Project 7: Difference-in-Differences and Synthetic Control - Jorge Morales

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

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

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

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

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

# set seed
set.seed(44)

# load data
medicaid_expansion <- read_csv('/Users/jama/Documents/GitHub/Computational-Social-Science-Projects/Proj
```

```
## 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.

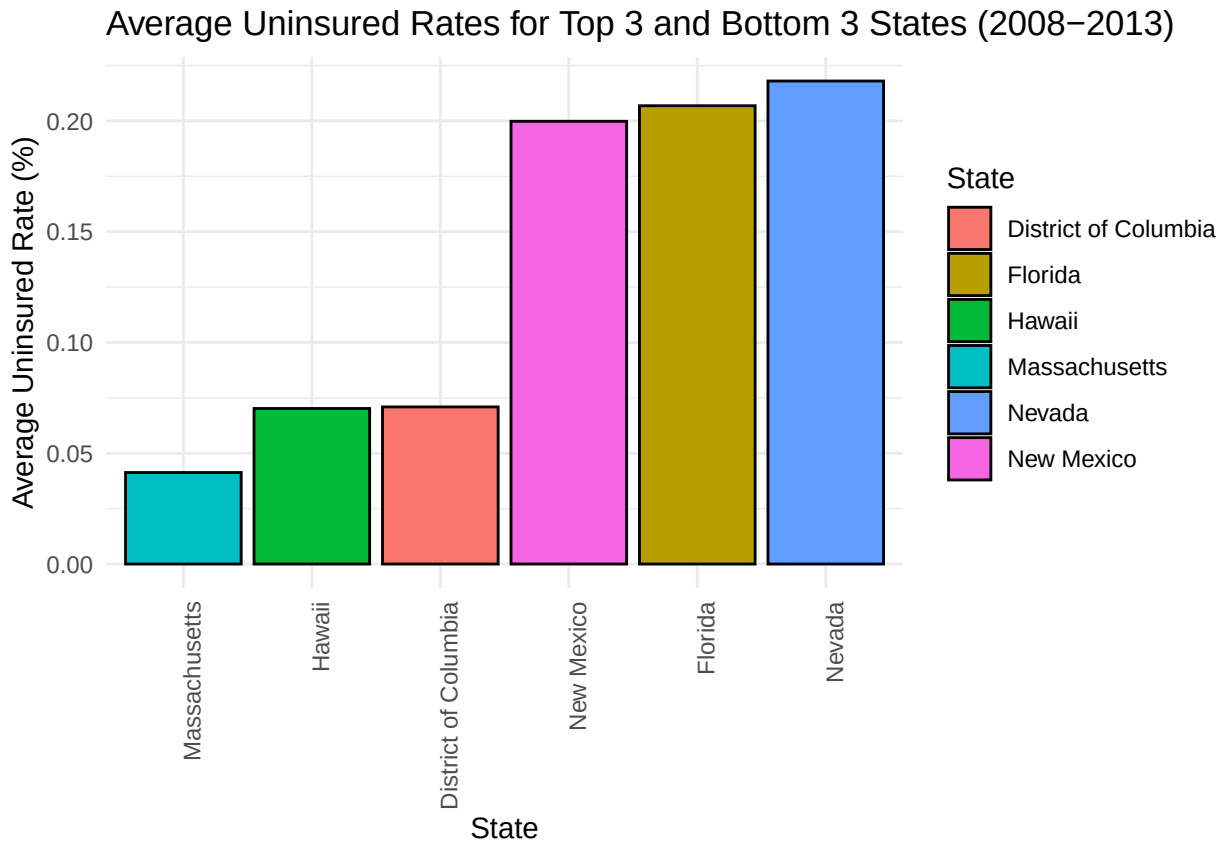
Exploratory Data Analysis

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.

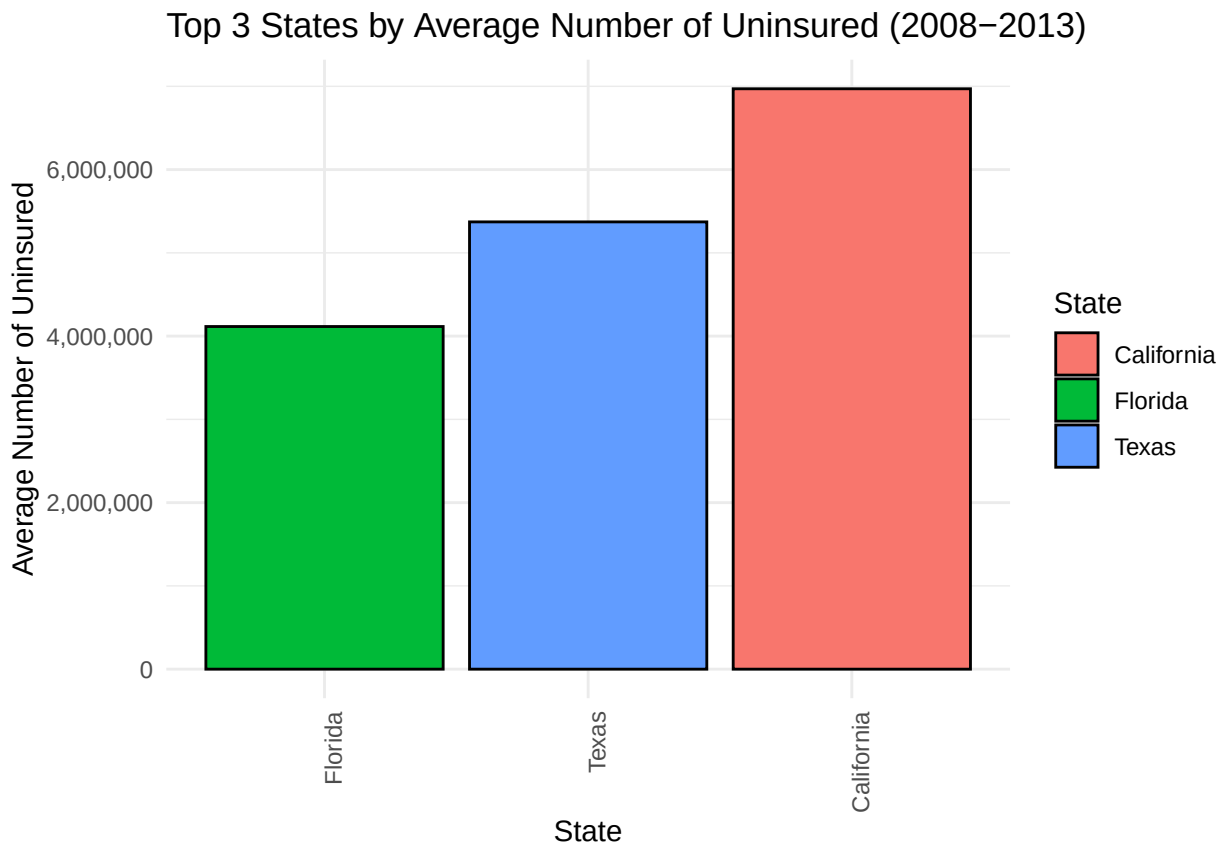
```
# highest and lowest uninsured rates


# Filter the data for the years 2008 to 2013, calculate average uninsured rates, then sort and select t
```

```r
top_bottom_states_avg <- medicaid_expansion %>%
  filter(year >= 2008, year <= 2013) %>%
  group_by(State) %>%
  summarise(avg_uninsured_rate = mean(uninsured_rate, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(avg_uninsured_rate)) %>%
  slice(c(1:3, (n()-2):n()))

# Create the bar graph
ggplot(top_bottom_states_avg, aes(x = reorder(State, avg_uninsured_rate), y = avg_uninsured_rate, fill =
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Average Uninsured Rates for Top 3 and Bottom 3 States (2008-2013)",
       x = "State",
       y = "Average Uninsured Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) # Rotate x-axis labels for better readabilit
```

