Project 7: Difference-in-Differences and Synthetic Control

Takun Wang

2024-04-26

1 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:

1. Require companies with more than 50 employees to provide health insurance.
2. Build state-run healthcare markets ("exchanges") for individuals to purchase health insurance.
3. Provide subsidies to middle income individuals and families who do not qualify for employer based coverage.
4. 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.

```r
## Install and load packages
library(tidyverse)
library(Synth)
#devtools::install_github("ebenmichael/augsynth")
library(augsynth)
```

2 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.
- **Population**: State population in 2010.

```
## Load data
med <- read_csv("data/medicaid_expansion.csv")
head(med)
```

State	Date_Adopted	year	uninsured_rate	population
Alabama	NA	2008	0.139716	4849377
Alaska	2015-09-01	2008	0.207716	737732
Arizona	2014-01-01	2008	0.187312	6731484
Arkansas	2014-01-01	2008	0.178883	2994079
California	2014-01-01	2008	0.178212	38802500
Colorado	2014-01-01	2008	0.170183	5355856

```
## Rename variables
med <- med %>% rename(state = State,
                      adoption = Date_Adopted,
                      uninsured = uninsured_rate)

## Explore data structure
med %>% distinct(state) %>% nrow()  # n = 51 (50 states + D.C.)
```

```
## [1] 51
```

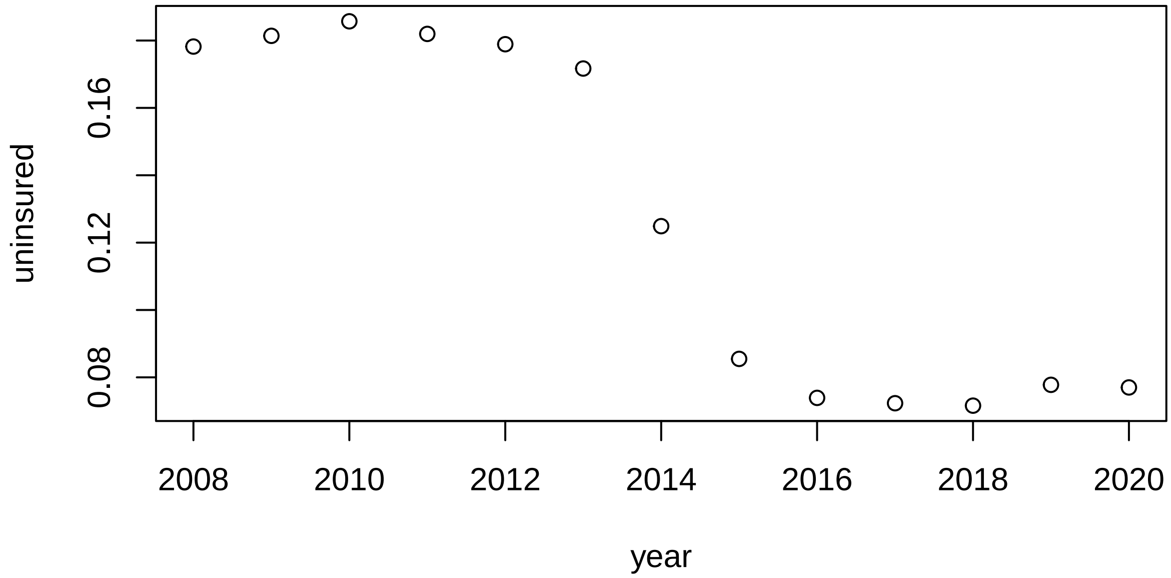
```
med %>% distinct(year) %>% nrow()   # n = 13
```

```
## [1] 13
```

```
med %>% nrow()                      # n = 51 * 13 = 663
```

```
## [1] 663
```

```
## A quick visualization
med %>% filter(state == "California") %>% select(year, uninsured) %>% plot()
```

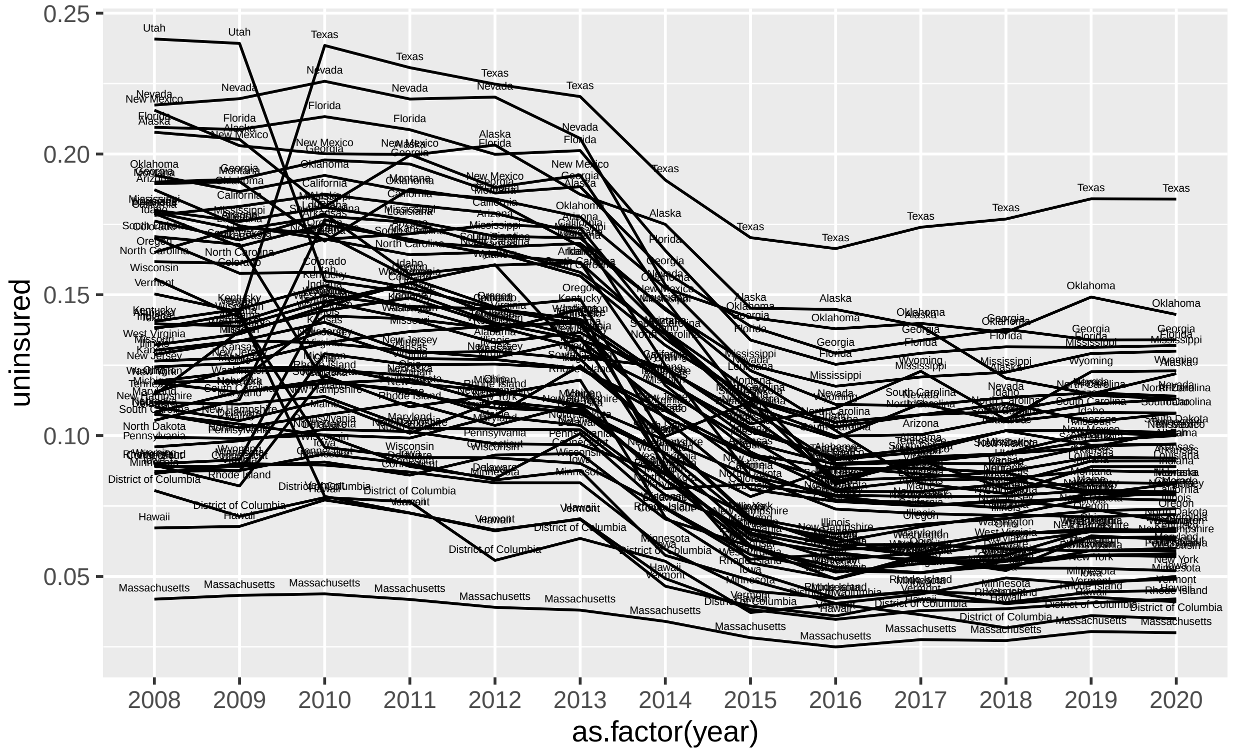


3 Exploratory Data Analysis

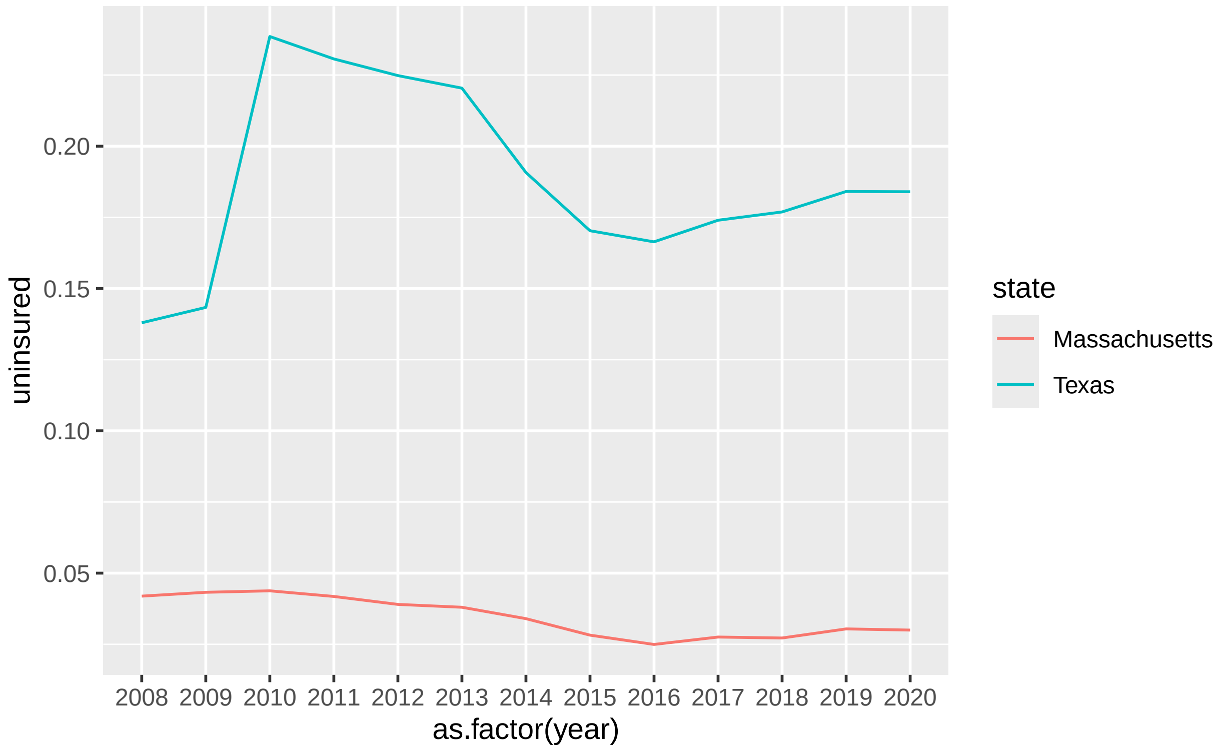
Create plots and provide 1-2 sentence analyses to answer the following questions:

1. Which states had the highest uninsured rates prior to 2014? The lowest?
    - Texas had the highest uninsured rates prior to 2014 and Massachusetts had the lowest.

```
## Plot all states
med %>% ggplot(aes(y = uninsured, x = as.factor(year), group = state)) +
  geom_line() +
  geom_text(aes(x = as.factor(year), y = uninsured, label = state), vjust = -1, size = 1.5)
```



