Project 7: Difference-in-Differences and Synthetic Control

Alex Schulte

April 21, 2024

```
# 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)


options(repos = c(CRAN = "https://cran.rstudio.com"))

# set seed
set.seed(44)

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

medicaid_expansion <- read.csv("~/git/Computational-Social-Science-Projects/My Project 7/data/medicaid_e

#View(medicaid_expansion)

summary(medicaid_expansion)
```

```
##     State           Date_Adopted            year      uninsured_rate
##  Length:663         Length:663         Min.   :2008   Min.   :0.02495
##  Class :character   Class :character   1st Qu.:2011   1st Qu.:0.07702
##  Mode  :character   Mode  :character   Median :2014   Median :0.10475
##                                        Mean   :2014   Mean   :0.10978
##                                        3rd Qu.:2017   3rd Qu.:0.13888
##                                        Max.   :2020   Max.   :0.24082
##
##    population
##  Min.   :  584153
##  1st Qu.: 1850326
##  Median : 4531566
##  Mean   : 6364343
##  3rd Qu.: 7061530
##  Max.   :38802500
##  NA's   :13
```

Introduction

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

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

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

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

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

Data

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

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

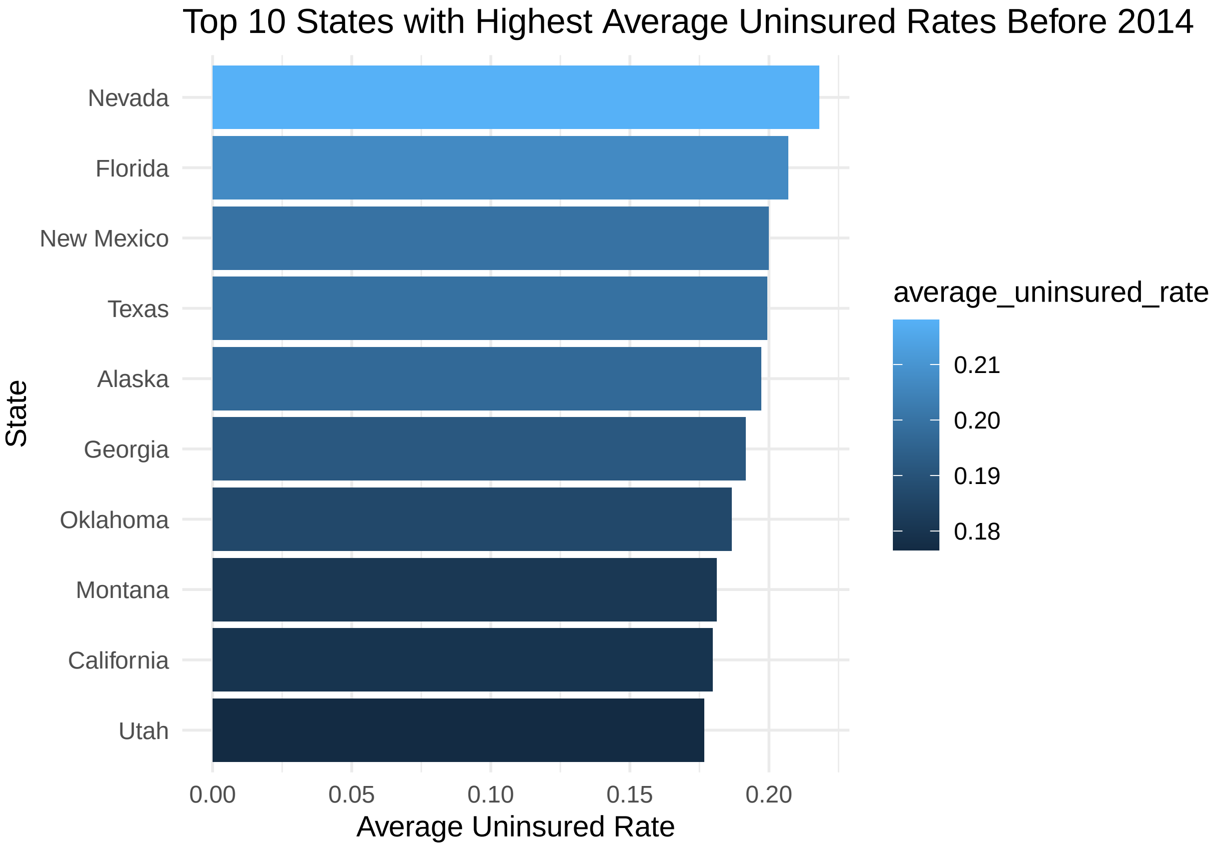
Exploratory Data Analysis

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

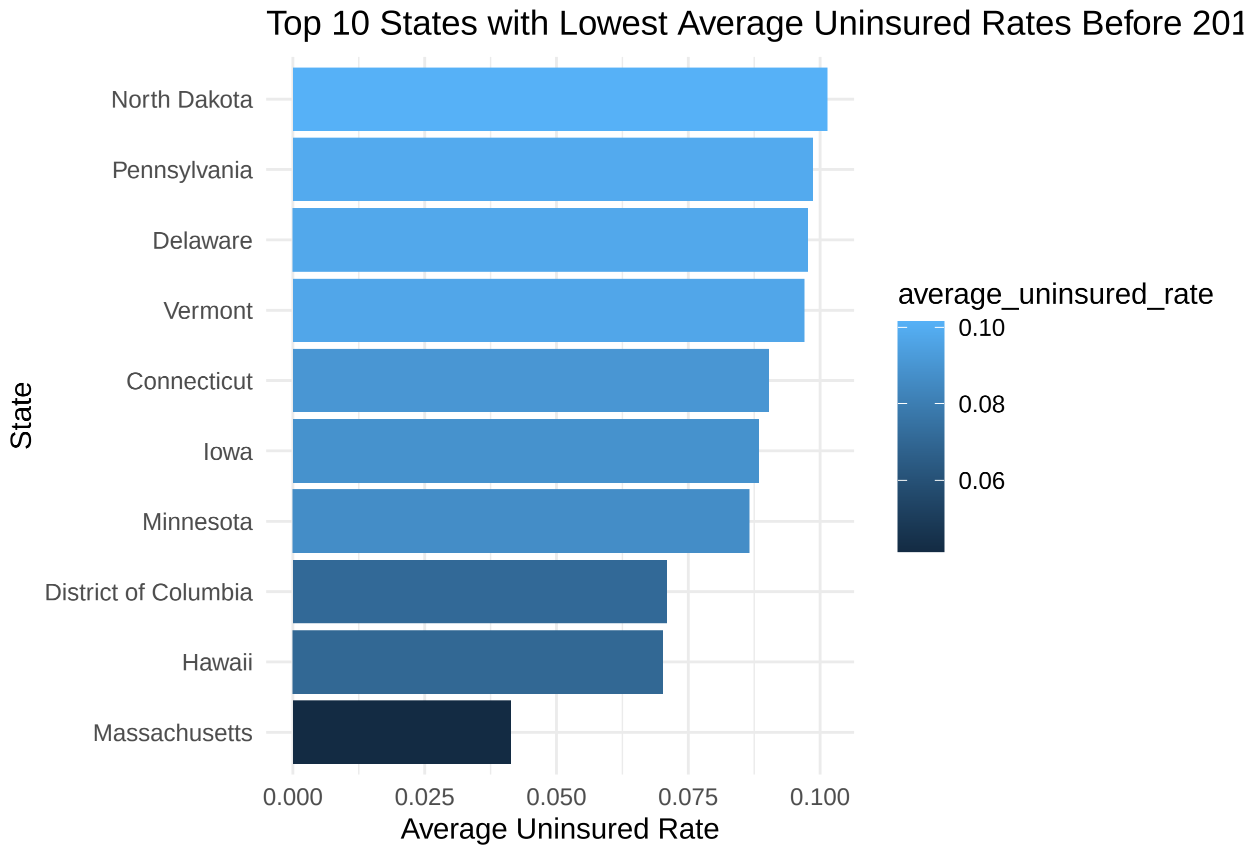
- Which states had the highest uninsured rates prior to 2014? The lowest?