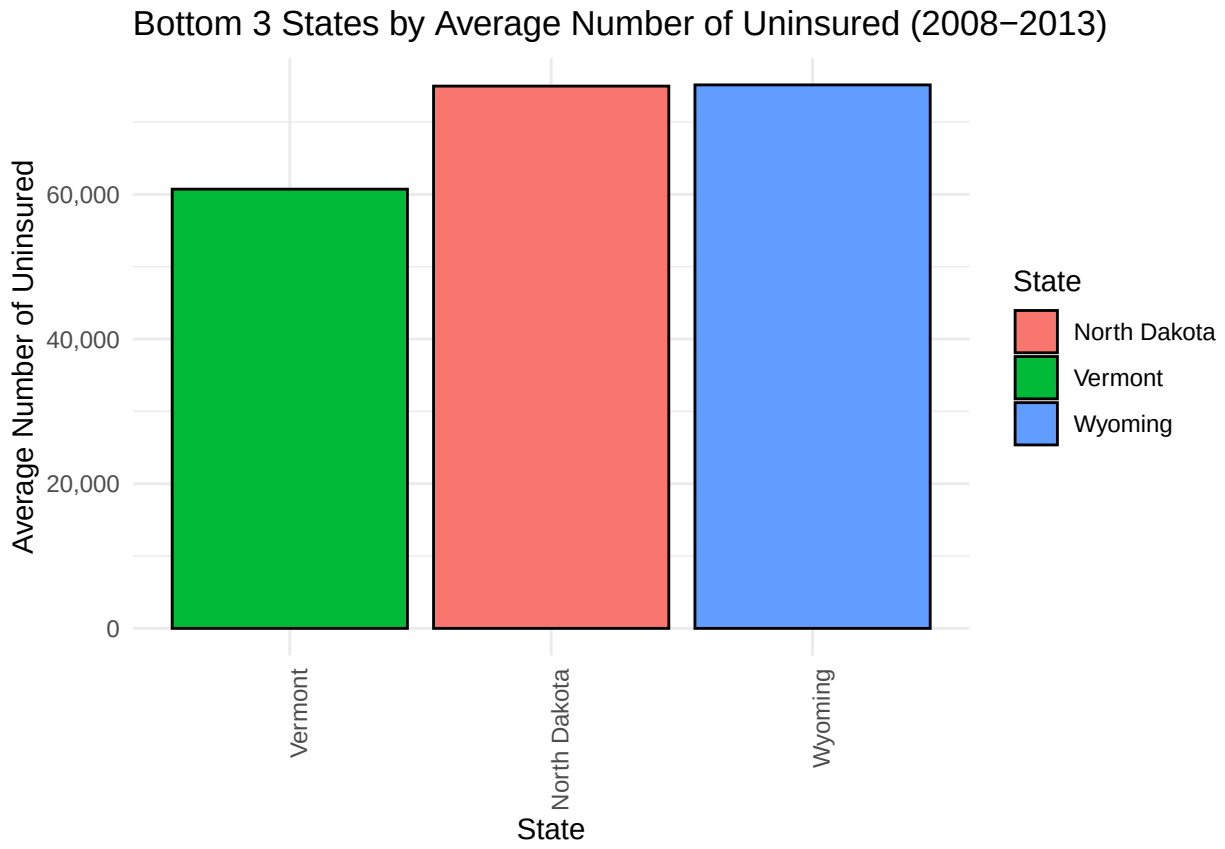



```r
# most uninsured Americans
# I will create a new variable for the number of Americans uninsured in a state per year
medicaid_expansion$uninsured_count <- medicaid_expansion$uninsured_rate * medicaid_expansion$population

medicaid_expansion_noDC <- medicaid_expansion %>%
  filter(State != "District of Columbia")

# Filter and calculate the averages as before, then arrange
```

```r
top_bottom_states_avg_count <- medicaid_expansion_noDC %>%
  filter(year >= 2008, year <= 2013) %>%
  group_by(State) %>%
  summarise(avg_uninsured_count = mean(uninsured_count, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(avg_uninsured_count)) %>%
  mutate(rank = row_number())  # Assign a rank after arranging

# Separate data for top 3 and bottom 3
top_3_states <- filter(top_bottom_states_avg_count, rank <= 3)
bottom_3_states <- filter(top_bottom_states_avg_count, rank >= (n() - 2))

# Create the bar graph for top 3 states
top_3_plot <- ggplot(top_3_states, aes(x = reorder(State, avg_uninsured_count), y = avg_uninsured_count
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Top 3 States by Average Number of Uninsured (2008-2013)",
       x = "State",
       y = "Average Number of Uninsured") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Create the bar graph for bottom 3 states
bottom_3_plot <- ggplot(bottom_3_states, aes(x = reorder(State, avg_uninsured_count), y = avg_uninsured
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Bottom 3 States by Average Number of Uninsured (2008-2013)",
       x = "State",
       y = "Average Number of Uninsured") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

print(top_3_plot)
```

