Project 7: Difference-in-Differences and Synthetic Control

Maggie Ye

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

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

```r
if (!require("devtools")) install.packages("devtools")
```

```
## Loading required package: devtools

## Loading required package: usethis
```

```r
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
```

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

# set seed
set.seed(44)

# load data
medicaid_expansion <- read.csv("medicaid_expansion.csv")

medicaid_expansion <- medicaid_expansion %>%
  mutate(uninsured_population = uninsured_rate / 100 * population)

# Note: I created the variable Adopted_Year directly in the csv file
```

Exploratory Data Analysis

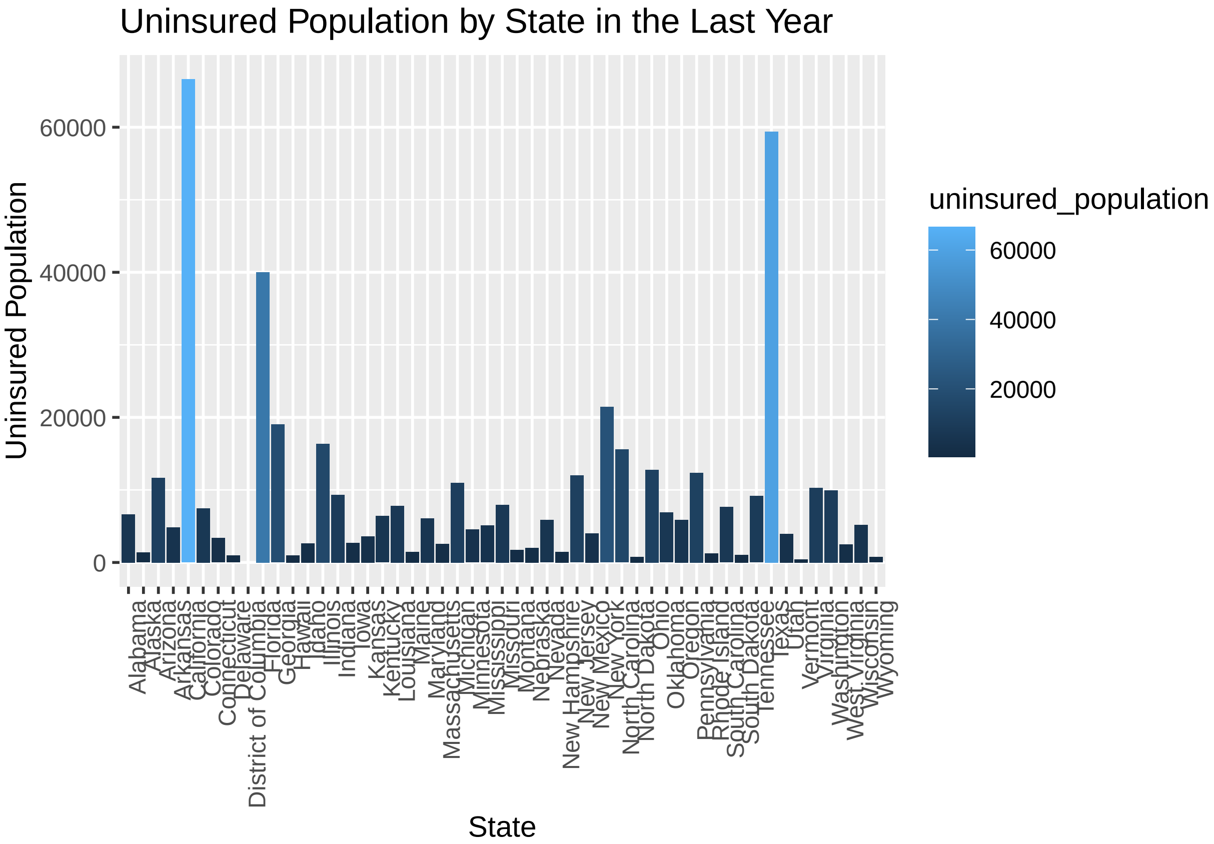
Create plots and provide 1-2 sentence analyses to answer the following questions:

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

```r
# highest and lowest uninsured rates

library(ggplot2)
library(dplyr)
```

```r
# Use 2013 data
data_2013 <- medicaid_expansion %>%
  filter(year == 2013)

