Project 7: Difference-in-Differences and Synthetic Control

```r
gc(); rm(list=ls())
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 473090 25.3    1021437 54.6   660497 35.3
## Vcells 880706  6.8    8388608 64.0  1770414 13.6
```

```r
# Install and load packages
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```r
devtools::install_github("ebenmichael/augsynth")
```

```
## Using GitHub PAT from the git credential store.
```

```
## Skipping install of 'augsynth' from a github remote, the SHA1 (0f4f1bcc) has not changed since last i
##   Use `force = TRUE` to force installation
```

```r
options(repos = c(CRAN = "https://cloud.r-project.org"))

pacman::p_load(# Tidyverse packages including dplyr and ggplot2
               tidyverse,
               ggthemes,
               augsynth,
               gsynth)

# set seed
set.seed(44)

# load data
medicaid_expansion <- read_csv('medicaid_expansion.csv')
```

```
## Rows: 663 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): State
## dbl  (3): year, uninsured_rate, population
## date (1): Date_Adopted
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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.

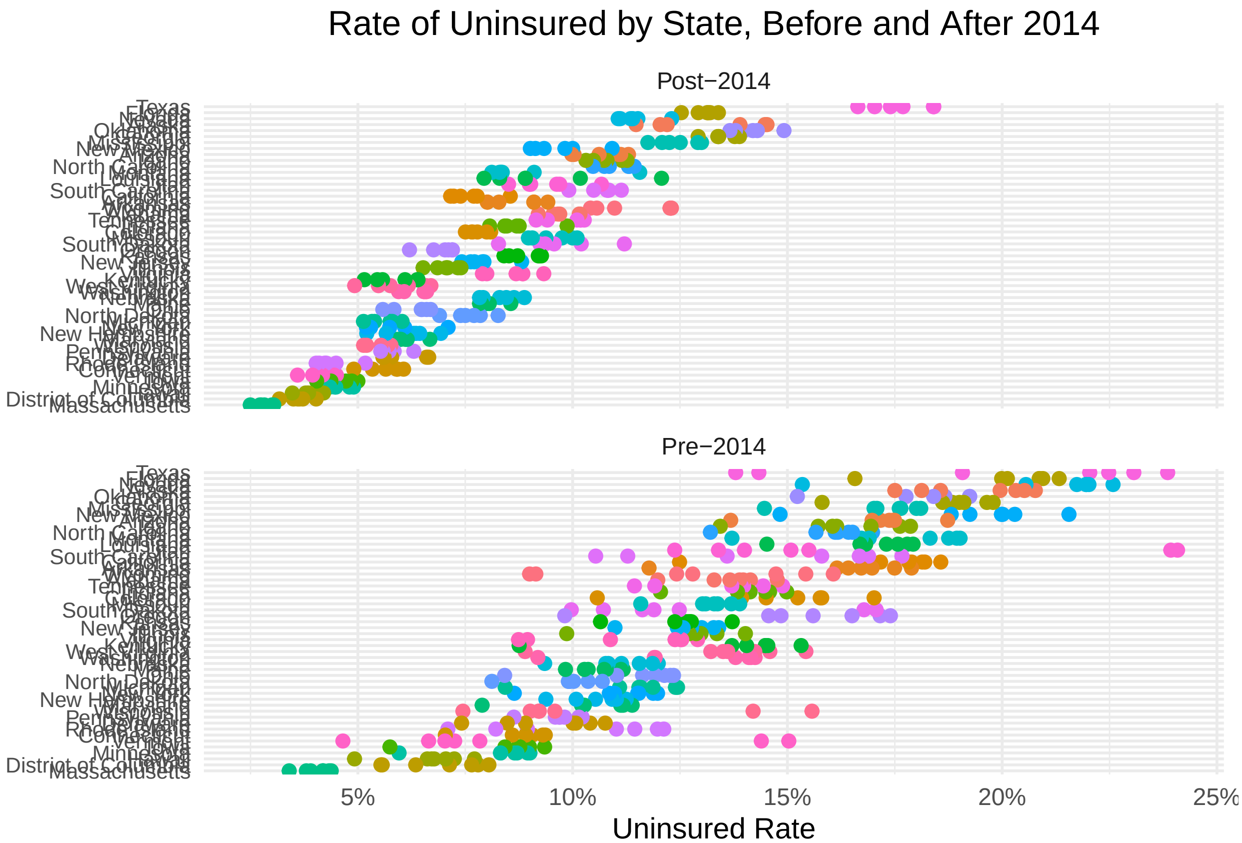
Exploratory Data Analysis

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?

- 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.

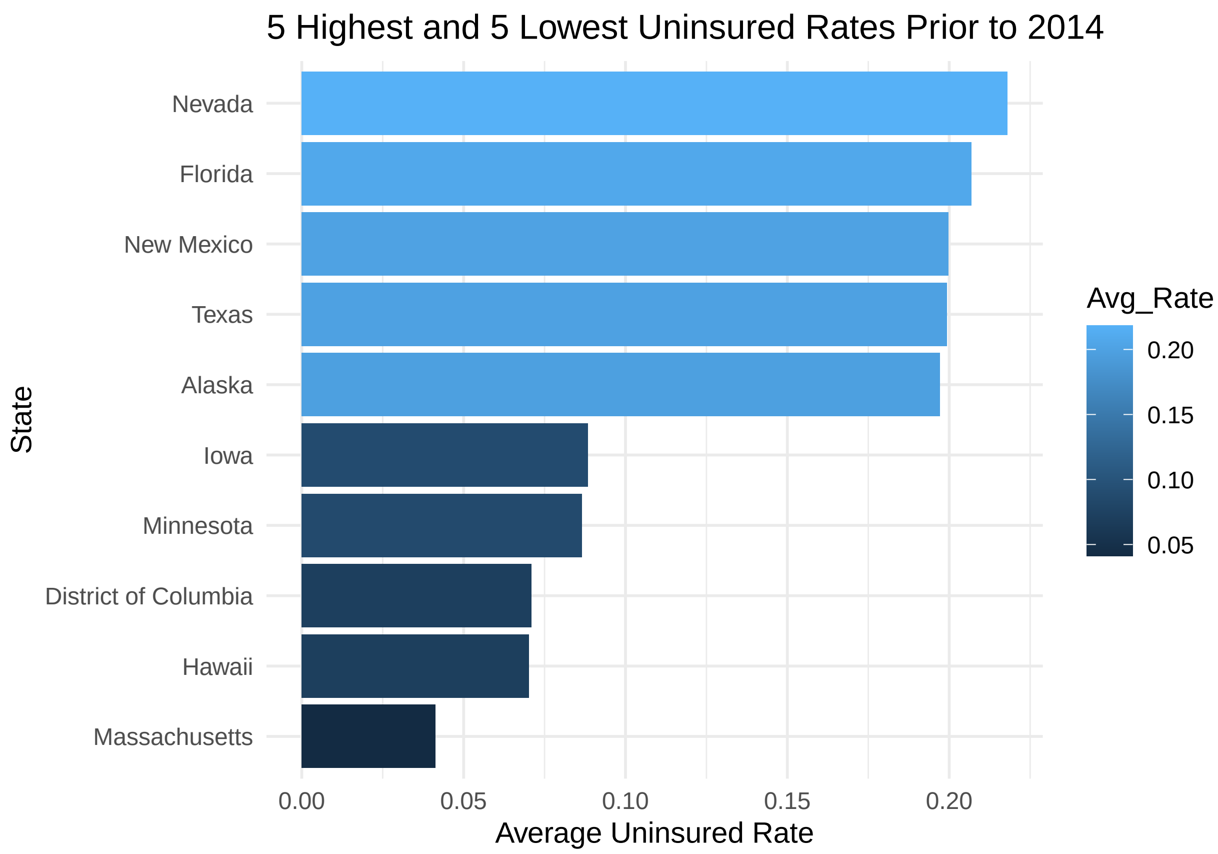
```r
# all states
medicaid_expansion <- medicaid_expansion %>%
  mutate(Period = ifelse(year > 2014, "Post-2014", "Pre-2014"))

