Project 7: Difference-in-Differences and Synthetic Control

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 474893 25.4    1026588 54.9   660497 35.3
## Vcells 886911  6.8    8388608 64.0  1770407 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
##   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,
               scales)

# 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.

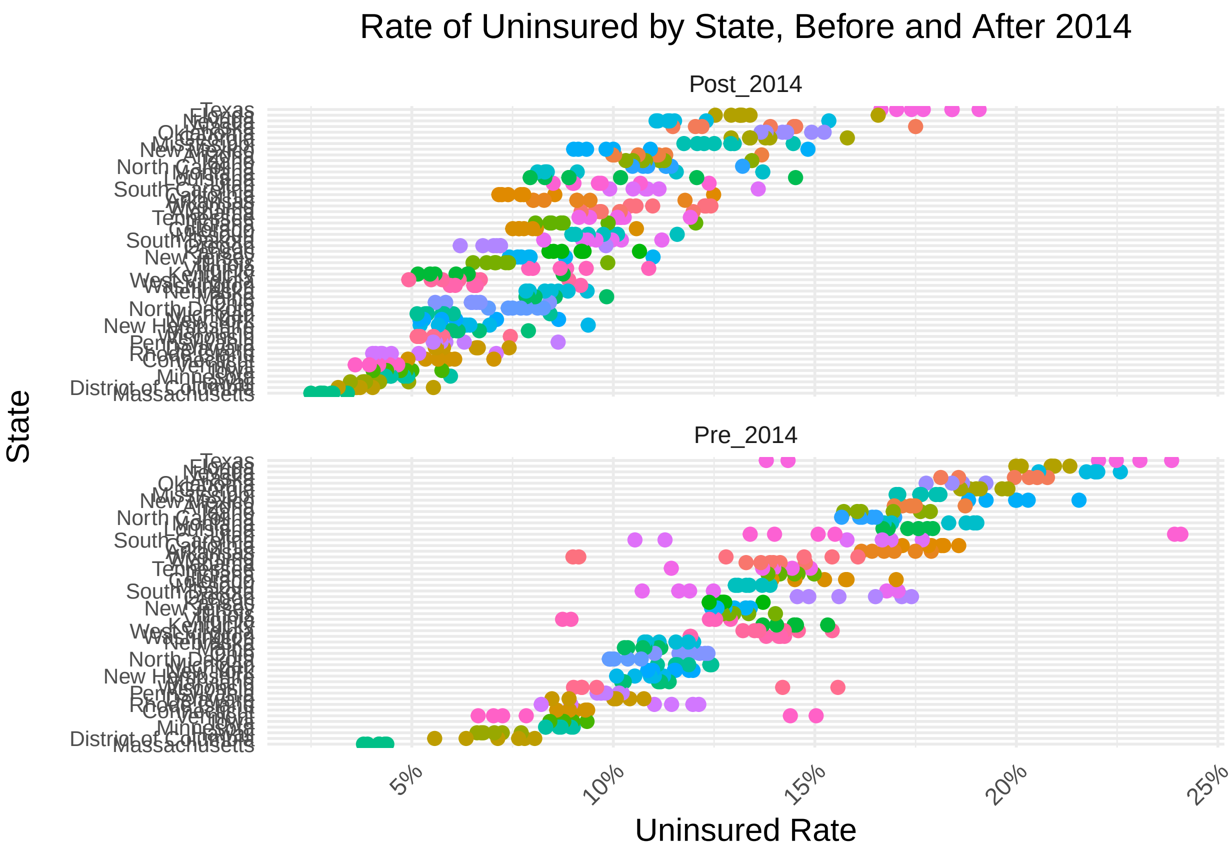
Exploratory Data Analysis

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? **Answer:** The states with the highest uninsured rates pre-2014 were Nevada, Florida, and New Mexico, while the states with the lowest uninsured rates were Massachusetts, Hawaii, and District of Columbia. These states were selected based on their average uninsured rates until 2013.

- 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. **Answer:** In the year of 2010, California, Texas, Florida, New York, and Georgia were the states that had the largest number of uninsured Americans. In the year 2020, Texas, California, Florida, Georgia, and North Carolina are the 5 states with the largest number of uninsured Americans.

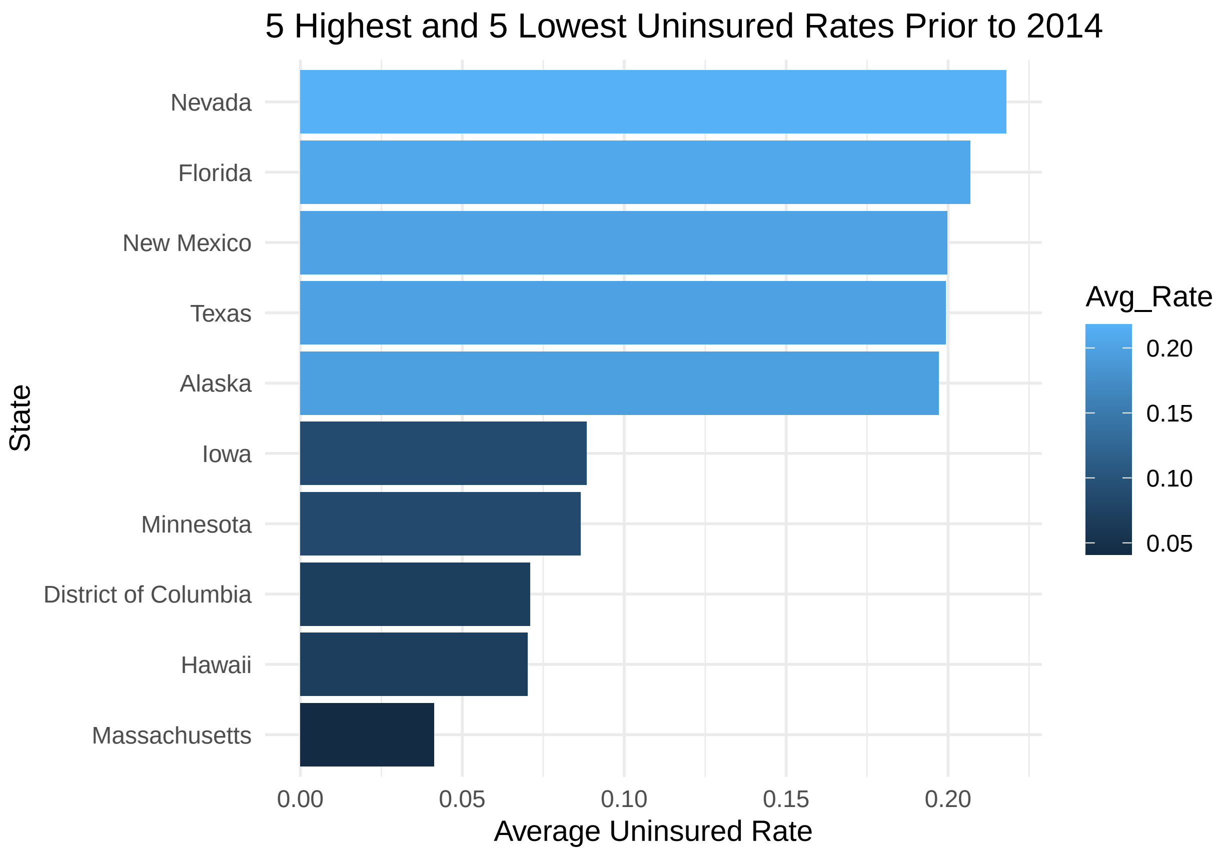
```r
# all states
medicaid_expansion <- medicaid_expansion %>%
  mutate(Period = ifelse(year < 2014, "Pre_2014", "Post_2014"))

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(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(size = 12),
        axis.text.x = element_text(hjust = 1, angle = 45),
        axis.text.y = element_text(size = 8,  margin = margin(t = 10, r = 2, b = 10, l = 10)),
        legend.position = 'None') +
  scale_x_continuous(labels = scales::percent)
```



