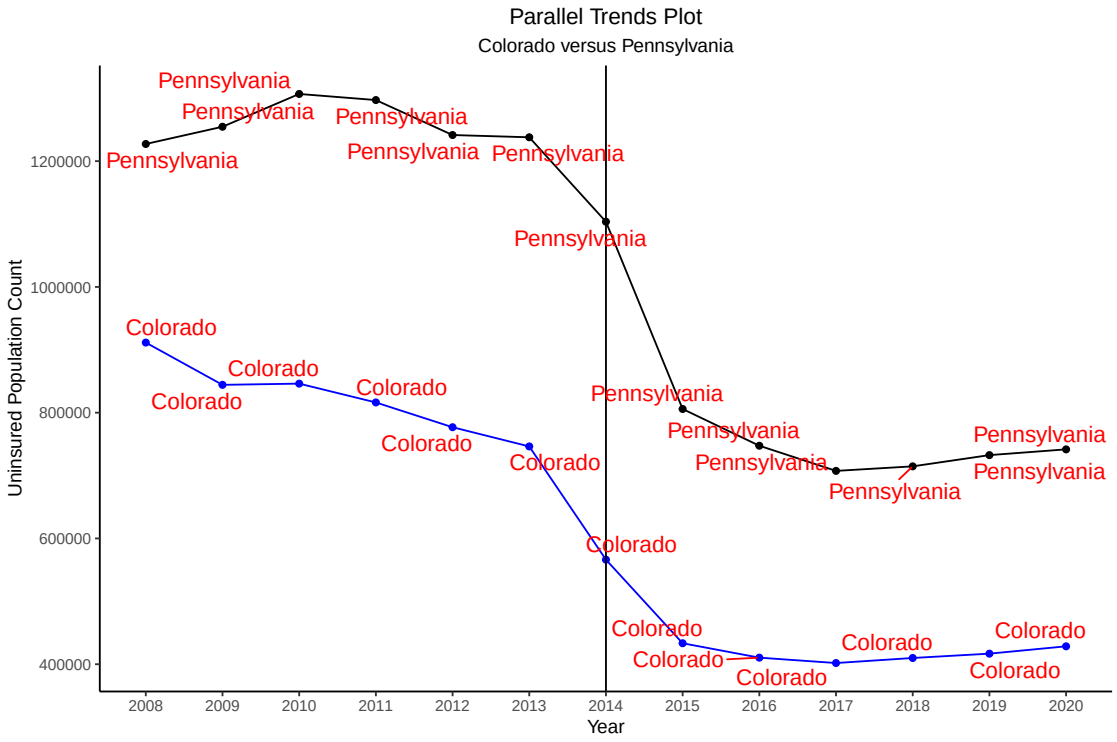
```
##  [1] "Arizona"              "Arkansas"             "California"
##  [4] "Colorado"             "Connecticut"          "Delaware"
##  [7] "District of Columbia" "Hawaii"               "Illinois"
## [10] "Iowa"                 "Kentucky"             "Maryland"
## [13] "Massachusetts"        "Minnesota"            "Nevada"
## [16] "New Jersey"           "New Mexico"           "New York"
## [19] "North Dakota"         "Ohio"                 "Oregon"
## [22] "Rhode Island"         "Vermont"              "Washington"
## [25] "West Virginia"
```

```
unique(medicaid_expansion[which(!medicaid_expansion$Date_Adopted=="2014-01-01"),]$State)
```

```
##  [1] "Alaska"        "Idaho"         "Indiana"       "Louisiana"
##  [5] "Michigan"      "Montana"       "Nebraska"      "New Hampshire"
##  [9] "Pennsylvania"  "Utah"          "Virginia"
```

```
# in-between-date states
unique(medicaid_expansion[which(medicaid_expansion$Date_Adopted > "2014-01-01" &
                                medicaid_expansion$Date_Adopted < "2015-01-01"),]$State)
```

```
## [1] "Michigan"      "New Hampshire"
```



The difference between the treatment and control units appear to be constant for the most part in the pre-treatment period, that is, before 2014. It is really difficult to pick states that meet this trend plot, but I believe these are two good states options. Others considered were the following: Illinois-Virginia -> Illinois-Indiana -> Illinois-Virginia -> Illinois-Pennsylvania (2nd) -> West Virginia-Louisiana -> West Virginia-Idaho -> West Virginia-Montana -> West Virginia-Virginia (4th) -> Colorado-Indiana (3rd) -> Colorado-Pennsylvania (BEST).