The plots show that Nevada, Florida, and New Mexico had the highest rates before 2014. Massachusetts, Hawaii, and DC (assuming that counts as a state) had the lowest.

```
library(ggplot2)
library(dplyr)

# create dataset with only years before 2014 & get avg annual rate per state
pre_2014_data <- subset(medicaid_expansion, year < 2014)

pre_2014_avg <- pre_2014_data %>%
  group_by(State) %>%
  summarise(average_uninsured_rate = mean(uninsured_rate, na.rm = TRUE))

# highest (note: only plotting top 10 for readability)
ggplot(head(pre_2014_avg %>% arrange(desc(average_uninsured_rate)), 10), aes(x = reorder(State, average_
  geom_bar(stat = "identity") +
  labs(title = "Top 10 States with Highest Average Uninsured Rates Before 2014", x = "State", y = "Avera
  coord_flip() +
  theme_minimal()
```



```
# lowest (note: only plotting top 10 for readability)
ggplot(head(pre_2014_avg %>% arrange(average_uninsured_rate), 10), aes(x = reorder(State, average_uninsu
  geom_bar(stat = "identity") +
  labs(title = "Top 10 States with Lowest Average Uninsured Rates Before 2014", x = "State", y = "Averag
  coord_flip() +
  theme_minimal()
```



- 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.

In 2014, the top 5 states with the most uninsured people were CA, TX, FL, NY, and GA. In 2020, it was TX, CA, Fl, CA and NC. I'm assuming most of the changes were due to which states expanded Medicaid and which did not.

```
medicaid_expansion<- medicaid_expansion %>%
  arrange(State, year)

#View(medicaid_expansion)
head(medicaid_expansion)
```

```
##     State Date_Adopted year uninsured_rate population
## 1 Alabama         <NA> 2008       0.139716    4849377
## 2 Alabama         <NA> 2009       0.138865    4849377
## 3 Alabama         <NA> 2010       0.147653    4849377
## 4 Alabama         <NA> 2011       0.141378    4849377
## 5 Alabama         <NA> 2012       0.132940    4849377
## 6 Alabama         <NA> 2013       0.136610    4849377
```

```
# Remove rows where State is "District of Columbia" bc missing population for all years
filtered_data <- medicaid_expansion %>%
  filter(State != "District of Columbia")

# Calculate uninsured individuals for data prior to 2014
uninsured_pre_2014 <- filtered_data %>%
```

```r
  filter(year < 2014) %>%
  mutate(uninsured_individuals = uninsured_rate * population) %>%
  group_by(State) %>%
  summarise(total_uninsured = sum(uninsured_individuals, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(total_uninsured))

# Calculate uninsured individuals for 2020 (last year in dataset)
uninsured_2020 <- filtered_data %>%
  filter(year == 2020) %>%
  mutate(uninsured_individuals = uninsured_rate * population) %>%
  group_by(State) %>%
  summarise(total_uninsured = sum(uninsured_individuals, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(total_uninsured))

# Print results, pre 2014
print(uninsured_pre_2014)
```

```
## # A tibble: 50 x 2
##    State          total_uninsured
##    <chr>                    <dbl>
##  1 California            41824710.
##  2 Texas                 32232489.
##  3 Florida               24690566.
##  4 New York              13572826.
##  5 Georgia               11611631.
##  6 Illinois              10163254.
##  7 North Carolina         9732267.
##  8 Ohio                   8271125.
##  9 Pennsylvania           7565335.
## 10 Arizona                7080801.
## # i 40 more rows
```

```r
# Print results, pre 2020
print(uninsured_2020)
```

```
## # A tibble: 50 x 2
##    State          total_uninsured
##    <chr>                    <dbl>
##  1 Texas                  4960080.
##  2 California             2987792.
##  3 Florida                2625915.
##  4 Georgia                1353044.
##  5 North Carolina         1123668.
##  6 New York               1026804.
##  7 Illinois                953163.
##  8 Ohio                    765215.
##  9 Arizona                 760658.
## 10 Pennsylvania            741658.
## # i 40 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.

```
# Creating a sample table with one row per state and the Date_Adopted
state_date_adopted_table <- medicaid_expansion %>%
  select(State, Date_Adopted) %>%
  distinct(State, .keep_all = TRUE)

#  Arrange by state and date to ensure the earliest date is picked first
state_date_adopted_table <- medicaid_expansion %>%
  arrange(State, Date_Adopted) %>%
  select(State, Date_Adopted) %>%
  distinct(State, .keep_all = TRUE)