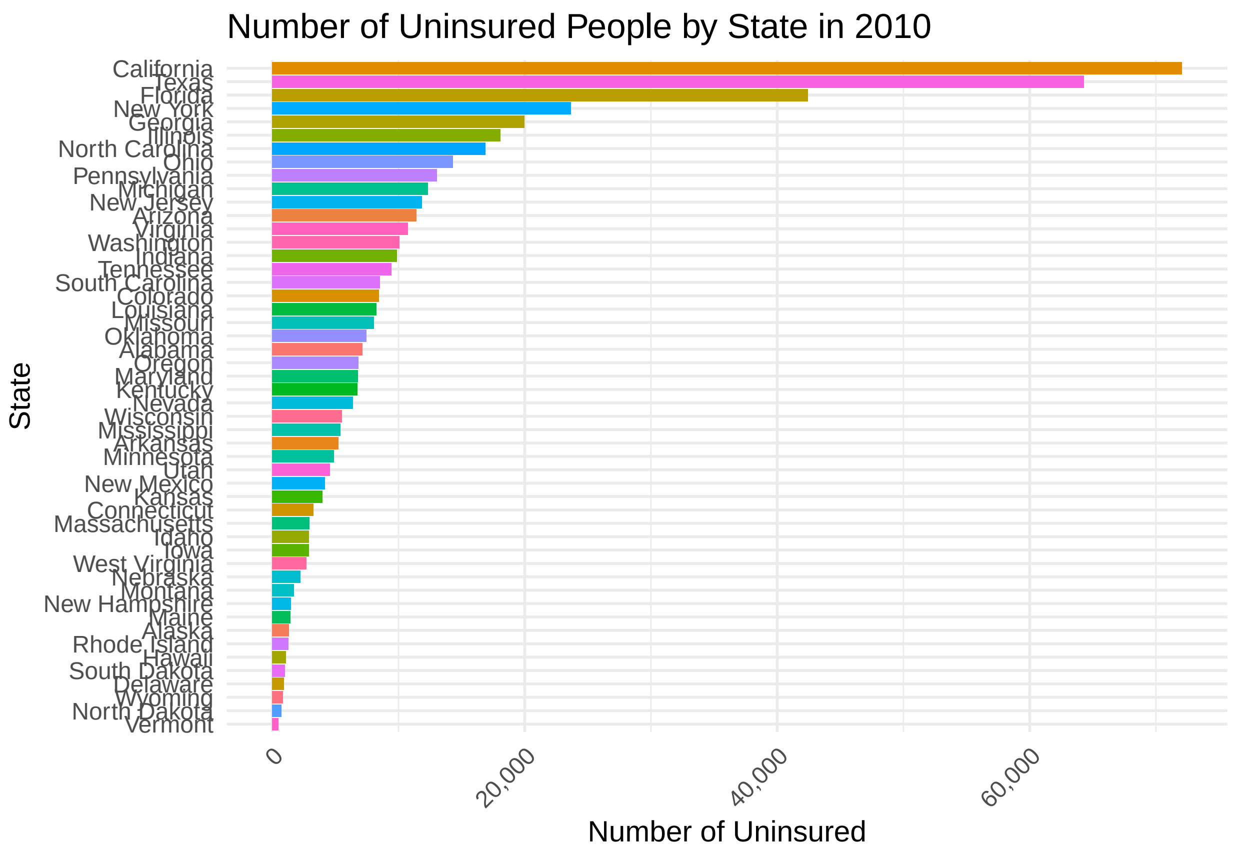
```
# highest and lowest uninsured rates
avg_uninsured_pre <- medicaid_expansion %>%
  filter(year < 2014) %>%
  group_by(State) %>%
  mutate(num_uninsured = uninsured_rate*population/100) %>%
  summarise(Avg_Rate = mean(uninsured_rate, na.rm = TRUE),
            Avg_num = mean(num_uninsured, na.rm =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

```
## # 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
```

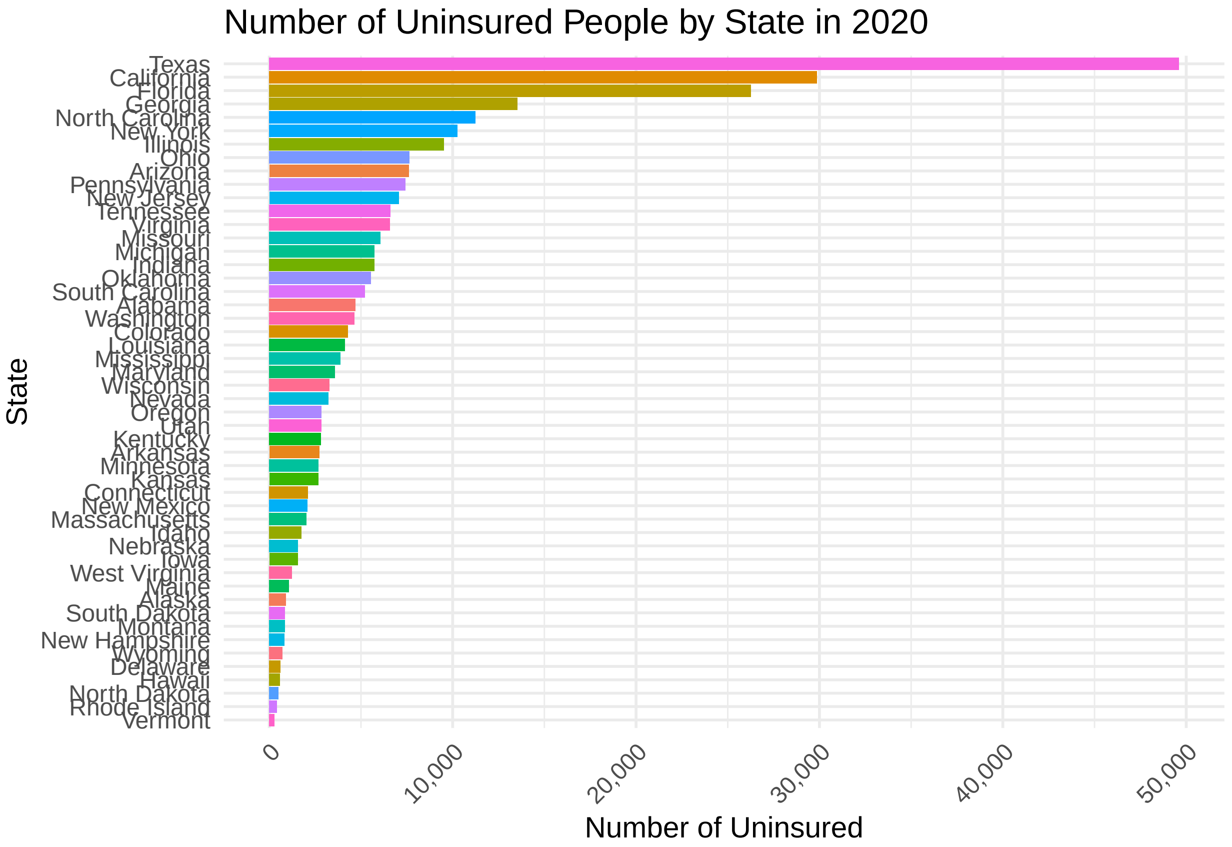
```
# Plotting 5 highest and 5 lowest uninsured rates
ggplot(avg_uninsured_pre[c(1:5,47:51), ], aes(x = reorder(State, Avg_Rate), y = Avg_Rate, fill = Avg_Rat
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "5 Highest and 5 Lowest Uninsured Rates Prior to 2014", x = "State", y = "Average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



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

ggplot(Num_uninsured_10, aes(x = num_uninsured, y = reorder(State, num_uninsured), fill = State)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  scale_x_continuous(labels = label_comma()) +
  labs(x = "Number of Uninsured", y = "State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Number of Uninsured People by State in 2010")
```



```
# most uninsured Americans 2020

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

ggplot(Num_uninsured_20, aes(x = num_uninsured, y = reorder(State, num_uninsured), fill = State)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  scale_x_continuous(labels = label_comma()) +
  labs(x = "Number of Uninsured", y = "State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Number of Uninsured People by State in 2020")
```


Number of Uninsured People by State in 2020

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).

Answer In order to identify two states with similar uninsured rates prior to 2014, the average uninsured rates before 2014, denoted as avg_Rate, were compared. Initially, when comparing the uninsured rates and the number of uninsured individuals (num_uninsured) between Nevada and Florida, it was observed that although their uninsured rates were comparable, there was a significant difference in the actual numbers of uninsured individuals (num_uninsured). Consequently, the average number of uninsured individuals prior to 2014, indicated as Avg_num, was also compared for a more accurate assessment. Based on this comprehensive comparison, Arizona and Mississippi were ultimately selected as the two states with the most similar uninsured statistics before 2014.

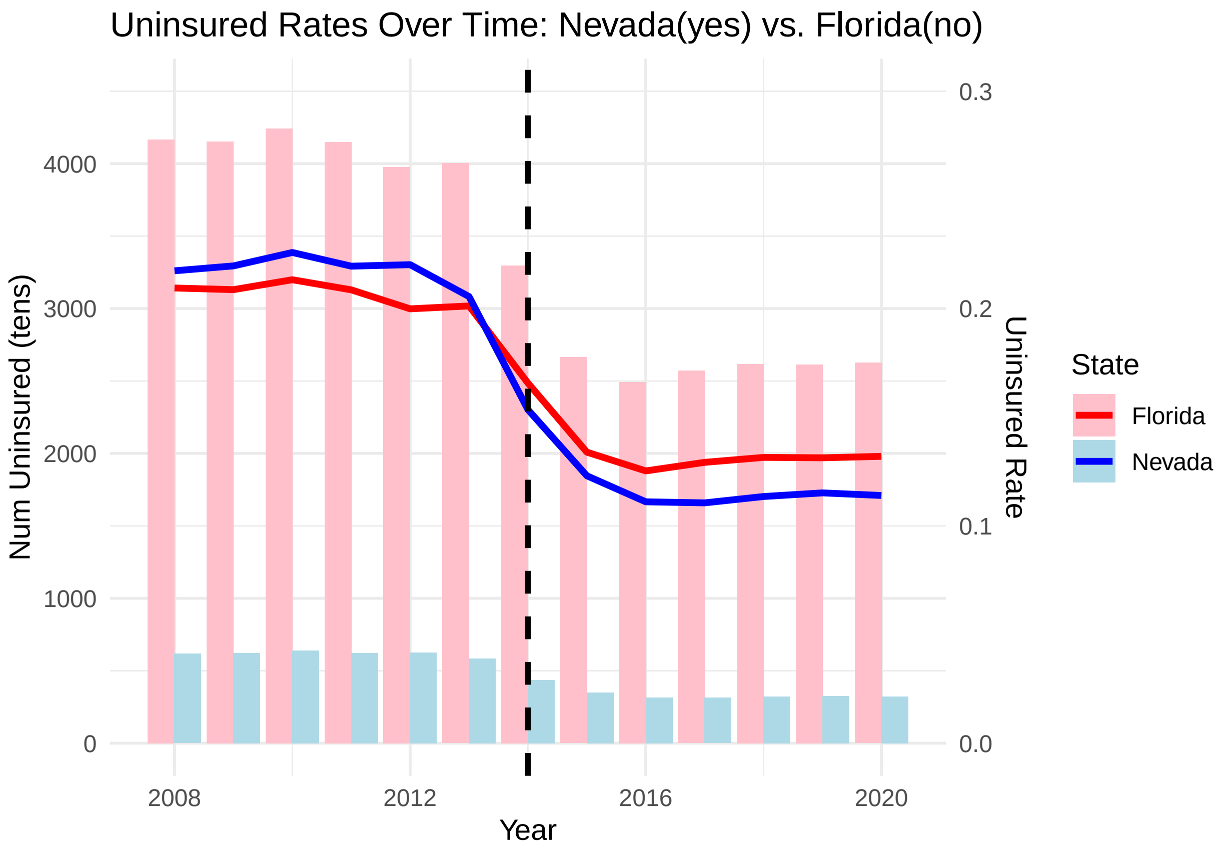
```r
adopt <- medicaid_expansion %>%
  filter(Date_Adopted == as.Date("2014-01-01")) %>%
  distinct(State) %>%
  pull(State)