- Estimate a difference-in-differences estimate of the effect of the Medicaid expansion on the uninsured share of the population. You may follow the lab example where we estimate the differences in one pre-treatment and one post-treatment period, or take an average of the pre-treatment and post-treatment outcomes

```
# Difference-in-Differences estimation

# Pennsylvania-Colorado
pc <- medicaid_expansion %>%
  filter(State %in% c("Pennsylvania","Colorado")) %>%
  mutate(uninsured_pop = uninsured_rate*population) %>%
  filter(year >= 2013 & year <= 2015)

# pre-treatment difference

pre_diff <- pc %>%
  filter(year == 2013) %>%
  select(State,
         uninsured_pop) %>%
  spread(State,
         uninsured_pop) %>%
  summarise(Colorado - Pennsylvania)
```

```r
# post-treatment difference

post_diff <- pc %>%
  filter(year == 2015) %>%
  select(State,
         uninsured_pop) %>%
  spread(State,
         uninsured_pop) %>%
  summarise(Colorado - Pennsylvania)

# diff-in-diffs

diff_in_diffs <- post_diff - pre_diff
diff_in_diffs
```

```
##   Colorado - Pennsylvania
## 1                119083.6
```

Discussion Questions

- Card/Krueger's original piece utilized the fact that towns on either side of the Delaware river are likely to be quite similar to one another in terms of demographics, economics, etc. Why is that intuition harder to replicate with this data?
- **Answer**: The political history that determines state-level political decision-making on healthcare and Medicaid specifically cannot be applied in the context of this intuition. Further, what determines which populations get access to insurance and which do not makes it difficult to use this intuition as well.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: The strength of the parallel trends assumption is that you can see change across time, helping one to make an informed decision on if the comparative units before treatment is satisfied well. The weakness is that there is no statistic that calculates accuracy of how informative the assumption is met–we simply use our eyes to make this argument, which might introduce bias.

Synthetic Control

Estimate Synthetic Control

Although several states did not expand Medicaid on January 1, 2014, many did later on. In some cases, a Democratic governor was elected and pushed for a state budget that included the Medicaid expansion, whereas in others voters approved expansion via a ballot initiative. The 2018 election was a watershed moment where several Republican-leaning states elected Democratic governors and approved Medicaid expansion. In cases with a ballot initiative, the state legislature and governor still must implement the results via legislation. For instance, Idaho voters approved a Medicaid expansion in the 2018 election, but it was not implemented in the state budget until late 2019, with enrollment beginning in 2020.