print(state_date_adopted_table)
```

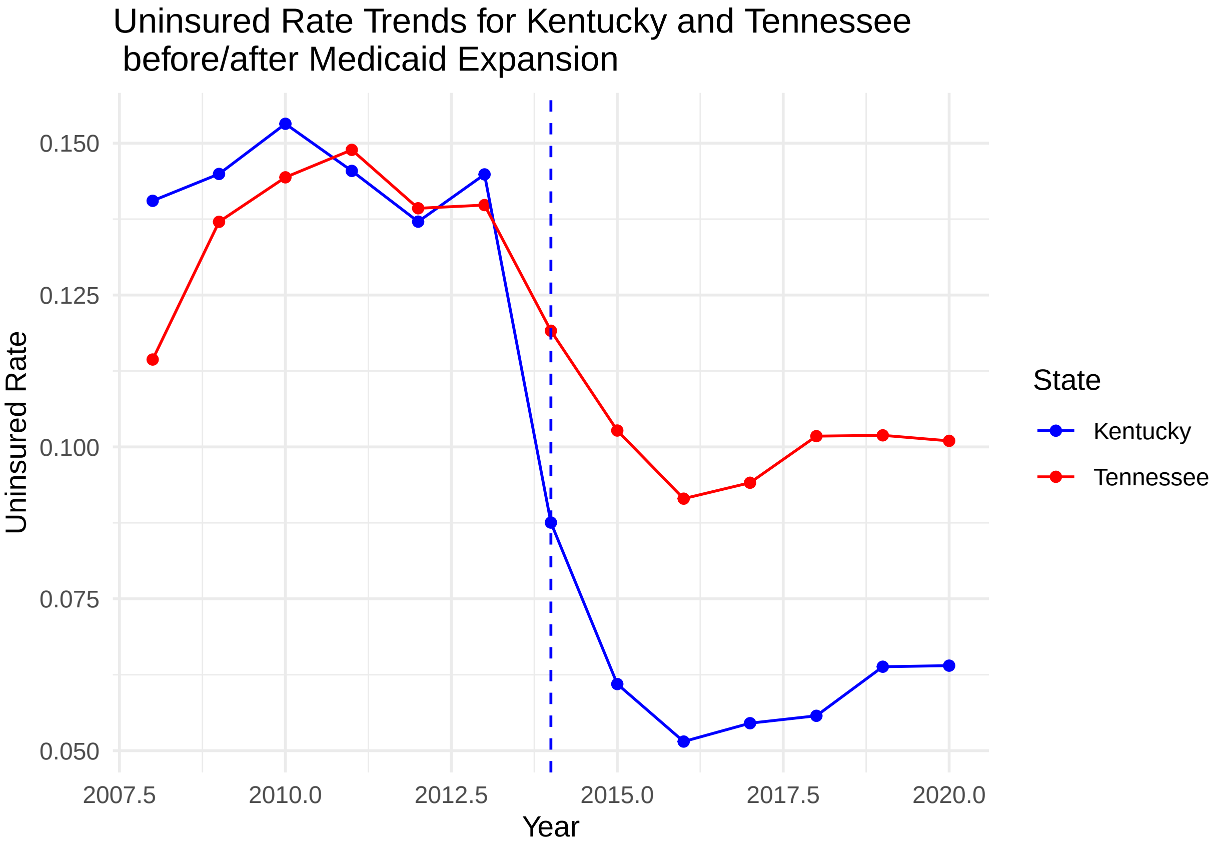
```
##                   State Date_Adopted
## 1               Alabama         <NA>
## 2                Alaska   2015-09-01
## 3               Arizona   2014-01-01
## 4              Arkansas   2014-01-01
## 5            California   2014-01-01
## 6              Colorado   2014-01-01
## 7           Connecticut   2014-01-01
## 8              Delaware   2014-01-01
## 9  District of Columbia   2014-01-01
## 10              Florida         <NA>
## 11              Georgia         <NA>
## 12               Hawaii   2014-01-01
## 13                Idaho   2020-01-01
## 14             Illinois   2014-01-01
## 15              Indiana   2015-02-01
## 16                 Iowa   2014-01-01
## 17               Kansas         <NA>
## 18             Kentucky   2014-01-01
## 19            Louisiana   2016-07-01
## 20                Maine         <NA>
## 21             Maryland   2014-01-01
## 22        Massachusetts   2014-01-01
## 23             Michigan   2014-04-01
## 24            Minnesota   2014-01-01
## 25          Mississippi         <NA>
## 26             Missouri         <NA>
## 27              Montana   2016-01-01
## 28             Nebraska   2020-10-01
## 29               Nevada   2014-01-01
## 30        New Hampshire   2014-08-15
## 31           New Jersey   2014-01-01
## 32           New Mexico   2014-01-01
## 33             New York   2014-01-01
## 34       North Carolina         <NA>
## 35         North Dakota   2014-01-01
## 36                 Ohio   2014-01-01
## 37             Oklahoma         <NA>
## 38               Oregon   2014-01-01
## 39         Pennsylvania   2015-01-01
## 40         Rhode Island   2014-01-01
## 41       South Carolina         <NA>
## 42         South Dakota         <NA>
## 43            Tennessee         <NA>
## 44                Texas         <NA>
## 45                 Utah   2020-01-01
## 46              Vermont   2014-01-01
## 47             Virginia   2019-01-01
## 48           Washington   2014-01-01
## 49        West Virginia   2014-01-01
## 50            Wisconsin         <NA>
## 51              Wyoming         <NA>
```