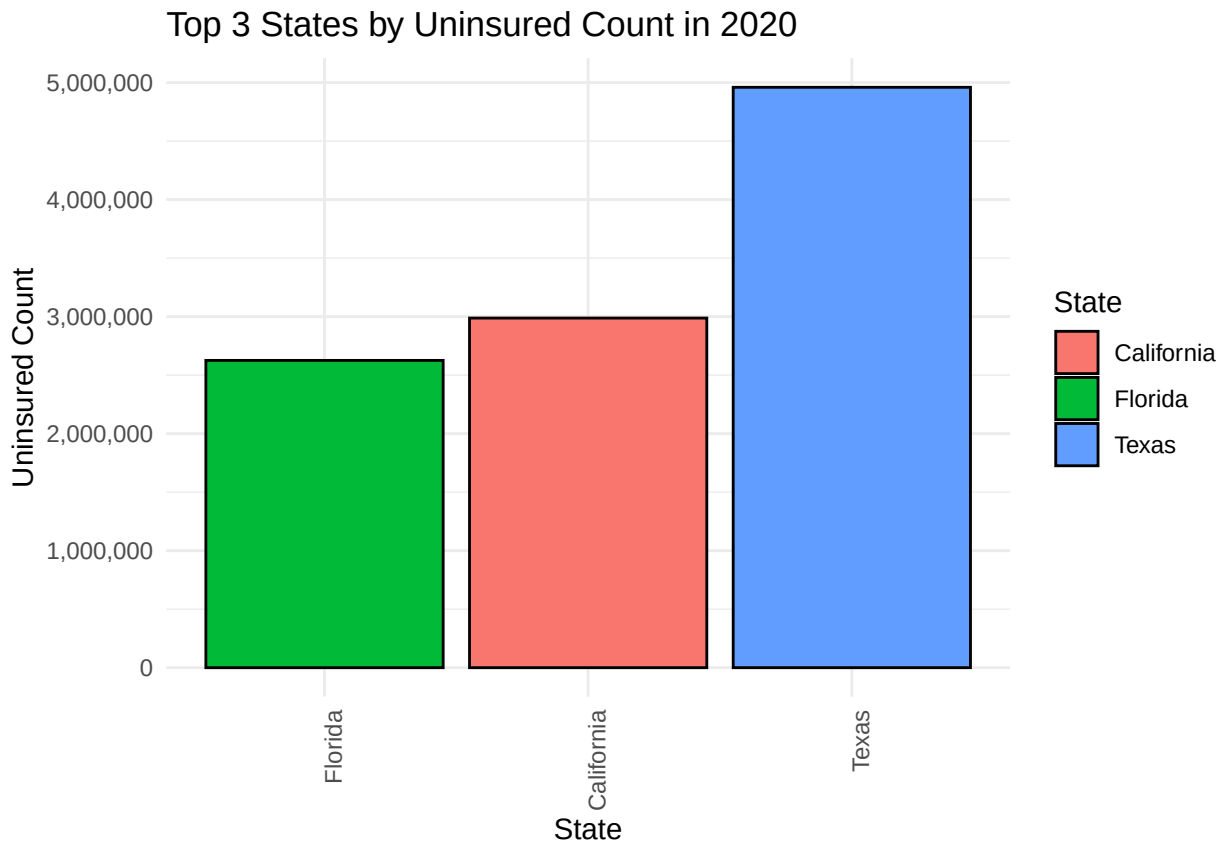



```
print(bottom_3_plot)
```

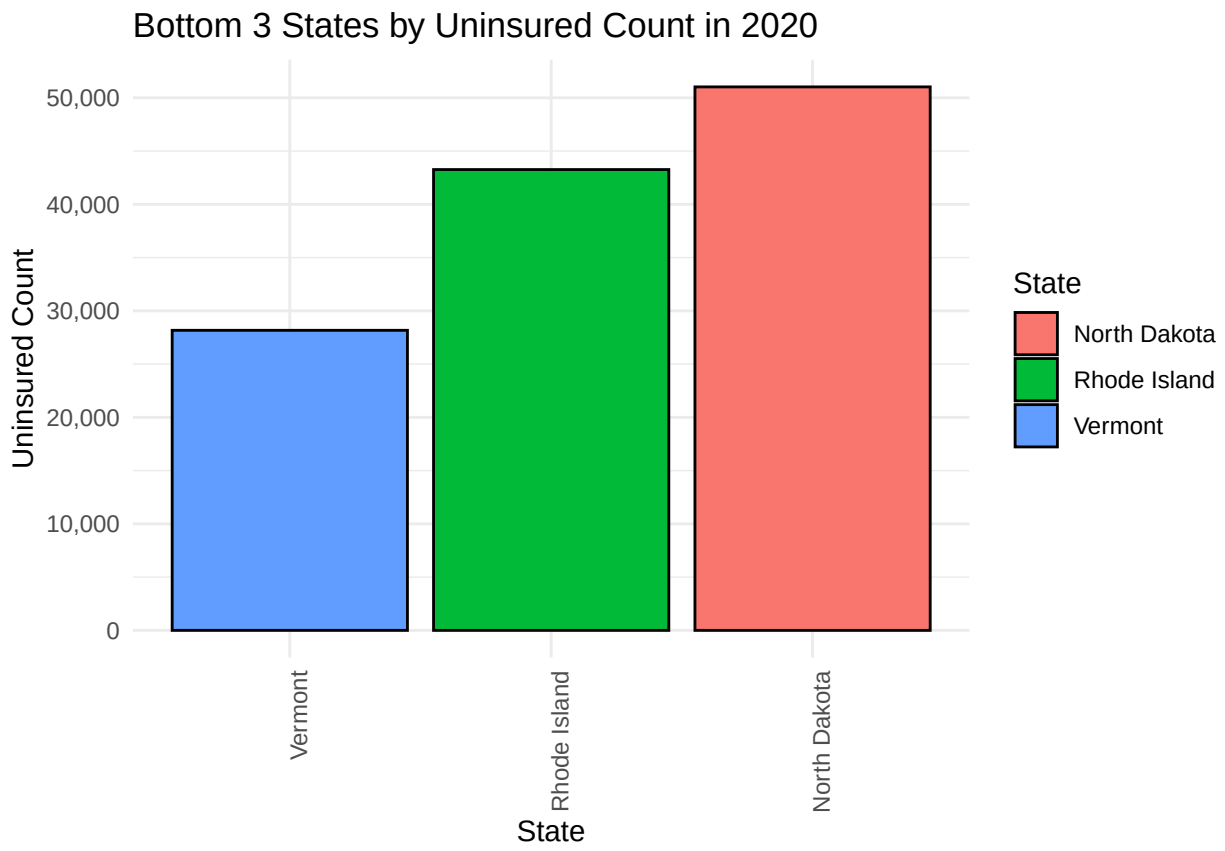



```
#Now Ill do it for the year 2020

# Filter for the year 2020
top_bottom_states_2020 <- medicaid_expansion_noDC %>%
  filter(year == 2020) %>%
  arrange(desc(uninsured_count)) %>%
  mutate(rank = row_number())  # Assign a rank after arranging

# Separate data for top 3 and bottom 3
top_3_states_2020 <- filter(top_bottom_states_2020, rank <= 3)
bottom_3_states_2020 <- filter(top_bottom_states_2020, rank >= (n() - 2))

# Create the bar graph for top 3 states
top_3_plot_2020 <- ggplot(top_3_states_2020, aes(x = reorder(State, uninsured_count), y = uninsured_cou
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Top 3 States by Uninsured Count in 2020",
       x = "State",
       y = "Uninsured Count") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Create the bar graph for bottom 3 states
bottom_3_plot_2020 <- ggplot(bottom_3_states_2020, aes(x = reorder(State, uninsured_count), y = uninsur
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Bottom 3 States by Uninsured Count in 2020",
```

```
       x = "State",
       y = "Uninsured Count") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

print(top_3_plot_2020)
```




```
print(bottom_3_plot_2020)
```



Difference-in-Differences Estimation

Estimate Model

Do the following:

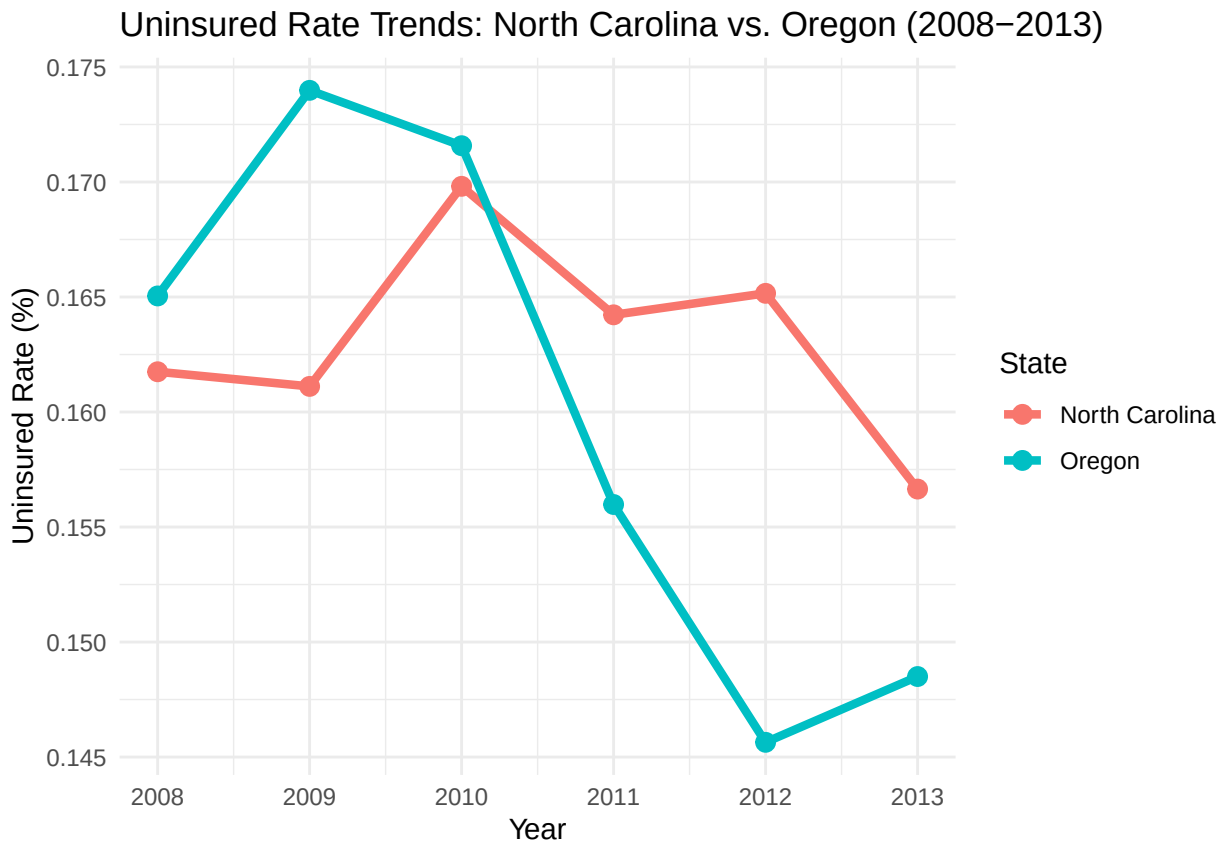
- Choose a state that adopted the Medicaid expansion on January 1, 2014 and a state that did not. **Hint**: Do not pick Massachusetts as it passed a universal healthcare law in 2006, and also avoid picking a state that adopted the Medicaid expansion between 2014 and 2015.
- Assess the parallel trends 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).

```
# Parallel Trends plot
#I'll chose Oregon vs NC


# Filter data for North Carolina and Oregon from 2008 to 2013
nc_or_data <- medicaid_expansion %>%
  filter(State %in% c("North Carolina", "Oregon") & year >= 2008 & year <= 2013)

# Plot uninsured rates for both states over the years
trend_plot <- ggplot(nc_or_data, aes(x = year, y = uninsured_rate, color = State, group = State)) +
  geom_line(size = 1.5) +  # Use thicker lines for better visibility
  geom_point(size = 3) +  # Add points to the line graph
```

```
  labs(title = "Uninsured Rate Trends: North Carolina vs. Oregon (2008-2013)",
       x = "Year",
       y = "Uninsured Rate (%)") +
  theme_minimal() +
  scale_x_continuous(breaks = 2008:2013)  # Ensure all years are included as x-axis ticks
```

```
## 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.
```

```
print(trend_plot)
```