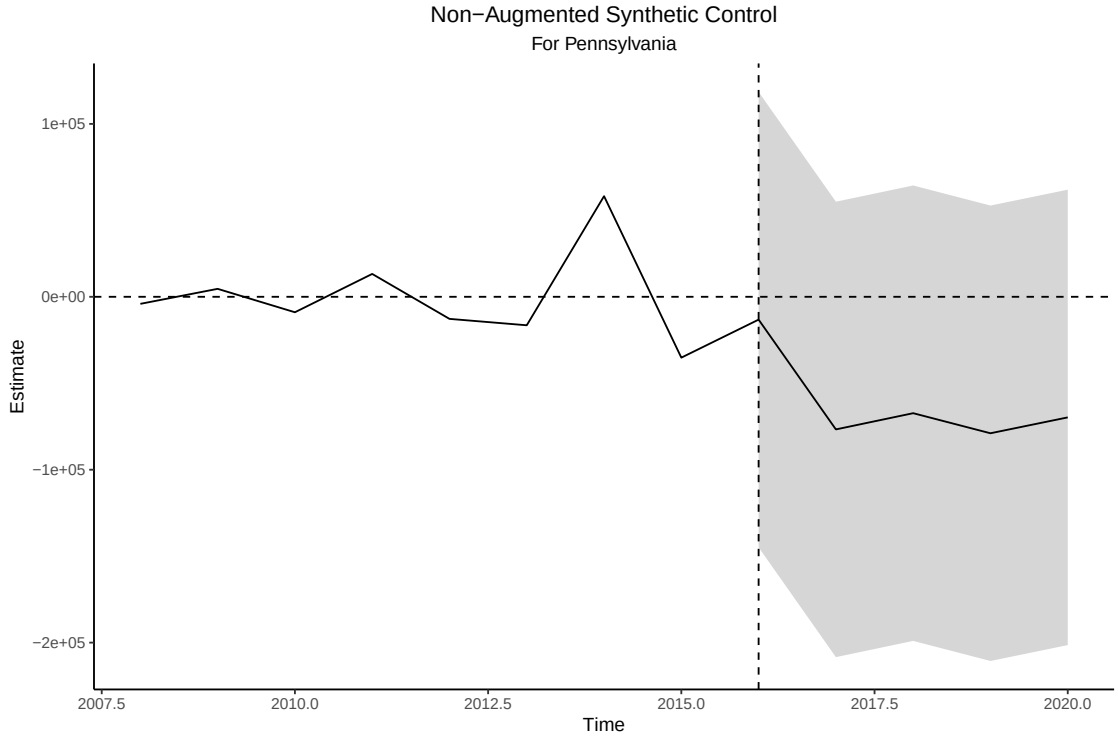
Do the following:

- Choose a state that adopted the Medicaid expansion after January 1, 2014. Construct a non-augmented synthetic control and plot the results (both pre-treatment fit and post-treatment differences). Also report the average ATT and L2 imbalance.

```
unique(medicaid_expansion[which(medicaid_expansion$Date_Adopted>"2014-01-01"),]$State)
```

```
##  [1] "Alaska"        "Idaho"         "Indiana"       "Louisiana"
##  [5] "Michigan"      "Montana"       "Nebraska"      "New Hampshire"
##  [9] "Pennsylvania"  "Utah"          "Virginia"
```

```
## One outcome and one treatment time found. Running single_augsynth.
```




```
print(syn)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "None", scm = ..2)
##
## Average ATT Estimate: -61172.270
```

The average ATT estimate is -61171.846.

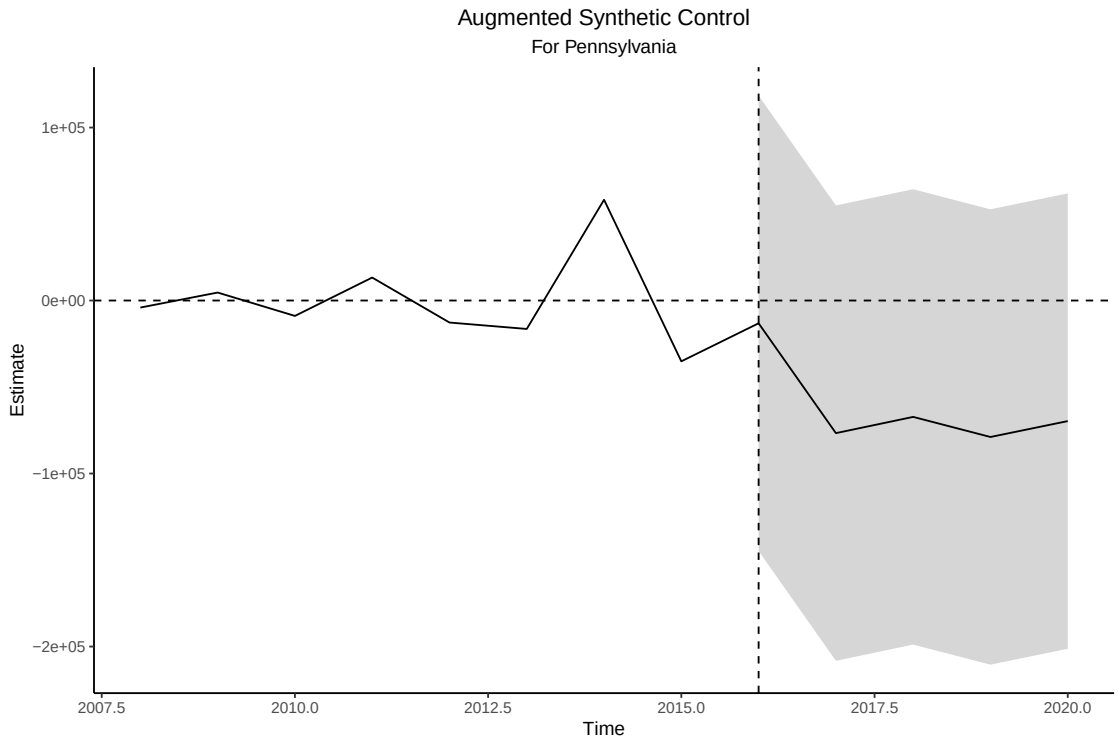
```
print(syn$l2_imbalance)
```

```
## [1] 73161.63
```