adopt_df <- avg_uninsured_pre %>%
  filter(State %in% adopt) %>%
  arrange(desc(Avg_Rate))
adopt_df
```

```
## # A tibble: 25 x 3
##    State         Avg_Rate Avg_num
##    <chr>            <dbl>   <dbl>
##  1 Nevada           0.218   6190.
##  2 New Mexico       0.200   4168.
##  3 California       0.180  69708.
##  4 Arizona          0.175  11801.
##  5 Arkansas         0.169   5072.
##  6 Oregon           0.160   6357.
##  7 Colorado         0.154   8236.
##  8 Kentucky         0.144   6370.
##  9 West Virginia    0.141   2609.
## 10 Washington       0.134   9429.
## # i 15 more rows
```

```
Nadopt <- medicaid_expansion %>%
  filter(is.na(Date_Adopted), ) %>%
  distinct(State) %>%
  pull(State)

Nadopt_df <- avg_uninsured_pre %>%
  filter(State %in% Nadopt) %>%
  arrange(desc(Avg_Rate))
Nadopt_df
```

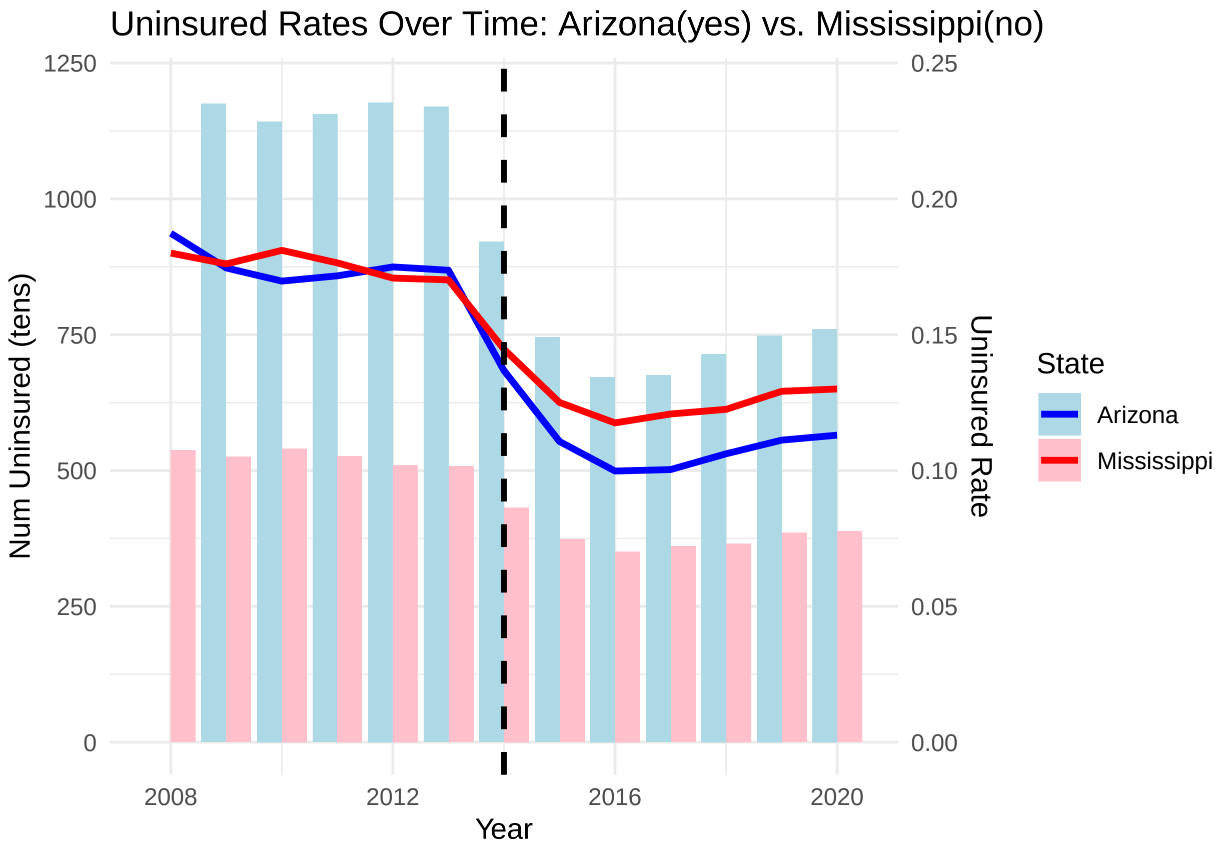
```
## # A tibble: 15 x 3
##    State          Avg_Rate Avg_num
##    <chr>             <dbl>   <dbl>
##  1 Florida           0.207  41151.
##  2 Texas             0.199  53721.
##  3 Georgia           0.192  19353.
##  4 Oklahoma          0.187   7237.
##  5 Mississippi       0.176   5246.
##  6 North Carolina    0.163  16220.
##  7 South Carolina    0.148   7156.
##  8 Alabama           0.140   6766.
##  9 Tennessee         0.137   8992.
## 10 South Dakota      0.134   1146.
## 11 Missouri          0.134   8123.
## 12 Wyoming           0.129    751.
## 13 Kansas            0.128   3724.
## 14 Wisconsin         0.111   6411.
## 15 Maine             0.107   1429.
```

```
# Parallel Trends plot: "Nevada","Florida"
NV_FL_data <- medicaid_expansion %>%
  mutate(num_uninsured = uninsured_rate*population/1000) %>%
  filter(State %in% c("Nevada","Florida")) %>%
  select(State, Date_Adopted, year, num_uninsured, uninsured_rate)


ggplot(NV_FL_data, aes(x = year)) +
  geom_col(aes(y = num_uninsured, fill = State), position = position_dodge(), show.legend = TRUE) +
  geom_line(aes(y = uninsured_rate*15000, group = State, color = State), size = 1.2, show.legend = TRUE
  scale_y_continuous(
    name = "Num Uninsured (tens)",
    limits = c(0, 4500),
    sec.axis = sec_axis(~ ./15000, name = "Uninsured Rate")
  ) +
  labs(title = "Uninsured Rates Over Time: Nevada(yes) vs. Florida(no)",
       x = "Year",
       y = "Uninsured Rate (%)") +
  theme_minimal()+
  scale_color_manual(values = c("Nevada" = "blue", "Florida" = "red")) +
  scale_fill_manual(values = c("Nevada" = "lightblue", "Florida" = "pink")) +
  geom_vline(xintercept = 2014, linetype = "dashed", color = "black", size = 1)
```



```
# Parallel Trends plot: "Arizona","Mississippi"
az_ms_data <- medicaid_expansion %>%
```

```r
  mutate(num_uninsured = uninsured_rate*population/1000) %>%
  filter(State %in% c("Arizona","Mississippi")) %>%
  select(State, Date_Adopted, 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*5000, group = State, color = State), size = 1.2, show.legend = TRUE)
  scale_y_continuous(
    name = "Num Uninsured (tens)",
    limits = c(0, 1200),
    sec.axis = sec_axis(~ ./5000, name = "Uninsured Rate")
  ) +
  labs(title = "Uninsured Rates Over Time: Arizona(yes) vs. Mississippi(no)",
       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)
```



- 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 AZ, MS
az_ms_did <- medicaid_expansion %>%
  mutate(num_uninsured = uninsured_rate * population / 100) %>%
  filter(State %in% c("Arizona", "Mississippi")) %>%
  mutate(period = ifelse(year < 2014, "Pre", "Post")) %>%
  select(State, Date_Adopted, year, num_uninsured, uninsured_rate, period)


# 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),
            avg_uninsured_num = mean(num_uninsured, na.rm = TRUE)) %>%
  ungroup()
```