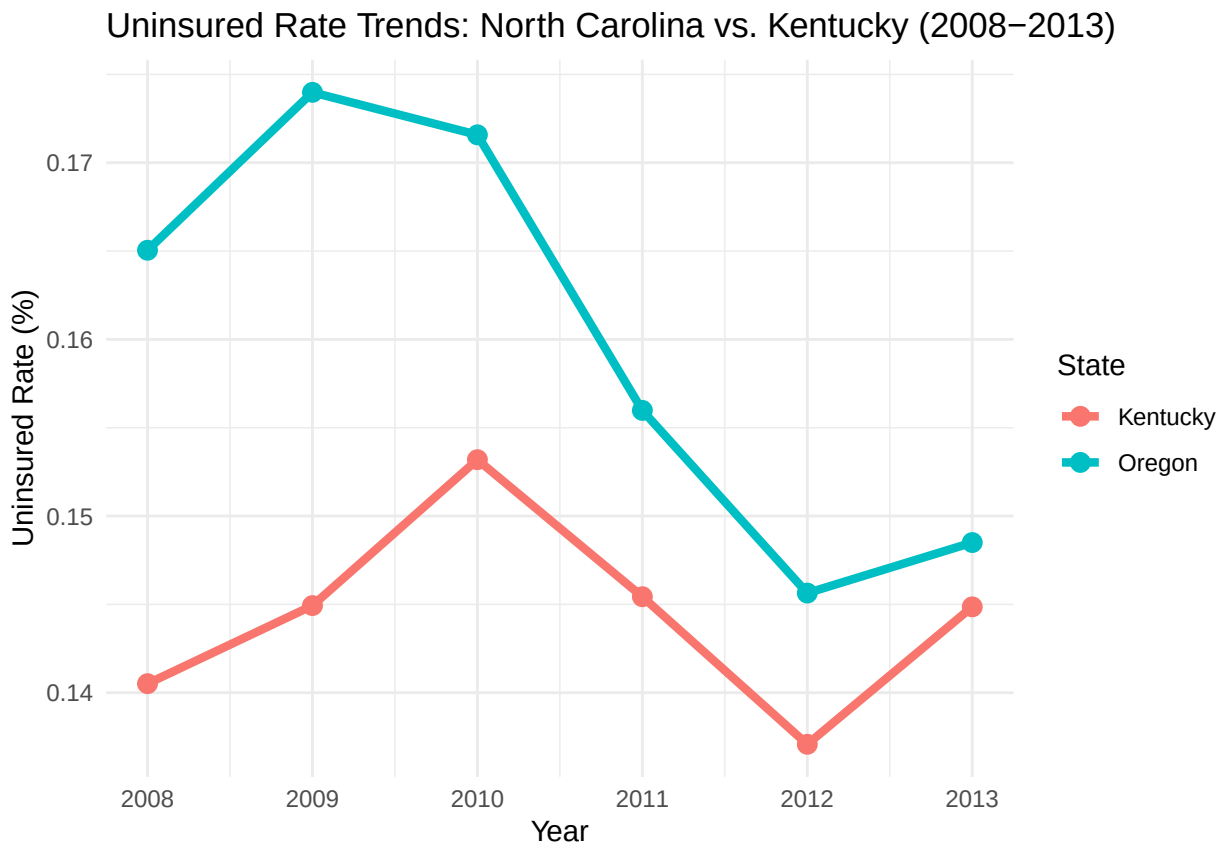



```
#Clearly not a good control group Oregon was already going down before

# Regression model to test for parallel trends
model <- lm(uninsured_rate ~ year * factor(State), data = nc_or_data)
summary(model)
```

```
##
## Call:
## lm(formula = uninsured_rate ~ year * factor(State), data = nc_or_data)
##
```

```
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0081724 -0.0033912 -0.0000728  0.0036413  0.0088374
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.2516032  3.0495419   0.410   0.6923
## year                      -0.0005414  0.0015168  -0.357   0.7304
## factor(State)Oregon        9.4412977  4.3127036   2.189   0.0600 .
## year:factor(State)Oregon -0.0046975  0.0021451  -2.190   0.0599 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006345 on 8 degrees of freedom
## Multiple R-squared:  0.614,  Adjusted R-squared:  0.4693
## F-statistic: 4.242 on 3 and 8 DF,  p-value: 0.04537
```

```r
#The test shows that we cant reject the null, so the trends could be similar, nontheless the graph make

#Lets see if Kentucky and NC are better to compare
nc_kfc_data <- medicaid_expansion %>%
  filter(State %in% c("Kentucky", "Oregon") & year >= 2008 & year <= 2013)

# Plot uninsured rates for both states over the years
trend_plot <- ggplot(nc_kfc_data, aes(x = year, y = uninsured_rate, color = State, group = State)) +
  geom_line(size = 1.5) +  # Use thicker lines for better visibility
  geom_point(size = 3) +  # Add points to the line graph
  labs(title = "Uninsured Rate Trends: North Carolina vs. Kentucky (2008-2013)",
       x = "Year",
       y = "Uninsured Rate (%)") +
  theme_minimal() +
  scale_x_continuous(breaks = 2008:2013)  # Ensure all years are included as x-axis ticks

print(trend_plot)
```




```
#This looks better, I think there is a case to be made by comaring the trends of both states.

model <- lm(uninsured_rate ~ year * factor(State), data = nc_kfc_data)
summary(model)
```

```
##
## Call:
## lm(formula = uninsured_rate ~ year * factor(State), data = nc_kfc_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0081724 -0.0050355  0.0006981  0.0026082  0.0088374
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.6941777  3.2926167   0.211   0.8383
## year                       -0.0002735  0.0016377  -0.167   0.8715
## factor(State)Oregon         9.9987232  4.6564631   2.147   0.0640 .
## year:factor(State)Oregon -0.0049654  0.0023161  -2.144   0.0644 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.006851 on 8 degrees of freedom
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.6783
## F-statistic:  8.73 on 3 and 8 DF,  p-value: 0.006649
```

```
#We cannot reject the hypothesis that the interaction is different so the trends might be similar enoug
```

- 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
# create a dataset for kansas and colorado
dnd_data <- medicaid_expansion %>%
  filter(State %in% c("Kentucky", "North Carolina") )

# pre-treatment difference
# ----------
pre_diff <-
  dnd_data %>%
  # filter out only the quarter we want
  filter(year == 2013) %>%
  # subset to select only vars we want
  select(State,
         uninsured_rate) %>%
  # make the data wide
  pivot_wider(names_from = State,
              values_from = uninsured_rate) %>%
  # subtract to make calculation
  summarise(Kentucky - `North Carolina`)