```
#View(state_date_adopted_table)
```

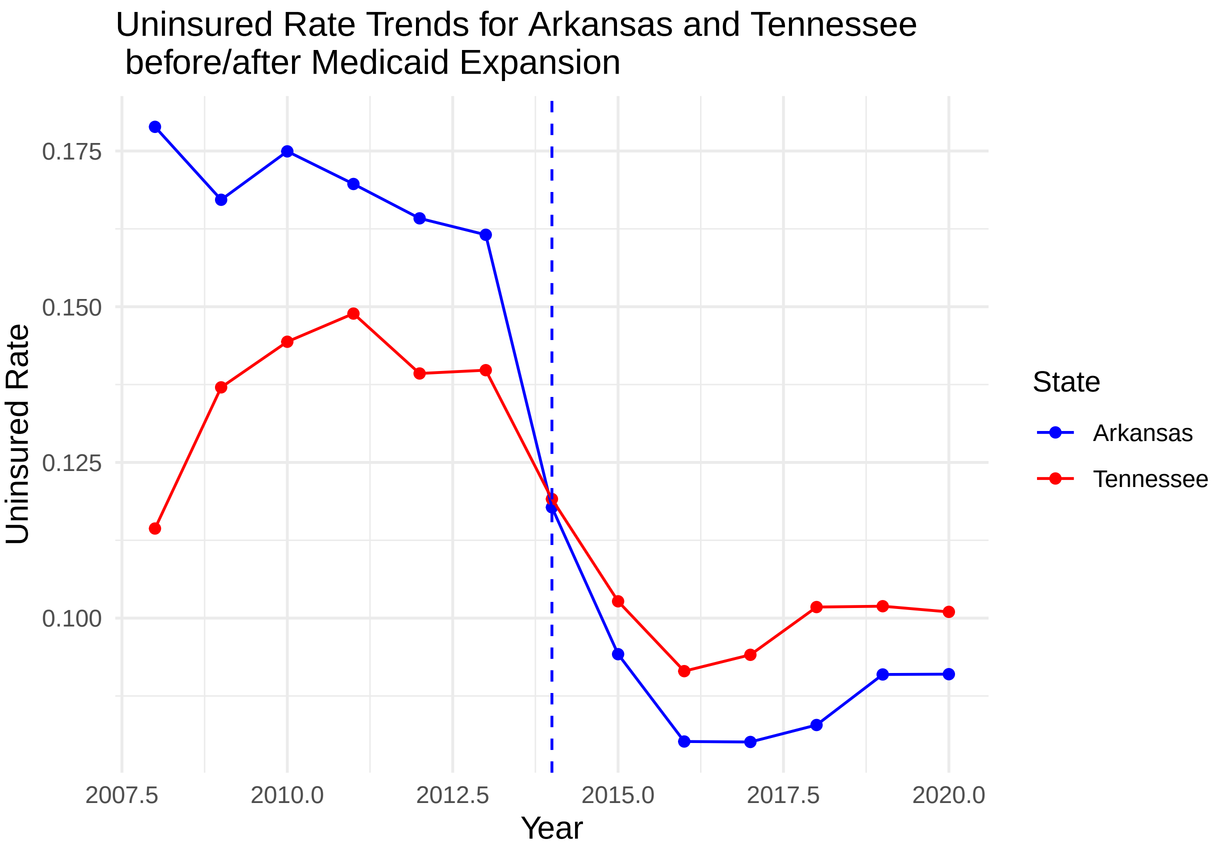
Feasible choices based on this table and other research: KY vs. TN, AR vs. TN, AR vs. MS, MN vs. WI

- 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).

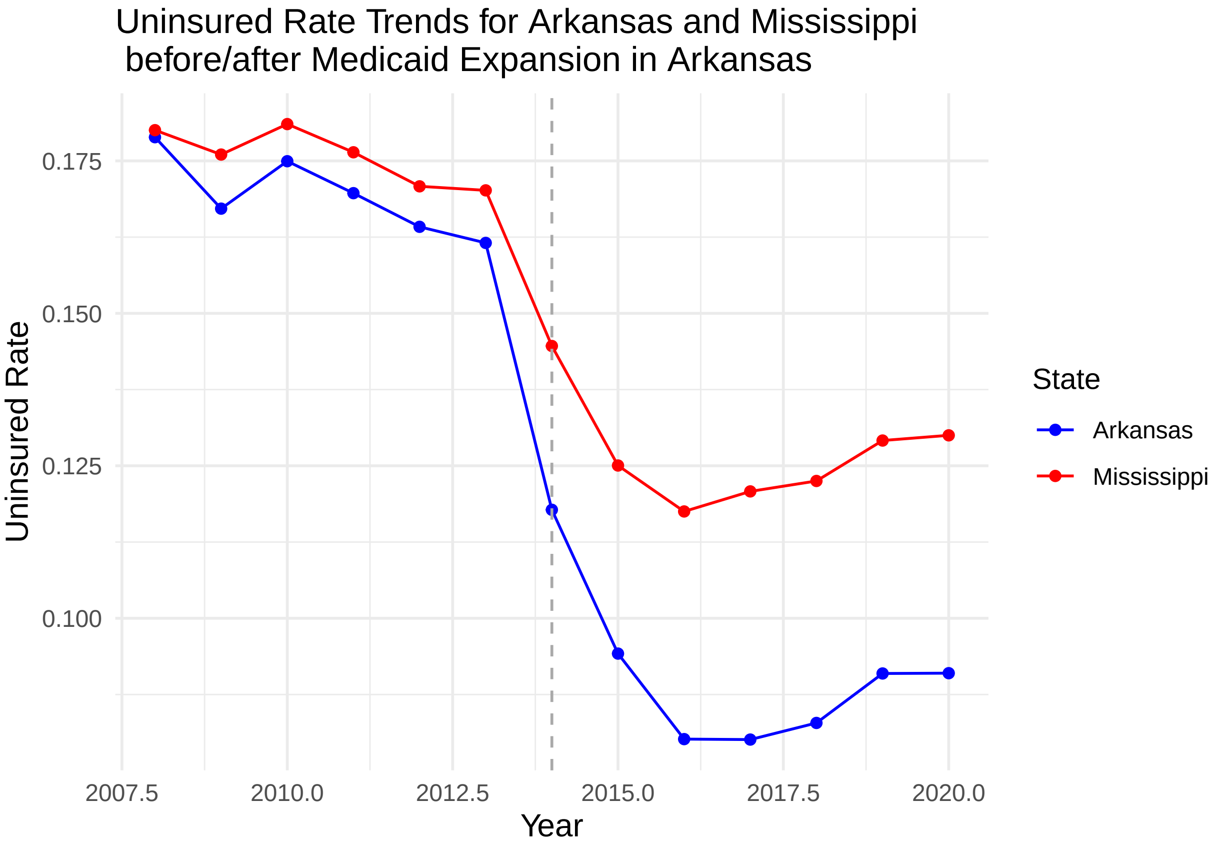
```
# KY vs. TN

medicaid_expansion %>%

  # Process
  #---------
  # Filter for Kentucky and Tennessee
  filter(State %in% c("Kentucky", "Tennessee")) %>%

  # Plot
  #---------
  ggplot() +
    # Add in point layer
    geom_point(aes(x = year,
                   y = uninsured_rate,
                   color = State)) +
    # Add in line layer
    geom_line(aes(x = year,
                  y = uninsured_rate,
                  color = State)) +
    # Add a vertical line at the start of 2014, when Kentucky expanded Medicaid
    geom_vline(aes(xintercept = 2014), linetype="dashed", color = "blue") +

    # Themes
    theme_minimal() +
    theme(axis.title = element_text()) +

    # Labels
    ggtitle('Uninsured Rate Trends for Kentucky and Tennessee \n before/after Medicaid Expansion') +
    xlab('Year') +
    ylab('Uninsured Rate') +
    scale_color_manual(values = c("Kentucky" = "blue", "Tennessee" = "red"))
```



```
# AR vs. TN

medicaid_expansion %>%

  # Process
  #---------
  # Filter for Arkansas and Tennessee
  filter(State %in% c("Arkansas", "Tennessee")) %>%

  # Plot
  #---------
  ggplot() +
    # Add in point layer
    geom_point(aes(x = year,
                   y = uninsured_rate,
                   color = State)) +
    # Add in line layer
    geom_line(aes(x = year,
                  y = uninsured_rate,
                  color = State)) +
    # Add a vertical line at the start of 2014, when Arkansas expanded Medicaid
    geom_vline(xintercept = 2014, linetype="dashed", color = "blue") +
    # Themes
    theme_minimal() +
    theme(axis.title = element_text(size = 12)) +

    # Labels
    ggtitle('Uninsured Rate Trends for Arkansas and Tennessee \n before/after Medicaid Expansion') +
    xlab('Year') +
    ylab('Uninsured Rate') +
    scale_color_manual(values = c("Arkansas" = "blue", "Tennessee" = "red"))
```



```
# AR (treatment) vs. MS (control)

medicaid_expansion %>%

  # Process
  #---------
  # Filter for Arkansas and Mississippi
  filter(State %in% c("Arkansas", "Mississippi")) %>%

  # Plot
  #---------
  ggplot() +
    # Add in point layer
    geom_point(aes(x = year,
                   y = uninsured_rate,
                   color = State)) +
    # Add in line layer
    geom_line(aes(x = year,
                  y = uninsured_rate,
                  color = State)) +
    # Add a vertical line at the start of 2014, when Arkansas expanded Medicaid
    geom_vline(xintercept = 2014, linetype="dashed", color = "darkgrey") +

    # Themes
    theme_minimal() +
    theme(axis.title = element_text(size = 12)) +

    # Labels
    ggtitle('Uninsured Rate Trends for Arkansas and Mississippi \n before/after Medicaid Expansion in A
    xlab('Year') +
    ylab('Uninsured Rate') +
    scale_color_manual(values = c("Arkansas" = "blue", "Mississippi" = "red"))
```



FINDING: AR vs. MS is best meets parallel trends assumption

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

```
# Difference-in-Differences estimation for AR vs. MS
# modeling after KS vs. CO code in lab

# Filter dataset for Arkansas and Mississippi from 2008 to 2020
am_data <- medicaid_expansion %>%
  filter(State %in% c("Arkansas", "Mississippi"))

# Calculate pre-treatment and post-treatment averages
# Splitting at 2014 for intervention
pre_treatment <- am_data %>%
  filter(year < 2014) %>%
  group_by(State) %>%
  summarise(average_pre = mean(uninsured_rate, na.rm = TRUE))

post_treatment <- am_data %>%
  filter(year >= 2014) %>%
  group_by(State) %>%
  summarise(average_post = mean(uninsured_rate, na.rm = TRUE))