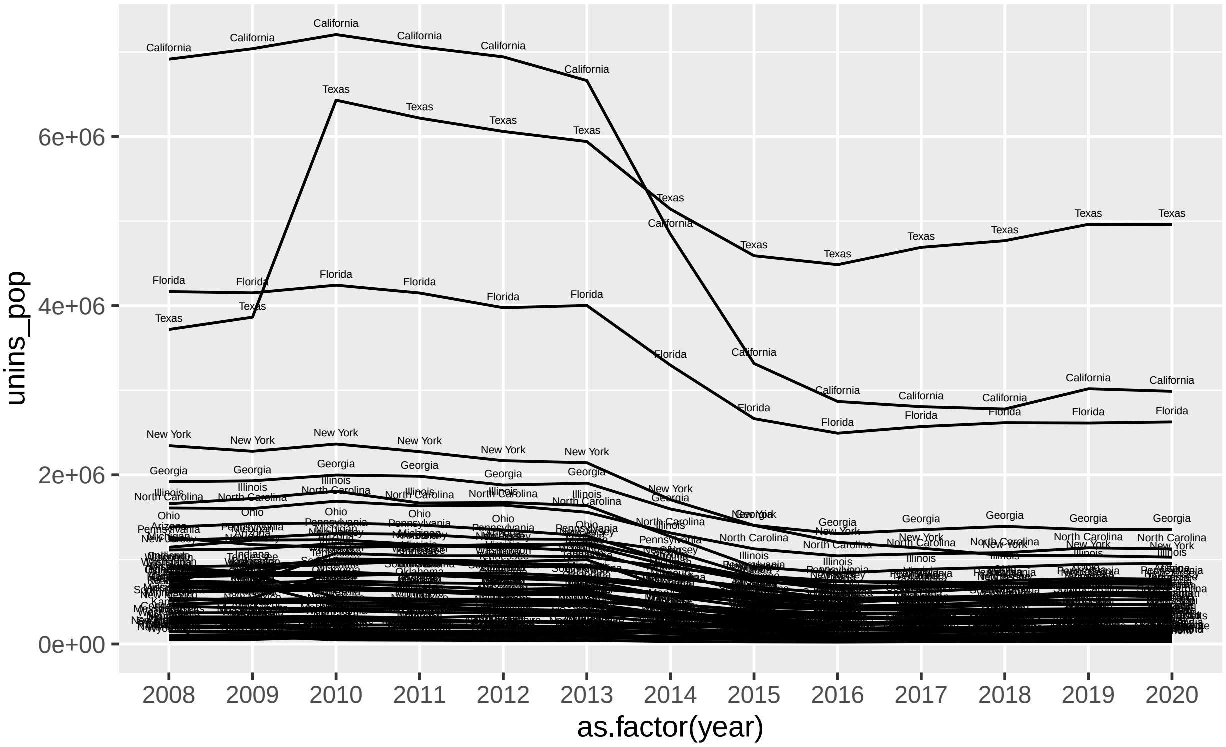
```r
## Focus on the states with highest and lowest uninsured rates
med %>% filter(state == "Massachusetts" | state == "Texas") %>%
  ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = state)) +
  geom_line()
```



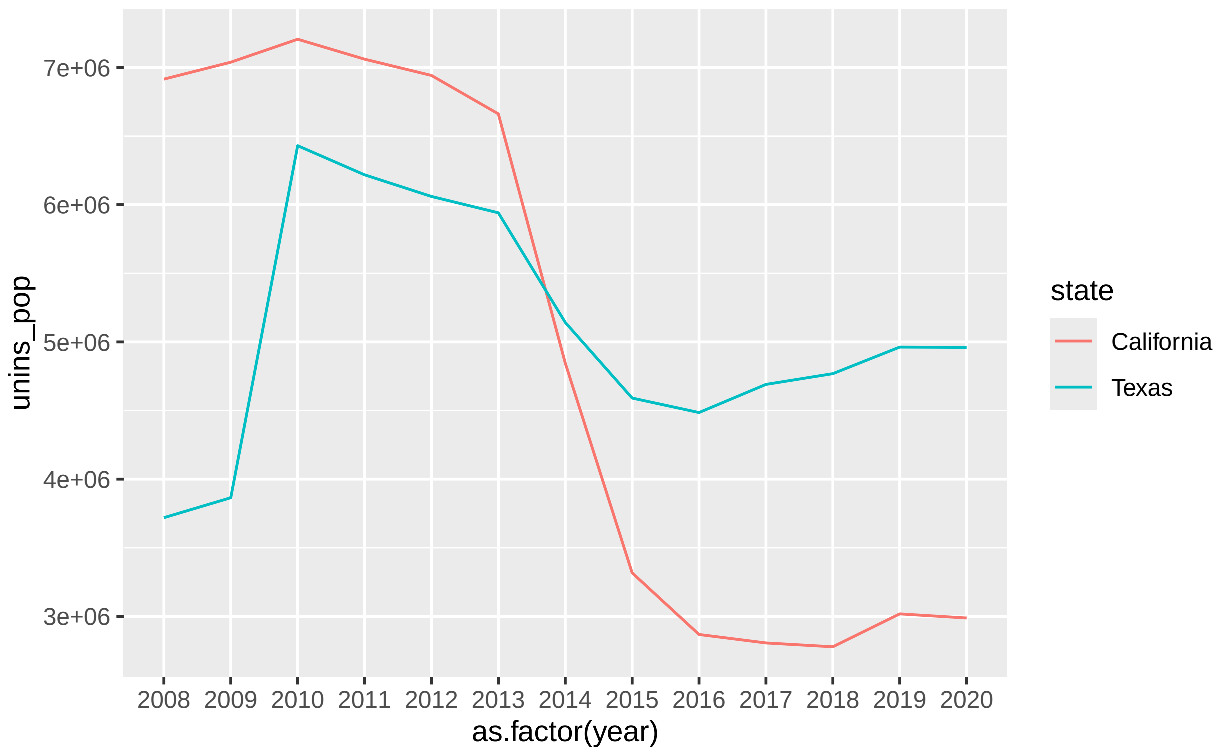
2. Which states were home to most uninsured Americans prior to 2014? How about in the last year in the data set? **Note**: 2010 state population is provided as a variable to answer this question. In an actual study you would likely use population estimates over time, but to simplify you can assume these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

    - California were home to most uninsured Americans prior to 2014 and in 2020, the last year in the data set, Texas were.

```r
## Create uninsured population variable
med <- med %>% mutate(unins_pop = population * uninsured)

## Plot all states
med %>% ggplot(aes(y = unins_pop, x = as.factor(year), group = state)) +
  geom_line() +
  geom_text(aes(x = as.factor(year), y = unins_pop, label = state), vjust = -1, size = 1.5)
```



```r
## Focus on the states with largest uninsured population
med %>% filter(state == "California" | state == "Texas") %>%
  ggplot(aes(y = unins_pop, x = as.factor(year), group = state, color = state)) +
  geom_line()
```



4 Difference-in-Differences Estimation

4.1 Estimation

Do the following:

1. 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.
   - For my selection, I opted for California due to its significant uninsured population in the United States before 2014 and its immediate adoption of Medicaid expansion on January 1, 2014.

```
## All adoption dates
med %>% distinct(adoption) %>% pull()
```

```
##  [1] NA           "2015-09-01" "2014-01-01" "2020-01-01" "2015-02-01"
##  [6] "2016-07-01" "2014-04-01" "2016-01-01" "2020-10-01" "2014-08-15"
## [11] "2015-01-01" "2019-01-01"
```

```
## List of states that did not throughout
state_not_adopted <- med %>% filter(is.na(adoption)) %>% distinct(state) %>% pull(state)
state_not_adopted
```

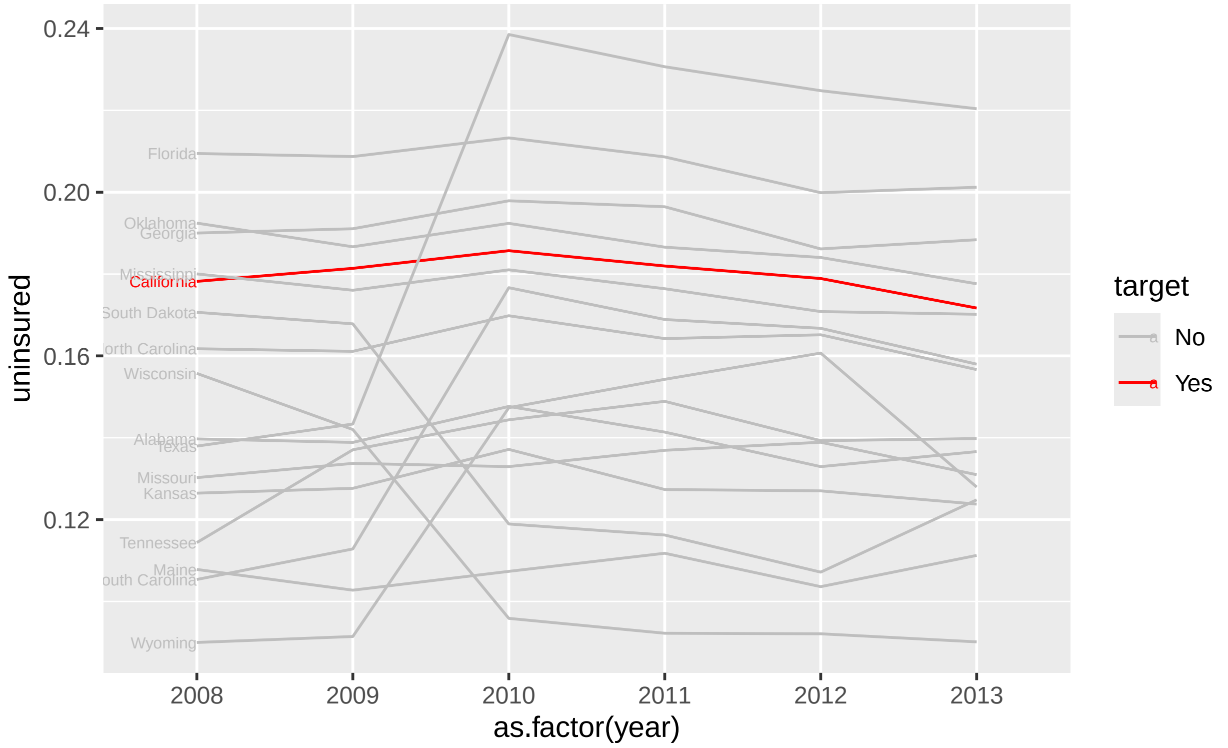
```
##  [1] "Alabama"        "Florida"        "Georgia"        "Kansas"
##  [5] "Maine"          "Mississippi"    "Missouri"       "North Carolina"
##  [9] "Oklahoma"       "South Carolina" "South Dakota"   "Tennessee"
## [13] "Texas"          "Wisconsin"      "Wyoming"
```

```r
## Create a dataframe with only the targeted state and states that did not adopt
target_state <- "California"
med_did <- med %>% filter(state %in% c(target_state, state_not_adopted)) %>%
  mutate(target = if_else(state == target_state, "Yes", "No"))
```