# post-treatment difference
# ----------

post_diff <-
  dnd_data %>%
  # filter out only the quarter we want
  filter(year == 2014) %>%
  # subset to select only vars we want
  select(State,
         uninsured_rate) %>%
  # make the data wide
  pivot_wider(names_from = State,
              values_from = uninsured_rate) %>%
  # subtract to make calculation
  summarise(Kentucky - `North Carolina`)

# diff-in-diffs
# ----------
diff_in_diffs <- post_diff - pre_diff
diff_in_diffs
```

```
##   Kentucky - `North Carolina`
## 1                   -0.03273
```

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**: It is hard to think that nothing else changed at the same time and that both groups are comparable. Specially, unobservable variables could be affecting the trends and could lead to incorrect definition of the control group, and thus the counterfactual.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: It is necessary to justify that we have a good counterfactual to compare and obtain causal inference. Strengths of parallel trends are that it is easy to see and partially (imperfectly as it relies on observable variables and trends) test if the trends between control and treatment groups are similar. Weaknesses are that we need enough periods of time before to identify trends, we depend on conditional expectations and thus the specifications could be wrong, there is not formal full test to be sure that the parallel trend would hold.

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


library(lubridate)
medicaid_expansion_noDC$Year_Adopted <- year(medicaid_expansion_noDC$Date_Adopted)

medicaid_expansion_noDC <- medicaid_expansion_noDC %>%
  mutate(
    Year_Adopted = coalesce(as.integer(Year_Adopted), Inf),  # Replace NA with 0 and convert to integer
    year = as.integer(year),
    prepost = if_else(Year_Adopted <= year, 1, 0)  # Set prepost based on the condition
  )

medicaid_expansion_noDC_filter <- medicaid_expansion_noDC %>%
  filter(Year_Adopted >= 2019)
```

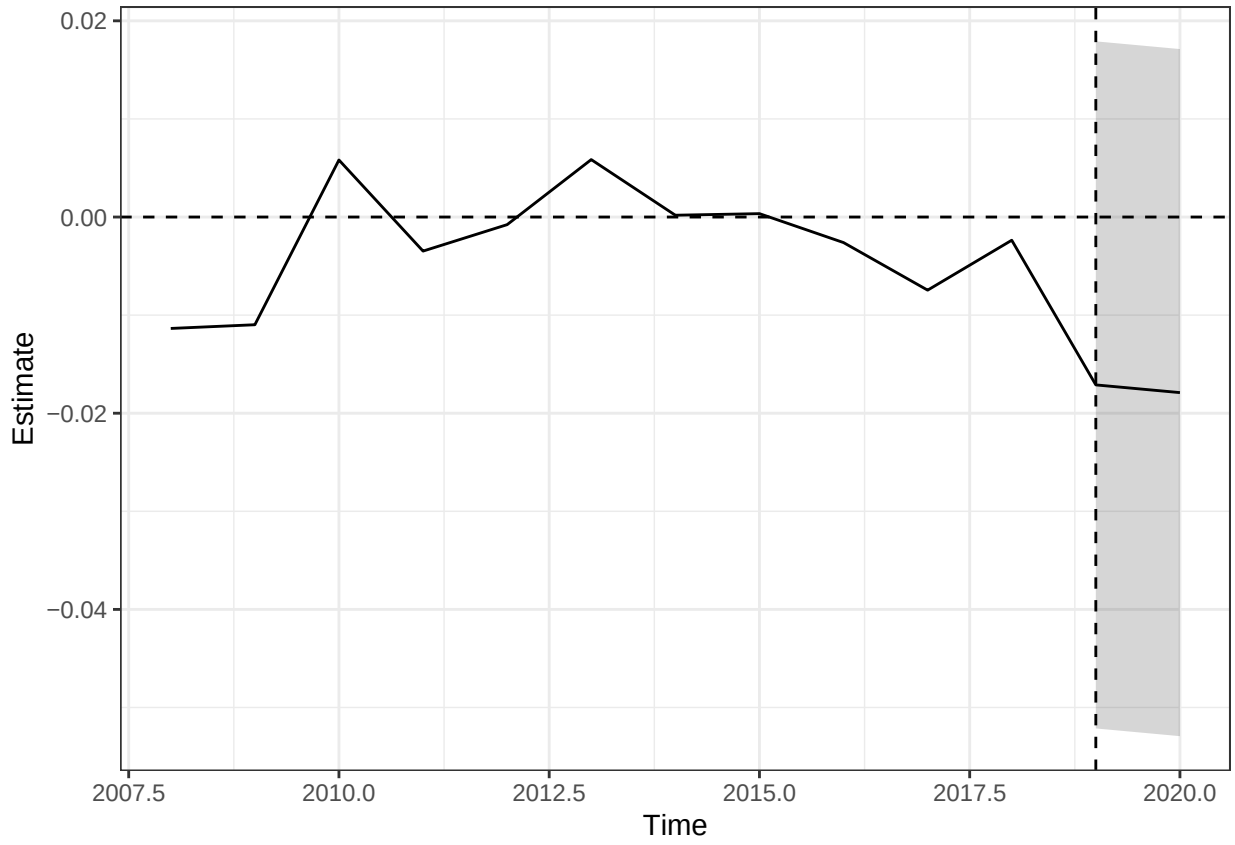
```r
medicaid_expansion_noDC_filter <- medicaid_expansion_noDC_filter %>%
  mutate(treatment = if_else(State =="Virginia" & year >= 2019, 1, 0))


syn <-                              # save object
  augsynth(uninsured_rate ~ treatment, # treatment - use instead of treated bc latter codes 2012.25 as
                          State,     # unit
                          year,  # time
                          medicaid_expansion_noDC_filter,    # data
           progfunc = "None",       # plain syn control
           scm = T)                 # synthetic control
```

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

```r
# summary
summary(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 (p Value for Joint Null):  -0.0175   ( 0.071 )
## L2 Imbalance: 0.020
## Percent improvement from uniform weights: 82.2%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2019   -0.017             -0.052              0.018   0.246
##  2020   -0.018             -0.053              0.017   0.239
```

```r
plot(syn)
```



- 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


syn2 <-                                 # save object
  augsynth(uninsured_rate ~ treatment, # treatment - use instead of treated bc latter codes 2012.25 as
                           State,     # unit
                           year,  # time
                           medicaid_expansion_noDC_filter,    # data
           progfunc = "ridge",       # plain syn control
           scm = T)                 # synthetic control
```

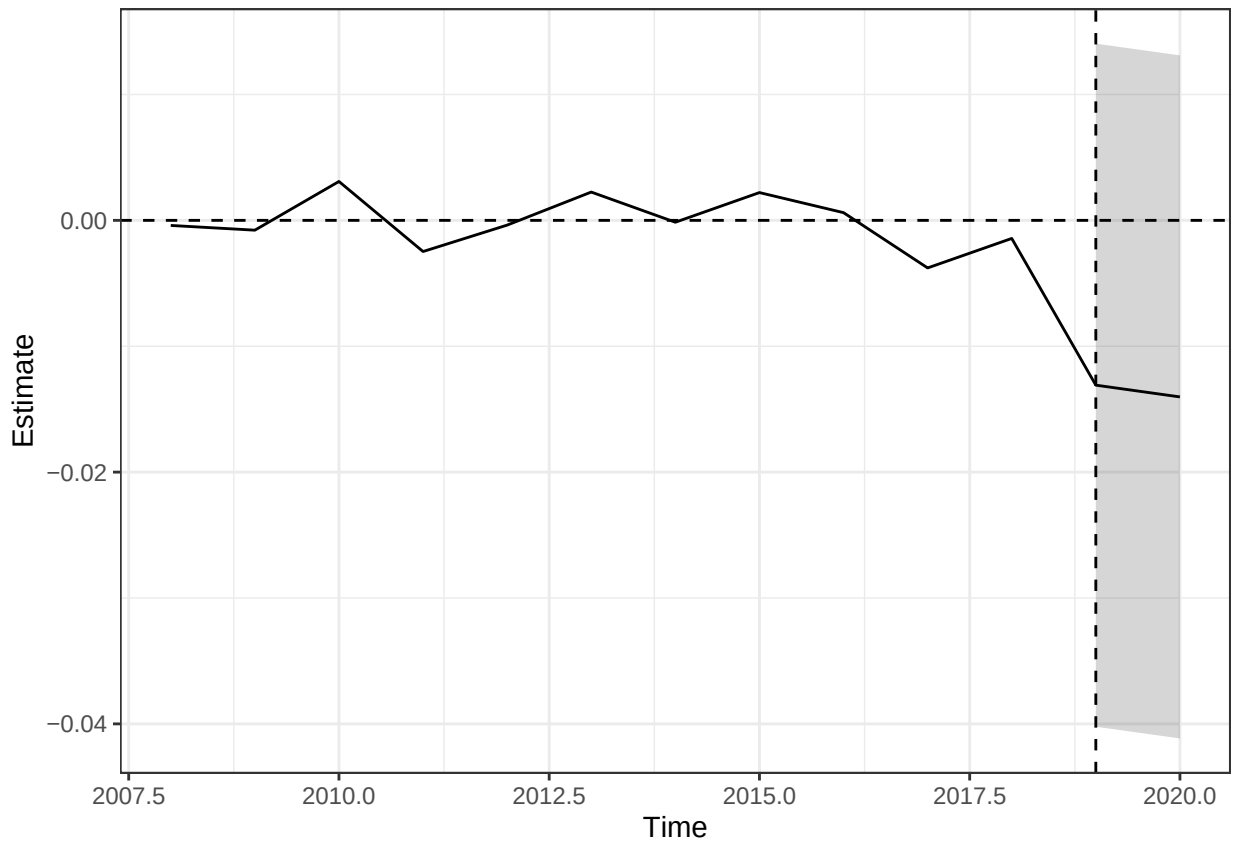
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
# summary
summary(syn2)
```