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

```
az_ms_sum
```

```
## # A tibble: 4 x 4
##   State       period avg_uninsured_rate avg_uninsured_num
##   <chr>       <chr>               <dbl>             <dbl>
## 1 Arizona     Post                0.111             7481.
## 2 Arizona     Pre                 0.175            11801.
## 3 Mississippi Post                0.127             3794.
## 4 Mississippi Pre                 0.176             5246.
```

```
# Define DiD Function
calculate_did <- function(pre_treatment, pre_control, post_treatment, post_control) {

  pre_diff = pre_treatment - pre_control
  post_diff = post_treatment - post_control

  did = post_diff - pre_diff
  return(did)
}

did_rate <- calculate_did(
  pre_treatment = az_ms_sum$avg_uninsured_rate[az_ms_sum$State == "Arizona" & az_ms_sum$period == "Pre"
  pre_control = az_ms_sum$avg_uninsured_rate[az_ms_sum$State == "Mississippi" & az_ms_sum$period == "Pr
  post_treatment = az_ms_sum$avg_uninsured_rate[az_ms_sum$State == "Arizona" & az_ms_sum$period == "Pos
  post_control = az_ms_sum$avg_uninsured_rate[az_ms_sum$State == "Mississippi" & az_ms_sum$period == "P
)

did_num <- calculate_did(
  pre_treatment = az_ms_sum$avg_uninsured_num[az_ms_sum$State == "Arizona" & az_ms_sum$period == "Pre"]
  pre_control = az_ms_sum$avg_uninsured_num[az_ms_sum$State == "Mississippi" & az_ms_sum$period == "Pre
  post_treatment = az_ms_sum$avg_uninsured_num[az_ms_sum$State == "Arizona" & az_ms_sum$period == "Post
  post_control = az_ms_sum$avg_uninsured_num[az_ms_sum$State == "Mississippi" & az_ms_sum$period == "Pos
)
```

```
# DiD estimate
print(paste("Difference-in-Differences for uninsured rate:", did_rate))
```

```
## [1] "Difference-in-Differences for uninsured rate: -0.0155223047619048"
```

```
print(paste("Difference-in-Differences for uninsured number:", did_num))
```

```
## [1] "Difference-in-Differences for uninsured number: -2867.83544557019"
```

```
### Difference-in-Differences estimation with all State
did_all <- medicaid_expansion %>%
   mutate(Period = ifelse(year < 2014, "Pre", "Post"),
          num_uninsured = uninsured_rate*population/100) %>%
   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"), "treatment", "control"))
  select(State, year, 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(Treatment, Period) %>%
   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 'Treatment'. You can override using the
## `.groups` argument.
```

```
did_all_sum
```

```
## # A tibble: 4 x 4
##   Treatment Period avg_uninsured_rate avg_num_uninsured
##   <chr>     <chr>               <dbl>             <dbl>
## 1 control   Post               0.116             11626.
## 2 control   Pre                0.157             15327.
## 3 treatment Post               0.0769             6499.
## 4 treatment Pre                0.141             12371.
```

```
# Compute differences
all_did_rate <- calculate_did(
  pre_treatment = did_all_sum$avg_uninsured_rate[did_all_sum$Treatment == "treatment" & did_all_sum$Peri
  pre_control = did_all_sum$avg_uninsured_rate[did_all_sum$Treatment == "control" & did_all_sum$Period
  post_treatment = did_all_sum$avg_uninsured_rate[did_all_sum$Treatment == "treatment" & did_all_sum$Pe
  post_control = did_all_sum$avg_uninsured_rate[did_all_sum$Treatment == "control" & did_all_sum$Period
)

all_did_num <- calculate_did(
  pre_treatment = did_all_sum$avg_num_uninsured[did_all_sum$Treatment == "treatment" & did_all_sum$Peri
  pre_control = did_all_sum$avg_num_uninsured[did_all_sum$Treatment == "control" & did_all_sum$Period =
  post_treatment = did_all_sum$avg_num_uninsured[did_all_sum$Treatment == "treatment" & did_all_sum$Per
  post_control = did_all_sum$avg_num_uninsured[did_all_sum$Treatment == "control" & did_all_sum$Period
)

# DiD estimate
print(paste("Difference-in-Differences for uninsured rate:", all_did_rate))
```

```
## [1] "Difference-in-Differences for uninsured rate: -0.023048991005291"
```

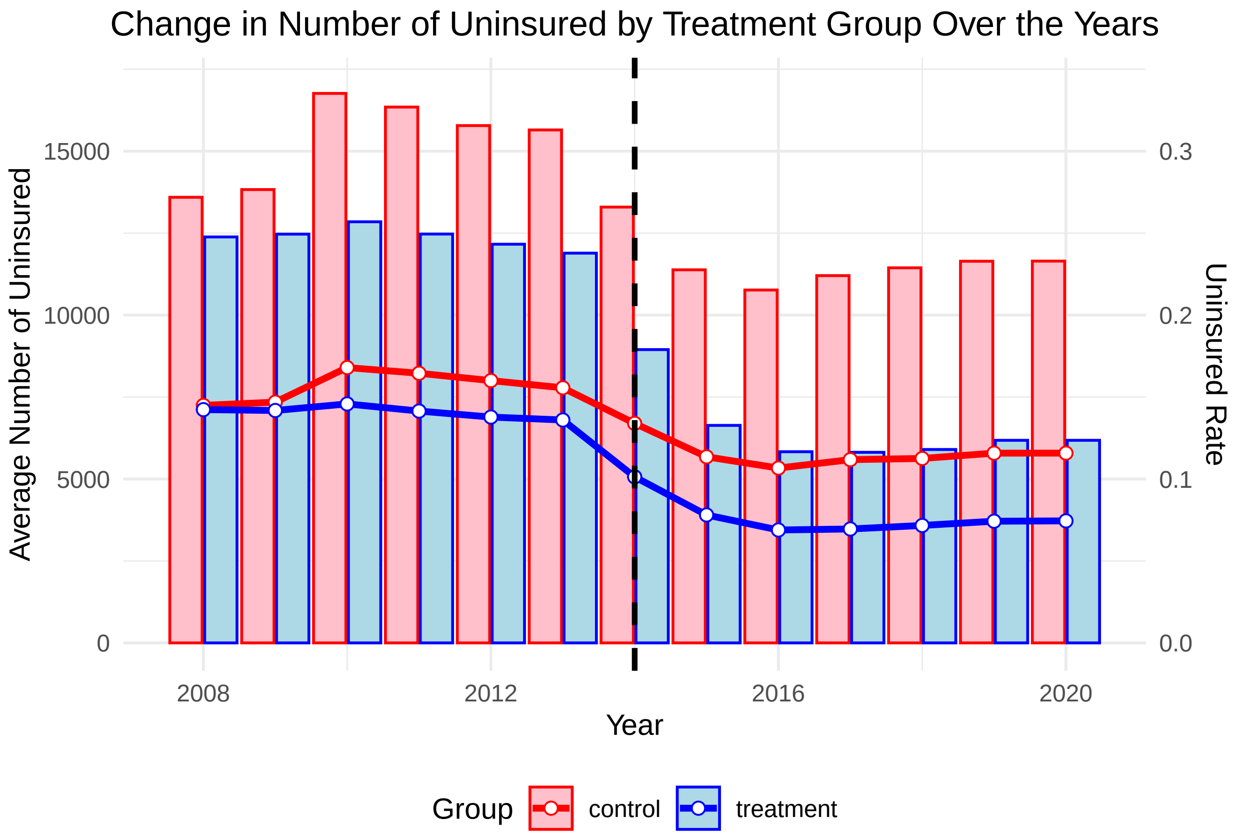
```
print(paste("Difference-in-Differences for uninsured number:", all_did_num))
```

```
## [1] "Difference-in-Differences for uninsured number: -2171.4107338306"
```

```
# Plot
plot_treat <- did_all %>%
  group_by(Treatment, year) %>%
  summarise(avg_uninsured_rate = mean(uninsured_rate),
            avg_num_uninsured = mean(num_uninsured))
```

```
## ‘summarise()‘ has grouped output by ’Treatment’. You can override using the
## ‘.groups‘ argument.
```

```
# Create the line plot
ggplot(plot_treat, aes(x = year, group = Treatment, color = as.factor(Treatment), fill = as.factor(Trea
  geom_col(aes(y = avg_num_uninsured), position = position_dodge(width = 0.97)) +
  geom_line(aes(y = avg_uninsured_rate * 50000), size = 1.2) +
  geom_point(aes(y = avg_uninsured_rate * 50000), size = 2, shape = 21, fill = "white") +
  scale_y_continuous(
    name = "Average Number of Uninsured",
    limits = c(0, 17000),
    sec.axis = sec_axis(~ ./50000, name = "Uninsured Rate")
  ) +
  scale_color_manual(values = c("red", "blue")) +
  scale_fill_manual(values = c("pink", "lightblue")) +
  labs(
    title = "Change in Number of Uninsured by Treatment Group Over the Years",
    x = "Year",
    y = "Average Number of Uninsured",
    color = "Group",
    fill = "Group"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "bottom"
  ) +
  geom_vline(xintercept = 2014, linetype = "dashed", color = "black", size = 1)
```



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**: Even in cities that are geographically close to each other, there can still be significant differences in a variety of factors in terms of demographics, SES, other covariates. Especially, if they are on the other side of a river they can have cultural differences, economic disparity, healthcare infrastructure, and political climate.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: The parallel trends assumption in DiD has some strengths. It can control for unobserved, time-invariant differences between treatment and control groups, simplifying the causal interpretation of changes due to the intervention. However, weaknesses include the challenge in verifying the assumption itself, as actual parallel trends cannot be definitively tested statistically. Additionally, DiD is sensitive to other time-varying factors that could differentially affect the groups, potentially violating the assumption and biasing the results.