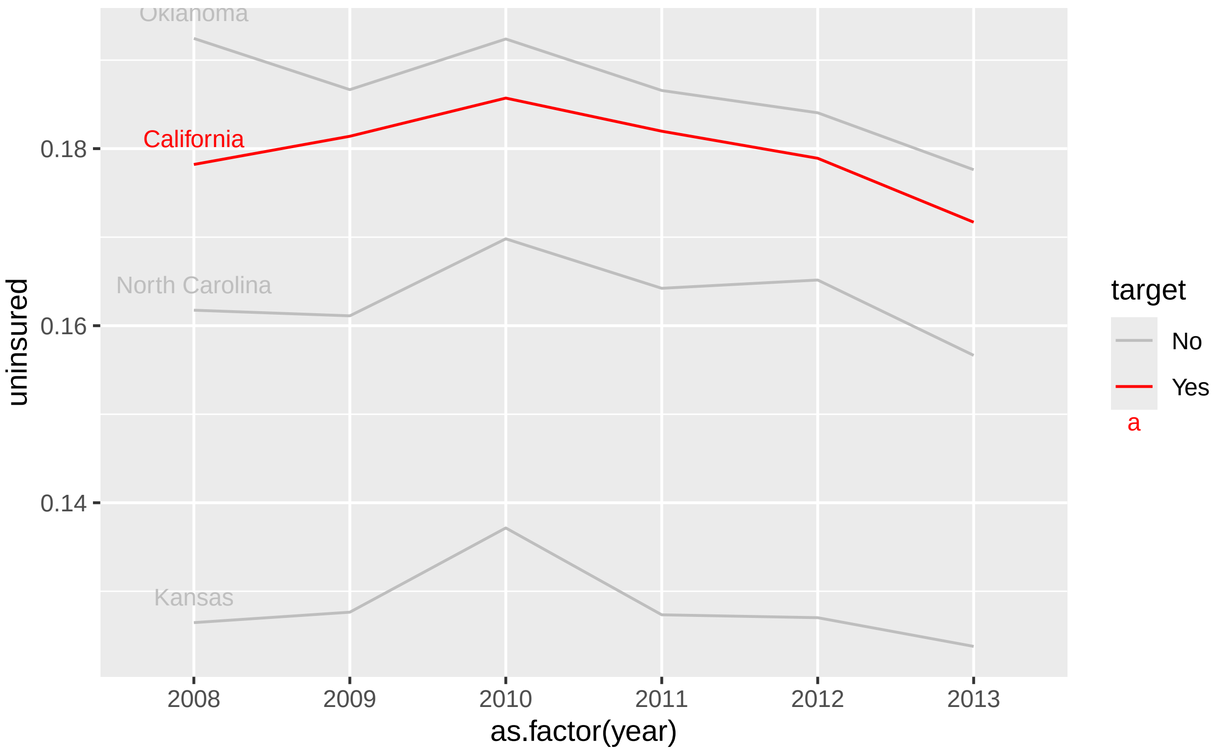
2. Assess the parallel trends assumption for your choices using a plot. If you are not satisfied that the assumption has been met, pick another state and try again (but detail the states you tried).

    - To assess the parallel trends assumption, I initially compared California's pre-adoption trend with that of the 15 states that did not adopt Medicaid expansion. Upon visual inspection, I identified three potential comparison states.
    - Subsequently, I generated another plot focusing on these three states. Ultimately, I selected Oklahoma as the final choice due to its closely aligned trend with California from 2009 to 2013 and its similar uninsured rate during this period.

```r
## Plot parallel trends
state_first_obs <- med_did %>% group_by(state) %>% filter(row_number() == 1)
med_did %>% filter(year <= 2013) %>%
  ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = target)) +
  geom_line() +
  scale_color_manual(values = c("Yes" = "red", "No" = "grey")) +
  geom_text(data = state_first_obs, aes(label = state), hjust = 1, size = 2)
```



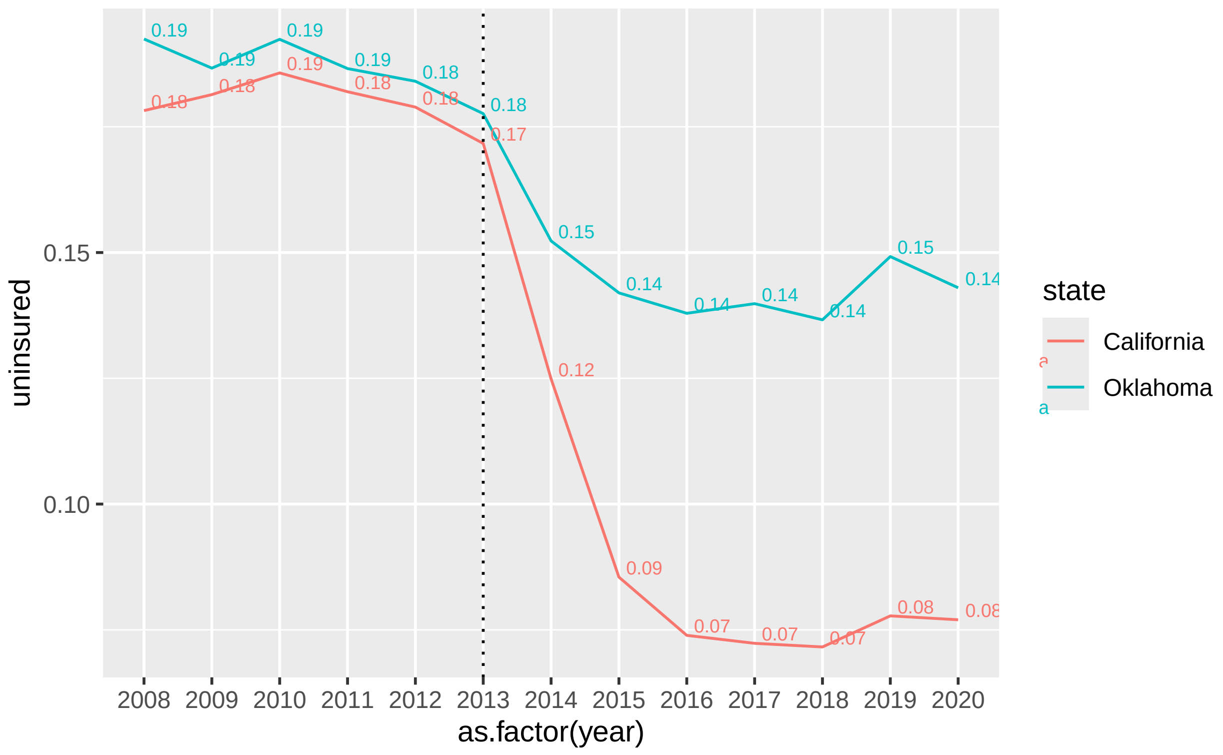
```r
## Focus on several candidate states
med_did <- med_did %>% filter(state %in% c("Kansas", "North Carolina", "Oklahoma", "California"))
state_first_obs <- med_did %>% group_by(state) %>% filter(row_number() == 1)

med_did %>% filter(year <= 2013) %>%
  ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = target)) +
  geom_line() +
  scale_color_manual(values = c("Yes" = "red", "No" = "grey")) +
  geom_text(data = state_first_obs, aes(label = state), vjust = -1, size = 3)
```



3. 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.

    - After extending the comparison plot to include the post-adoption period, an intriguing pattern emerged: while the uninsured rates in the two states closely aligned from 2009 to 2013, California experienced a significantly larger drop in the subsequent two years.
    - Utilizing the difference-in-differences estimator with 2014 designated as the post-adoption and 2013 as the pre-adoption reference points, the estimated effect of Medicaid expansion reveals **a 2.15% decrease in the uninsured rate**.

```
## Plot the whole trends
med_did <- med_did %>% filter(state %in% c("Oklahoma", "California"))
med_did %>% ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = state)) +
  geom_line() +
  geom_vline(xintercept = "2013", linetype = "dotted", color = "black") +
  geom_text(aes(label = round(uninsured, 2)), hjust = -0.2, vjust = -0.2, size = 2.5)
```



```r
## Transpose state data
cal <- med_did %>% filter(state == "California") %>% select(uninsured) %>% t() %>% data.frame()
colnames(cal) <- 2008:2020

okl <- med_did %>% filter(state == "Oklahoma") %>% select(uninsured) %>% t() %>% data.frame()
colnames(okl) <- 2008:2020

## Diff-in-Diff estimation
diff_cal <- cal$"2014" - cal$"2013"
diff_okl <- okl$"2014" - okl$"2013"

diff_in_diff <- diff_cal - diff_okl
diff_in_diff
```

```
## [1] -0.02149
```

4.2 Discussion Questions

1. 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?

    - As demonstrated by the graph of pre-adoption trends across all states, there are significant differences in the uninsured rate even among geographically adjacent states. Unlike the localized context of Card and Krueger's study, Medicaid coverage is influenced by state-level policies and political climates, factors that cannot be explained solely by geographic adjacency.

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

- The parallel trends assumption provides an intuitive visualization of the treatment effect. When parallel trends exist, the visual divergence after the intervention (or lack thereof) is compelling, especially when multiple time points are considered, a design known as comparative interrupted time series (CITS).
- However, this strength is accompanied by a significant weakness: in real-world settings, it is often challenging to identify a comparison group with an exact or sufficiently close parallel trend. This limitation has led to the development of alternative methods such as synthetic control.

5 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.

5.1 Non-Augmented Synthetic Control

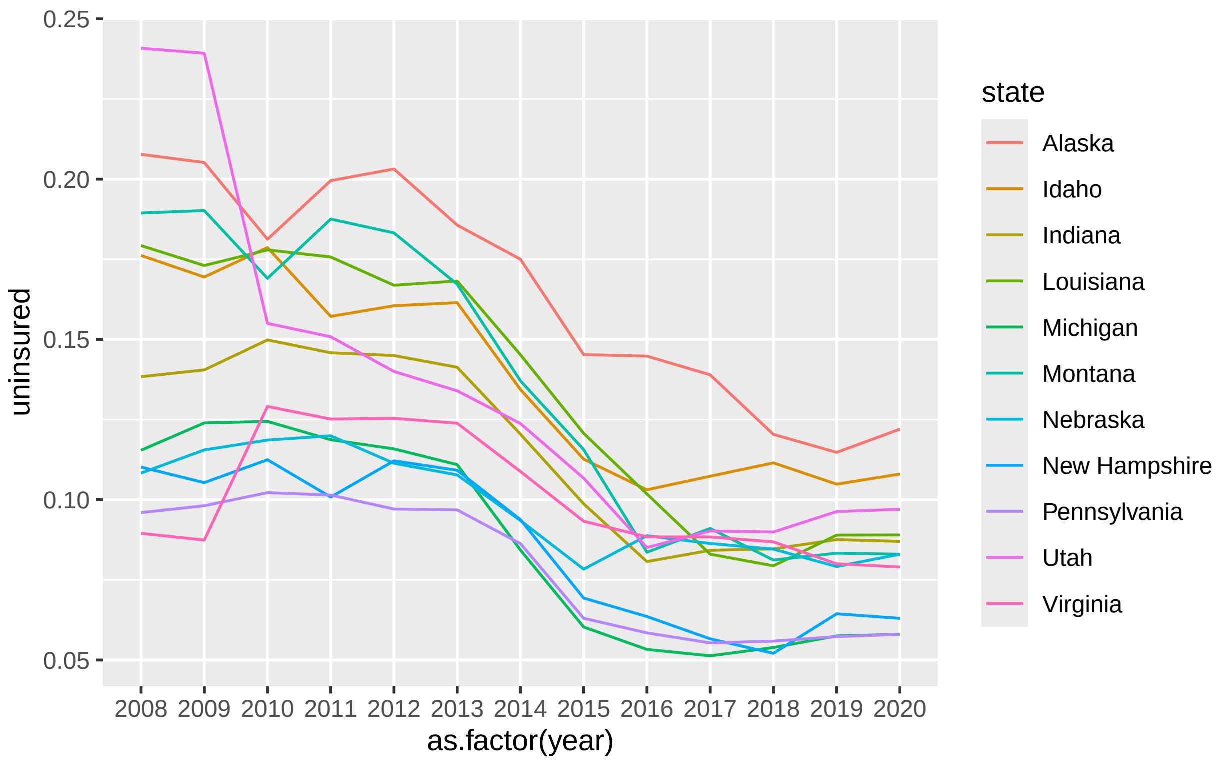
Choose a state that adopted the Medicaid expansion after January 1, 2014. Construct a non-augmented synthetic control and plot the results (both pre-treatment fit and post-treatment differences). Also report the average ATT and L2 imbalance.