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

```
## Average ATT Estimate (p Value for Joint Null):  -0.0136   ( 0.027 )
## L2 Imbalance: 0.007
## Percent improvement from uniform weights: 94.2%
##
## Avg Estimated Bias: -0.004
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2019   -0.013             -0.040              0.014   0.082
##  2020   -0.014             -0.041              0.013   0.092
```

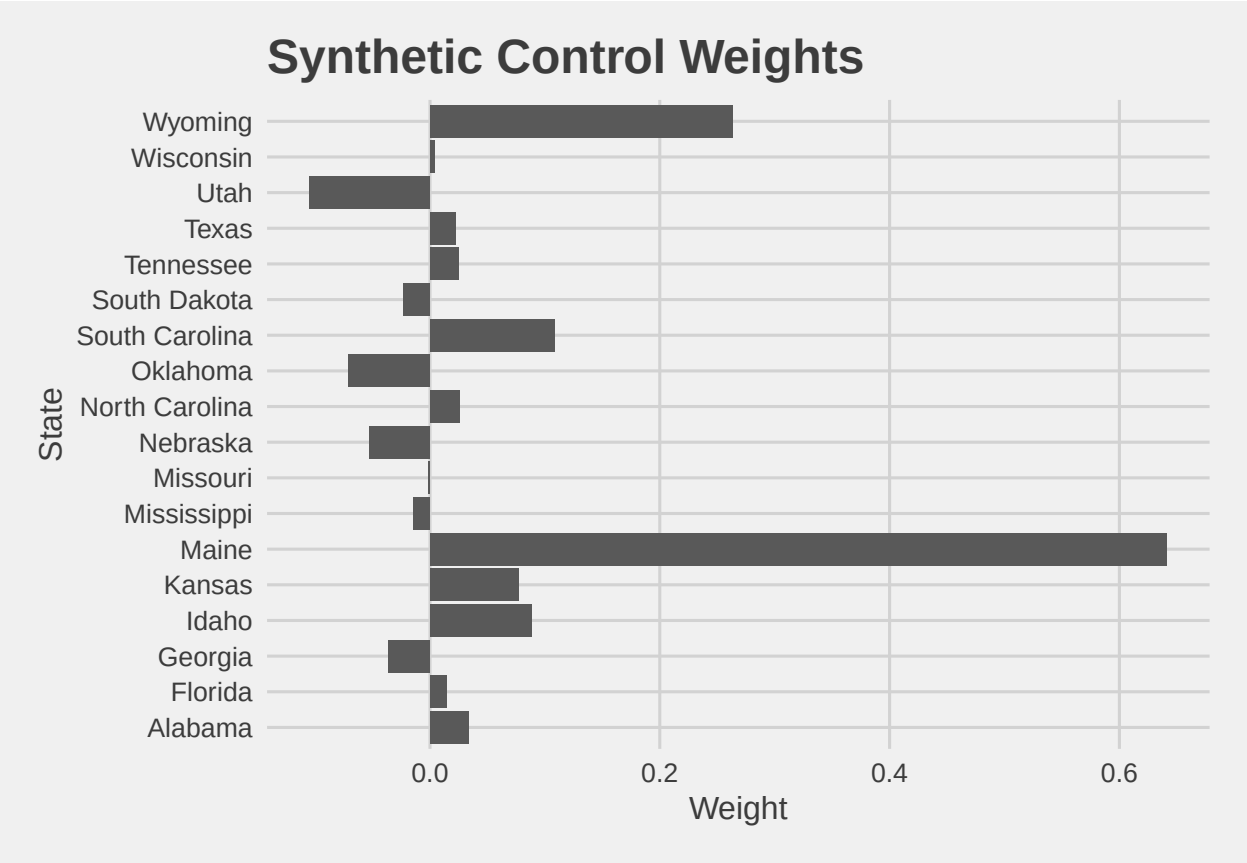
```
plot(syn2)
```



- Plot barplots to visualize the weights of the donors.

```
data.frame(syn2$weights) %>% # coerce to data frame since it's in vector form
  # process
  # ---------
  # change index to a column
  tibble::rownames_to_column('State') %>% # move index from row to column (similar to index in row as i
  # plot
  # ---------
  ggplot() +
  # stat = identity to take the literal value instead of a count for geom_bar()
```

```
geom_bar(aes(x = State,
             y = syn2.weights),
         stat = 'identity') +  # override count() which is default of geom_bar(), could use geom_col(
coord_flip() +   # flip to make it more readable
# themes
theme_fivethirtyeight() +
theme(axis.title = element_text()) +
# labels
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**: Syntehtic controls helps to not over rely on the researchers inspection and decision of potential control groups. Thus instead of finding a right state or city to serve as counterfactual, it will help to find the weighted combination of states that will approximate better the pre-trends and thus construct a much better counterfactual case. This clearly helps to strengthen the case for the parallel trends assumption needed for estimation. The disadvantage is that more data is needed, and that the weighted

control might still be limited to be credible as a parallel trend. If there were unobservable variables explainig some of the trends of the treatment group, synthethic controls would still be limited in building a good comparison. At least with DnD the rethorical part of convinincing that the states serve as controls is easier to be explained and that can help give robustness to the results.

- 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 it can create a problem as it is hard to explain how it works and how it is adjusting for the overweights. Negative weight discount some observations-period and that is hard to identify and explain. So the narrative about the counterfactual is not clear at all, it is like a black box in certain way, maybe gray. The balance it provides it is good, but it would be ideal to contrast with the tradiotional synthetic control and if there is not a big difference, simple is better to explain. dw