# Join pre and post data for easier calculations
did_data <- left_join(pre_treatment, post_treatment, by = "State")

# Calculate DiD
# Making data wide to perform the subtraction
did_data_wide <- did_data %>%
  pivot_wider(names_from = State, values_from = c(average_pre, average_post))

# Calculate differences and Diff-in-Diff
diffs <- did_data_wide %>%
  summarise(
    pre_diff = `average_pre_Arkansas` - `average_pre_Mississippi`,
    post_diff = `average_post_Arkansas` - `average_post_Mississippi`,
    DiD = (post_diff - pre_diff)
  )

print(diffs)
```

```
## # A tibble: 1 x 3
##   pre_diff post_diff     DiD
##      <dbl>     <dbl>   <dbl>
## 1 -0.00635   -0.0361 -0.0297
```

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**: Although AR and TN do share a small border along the Mississippi river, I think they are not as intuitively similar as the states in Card/Krueger's original piece. In addition, we're looking at broader geographic areas (states vs. towns) – States are much larger and more diverse than towns. This diversity across states includes urban and rural populations, economic diversity (such as the presence of different industries), and racial and ethnic diversity, all of which can affect uninsured rates differently. These factors can create varying starting points and different trajectories post-policy implementation, complicating the validity of the parallel trends assumption and direct comparisons across treatment vs. control states.

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

- **Answer**: Key strengths are that if researchers (and their audiences) are confident parallel trends hold, then they make some causal claims (which will be stronger with several sensitivity analyses, etc.) This is true because the DID design and assumption of parallel trends effectively controls for unobserved confounders that are constant over time. Another strength is that this method is relatively easy to implement.

Key weaknesses are the impossibility of 100% proving the parallel trends assumption because not not directly testable. Additionally, DID assumes that the treatment effect is constant over time after the policy is implemented. If the treatment effects grow or diminish over time, standard DID approaches may not accurately capture these dynamics. Last, DID typically does not control for time-varying confounders.

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.

I am choosing Montana, which adopted 1/1/2016.

```
# changing to Date format for code below

class(medicaid_expansion$Date_Adopted)
```

```
## [1] "character"
```

```
medicaid_expansion$Date_Adopted <- as.Date(medicaid_expansion$Date_Adopted, format = "%Y-%m-%d")
class(medicaid_expansion$Date_Adopted)
```

```
## [1] "Date"
```

```
#View(medicaid_expansion)
```

```
# create treatment indicator for Montana

medicaid_expansion_mt <-
  medicaid_expansion %>%
  # select subset of variables
  select(State, Date_Adopted, year, uninsured_rate, population) %>%
  # create treatment flag
  mutate(treatment = case_when(State == "Montana" & year >= 2016 ~ 1,
                               TRUE ~ 0))
# view head
#View(medicaid_expansion_mt)
```

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

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

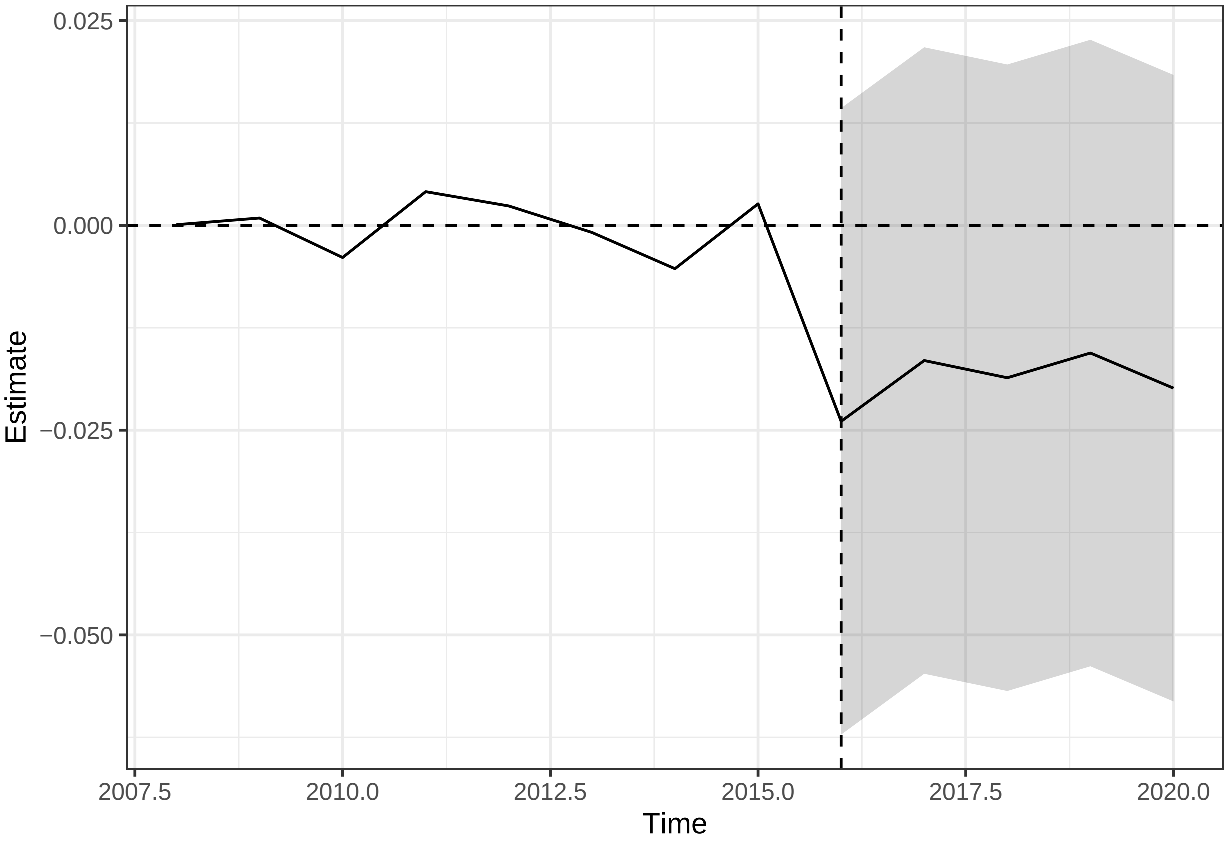
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
# 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):  -0.0189   ( 0.52 )
## L2 Imbalance: 0.009
## Percent improvement from uniform weights: 92.8%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.024             -0.062              0.014   0.110
##  2017   -0.016             -0.055              0.022   0.194
##  2018   -0.019             -0.057              0.020   0.109
##  2019   -0.016             -0.054              0.023   0.113
##  2020   -0.020             -0.058              0.018   0.118
```