1. In the first step, I plot the trends of uninsured rates for states that adopted Medicaid expansion after January 1, 2014, and states that did not adopt. The state of Montana, with its relatively higher uninsured rate and late adoption in 2016, is chosen as the target state.

```
## States with lagged adoption
state_lag_adopted <- med %>% filter(adoption != "2014-01-01") %>% distinct(state) %>% pull(state)
state_lag_adopted
```

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

```
med %>% filter(state %in% state_lag_adopted) %>%
  ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = state)) +
  geom_line()
```



```
## Create a dataframe with only the targeted state and states that did not adopt
target_state <- "Montana"
med_syn <- med %>% filter(state %in% c(target_state, state_not_adopted)) %>%
  mutate(target = if_else(state == target_state, "Yes", "No"))
```

2. Using only states that did not adopt Medicaid expansion as potential donors, I run a synthetic control analysis to construct a non-augmented synthetic control for the target state (Montana).

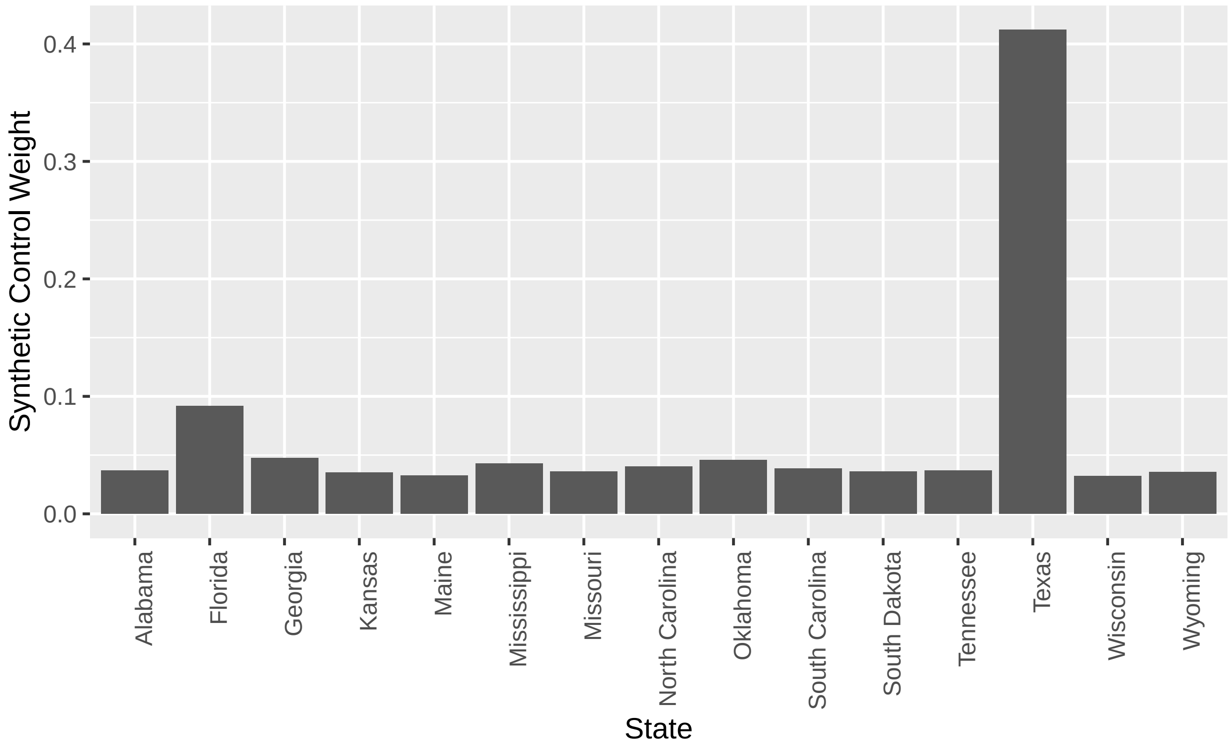
```
## Run synthetic control
med_syn <- med_syn %>% mutate(unit = as.factor(state)) %>%
  mutate(unit = as.numeric(unit))

dataprep_out <-dataprep(
  foo = as.data.frame(med_syn),
  predictors = "uninsured",
  dependent = "uninsured",
  unit.variable = "unit",
  time.variable = "year",
  treatment.identifier = 8,  # Montana as 8
  controls.identifier = unique(med_syn$unit)[-8],
  time.predictors.prior = 2008:2015,
  time.optimize.ssr = 2008:2015,
  unit.names.variable = "state",
  time.plot = 2008:2020)

synth_out <- synth(dataprep_out)
```

```
##
```

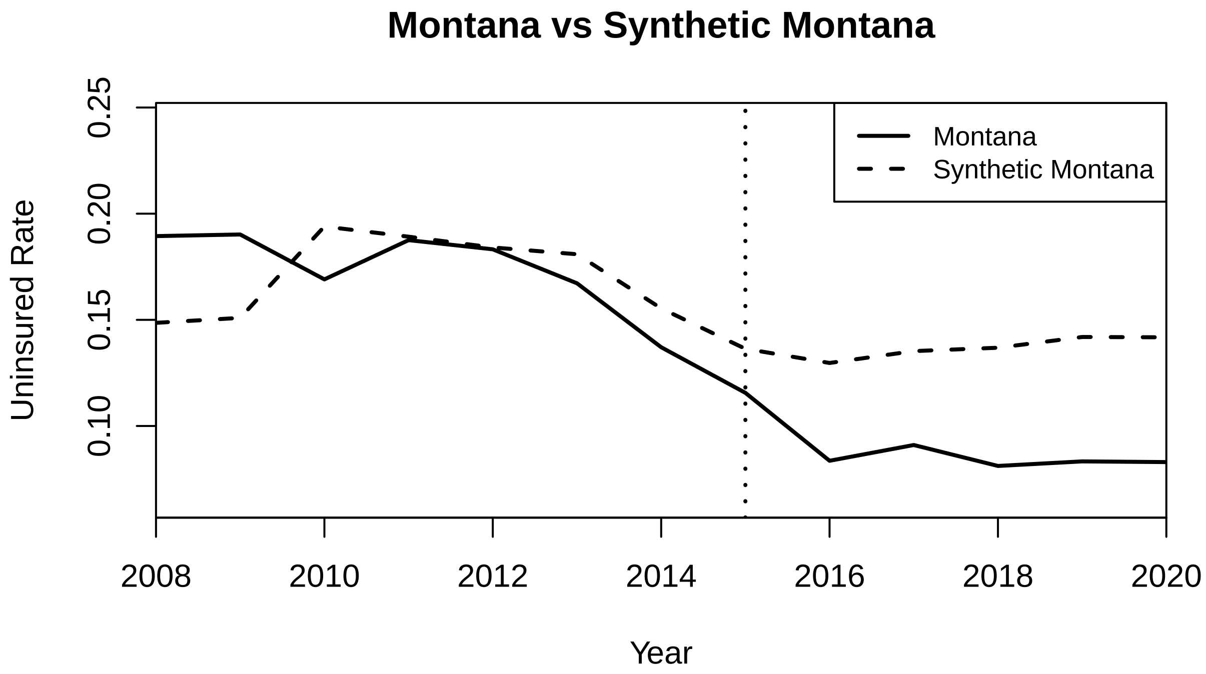
```
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  optimization over w weights: computing synthtic control unit
##
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.0005985629
##
## solution.v:
##  1
##
## solution.w:
##  0.03690677 0.09175621 0.04757687 0.0351374 0.03247667 0.04288505 0.03608387 0.04040613 0.04589153 0
```

3. After synthesizing the control, I plot the synthetic control weights assigned to each donor state, illustrating their relative contributions to the construction of the synthetic control for Montana. Additionally, I plot the parallel trends of uninsured rates between Montana and the synthetic Montana, visually assessing the pre-treatment fit and post-treatment differences between the two series.

```r
## Plot weights
unit <- med_syn %>% group_by(unit) %>% filter(row_number() == 1) %>% select(state, unit)
synweight <- data.frame(unit = c(1:16)[-8], synth_out$solution.w) %>% left_join(unit)

ggplot(synweight, aes(x = state, y = w.weight)) +
  geom_bar(stat = "identity") +
  labs(x = "State", y = "Synthetic Control Weight") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
## Plot parallel trends
path.plot(synth.res = synth_out,
          dataprep.res = dataprep_out,
          tr.intake = 2015,
          Ylab = "Uninsured Rate",
          Xlab = "Year",
          Legend = c("Montana", "Synthetic Montana"),
          Main = "Montana vs Synthetic Montana")
```



4. Finally, I present the results of the synthetic control analysis, organizing the uninsured rate data for Montana and the synthetic Montana by year and calculating the difference-in-differences estimate. The estimated effect of Medicaid expansion reveals **a 2.53% decrease in the uninsured rate**.

```r
## Organize results
synres <- data.frame(dataprep_out$Y1plot, dataprep_out$Y0plot %*% synth_out$solution.w) %>%
  rename(Montana = X8, Syn_Montana = w.weight) %>%
  rownames_to_column(var = "year")

## Transpose state data
mon <- synres %>% select(Montana) %>% t() %>% data.frame()
colnames(mon) <- 2008:2020

syn <- synres %>% select(Syn_Montana) %>% t() %>% data.frame()
colnames(syn) <- 2008:2020

## Diff-in-Diff estimation
diff_mon <- mon$"2016" - mon$"2015"
diff_syn <- syn$"2016" - syn$"2015"

diff_in_diff <- diff_mon - diff_syn
diff_in_diff
```