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.

```r
# adopted after 2014-01-01
late <- medicaid_expansion %>%
  filter(Date_Adopted > "2014-01-01") %>%
  unique()
late
```

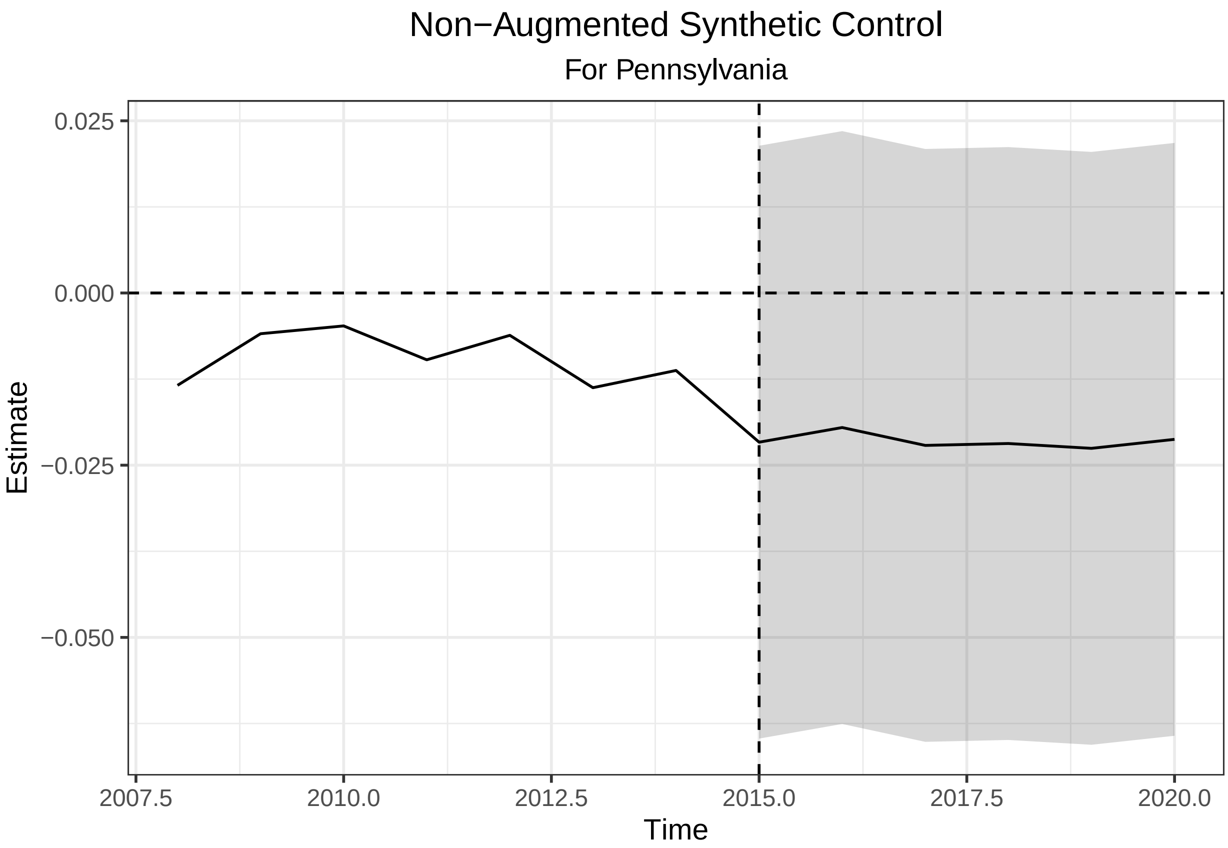
```
## # A tibble: 143 x 6
##    State         Date_Adopted  year uninsured_rate population Period
##    <chr>         <date>       <dbl>          <dbl>      <dbl> <chr>
##  1 Alaska        2015-09-01    2008         0.208      737732 Pre_2014
##  2 Idaho         2020-01-01    2008         0.176     1634464 Pre_2014
##  3 Indiana       2015-02-01    2008         0.138     6596855 Pre_2014
##  4 Louisiana     2016-07-01    2008         0.179     4649676 Pre_2014
##  5 Michigan      2014-04-01    2008         0.115     9909877 Pre_2014
##  6 Montana       2016-01-01    2008         0.189     1023579 Pre_2014
##  7 Nebraska      2020-10-01    2008         0.108     1881503 Pre_2014
##  8 New Hampshire 2014-08-15    2008         0.110     1326813 Pre_2014
##  9 Pennsylvania  2015-01-01    2008         0.0960   12787209 Pre_2014
## 10 Utah          2020-01-01    2008         0.241     2942902 Pre_2014
## # i 133 more rows
```

```r
donor_pool <- medicaid_expansion %>%
  filter(is.na(Date_Adopted)) %>%
  select(State) %>%
  distinct(State) %>%
  pull(State)




# non-augmented synthetic control
aug_df <- medicaid_expansion %>%
  mutate(num_uninsured = uninsured_rate*population/100) %>%
  filter(State=="Pennsylvania"|State %in% donor_pool) %>%
  mutate(Treatment = ifelse(lubridate::year(Date_Adopted) > year | is.na(Date_Adopted), 0, 1)) %>%
  filter(!is.na(uninsured_rate))
```

```
syn <- augsynth(uninsured_rate ~ Treatment, State, year, aug_df,
                progfunc = "None", scm = T)
```

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

```
plot(syn) +
  labs(title = "Non-Augmented Synthetic Control",
       subtitle = "For Pennsylvania") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



```
# save the l2 imbalance and ATT estimates
syn_summary <- summary(syn)

syn_ATT <- syn_summary[["average_att"]][["Estimate"]] %>% round(5)
syn_L2 <- syn_summary[["l2_imbalance"]] %>% round(5)

print(paste("The average ATT estimate:", syn_ATT))
```

```
## [1] "The average ATT estimate: -0.0215"
```

```
print(paste("The L2 imbalance:", syn_L2))
```

```
## [1] "The L2 imbalance: 0.02617"
```

- 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
re_syn <- augsynth(uninsured_rate ~ Treatment, State, year, aug_df,
                   progfunc = "Ridge", scm = T)
```

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

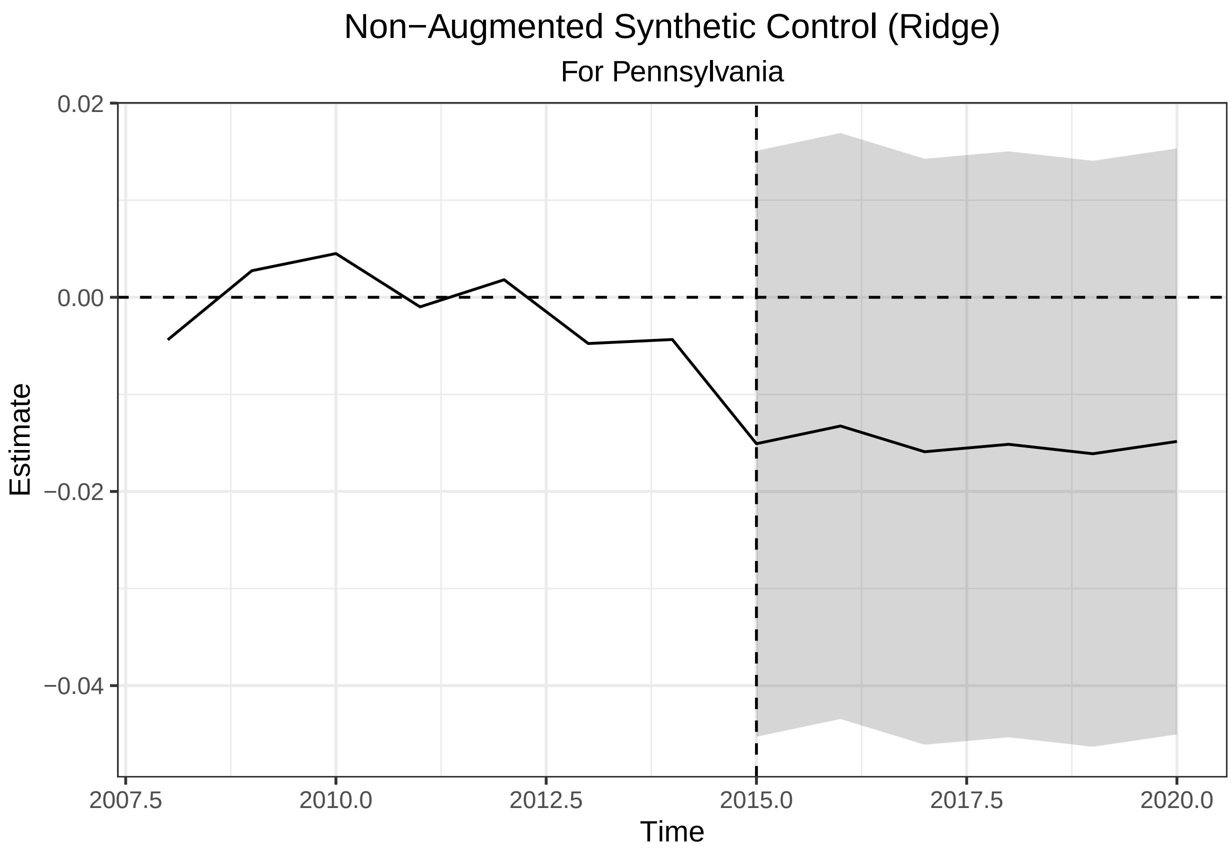
```
# save the l2 imbalance and ATT estimates
resyn_summary <- summary(re_syn)

resyn_ATT <- resyn_summary[["average_att"]][["Estimate"]] %>% round(5)
resyn_L2 <- resyn_summary[["l2_imbalance"]] %>% round(5)

print(paste("The average ATT estimate:", resyn_ATT))
```

```
## [1] "The average ATT estimate: -0.01506"
```

```
print(paste("The L2 imbalance:", resyn_L2))
```