Staggered Adoption Synthetic Control

Estimate Multisynth

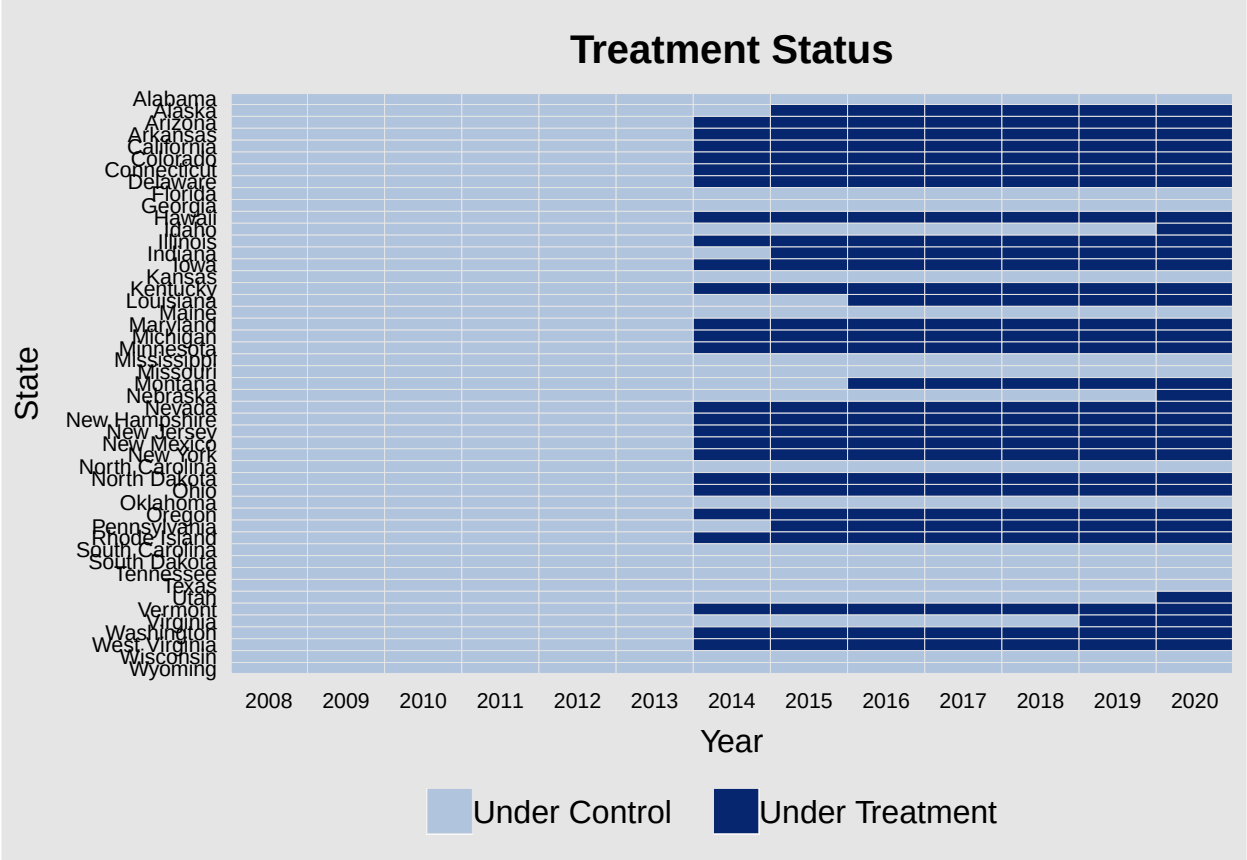
Do the following:

- 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

library(panelView)
```

```
## ## See bit.ly/panelview4r for more info.
## ## Report bugs -> yiqingxu@stanford.edu.
```

```
#Take out Massachusetts
 medicaid_expansion_noDC <- medicaid_expansion_noDC  %>%
    filter(!State %in% c("Massachusetts"))

panelview(formula = uninsured_rate ~ prepost,
          data = medicaid_expansion_noDC, index = c("State", "year"),
          xlab = "Year", ylab = "State")
```




```
#
# specifying a synthetic control using tidysynth
# ------------------------------------------------------------------
```

```
# load library



# multisynth model time cohorts
ppool_syn <- multisynth(uninsured_rate ~ prepost,
                        State,                        # unit
                        year,                         # time
                        medicaid_expansion_noDC, # data
                        n_leads = 6)                  # post-treatment periods to estimate


# view results
print(ppool_syn$nu)
```

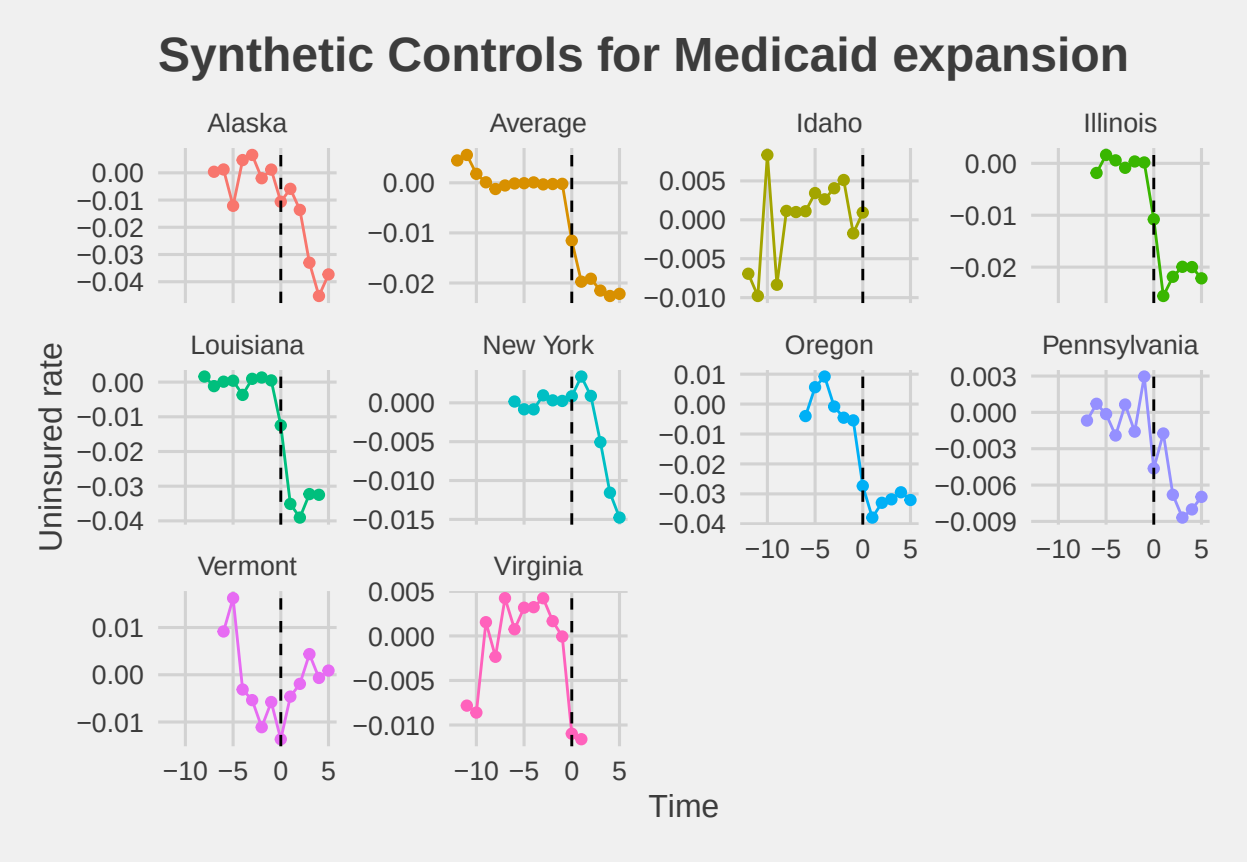
```
## [1] 0.2998225
```

```
ppool_syn_summ<- summary(ppool_syn)


set.seed(1233)
```

```r
# Randomly select 5 states from the dataset
random_states <- ppool_syn_summ$att %>%
  distinct(Level) %>%
  slice_sample(n = 10) %>%
  pull(Level)  # Extracting the state names

# Filter data to include only the randomly selected states
filtered_data <- ppool_syn_summ$att %>%
  filter(Level %in% random_states)

# Plotting the synthetic control estimates for the randomly selected group of states
filtered_data %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0, linetype="dashed") +  # Adding a dashed line at x = 0 for visual reference
  facet_wrap(~Level, scales = "free_y") +  # Facet by state (Level) and allow Y scales to vary
  labs(title = 'Synthetic Controls for Medicaid expansion',
       x = 'Time',
       y = 'Uninsured rate') +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        legend.position = 'None')  # Removing the legend
```

```
## Warning: Removed 59 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 59 rows containing missing values or values outside the scale range
## (`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.