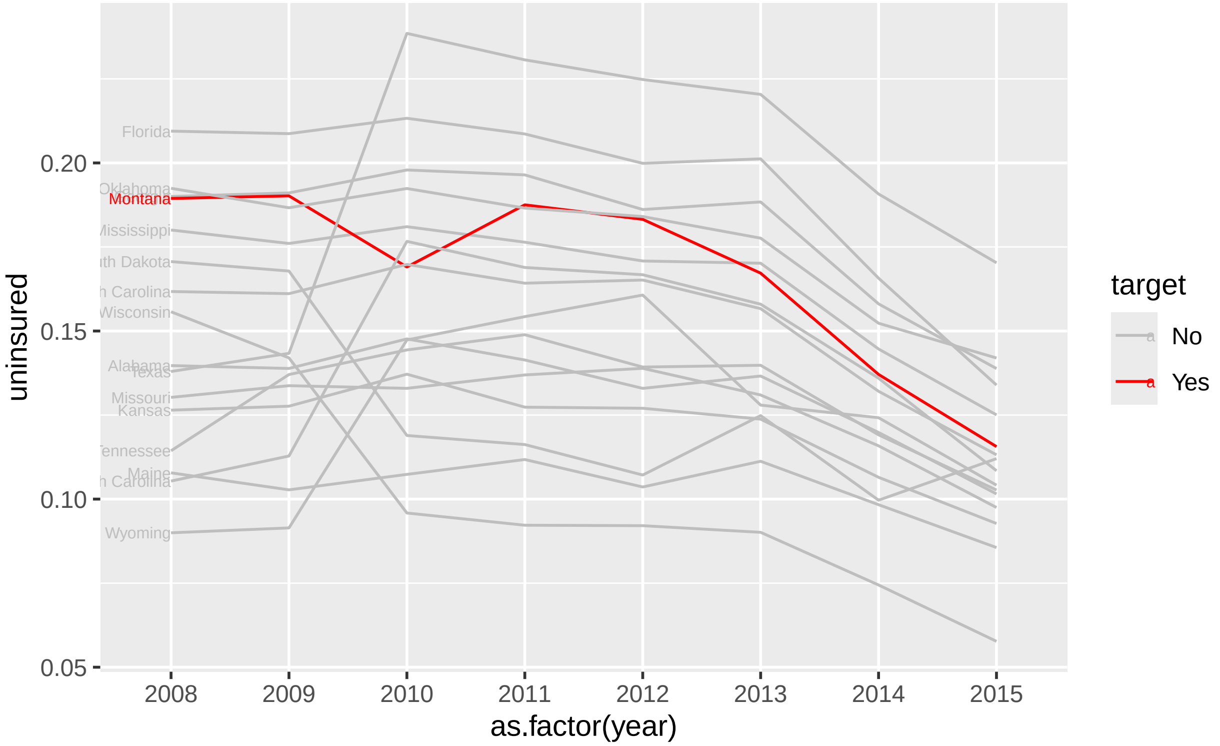
```
## [1] -0.02532886
```

5.2 CITS as Comparison

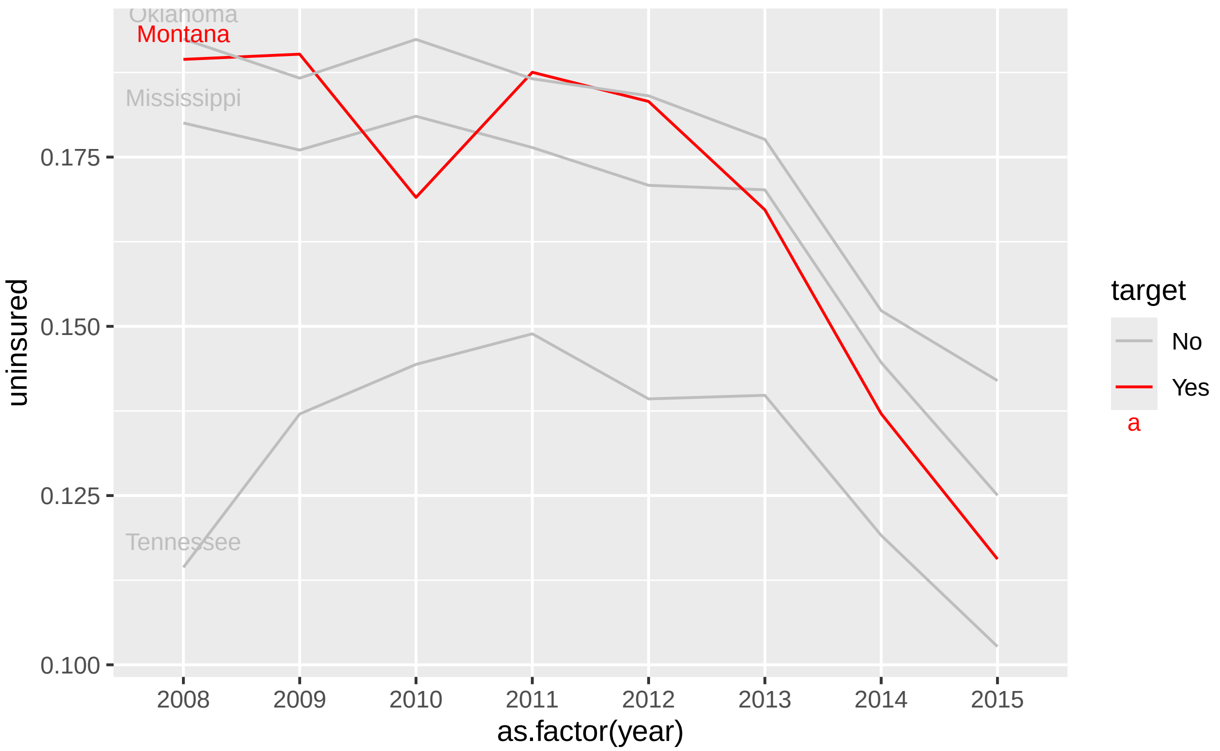
In this section, I compare the DID estimate obtained using a non-augmented synthetic control with that derived using CITS analysis (same as the one in section 4).

1. First, I plot the uninsured rate trends for states with lagged Medicaid expansion and those that did not expand it, focusing on the pre-treatment period up to 2015 (since Montana adopted it at 2016-01-01). Initially, all states are included to visualize the overall trends and assess parallelism. Subsequently, the plot focuses on several candidate states, including Mississippi, Oklahoma, Tennessee.

```
## Plot parallel trends
state_first_obs <- med_syn %>% group_by(state) %>% filter(row_number() == 1)
med_syn %>% filter(year <= 2015) %>%
  ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = target)) +
  geom_line() +
  scale_color_manual(values = c("Yes" = "red", "No" = "grey")) +
  geom_text(data = state_first_obs, aes(label = state), hjust = 1, size = 2)
```

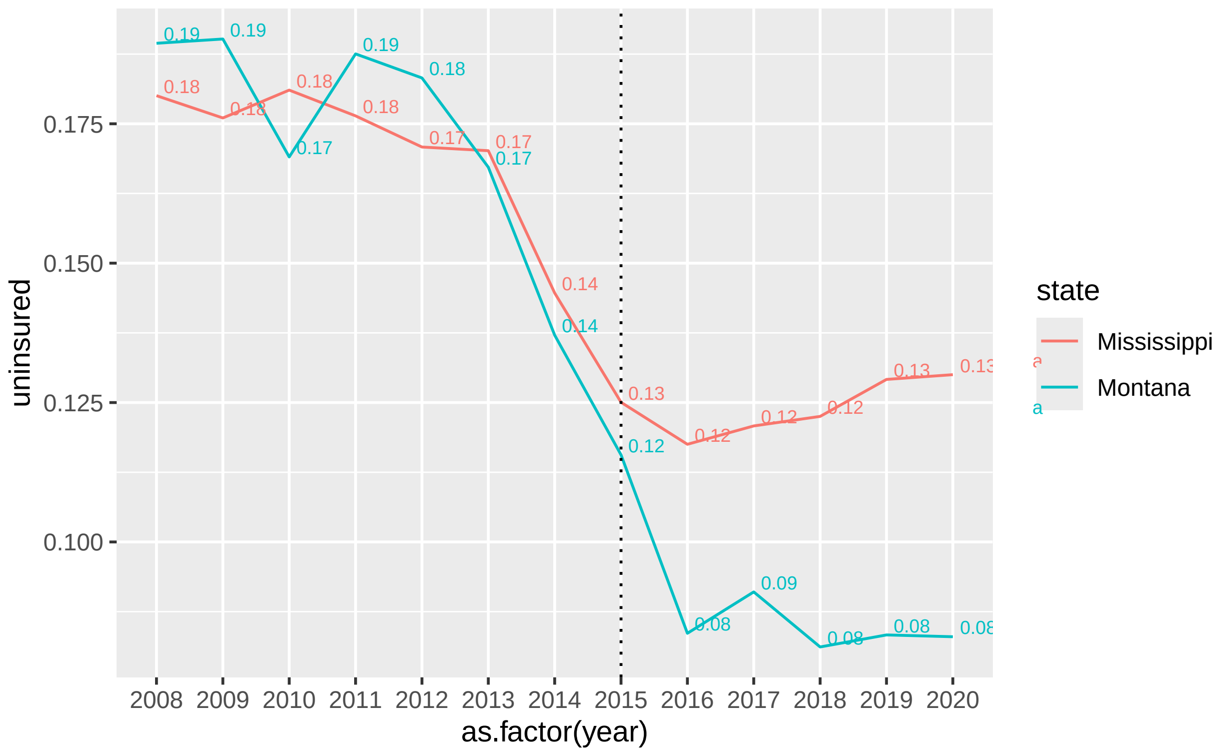


```
## Focus on several candidate states
med_syn <- med_syn %>% filter(state %in% c("Mississippi", "Oklahoma", "Tennessee", "Montana"))
state_first_obs <- med_syn %>% group_by(state) %>% filter(row_number() == 1)

med_syn %>% filter(year <= 2015) %>%
  ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = target)) +
  geom_line() +
  scale_color_manual(values = c("Yes" = "red", "No" = "grey")) +
  geom_text(data = state_first_obs, aes(label = state), vjust = -1, size = 3)
```