Average ATT Estimate (p Value for Joint Null): -0.0189 (0.49) and L2 Imbalance: 0.009

```
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 control - Ridge

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

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

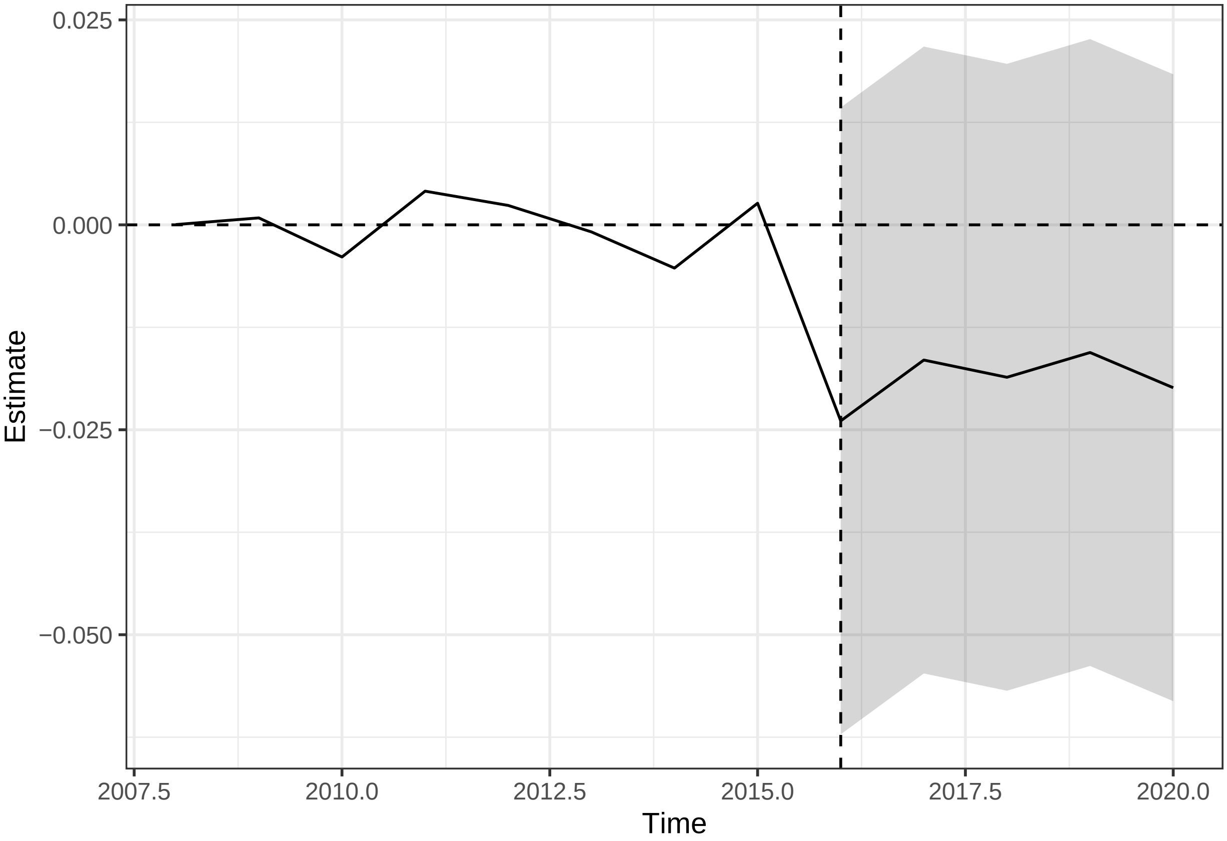
```
# summary
summary(ridge_syn)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "ridge", scm = ..2)
##
## Average ATT Estimate (p Value for Joint Null):  -0.0189   ( 0.51 )
## L2 Imbalance: 0.009
## Percent improvement from uniform weights: 92.8%
##
## Avg Estimated Bias: 0.000
##
```

```
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.024             -0.062              0.014   0.116
##  2017   -0.016             -0.055              0.022   0.223
##  2018   -0.019             -0.057              0.020   0.098
##  2019   -0.016             -0.054              0.023   0.115
##  2020   -0.020             -0.058              0.018   0.100
```

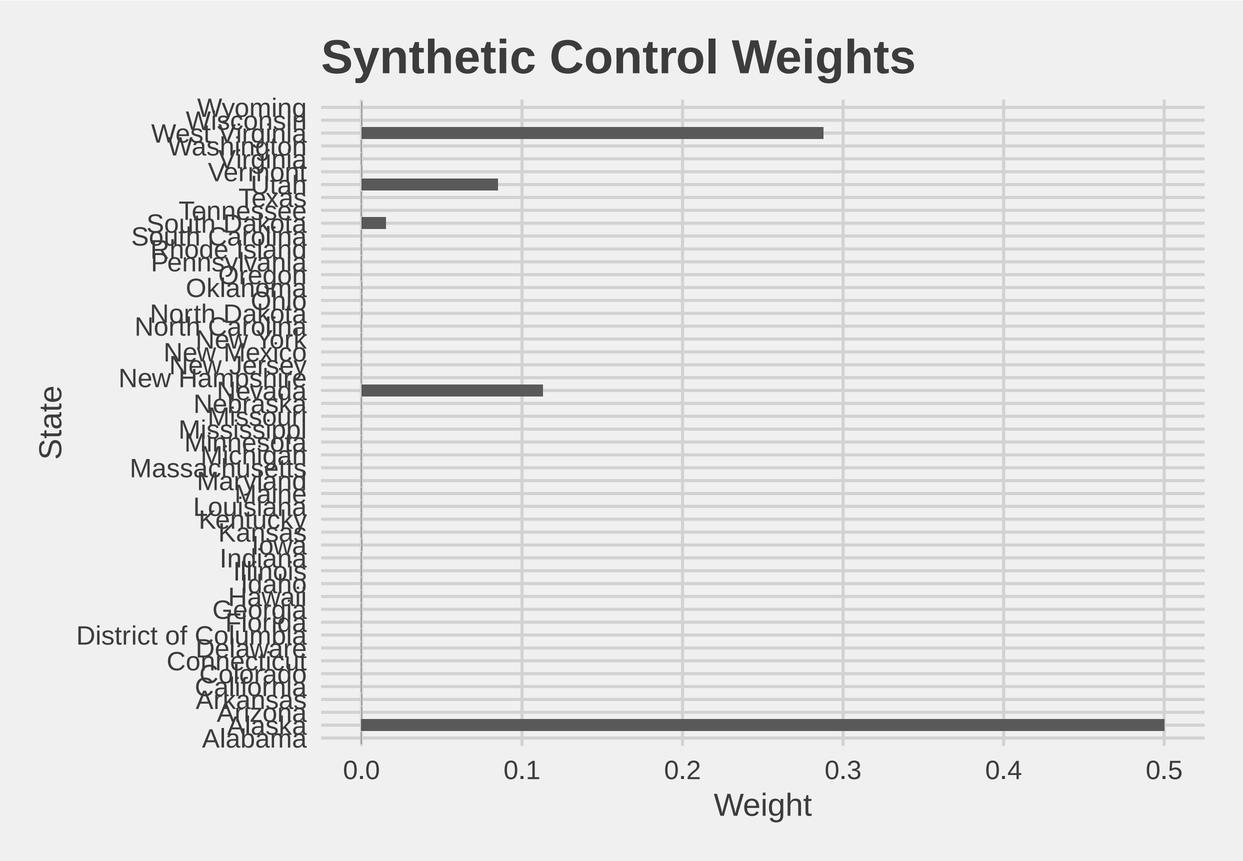
Average ATT Estimate (p Value for Joint Null): -0.0189 (0.5) and L2 Imbalance: 0.009

```r
plot(ridge_syn)
```



- Plot barplots to visualize the weights of the donors.