medicaid_expansion %>%
  arrange(desc(uninsured_rate)) %>%
  ggplot(aes(x = uninsured_rate, y = reorder(State, uninsured_rate), color = State)) +
  geom_point(size = 2) +
  labs(
    title = 'Rate of Uninsured by State, Before and After 2014',
    x = 'Uninsured Rate',
    y = 'State'
  ) +
  facet_wrap(~Period, scales = "free_y", ncol = 1) +
  theme_minimal() +
  theme(
    legend.position = 'none',
    axis.text.y = element_text(size = 8),
    axis.title.y = element_blank(),
    plot.title = element_text(hjust = 0.5)
  ) +
  scale_x_continuous(labels = scales::percent)
```




```r
# highest and lowest uninsured rates
avg_uninsured_rate <- medicaid_expansion %>%
  filter(year < 2014) %>%
  group_by(State) %>%
  mutate(num_uninsured = (uninsured_rate / 100) * population) %>%
  summarise(Avg_Rate = mean(uninsured_rate, na.rm = TRUE),
            Avg_num = mean(num_uninsured, na.rm = TRUE)) %>%
  arrange(desc(Avg_Rate))
avg_uninsured_rate
```

```
## # A tibble: 51 x 3
##    State      Avg_Rate Avg_num
##    <chr>         <dbl>   <dbl>
##  1 Nevada        0.218   6190.
##  2 Florida       0.207  41151.
##  3 New Mexico    0.200   4168.
##  4 Texas         0.199  53721.
##  5 Alaska        0.197   1454.
##  6 Georgia       0.192  19353.
##  7 Oklahoma      0.187   7237.
##  8 Montana       0.181   1854.
##  9 California    0.180  69708.
## 10 Utah          0.177   5198.
## # i 41 more rows
```

```r
# Plotting 5 highest and 5 lowest uninsured rates
ggplot(avg_uninsured_rate[c(1:5,47:51), ], aes(x = reorder(State, Avg_Rate), y = Avg_Rate, fill = Avg_Ra
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "5 Highest and 5 Lowest Uninsured Rates Prior to 2014", x = "State", y = "Average Uninsu
  theme_minimal()
```




```r
# most uninsured Americans
Num_uninsured_10  <- medicaid_expansion %>%
  mutate(num_uninsured = (uninsured_rate / 100) * population) %>%
  filter(year == 2010) %>%
  arrange(desc(num_uninsured))
Num_uninsured_10
```

```
## # A tibble: 51 x 7
##    State        Date_Adopted  year uninsured_rate population Period num_uninsured
##    <chr>        <date>       <dbl>          <dbl>      <dbl> <chr>          <dbl>
##  1 California   2014-01-01    2010          0.186   38802500 Pre-2~        72060.
##  2 Texas        NA            2010          0.239   26956958 Pre-2~        64301.
##  3 Florida      NA            2010          0.213   19893297 Pre-2~        42427.
##  4 New York     2014-01-01    2010          0.120   19746227 Pre-2~        23649.
##  5 Georgia      NA            2010          0.198   10097343 Pre-2~        19983.
##  6 Illinois     2014-01-01    2010          0.140   12880580 Pre-2~        18062.
##  7 North Caro~  NA            2010          0.170    9943964 Pre-2~        16886.
##  8 Ohio         2014-01-01    2010          0.123   11594163 Pre-2~        14316.
```

```
## 9 Pennsylvan~ 2015-01-01    2010          0.102   12787209 Pre-2~        13068.
## 10 Michigan    2014-04-01    2010          0.124    9909877 Pre-2~        12330.
## # i 41 more rows
```