2. I choose Mississippi as the comparison group, for its pre-adoption trend in 2013-2015 parallels with that of Montana. I plot their trends over the entire observation period and then calculate the DID estimate. **The estimated effect is -2.45%, which is pretty close with that of synthetic control (-2.53%).**

```
## Use Mississippi as comparison group
med_syn <- med_syn %>% filter(state %in% c("Montana", "Mississippi"))

## Plot the whole trends
med_syn %>% ggplot(aes(y = uninsured, x = as.factor(year), group = state, color = state)) +
  geom_line() +
  geom_vline(xintercept = "2015", linetype = "dotted", color = "black") +
  geom_text(aes(label = round(uninsured, 2)), hjust = -0.2, vjust = -0.2, size = 2.5)
```



```r
## Transpose state data
mon <- med_syn %>% filter(state == "Montana") %>% select(uninsured) %>% t() %>% data.frame()
colnames(mon) <- 2008:2020

mis <- med_syn %>% filter(state == "Mississippi") %>% select(uninsured) %>% t() %>% data.frame()
colnames(mis) <- 2008:2020

## Estimation
diff_mon <- mon$"2016" - mon$"2015"
diff_mis <- mis$"2016" - mis$"2015"

diff_in_diff <- diff_mon - diff_mis
diff_in_diff
```

```
## [1] -0.0244716
```

5.3 Augmented Synthetic Control

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

1. First, I replicated the previous approach using the *augsynth* package without augmentation. The point estimate yielded **an effect -2.4%**, akin to the previous estimate using *Synth* package. However, notable differences emerged: (1) A 95% confidence interval and a p-value of 0.109 were computed. (2) A L2 imbalance of 0.021 was calculated, quantifying the level of discrepancy between the treatment and control units. (3) Many donor weights were shrunken close to zero, and the composition of the synthetic control group differed considerably.

```r
## Create dataframe and treatment indicator
med_syn <- med %>% filter(state %in% c(target_state, state_not_adopted)) %>%
  mutate(treat = if_else(state == "Montana" & year >= 2016, 1, 0)) %>%
  mutate(unit = as.numeric(as.factor(state)))

## Non-augmented synthetic control using augsynth
syn <- augsynth(uninsured ~ treat, unit, year, med_syn,
              progfunc = "None", scm = T)
```

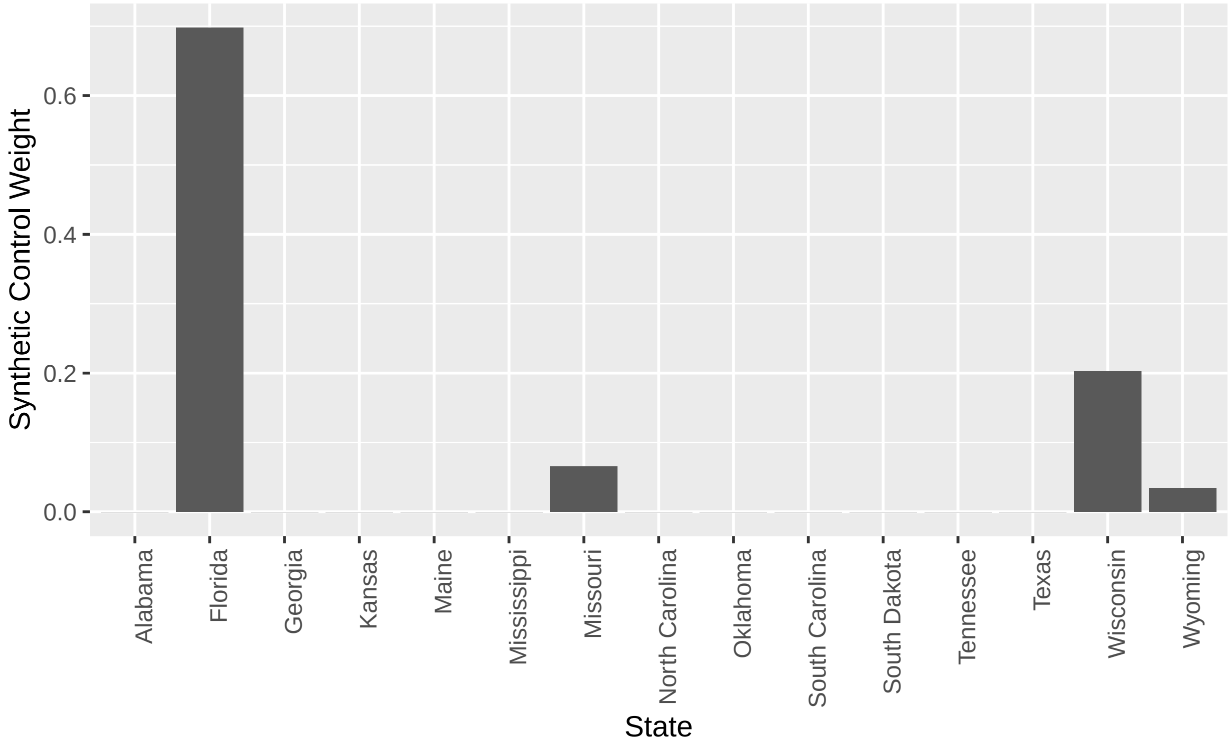
```
## One outcome and one treatment time found. Running single_augsynth.
```

```r
summary(syn)
```

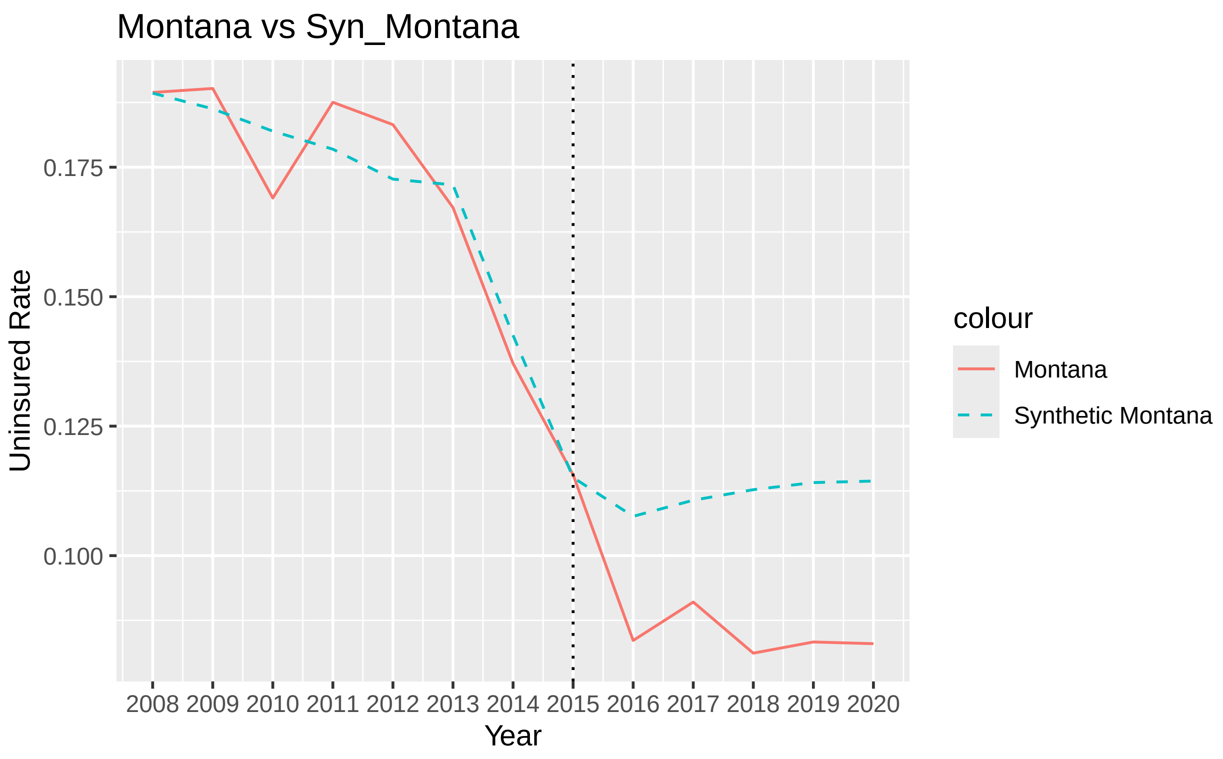
```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "None", scm = ..2)
##
## Average ATT Estimate (p Value for Joint Null):  -0.0275   ( 0.028 )
## L2 Imbalance: 0.021
## Percent improvement from uniform weights: 72.9%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.024             -0.080              0.032   0.115
##  2017   -0.020             -0.075              0.036   0.114
##  2018   -0.032             -0.087              0.024   0.111
##  2019   -0.031             -0.087              0.025   0.109
##  2020   -0.031             -0.087              0.024   0.098
```

```r
## Plot weights
data.frame(unit = c(1:16)[-8], syn$weights) %>% left_join(unit) %>%
  ggplot(aes(x = state, y = syn.weights)) +
  geom_bar(stat = "identity") +
  labs(x = "State", y = "Synthetic Control Weight") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Joining with ‘by = join_by(unit)‘
```



```
## Organize results
synres <- data.frame(syn$data$synth_data$Y1plot, syn$data$synth_data$Y0plot %*% syn$weights) %>%
  rownames_to_column(var = "year")
names(synres)[2:3] <- c("Montana", "Syn_Montana")

## Plot parallel trends
synres %>% ggplot(aes(x = as.numeric(year))) +
  geom_line(aes(y = Montana, color = "Montana"), linetype = "solid") +
  geom_line(aes(y = Syn_Montana, color = "Synthetic Montana"), linetype = "dashed") +
  geom_vline(xintercept = 2015, linetype = "dotted", color = "black") +
  labs(x = "Year", y = "Uninsured Rate", title = "Montana vs Syn_Montana") +
  scale_x_continuous(breaks = seq(2008, 2020, by = 1))
```



2. Then, I extended the analysis to include augmentation using the *augsynth* package. The point estimate for the augmented synthetic control approach was **-2.7%**, aligning closely with the previous estimate. However, augmentation introduced several key differences: (1) Weights were permitted to be negative, altering the composition of the synthetic control group. (2) The L2 imbalance was notably reduced to 0.013, indicating improved balance between the treatment and control groups. (3) Notably, the graph of parallel trends illustrated a synthetic Montana that closely followed the pre-adoption trend, suggesting better alignment with the untreated units.

```
## Augmented synthetic control
syn <- augsynth(uninsured ~ treat, unit, year, med_syn,
                progfunc = "Ridge", scm = T)
```