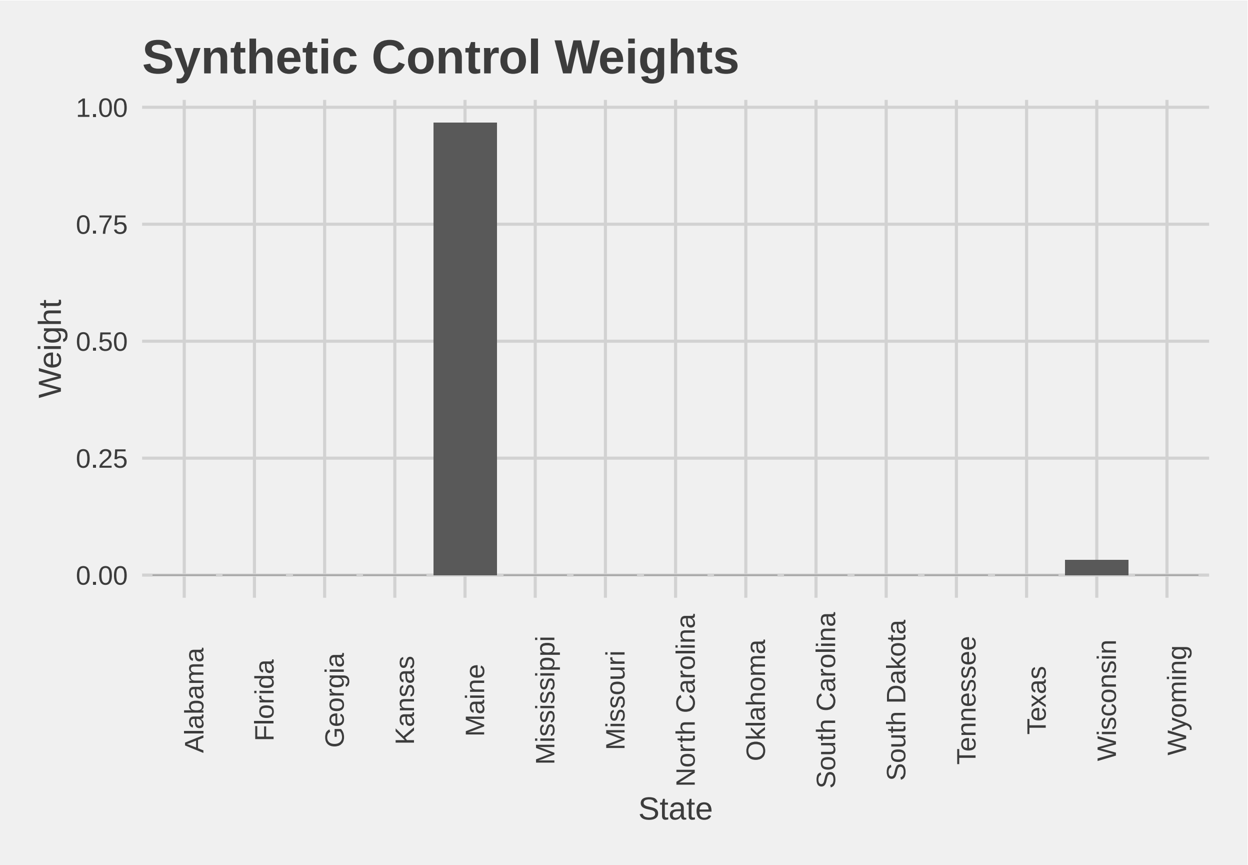
```
## [1] "The L2 imbalance: 0.00963"
```

```
plot(re_syn) +
  labs(title = "Non-Augmented Synthetic Control (Ridge)",
       subtitle = "For Pennsylvania") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```

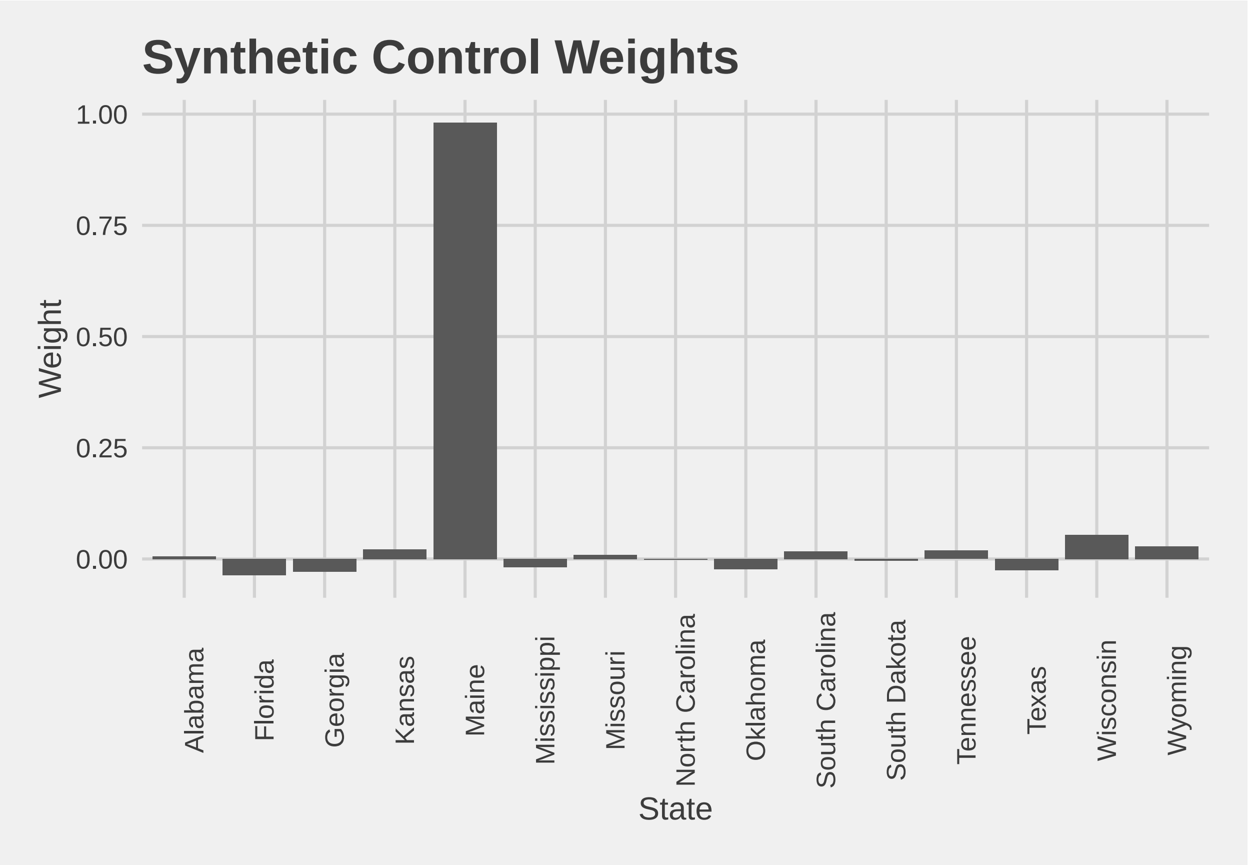


- Plot barplots to visualize the weights of the donors.

```r
# barplots of weights
data.frame(syn$weights) %>%
  tibble::rownames_to_column('State') %>%
  ggplot() +
  geom_bar(aes(x = State,
               y = syn.weights),
           stat = 'identity') +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        axis.text.x = element_text(angle = 90)) +
  ggtitle('Synthetic Control Weights') +
  xlab('State') +
  ylab('Weight')
```