```
Num_uninsured_20  <- medicaid_expansion %>%
  mutate(num_uninsured = (uninsured_rate / 100) * population) %>%
  filter(year == 2020) %>%
  arrange(desc(num_uninsured))
Num_uninsured_20
```

```
## # A tibble: 51 x 7
##    State       Date_Adopted  year uninsured_rate population Period num_uninsured
##    <chr>       <date>       <dbl>          <dbl>      <dbl> <chr>          <dbl>
##  1 Texas       NA            2020          0.184   26956958 Post--         49601.
##  2 California  2014-01-01    2020          0.077   38802500 Post--         29878.
##  3 Florida     NA            2020          0.132   19893297 Post--         26259.
##  4 Georgia     NA            2020          0.134   10097343 Post--         13530.
##  5 North Caro~ NA            2020          0.113    9943964 Post--         11237.
##  6 New York    2014-01-01    2020          0.052   19746227 Post--         10268.
##  7 Illinois    2014-01-01    2020          0.074   12880580 Post--          9532.
##  8 Ohio        2014-01-01    2020          0.066   11594163 Post--          7652.
##  9 Arizona     2014-01-01    2020          0.113    6731484 Post--          7607.
## 10 Pennsylvan~ 2015-01-01    2020          0.058   12787209 Post--          7417.
## # i 41 more rows
```

Difference-in-Differences Estimation

Estimate Model

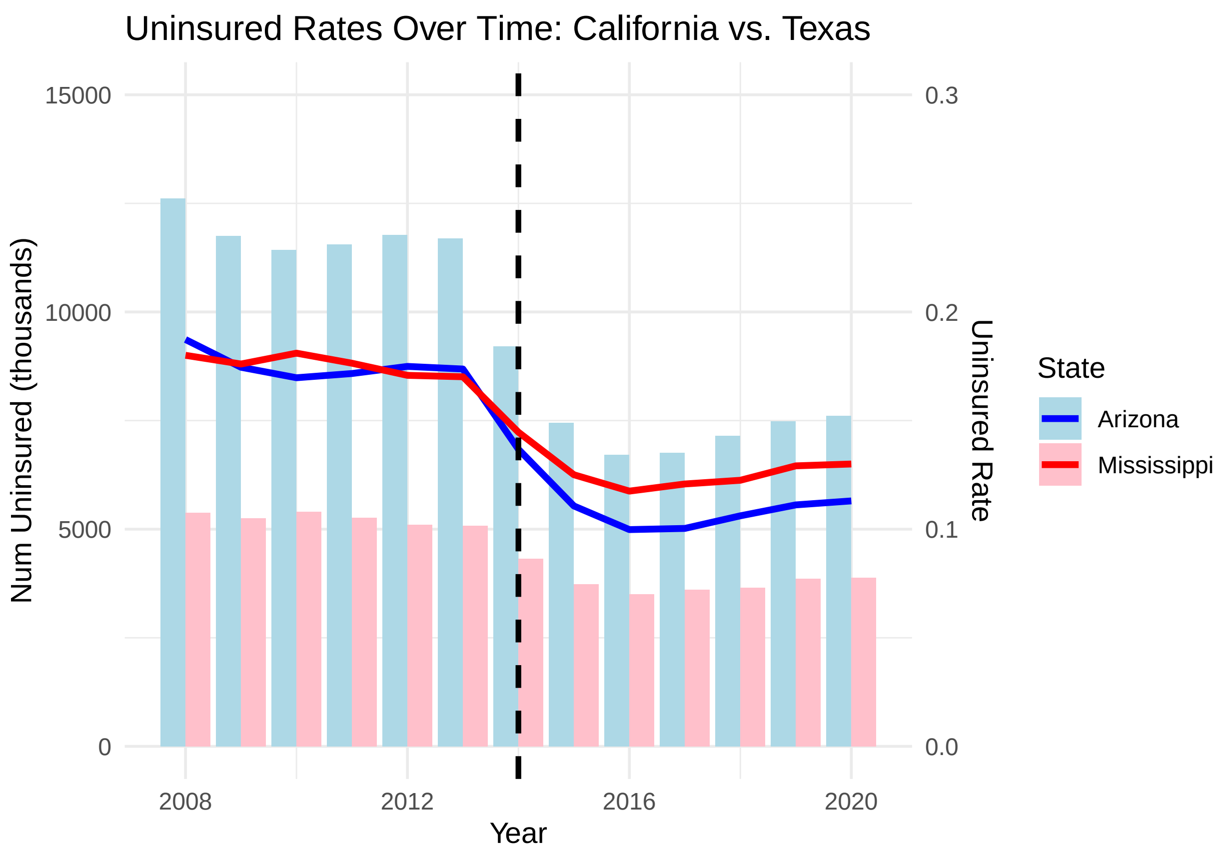
Do the following:

- 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).