```
# break observations into time cohorts
# ------------------------------------------------------------------
ppool_syn_time <- multisynth(uninsured_rate ~ prepost,
                         State,                          # unit
                         year,                           # time
                         medicaid_expansion_noDC,
                         n_leads =6,
                         time_cohort = TRUE)             # time cohort set to TRUE

# save summary
ppool_syn_time_summ <- summary(ppool_syn_time)

# view
ppool_syn_time_summ
```

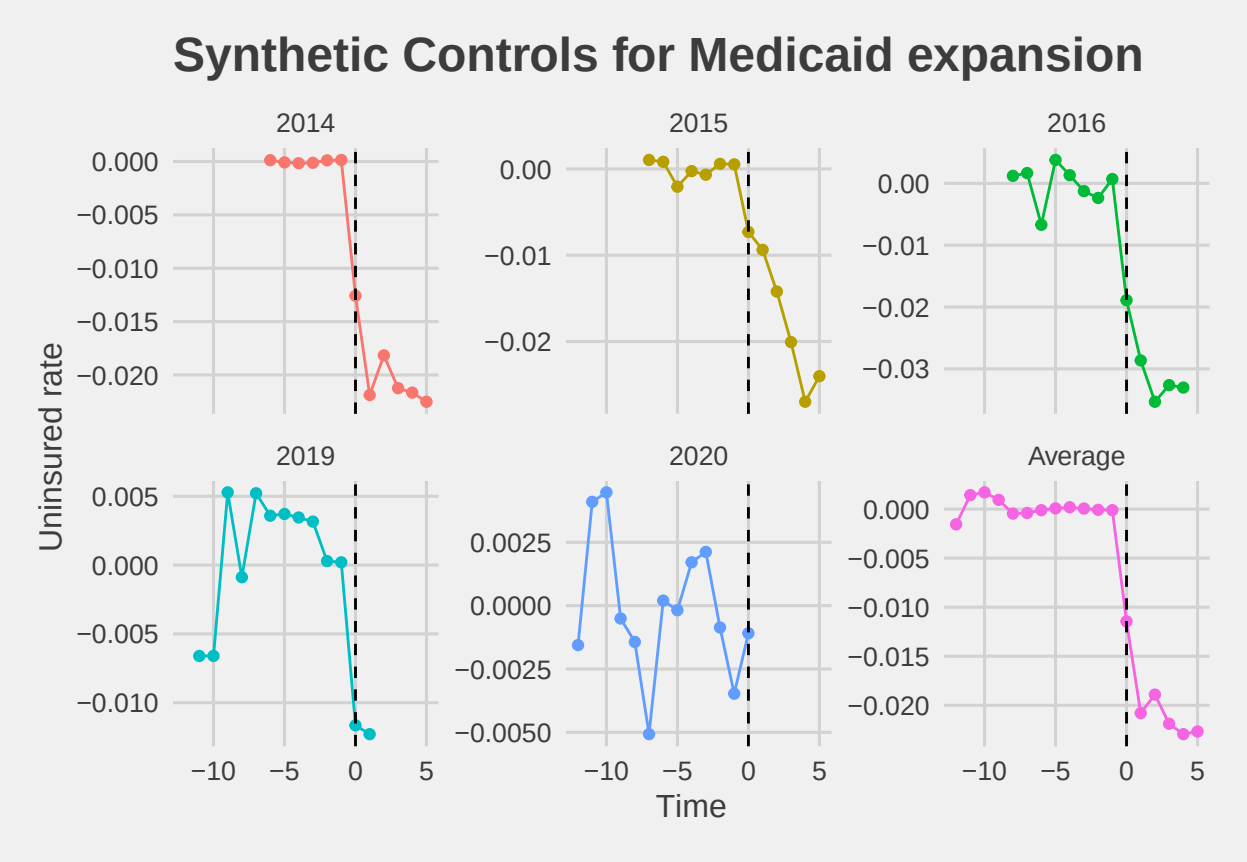
```
##
## Call:
## multisynth(form = uninsured_rate ~ prepost, unit = State, time = year,
##     data = medicaid_expansion_noDC, n_leads = 6, time_cohort = TRUE)
##
```

```
## Average ATT Estimate (Std. Error): -0.018  (0.005)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.009
## Percent improvement from uniform global weights: 99.1
##
## Individual L2 Imbalance: 0.005
## Scaled Individual L2 Imbalance: 0.020
## Percent improvement from uniform individual weights: 98
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01144842 0.004575223 -0.02098951 -0.003627693
##                     1 Average -0.02079388 0.005953941 -0.03263775 -0.008456091
##                     2 Average -0.01890875 0.005666561 -0.02987306 -0.008668575
##                     3 Average -0.02187971 0.005939081 -0.03412372 -0.011713997
##                     4 Average -0.02294204 0.005805471 -0.03453523 -0.012641986
##                     5 Average -0.02266974 0.005750122 -0.03421260 -0.012434464
```

```r
set.seed(1233)


# Plotting the synthetic control estimates for the randomly selected group of states
ppool_syn_time_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0, linetype="dashed") +  # Adding a dashed line at x = 0 for visual reference
  facet_wrap(~Level, scales = "free_y") +  # Facet by state (Level) and allow Y scales to vary
  labs(title = 'Synthetic Controls for Medicaid expansion',
       x = 'Time',
       y = 'Uninsured rate') +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        legend.position = 'None')  # Removing the legend
```

```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`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**: Yes, it seems that the effect size varied by states and probably some will be explained by the heterogenity of the implementation, it is also hard to tell due to the limited time periods after the adoption of certain states that could also have self selected on unobservable variables.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: It is not very clear that the idea is supported by my results. For adopters in 2014 and 2015 the reduction seems very similar, but for adopters in 2016 the reduction was higher and for 2019 it went the opposite way to something way lower, but with only a couple of years to compare. I think the idea would not be supported but also I don't think the data is conclusive.

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?

- **Answer**: DiD and SC help to take advantage of policy shocks, or interventions at aggragate levels to create counterfactuals that might be easier to justify in the aggregate level. Leveraging the fact that policies might vary by cities, states or countries, and not necessarily by individuals, this also helps to control for some of the unobservables and heterogenous factors at the individual level in some cases. It is harder to make the case that two individuals might be identical without a super complete dataset, but it is easier with groups. Both this metodologies rely in rethoric in certain way, and knowing about characteristics of the groups could be easier to justify than at the individual level.

- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?

- **Answer**: Endogeneity plays a very important role in DiD and SC, as it might not be easily identified or even observable. In a paper that I worked for a long time, I was using DiD to analyze the effect of a Criminal Justice Reform in Mexico. States decided when to start it baased on their resources, and capabilities, and political reasons so it was hard to justify that starting earlier or not even starting, was endogenous, thus weakening the parallel trends assumptions. As the control group was allegedly very different. The paper never came to light :(So yes, the lack of a full test of parallel trends will always leave the endogeneity quesiton open to rethoric and expertise on the field. In RDD it might be less important as usually the discontinuity needs to be explained as a very random threshold. Obviously, endogeneity can also happen. For example, in another paper I worked on, I was trying to use the time of the first pregnancy of women as a discontinuity for the wage trajectories. Nonetheless, the decision of when to have a baby is not random, and it might be endogenous to wages. So you need to justify the exogeneity of the shock too, from the beginning. One of the coolest papers that I heard recently, uses the threshold of high blood pressure and treatment below and above a blood pressure number, as one might not be very aware the threshold is kind of imprecise, so people near the cut might be considered very similar, but some will get medication and treatment and others don't. Regression discontinuity, might be better equipped when units around the cut are very similar, and you need data that has that type of granularity and detail around the cut. Whereas DiD and SC can have longer time periods analyzed, and can be better equipped to lead with heterogenity of other kinds.