The L2 imbalance is 73161.63.

- Re-run the same analysis but this time use an augmentation (default choices are Ridge, Matrix Completion, and GSynth). Create the same plot and report the average ATT and L2 imbalance.

```
## One outcome and one treatment time found. Running single_augsynth.
```




```
print(aug_syn)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "ridge", scm = ..2)
##
## Average ATT Estimate: -61125.700
```

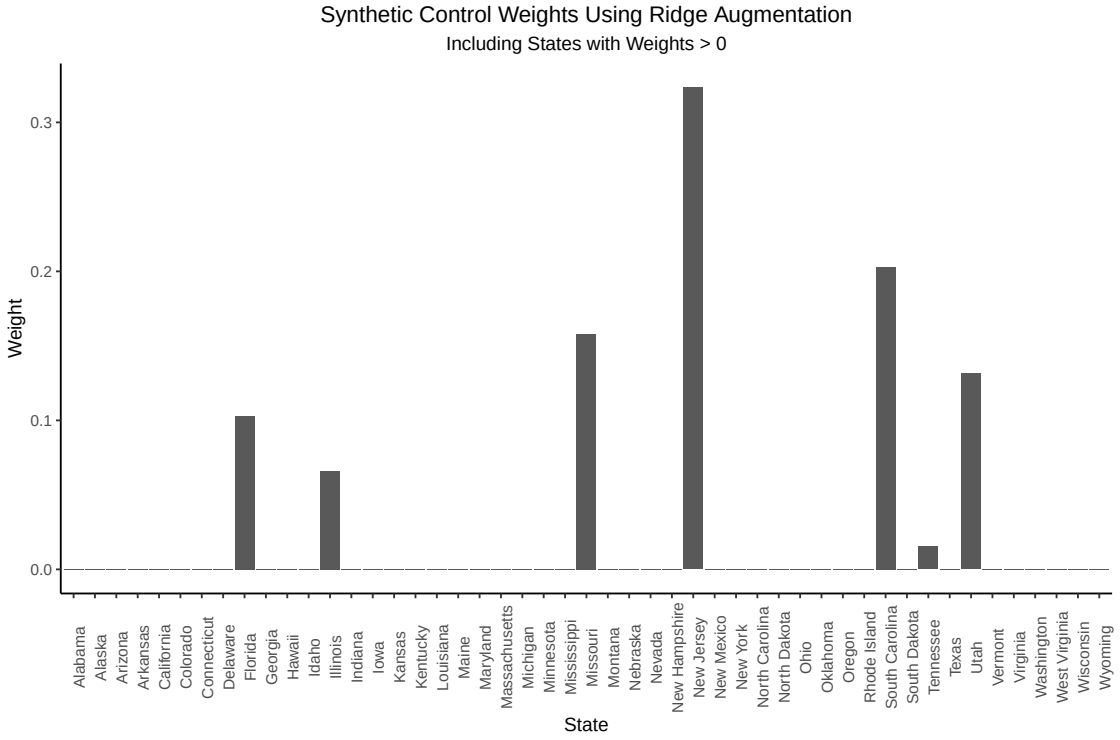
The average ATT estimate is -61126.124, which did not improve as much as without augmentation.

```
print(aug_syn$l2_imbalance)
```

```
## [1] 73158.75
```

The L2 imbalance is 73158.75, which also did not improve as much as without augmentation.