```
no_HE <- avg_uninsured_rate %>%
  filter(State %in% c("Wisconsin", "Wyoming", "Kansas", "Texas", "Tennessee",
                      "Mississippi", "Alabama", "Georgia", "South Carolina", "Florida")) %>%
  arrange(desc(Avg_Rate))

yes14_HE <- avg_uninsured_rate %>%
  filter(State %in% c("Washington", "Oregon", "California", "Nevada", "Arizona",
                      "New Mexico", "Colorado", "North Dakota", "Arkansas", "Iowa",
                      "Minnesota", "Illinois", "Michigan", "Indiana", "Kentucky",
                      "Ohio", "New York", "Pennsylvania")) %>%
  arrange(desc(Avg_Rate))
```

```r
# Parallel Trends plot
az_ms_data <- medicaid_expansion %>%
  mutate(num_uninsured = (uninsured_rate / 100) * population) %>%
  filter(State %in% c("Arizona","Mississippi")) %>%
  select(State, year, num_uninsured, uninsured_rate)


ggplot(az_ms_data, aes(x = year)) +
  geom_col(aes(y = num_uninsured, fill = State), position = position_dodge(), show.legend = TRUE) +
  geom_line(aes(y = uninsured_rate*50000, group = State, color = State), size = 1.2, show.legend = TRUE)
  scale_y_continuous(
    name = "Num Uninsured (thousands)",
    limits = c(0, 15000),  # Adjust based on your data range
    sec.axis = sec_axis(~ ./50000, name = "Uninsured Rate")  # Use a simple transformation if scaling i
  ) +
  labs(title = "Uninsured Rates Over Time: California vs. Texas",
       x = "Year",
       y = "Uninsured Rate (%)") +
  theme_minimal()+
  scale_color_manual(values = c("Arizona" = "blue", "Mississippi" = "red")) +
  scale_fill_manual(values = c("Arizona" = "lightblue", "Mississippi" = "pink")) +
  geom_vline(xintercept = 2014, linetype = "dashed", color = "black", size = 1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



- Estimates 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 between CA, TX
az_ms_did <- az_ms_data %>%
  mutate(period = ifelse(year < 2014, "pre", "post"),
         treatment = ifelse(State == "Arizona", 1, 0))


# Calculate average uninsured rate for pre and post periods for each group
az_ms_sum <- az_ms_did %>%
  group_by(State, period) %>%
  summarise(avg_uninsured_rate = mean(uninsured_rate, na.rm = TRUE)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'State'. You can override using the
## `.groups` argument.
```

```
# Compute differences
diffs <- az_ms_sum %>%
  spread(key = period, value = avg_uninsured_rate) %>%
  mutate(difference = post - pre)
```

```r
# DiD estimate
did_estimate <- diffs$difference[diffs$State == "Arizona"] - diffs$difference[diffs$State == "Mississip
did_estimate
```