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

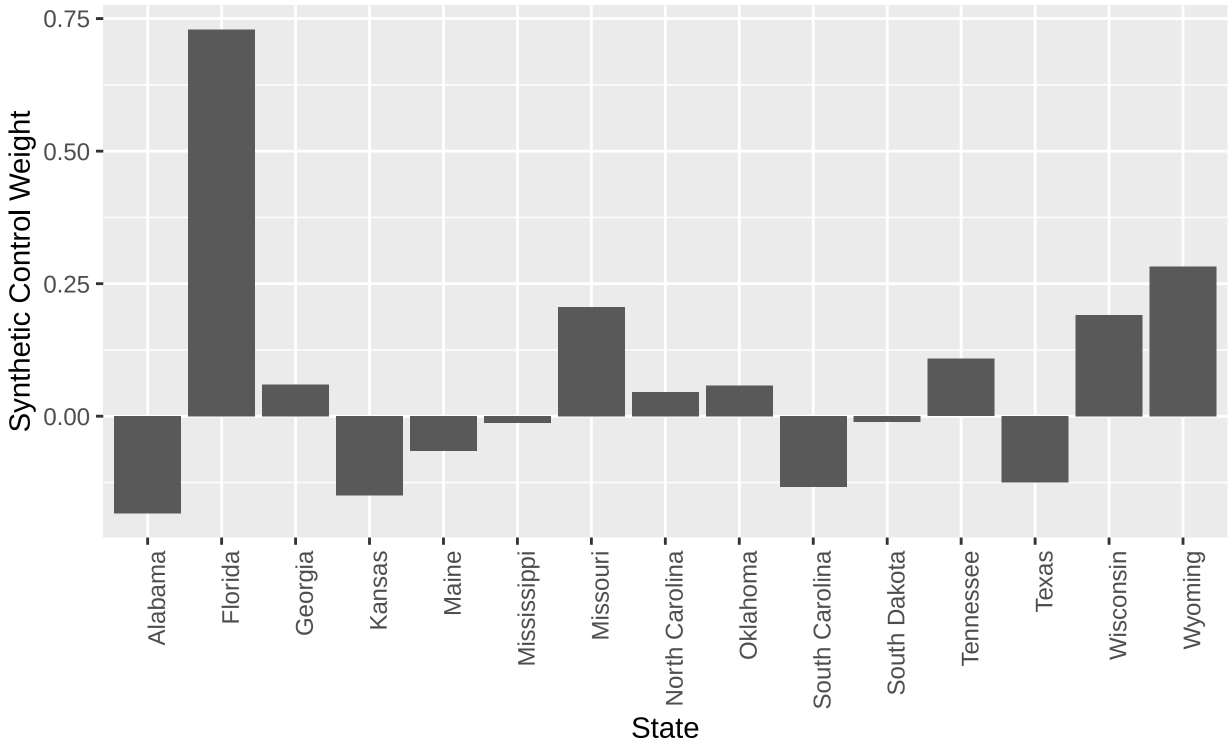
```
summary(syn)
```

```
##
## 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):  -0.0301   ( 0.57 )
## L2 Imbalance: 0.013
## Percent improvement from uniform weights: 82.4%
##
## Avg Estimated Bias: 0.003
##
```

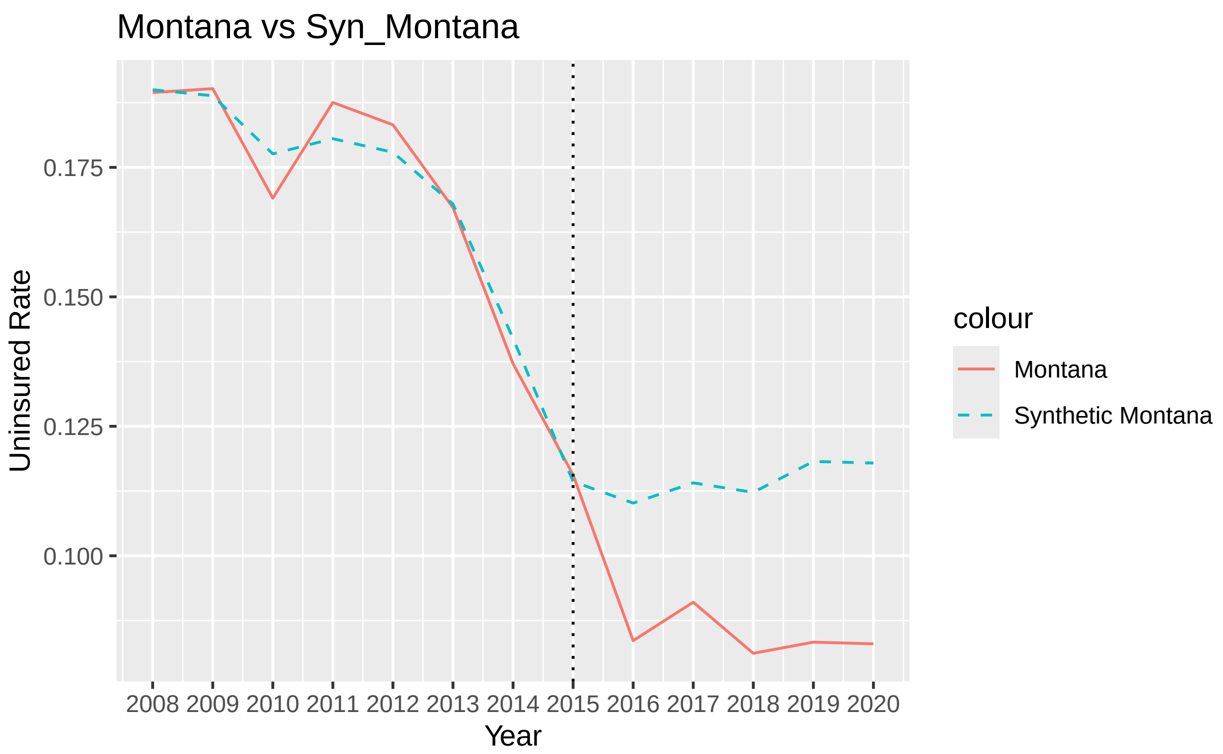
```
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.027             -0.087              0.034   0.119
##  2017   -0.023             -0.084              0.038   0.110
##  2018   -0.031             -0.092              0.030   0.120
##  2019   -0.035             -0.096              0.026   0.106
##  2020   -0.035             -0.096              0.026   0.104
```

```r
## Plot weights
data.frame(unit = c(1:16)[-8], syn$weights) %>% left_join(unit) %>%
  ggplot(aes(x = state, y = syn.weights)) +
  geom_bar(stat = "identity") +
  labs(x = "State", y = "Synthetic Control Weight") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Joining with ‘by = join_by(unit)‘
```



```r
## Organize results
synres <- data.frame(syn$data$synth_data$Y1plot, syn$data$synth_data$Y0plot %*% syn$weights) %>%
  rownames_to_column(var = "year")
names(synres)[2:3] <- c("Montana", "Syn_Montana")

## Plot parallel trends
synres %>% ggplot(aes(x = as.numeric(year))) +
  geom_line(aes(y = Montana, color = "Montana"), linetype = "solid") +
  geom_line(aes(y = Syn_Montana, color = "Synthetic Montana"), linetype = "dashed") +
  geom_vline(xintercept = 2015, linetype = "dotted", color = "black") +
  labs(x = "Year", y = "Uninsured Rate", title = "Montana vs Syn_Montana") +
  scale_x_continuous(breaks = seq(2008, 2020, by = 1))
```



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

5.4 Discussion Questions

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

    - **Advantage**: Synthetic control can create a synthetic state as a comparison group that closely mirrors the pre-adoption trend, addressing the challenge of finding a suitable comparative state in real-world settings.
    - **Disadvantage**: The synthetic state, derived from a combination of donor states with varying weights, poses interpretation challenges. Determining the significance of individual donor states in the synthetic control can be difficult.
    - **Further Consideration**: The variability in donor weights across different estimation methods complicates interpretation. For instance, the *Synth* package assigns a weight of over 40% to Texas, while the *augsynth* package allocates approximately 70% weight to Florida and minimal weight to Texas. Deciding between these contrasting donor weight distributions presents a challenge akin to selecting a single comparison state based on the graph of parallel trends.

2. 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?

    - Because of the inherent problem of interpreting donor weights, one can rather focus on the effectiveness of the synthetic control in tracking the actual treatment unit's pre-adoption trend. In this regard, the allowance of negative weights can be beneficial, as it often enhances the tracking effectiveness by enabling a more flexible adjustment of the synthetic control's composition.

- Besides, we can put more emphasis on the robustness of the results across different donor pools and estimation algorithms. Consistent estimates across varied donor compositions lend credibility to the estimated effects, irrespective of the specific interpretation of individual weights.

6 Staggered Adoption Synthetic Control

6.1 Estimate Multisynth

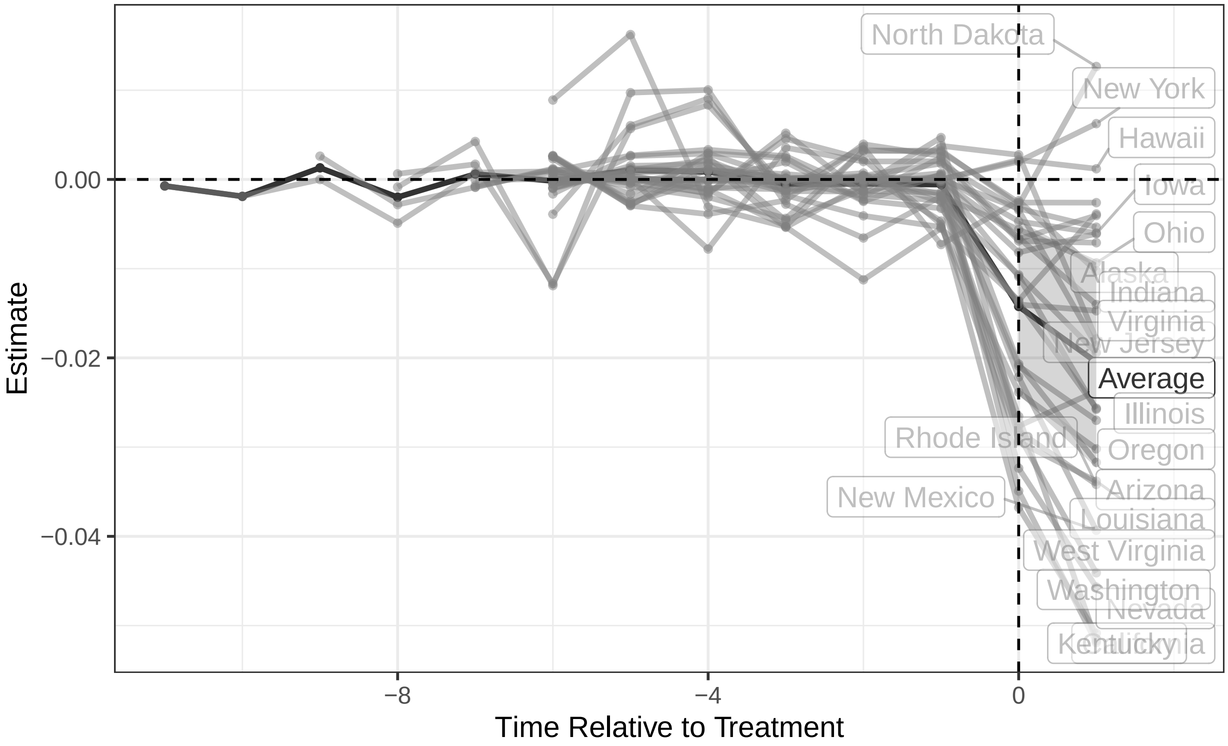
1. Estimate a multisynth model that treats each state individually. Choose a fraction of states that you can fit on a plot and examine their treatment effects.

```r
## Set seed
set.seed(224)  # to replicate SEs estimated by multisynth

## States to filter
state_to_filter <- c("Nebraska", "Idaho", "Utah",  # late adoption (2020)
                     "District of Columbia", "Massachusetts")

## Create dataframe
med_stag <- med %>% mutate(adop_yr = year(adoption)) %>%
  filter(!state %in% state_to_filter) %>%
  mutate(treat = case_when(year >= adop_yr ~ 1, .default = 0))

## Adoption later in the year (set treat=1 in the next year)
med_stag <- med_stag %>% mutate(treat = case_when(
  state == "New Hampshire" & year == 2014 ~ 0,
  state == "Alaska" & year == 2015 ~ 0,
  state == "Louisiana" & year == 2016 ~ 0,
  .default = treat))

# Multisynth
msyn <- multisynth(uninsured ~ treat, state, year,
                   nu = 0, med_stag)

## Results
msyn_sum <- summary(msyn)
msyn_sum
```

```
##
## Call:
## multisynth(form = uninsured ~ treat, unit = state, time = year,
##     data = med_stag, nu = 0)
##
## Average ATT Estimate (Std. Error): -0.017  (0.005)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.037
## Percent improvement from uniform global weights: 96.3
##
## Individual L2 Imbalance: 0.003
## Scaled Individual L2 Imbalance: 0.075
## Percent improvement from uniform individual weights: 92.5
##
##   Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                      0 Average -0.01423311 0.004894505 -0.02409366 -0.004588311
##                      1 Average -0.02047729 0.005884370 -0.03230657 -0.009333664
```