```r
# barplots of weights

#View(ridge_syn)

data.frame(ridge_syn$weights) %>%
  tibble::rownames_to_column('State') %>%
  ggplot() +
  geom_bar(aes(x = State, y = ridge_syn.weights),
           stat = 'identity') +
  coord_flip() + # coord flip
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  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?

In lab, we saw that some weights became negative when we did ridge augmentation with Kansas, but that is not the case in this Montana example, so I don't think I need to do additional preprocessing.

Discussion Questions

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

- **Answer**: Key advantages of SC over DID are that it provides more flexibility when finding (or constructing) a control that meets the parallel trends assumption. It is often difficult or impossible to find a control unit that has parallel trends to the treated unit and therefore makes sense for the counterfactual in the post period. However, SC methods by design select a weighted average of control units that best resemble the treated units prior to the intervention, thereby potentially providing a better and more tailored comparison group than DID. Additionally, SC better handles time varying unobserved confounders.

A key disadvantage of SC vs DID is SC can be a bit strange in that it doesn't actually (and will never) exist. The 'synthetic Montana' doesn't exist, and maybe there's a reason for that (eg the combination doesn't make sense or isn't realistic). Another disadvantage is there is limited applicability with many treated units (like we'll do in the next section). Constructing a separate synthetic control for each can become infeasible, particularly if the pool of potential control units is limited.

- One of the benefits of synthetic control is that the weights are bounded between [0,1]

and the weights must sum to 1. Augmentation might relax this assumption by allowing for negative weights. Does this create an interpretation problem, and how should we balance this consideration against the improvements augmentation offers in terms of imbalance in the pre-treatment period?

- **Answer**: Yes, this can create an interpretation problem. Negative weights are non-intuitive, but basically mean “short selling” or “betting against” a particular control unit, which does not have a clear real-world counterpart in this context. It can create issues with overfitting (especially with small sample sizes) and could affect the variance of the estimator and create more volatility in the post-treatment effect estimates.

A few important things should be kept in mind when balancing this consideration: sample size, number of predictors, and magnitude of the pre-treatment imbalance. In cases with larger pools of control units and more extensive pre-treatment period data, the traditional constraints (non-negative, summing to one) might be too restrictive, potentially leading to poor fits. In these cases, augmentation will likely provide a more flexible and better-suited model. Additionally, if the traditional synthetic control method fails to achieve a satisfactory pre-treatment fit, this would likely lead to substantial bias in estimating the treatment effect. Therefore, augmentation would be more justified.

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.

```r
# cleaning data (removing DC) and adding treatment indicator
# note that treatment = 1 for the year of Date_Adopted, even if it was not for the full year

# make sure Date_Adopted is in the correct date format
medicaid_expansion$Date_Adopted <- as.Date(medicaid_expansion$Date_Adopted, format = "%Y-%m-%d")

# Clean the dataset
cleaned_medicaid_expansion <- medicaid_expansion %>%
  # Remove "District of Columbia"
  filter(State != "District of Columbia") %>%
  # Create 'treated' variable based on Date_Adopted
  mutate(treated = if_else(year >= as.integer(format(Date_Adopted, "%Y")), 1, 0))

# View the cleaned data
#View(cleaned_medicaid_expansion)
```

```r
# multisynth model states
# implementing staggered adoption
# ----------------------------------------------------------------

#

# with default nu
# ---------
ppool_syn <- multisynth(uninsured_rate ~ treated,
                        State,                        # unit
                        year,                         # time
                        cleaned_medicaid_expansion, # data
                        n_leads = 4)                # post-treatment periods to estimate

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

```
## [1] 0.3058856
```

```r
ppool_syn
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
##     data = cleaned_medicaid_expansion, n_leads = 4)
##
## Average ATT Estimate: -0.015
```

```r
# setting nu to 0.5
# ---------
ppool_syn_0.5 <- multisynth(uninsured_rate ~ treated,
                        State,                       # unit
                        year,                        # time
                        nu = 0.5,                    # varying degree of pooling
                        cleaned_medicaid_expansion, # data
                        n_leads = 4)                # post-treatment periods to estimate




# view results
print(ppool_syn_0.5$nu)
```

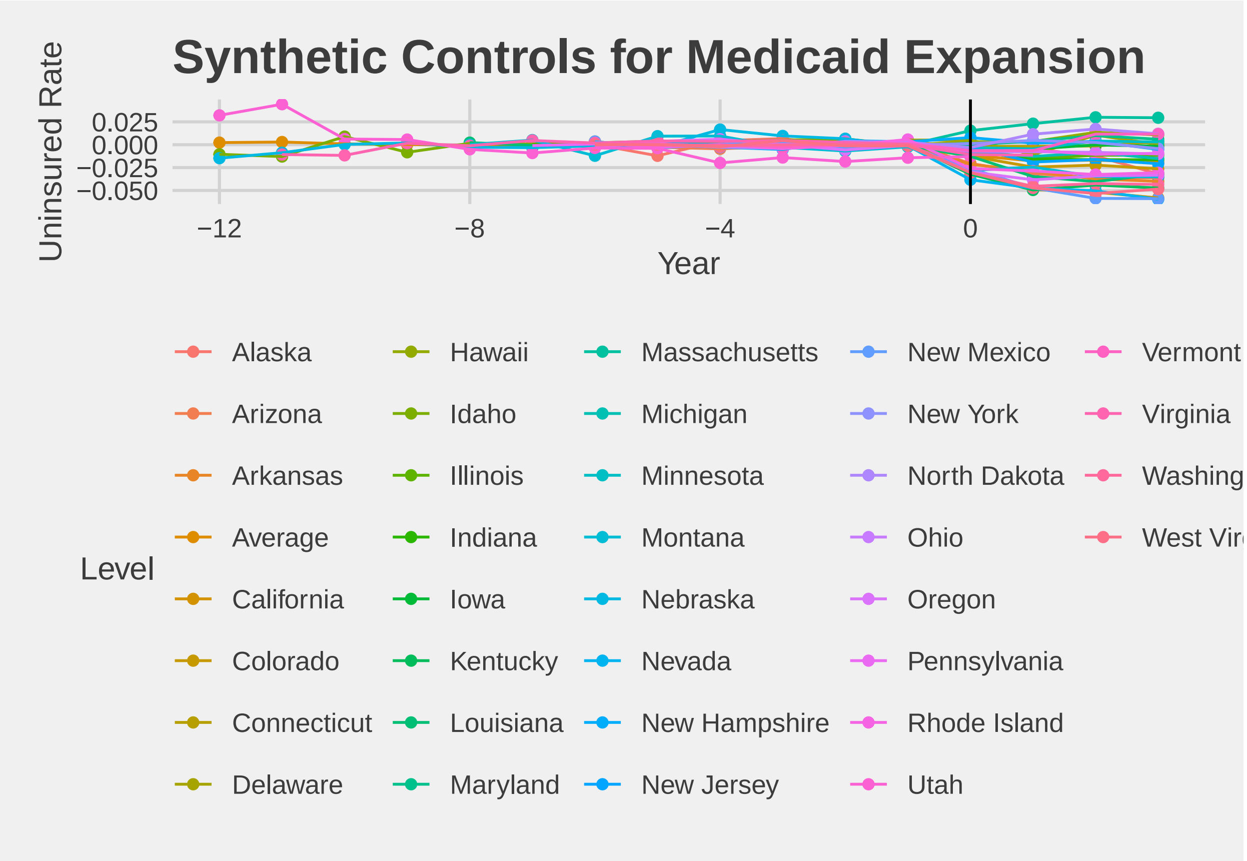
```
## [1] 0.5
```

```r
ppool_syn_summ <- summary(ppool_syn_0.5)

#View(ppool_syn_0.5)
```

```r
# Plotting treatment effects
# I'm just going to do all states like we did in lab

ppool_syn_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        legend.position = "bottom") +
  ggtitle('Synthetic Controls for Medicaid Expansion') +
  xlab('Year') +
  ylab('Uninsured Rate')
```

```
## Warning: Removed 227 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 227 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.