- Plot barplots to visualize the weights of the donors.



Discussion Questions

- What are the advantages and disadvantages of synthetic control compared to difference-in-differences estimators?
- **Answer**: Using diff-in-diff estimators is disadvantaged to synthetic control when it is hard to justify selecting good comparative units. Further, synthetic control has weights bounded between 0 and 1, where they sum to 1, and this is helpful to determine which states contribute most effectively to the synthetic control estimator. However, they can be invalid if pre-treatment balance outcomes are poor.
- One of the benefits of synthetic control is that the weights are bounded between [0,1] and the weights must sum to 1. Augmentation might relax this assumption by allowing for negative weights. Does this create an interpretation problem, and how should we balance this consideration against the improvements augmentation offers in terms of imbalance in the pre-treatment period?
- **Answer**: Yes, this creates an interpretation problem because it doesn't make sense to say that 'X' state contributed "negatively" to the construction of the synthetic control. There should be a penalization term to interpret the improvement.

Staggered Adoption Synthetic Control

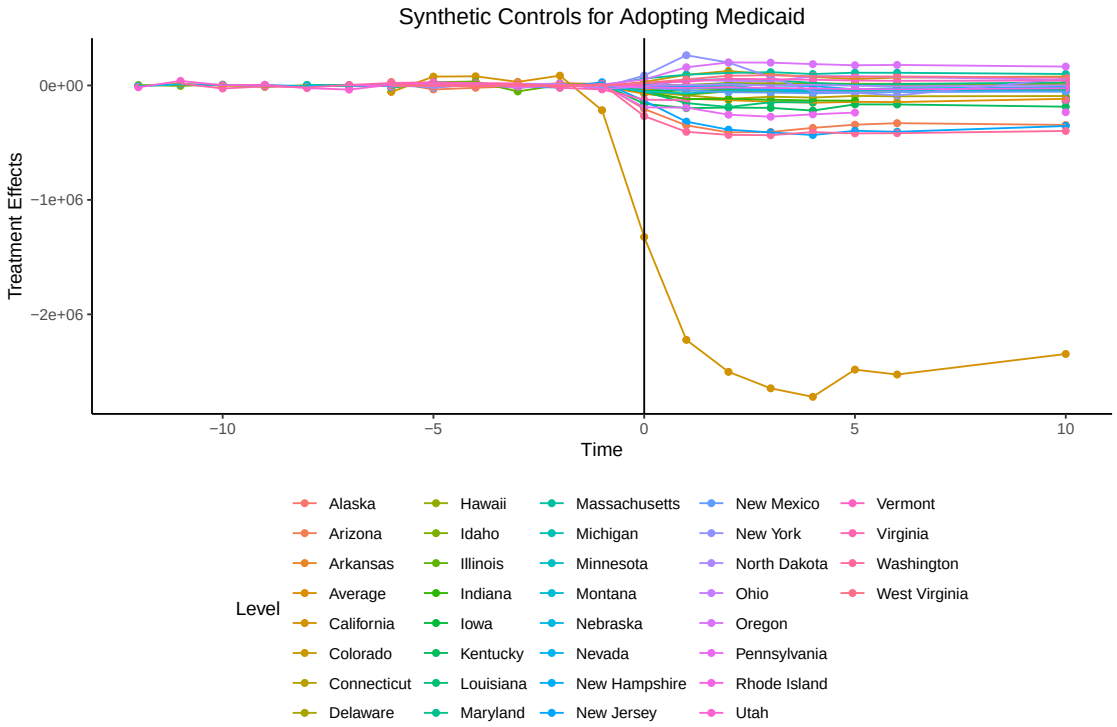
Estimate Multisynth

Do the following:

- Estimate a multisynth model that treats each state individually. Choose a fraction of states that you can fit on a plot and examine their treatment effects.

```
## Warning: Removed 210 rows containing missing values (`geom_point()`).
```