```
## [1] -0.0155223
```

```r
### Difference-in-Differences estimation with all State
did_all <- medicaid_expansion %>%
   mutate(period = ifelse(year < 2014, "pre", "post"),
          num_uninsured = (uninsured_rate / 100) * population) %>%
   filter(State %in% c("Wisconsin", "Wyoming", "Kansas", "Texas", "Tennessee",
                       "Mississippi", "Alabama", "Georgia", "South Carolina",
                       "Florida", "Washington", "Oregon", "California", "Nevada", "Arizona",
                       "New Mexico", "Colorado", "North Dakota", "Arkansas", "Iowa",
                       "Minnesota", "Illinois", "Michigan", "Indiana", "Kentucky",
                       "Ohio", "New York", "Pennsylvania")) %>%
   mutate(treatment = ifelse(State %in% c("Washington", "Oregon", "California", "Nevada", "Arizona",
                       "New Mexico", "Colorado", "North Dakota", "Arkansas", "Iowa",
                       "Minnesota", "Illinois", "Michigan", "Indiana", "Kentucky",
                       "Ohio", "New York", "Pennsylvania"), 1, 0)) %>%
   select(State, uninsured_rate, num_uninsured, period, treatment)

# Calculate average uninsured rate for pre and post periods for each group
did_all_sum <- did_all %>%
   group_by(State, period, treatment) %>%
   summarise(avg_uninsured_rate = mean(uninsured_rate, na.rm = TRUE),
             avg_num_uninsured = mean(num_uninsured, na.rm = TRUE)) %>%
   ungroup()
```

```
## `summarise()` has grouped output by 'State', 'period'. You can override using
## the `.groups` argument.
```

```r
# Compute differences
diffs_all <- did_all_sum %>%
   group_by(State, treatment) %>%
   summarise(diff_rate = avg_uninsured_rate[period == "post"] - avg_uninsured_rate[period == "pre"],
             diff_num = avg_num_uninsured[period == "post"] - avg_num_uninsured[period == "pre"])
```

```
## `summarise()` has grouped output by 'State'. You can override using the
## `.groups` argument.
```

```r
diffs_all
```

```
## # A tibble: 28 x 4
## # Groups:   State [28]
##    State      treatment diff_rate diff_num
##    <chr>          <dbl>     <dbl>    <dbl>
##  1 Alabama            0   -0.0389   -1888.
##  2 Arizona            1   -0.0642   -4320.
##  3 Arkansas           1   -0.0784   -2347.
##  4 California         1   -0.0964  -37393.
```

```
##  5 Colorado          1   -0.0720   -3854.
##  6 Florida           0   -0.0713  -14180.
##  7 Georgia           0   -0.0537   -5424.
##  8 Illinois          1   -0.0571   -7351.
##  9 Indiana           1   -0.0516   -3402.
## 10 Iowa              1   -0.0403   -1251.
## # i 18 more rows
```

```
# DiD estimate
did_rate <- mean(diffs_all$diff_rate[diffs_all$treatment == 1]) - mean(diffs_all$diff_rate[diffs_all$tre
did_num <- mean(diffs_all$diff_num[diffs_all$treatment == 1]) - mean(diffs_all$diff_num[diffs_all$treat

did_rate
```

```
## [1] -0.02304899
```

```
did_num
```

```
## [1] -2171.411
```

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**:
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**:

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.

```
# non-augmented synthetic control
```

- 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.

```
# augmented synthetic control
```

- Plot barplots to visualize the weights of the donors.

```
# barplots of weights
```

HINT: Is there any preprocessing you need to do before you allow the program to automatically find weights for donor states?

Discussion Questions

- What are the advantages and disadvantages of synthetic control compared to difference-in-differences estimators?
- **Answer**:
- 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**:

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.

```
# multisynth model states
```

- 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.

```
# multisynth model time cohorts
```

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**:
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**:

General Discussion Questions

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