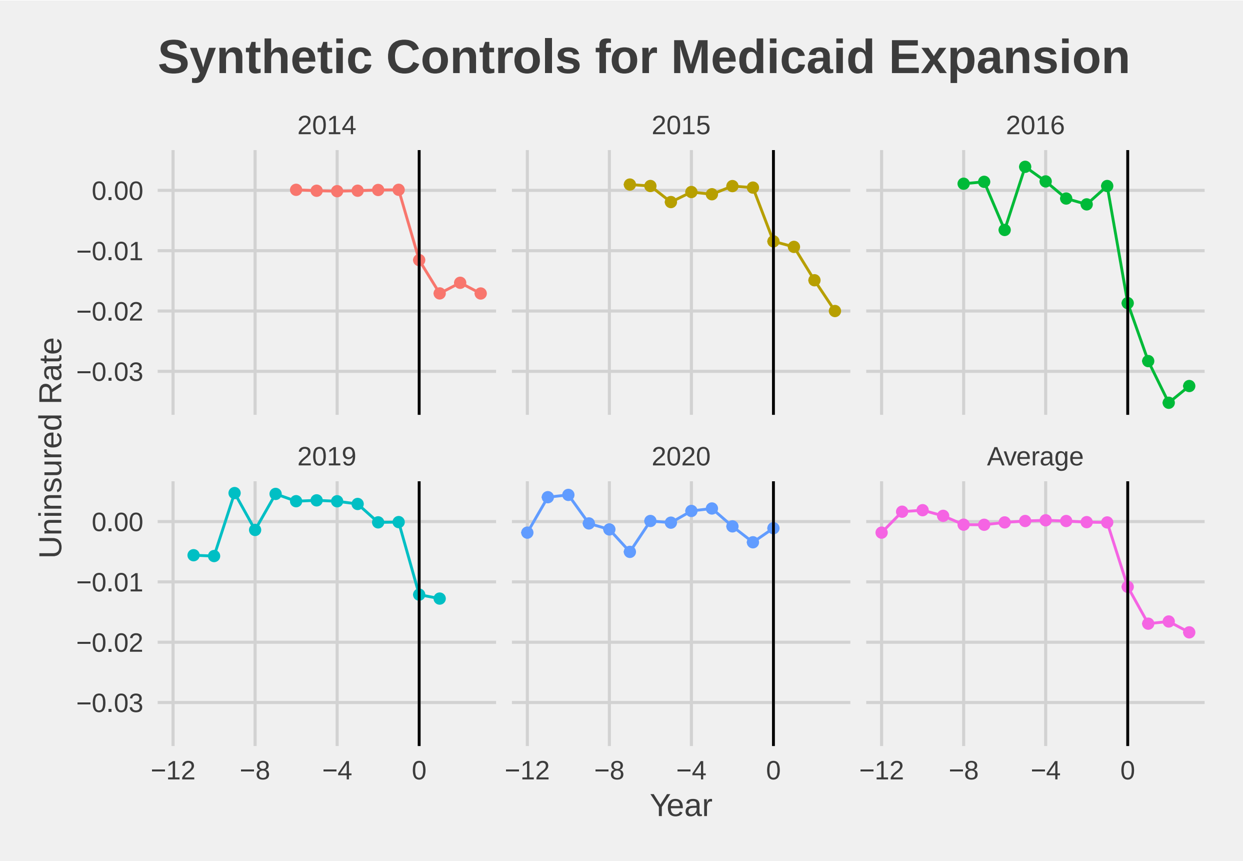
```
# multisynth model time cohorts

#
# break observations into time cohorts
# ----------------------------------------------------------------

ppool_syn_time  <- multisynth(uninsured_rate ~ treated,
                        State,                            # unit
                        year,                             # time
                        nu = 0.5,                         # varying degree of pooling
                        cleaned_medicaid_expansion, # data
                        n_leads = 4,
                        time_cohort = TRUE)                     # post-treatment periods to estimate





# save summary
ppool_syn_time_summ <- summary(ppool_syn_time)

# view
ppool_syn_time_summ
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
##     data = cleaned_medicaid_expansion, n_leads = 4, nu = 0.5,
##     time_cohort = TRUE)
##
## Average ATT Estimate (Std. Error): -0.015  (0.006)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.010
## Percent improvement from uniform global weights: 99
##
## Individual L2 Imbalance: 0.005
## Scaled Individual L2 Imbalance: 0.018
## Percent improvement from uniform individual weights: 98.2
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01081672 0.004683183 -0.02014133 -0.001814551
##                     1 Average -0.01692476 0.006013035 -0.02908437 -0.005907877
##                     2 Average -0.01656187 0.006401752 -0.02903309 -0.004379997
##                     3 Average -0.01836562 0.006532837 -0.03108187 -0.005786763
```

```
# plot effect for each time period (local treatment effects)
# ---------
ppool_syn_time_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        legend.position = 'None') +
  ggtitle('Synthetic Controls for Medicaid Expansion') +
  xlab('Year') +
  ylab('Uninsured Rate') +
  facet_wrap(~Level)
```

```
## Warning: Removed 27 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 27 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, there is clear evidence for this given the graphs above. Although it is outside the scope of this project, it would be interesting to research which types of waivers were associated with larger (or smaller) changes in the uninsurance rates and compare across states

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

- **Answer**: Yes, there is clear evidence of that in the graph above

General Discussion Questions

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

- **Answer**: A few different reasons. For one, these methods do well with aggregated data and heterogeneity. This makes them ideal for analyzing policies implemented at broader geographical levels like states or countries. Another reason is feasibility. Data for cities, states, and countries are often available from governmental or international sources and tend to be consistent and systematically collected, which is well suited for both DiD and SCM. Additionally, both can extend over multiple units and time periods.

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

- **Answer**: A key assumption of DID is NO selection into treatment, although it is not difficult to imagine situations in the real world where this assumption may be violated. On the other hand, it is harder to select into treatment in RDD designs, which are quasi experimental and make more use of randomness. For example, a clear cutoff is low birthweight babies $<$ 2500 g (which get more intense treatment post birth). It is hard to imagine a baby “selecting” into being just below the LBW cutoff.