```
data.frame(re_syn$weights) %>%
  tibble::rownames_to_column('State') %>%
  ggplot() +
  geom_bar(aes(x = State,
               y = re_syn.weights),
           stat = 'identity') +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        axis.text.x = element_text(angle = 90)) +
  ggtitle('Synthetic Control Weights') +
  xlab('State') +
  ylab('Weight')
```



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**: DiD is simple to understand, but it has several limitations, such as not being able to compare more than two units and the difficulty in establishing appropriate comparison units. In order to produce a more accurate counterfactual, synthetic control simulates pre-treatment features. It allows for differing historical data across units, accounts for non-parallel pre-treatment trends, and makes comparisons simple. However, it is harder to interpret, and the selection of predictors and the construction technique employed have a significant impact on the outcomes.
- 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**: It may become more difficult to interpret these weights as realistic combinations of control units because augmentation permits negative weights or sums greater than one. The model is more likely to overfit when weight constraints are loosened, especially in cases where there are lots of predictors or little pre-treatment data. While augmentation can increase pre-treatment fit, it can also make the synthetic control harder to perceive. Sensitivity analysis and validation are crucial for handling these trade-offs.

Staggered Adoption Synthetic Control

Estimate Multisynth

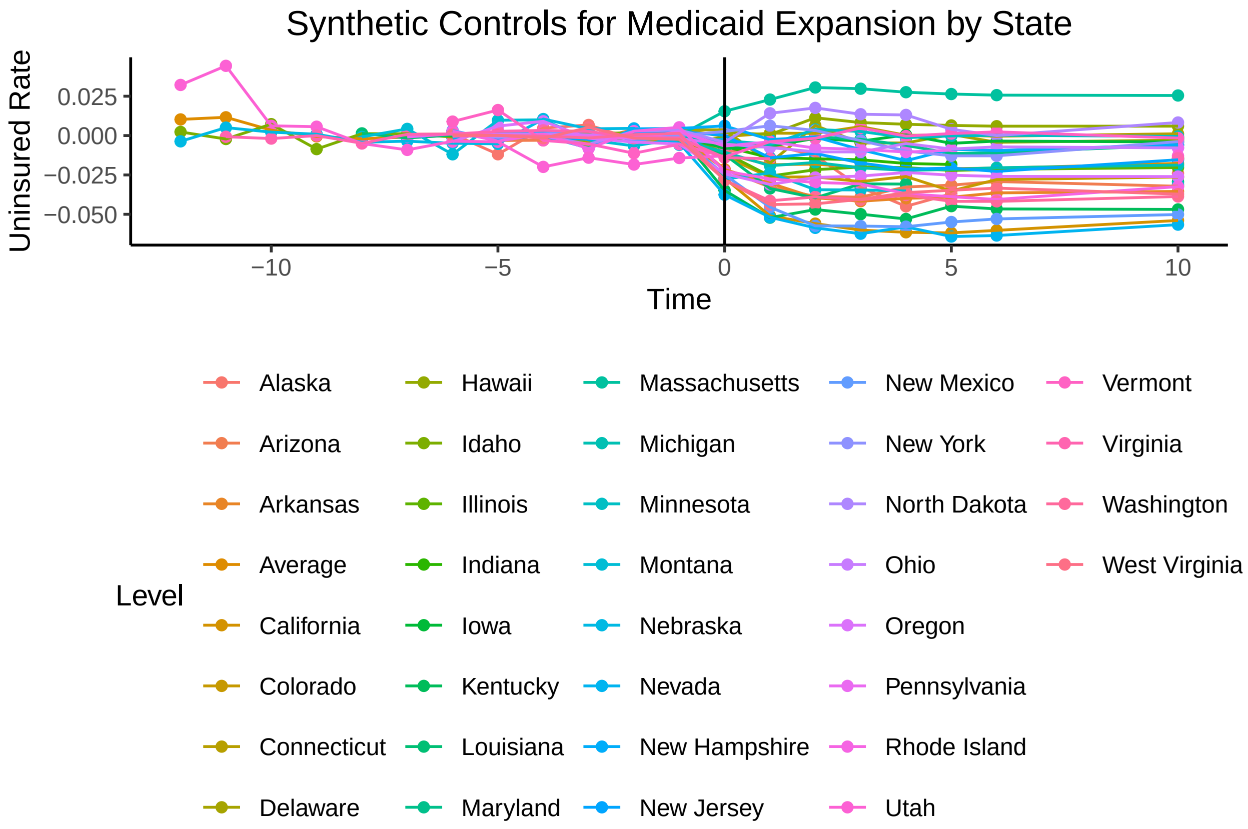
Do the following:

- 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.

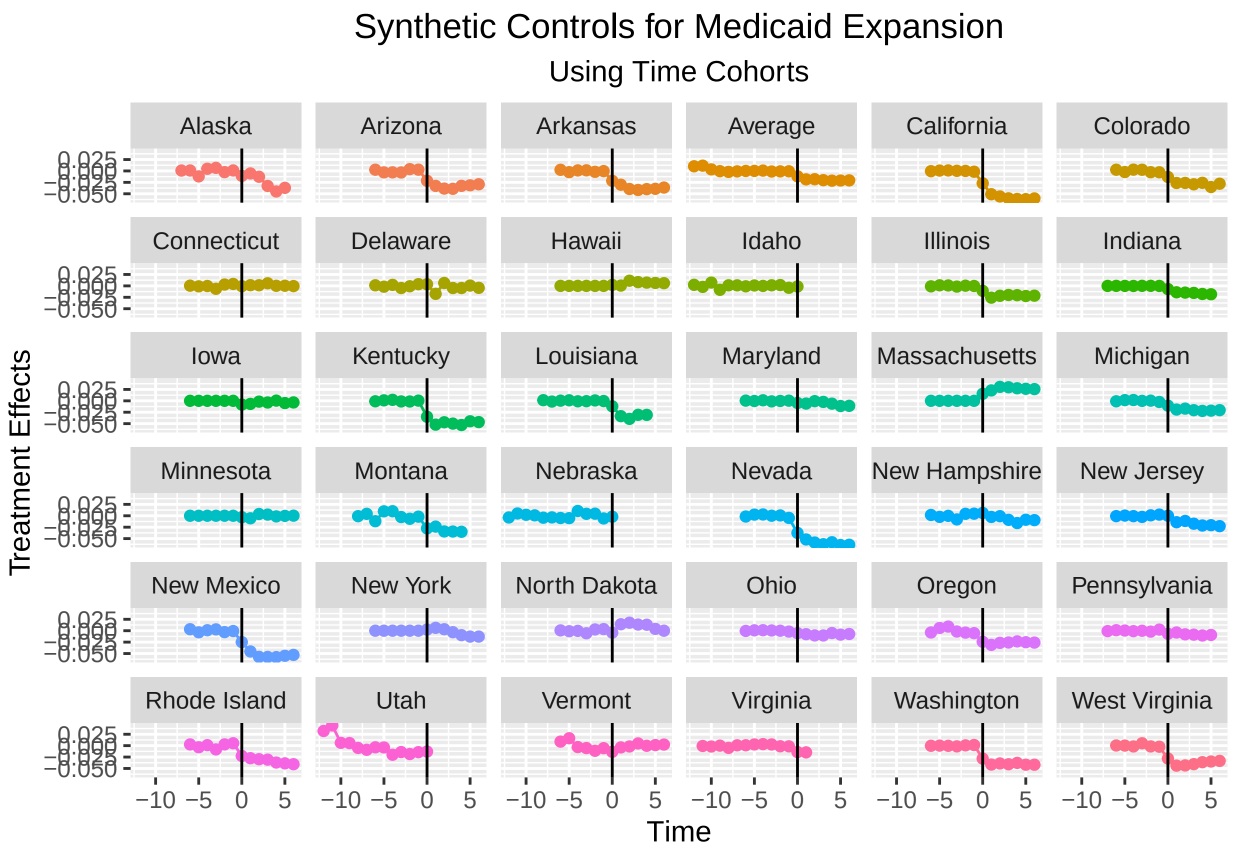
```
# multisynth model states
syn_C <- medicaid_expansion %>%
  mutate(uninsured_num = uninsured_rate*population/100,
         YearAdopted = ifelse(is.na(Date_Adopted), Inf, substr(Date_Adopted, 1, 4)),
         treated = 1 * (year >= YearAdopted)) %>%
  filter(!is.na(uninsured_num))


synC <- multisynth(uninsured_rate ~ treated, State, year,
                    syn_C, n_leads = 10, nu = 0)

syn_sum_C <- summary(synC)
syn_sum_C
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
##     data = syn_C, n_leads = 10, nu = 0)
##
## Average ATT Estimate (Std. Error): -0.018  (0.006)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.046
## Percent improvement from uniform global weights: 95.4
##
## Individual L2 Imbalance: 0.005
## Scaled Individual L2 Imbalance: 0.100
## Percent improvement from uniform individual weights: 90
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01170850 0.004594818 -0.02066224 -0.002979491
##                     1 Average -0.01851469 0.006004641 -0.03051885 -0.007094394
##                     2 Average -0.01814742 0.006261262 -0.03034271 -0.006112576
##                     3 Average -0.02019595 0.006520484 -0.03292365 -0.007384948
##                     4 Average -0.02132103 0.006240469 -0.03378202 -0.008908735
##                     5 Average -0.02098060 0.006121791 -0.03287666 -0.009364332
##                     6 Average -0.02067754 0.006617045 -0.03387751 -0.008008265
```