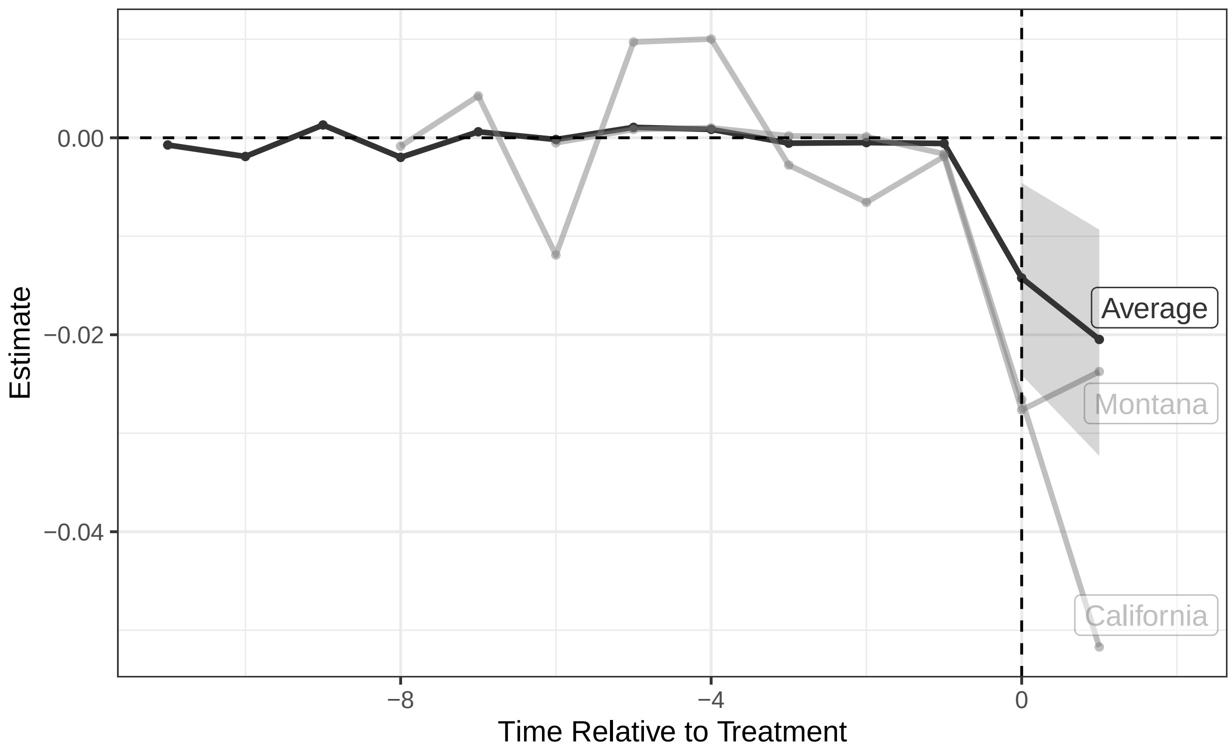
```
## Plot
plot(msyn_sum)
```

```
## Joining with ‘by = join_by(Level)‘
```



```
plot(msyn_sum, levels = c("Average", "California", "Montana"))
```

```
## Joining with ‘by = join_by(Level)‘
```



2. 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 with time cohorts
msyn <- multisynth(uninsured ~ treat, state, year,
                   med_stag, time_cohort = TRUE)
print(msyn$nu)
```

```
## [1] 0.5907676
```

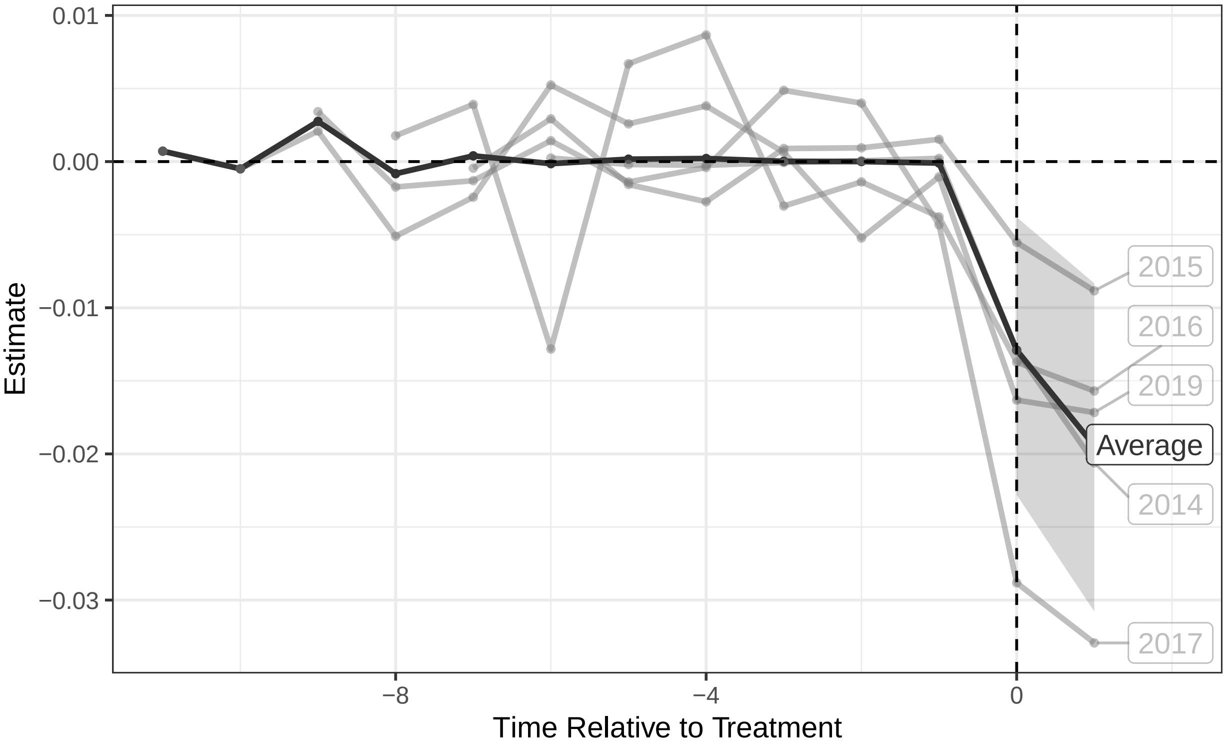
```
## Results
msyn_sum <- summary(msyn)
msyn_sum
```

```
##
## Call:
## multisynth(form = uninsured ~ treat, unit = state, time = year,
##     data = med_stag, time_cohort = TRUE)
##
## Average ATT Estimate (Std. Error): -0.016  (0.005)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.009
## Percent improvement from uniform global weights: 99.1
##
```

```
## Individual L2 Imbalance: 0.007
## Scaled Individual L2 Imbalance: 0.028
## Percent improvement from uniform individual weights: 97.2
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01290202 0.004843688 -0.02279835 -0.003806612
##                     1 Average -0.01944078 0.005979704 -0.03078662 -0.008326726
```

```
## Plot
plot(msyn_sum)
```

```
## Joining with ‘by = join_by(Level)‘
```



6.2 Discussion Questions

1. 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?

   - The first plot illustrates a significant variance in the effect sizes across different states. Specifically, at the year of adoption, effect sizes range from slightly above 0 to almost -4%.
   - This variance becomes even more pronounced in the second year after adoption, where some states experience a further drop in uninsured rates (such as the California case analyzed), while others witness a “bounce back” in uninsured rates (as observed in the Montana case). This

variability poses a challenge to DID estimation, as **the choice of reference time points in the post-adoption period can significantly impact the estimates**.

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

    - Analysis of the third graph reveals a considerable variance in effect sizes across different time cohorts, albeit smaller than that observed among individual states. Since most states adopted Medicaid expansion on 2014-01-01, the average estimates largely align with the 2014 cohort.
    - Notably, compared to the 2014 cohort, the 2015 cohort (Indiana, Pennsylvania, and New Hampshire) appears to benefit less from Medicaid expansion, potentially supporting the notion of early-adopter benefits. However, caution must be exercised in interpreting this, as **subsequent cohorts have fewer cases** (i.e., three cases in the 2015 cohort, only two cases in the 2016 cohort and one case each in the 2017 and 2019 cohorts), necessitating careful interpretation of the results.

7 General Discussion Questions

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

    - Policy interventions often occur at aggregated levels, such as the state level, leading some states to become "treated" while others serve as "control" units. When states in these two groups exhibit parallel trends in the pre-treatment period, making a comparison becomes straightforward and intuitive.

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

    - In DID and synthetic control, the parallel trends assumption assumes that states or units that underwent treatment and those that did not are comparable, implying a somewhat "random" assignment. However, depending on the research scenarios, this assumption may be violated due to various factors such as political climate, resulting in potential selection bias. In this analysis, for instance, states that opted for Medicaid expansion may have done so due to underlying preferences or political motivations, leading to differences that affect the treatment effect.
    - On the other hand, regression discontinuity (RD) leverages arbitrary assignment into treatment based on a cutoff at the individual level, which theoretically mitigates selection issues by approximating a randomized experiment. However, RD primarily estimates the local treatment effect at the cutoff, making its generalizability to other contexts less straightforward.