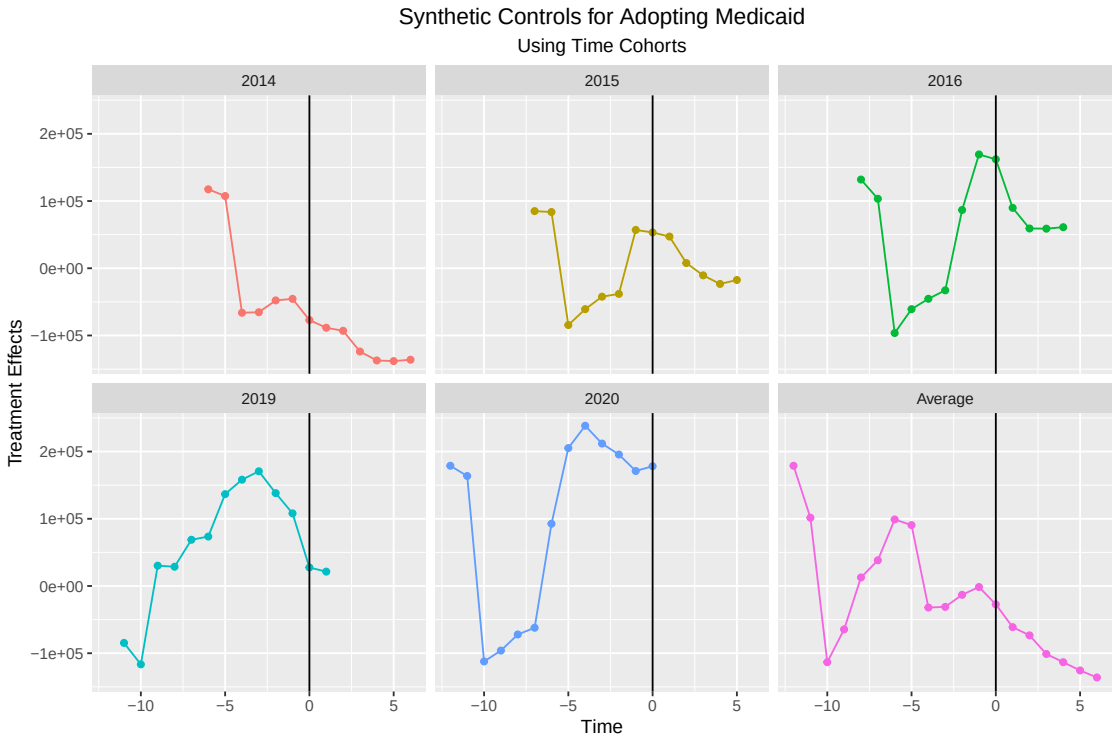
```
## Warning: Removed 180 rows containing missing values (`geom_line()`).
```



- Estimate a multisynth model using time cohorts. For the purpose of this exercise, you can simplify the treatment time so that states that adopted Medicaid expansion within the same year (i.e. all states that adopted epxansion in 2016) count for the same cohort. Plot the treatment effects for these time cohorts.

```
## Warning: Removed 36 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 36 rows containing missing values (`geom_line()`).
```



Discussion Questions

- One feature of Medicaid is that it is jointly administered by the federal government and the states, and states have some flexibility in how they implement Medicaid. For example, during the Trump administration, several states applied for waivers where they could add work requirements to the eligibility standards (i.e. an individual needed to work for 80 hours/month to qualify for Medicaid). Given these differences, do you see evidence for the idea that different states had different treatment effect sizes?
- **Answer**: For the most part, we see that California differs much more than any other states, so yes there is evidence that different states had different treatment effect sizes.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: Yes, focusing on the post-treatment area (right of vertical line at 0), the time cohorts that contribute the most tend to be the ones from later cohorts, therefore the earlier adopters had a larger decrease in the uninsured population.

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?
- **Answer**: Because these units have comparative units perfect for identifying the counterfactual.
- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?

- **Answer**: The whole point of regression discontinuity is to eliminate selection bias, while this can be a problem in DiD/synthetic control. Use regression discontinuity of the variation in treatment assignment at the cutoff is random, but you lose generalizability in the identification of those near the cutoff. Don't use it and condider DiD/synthetic control when the cutoff is not an issue.