```
# Plot
syn_sum_C$att %>% mutate(Time = ifelse(is.na(Time), 10, Time)) %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme_classic() +
  theme(axis.title = element_text(),
        plot.title = element_text(hjust = 0.5),
        legend.position = "bottom") +
  labs(title = "Synthetic Controls for Medicaid Expansion by State",
       x = "Time", y = "Uninsured Rate")
```

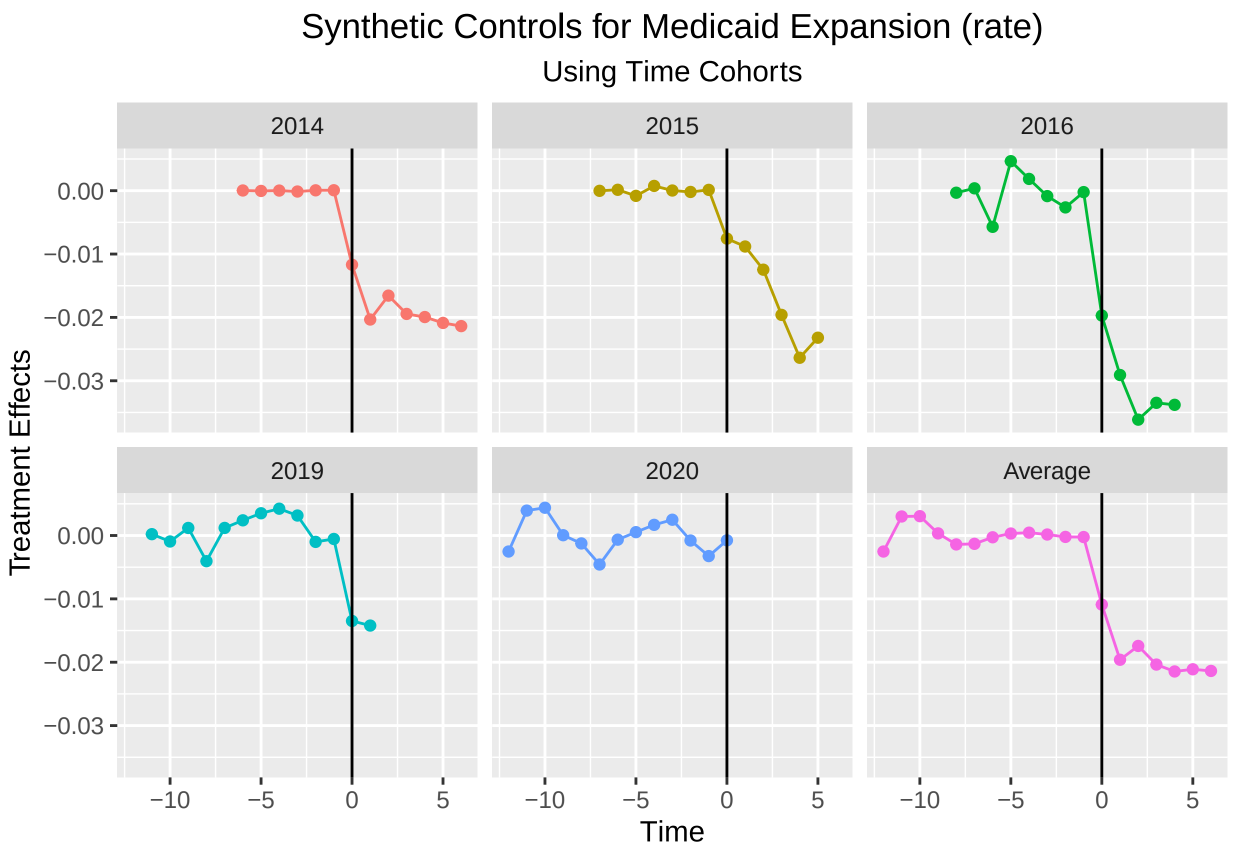


```
syn_sum_C$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme(axis.title = element_text(),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'None') +
  labs(title = "Synthetic Controls for Medicaid Expansion",
       subtitle = "Using Time Cohorts",
       x = "Time", y = "Treatment Effects") +
  facet_wrap(~Level)
```

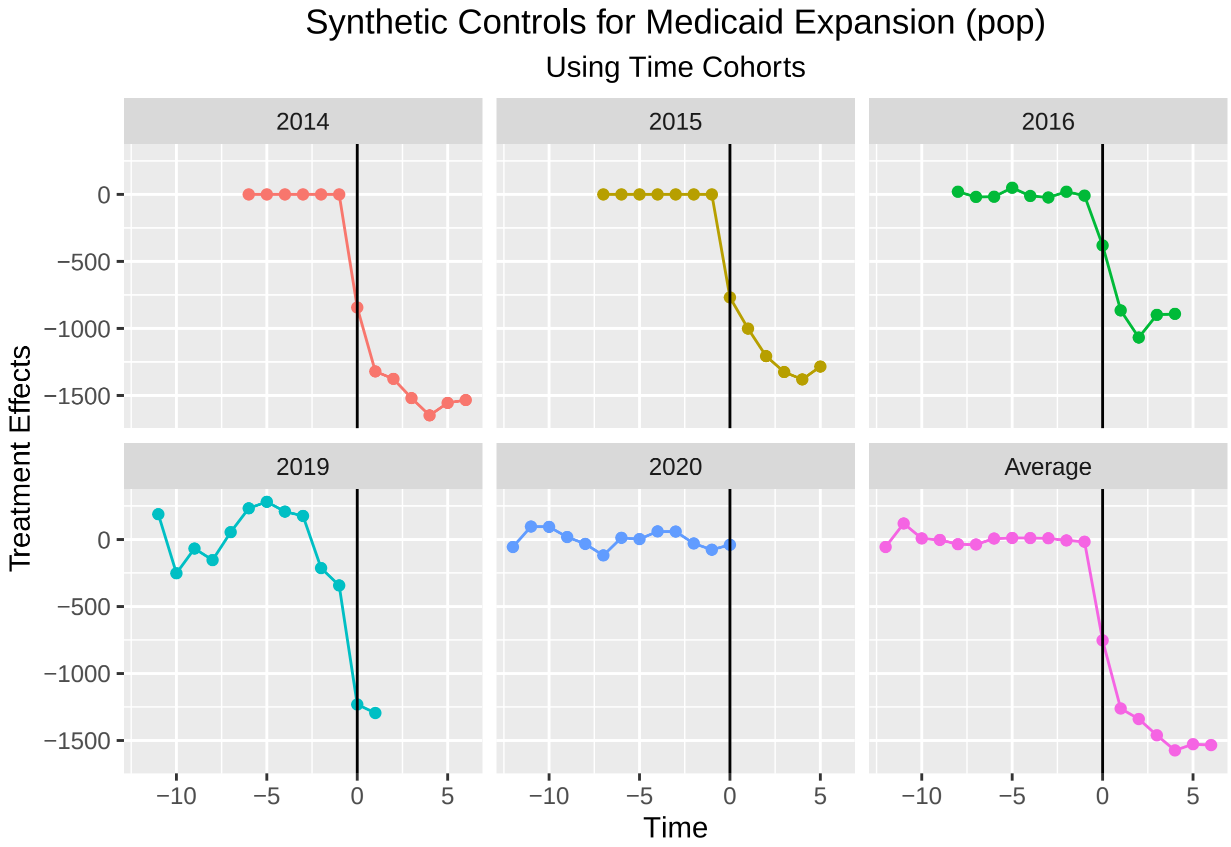


- Estimate a multisynth model using time cohorts. For the purpose of this exercise, you can simplify the treatment time so that states that adopted Medicaid expansion within the same year (i.e. all states that adopted epxansion in 2016) count for the same cohort. Plot the treatment effects for these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

```
# multisynth model time cohorts: rate
syn_CT <- multisynth(uninsured_rate ~ treated, State, year, syn_C,
                        n_leads = 10, nu = 0, time_cohort = TRUE)
syn_sum_CT <- summary(syn_CT)


syn_sum_CT$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme(axis.title = element_text(),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'None') +
  labs(title = "Synthetic Controls for Medicaid Expansion (rate)",
       subtitle = "Using Time Cohorts",
       x = "Time", y = "Treatment Effects") +
  facet_wrap(~Level)
```



```
# multisynth model time cohorts: population
syn_Cnum <- multisynth(uninsured_num ~ treated, State, year, syn_C,
                        n_leads = 10, nu = 0, time_cohort = TRUE)
syn_sum_Cnum <- summary(syn_Cnum)


syn_sum_Cnum$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme(axis.title = element_text(),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'None') +
  labs(title = "Synthetic Controls for Medicaid Expansion (pop)",
       subtitle = "Using Time Cohorts",
       x = "Time", y = "Treatment Effects") +
  facet_wrap(~Level)
```



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**: Each state shows different treatment effects. Many states, including Arizona, California, and Kentucky, had a significant decline in the year of adoption. Other states saw little change in their uninsured rates, and some even saw rises after Medicaid expansion. While there is evidence of varied treatment effects among states, it is unclear if this is related to the eligibility standards.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: The data suggest that the sooner states extend Medicaid, the greater the decline in the uninsured population. These early adopters took advantage of the federal incentives and infrastructure support available at the time, resulting in more advantages from the expansion than states that embraced later.

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?
- **Answer**: Because both DiD and synthetic control methods can handle common challenges in such data sets, such as controlling for unobserved confounders that vary across entities but are constant over time. DiD makes use of the parallel trends assumption, which postulates that over time, the difference between the treatment and control groups would not change in the absence of therapy. This allows the influence of the treatment to be separated from other potentially confusing patterns. Synthetic control further refines this approach by constructing a weighted combination of control units that best replicate the pre-treatment characteristics of the treated unit, offering a more precise estimation of what would have happened in the absence of treatment. This is especially useful when a clear comparison group is not readily apparent, or when treatment effects need to be estimated with greater accuracy.
- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?
- **Answer**: Selection for treatment influences the decision between DiD/synthetic control and regression discontinuity. DiD and synthetic control are employed when treatment is not determined by a clear cutoff, making them appropriate for evaluating policies that fluctuate among groups and over time. On the other hand, regression discontinuity is utilized when treatment is determined by a specified threshold, and it is perfect for analyzing the immediate impacts at that level. Regression discontinuity focuses on units near the cutoff, presuming they are comparable save for the treatment, resulting in clear causal inference at the threshold. In contrast, DiD and synthetic control capture wider policy effects over time and across regions or populations.