# Plot of uninsured populations for each state in the last year
ggplot(data_2013, aes(x = State, y = uninsured_population, fill = uninsured_population)) +
  geom_col() +
  labs(title = "Uninsured Population by State in the Last Year", x = "State", y = "Uninsured Population
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

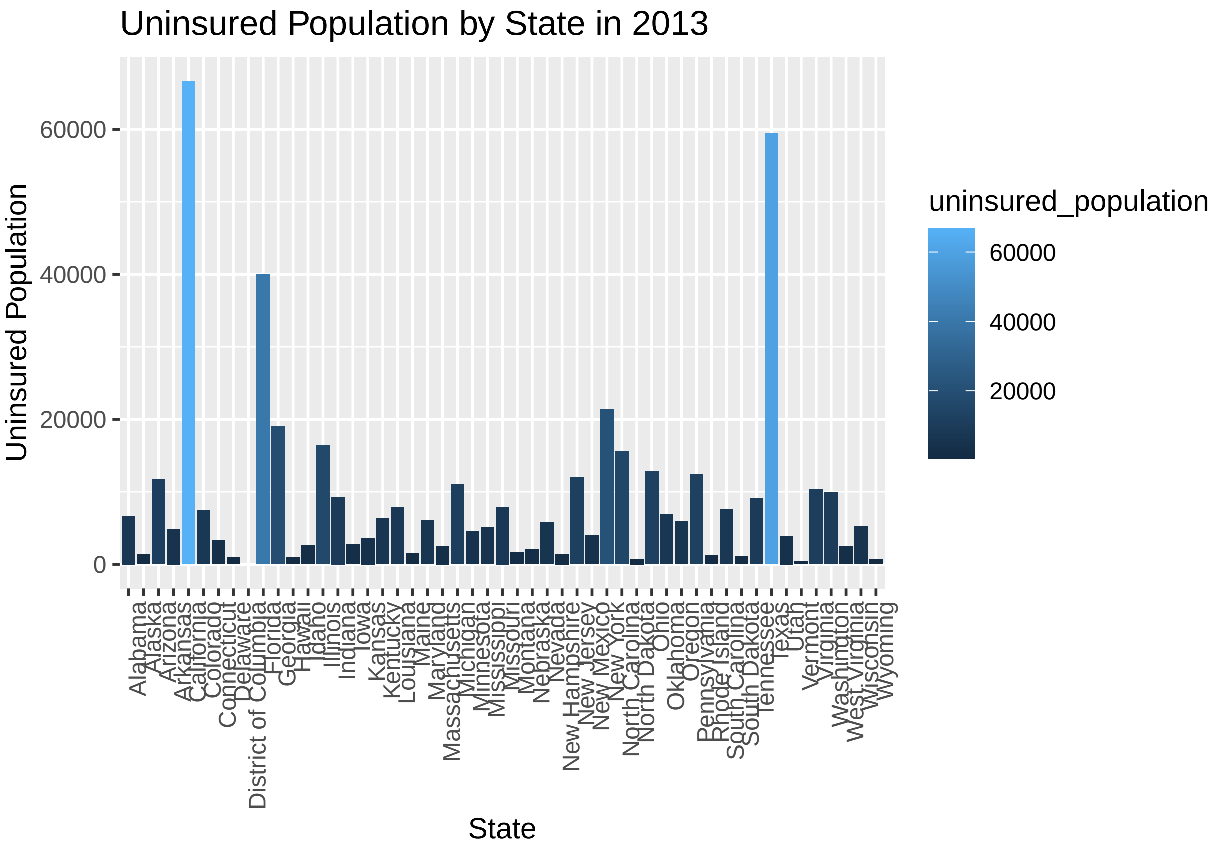
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_col()`).
```



```r
# Lowest: Vermont
# Highest: California
```

```r
# most uninsured Americans

# Plot of uninsured populations for each state in 2013
ggplot(data_2013, aes(x = State, y = uninsured_population, fill = uninsured_population)) +
  geom_col() +
  labs(title = "Uninsured Population by State in 2013", x = "State", y = "Uninsured Population") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

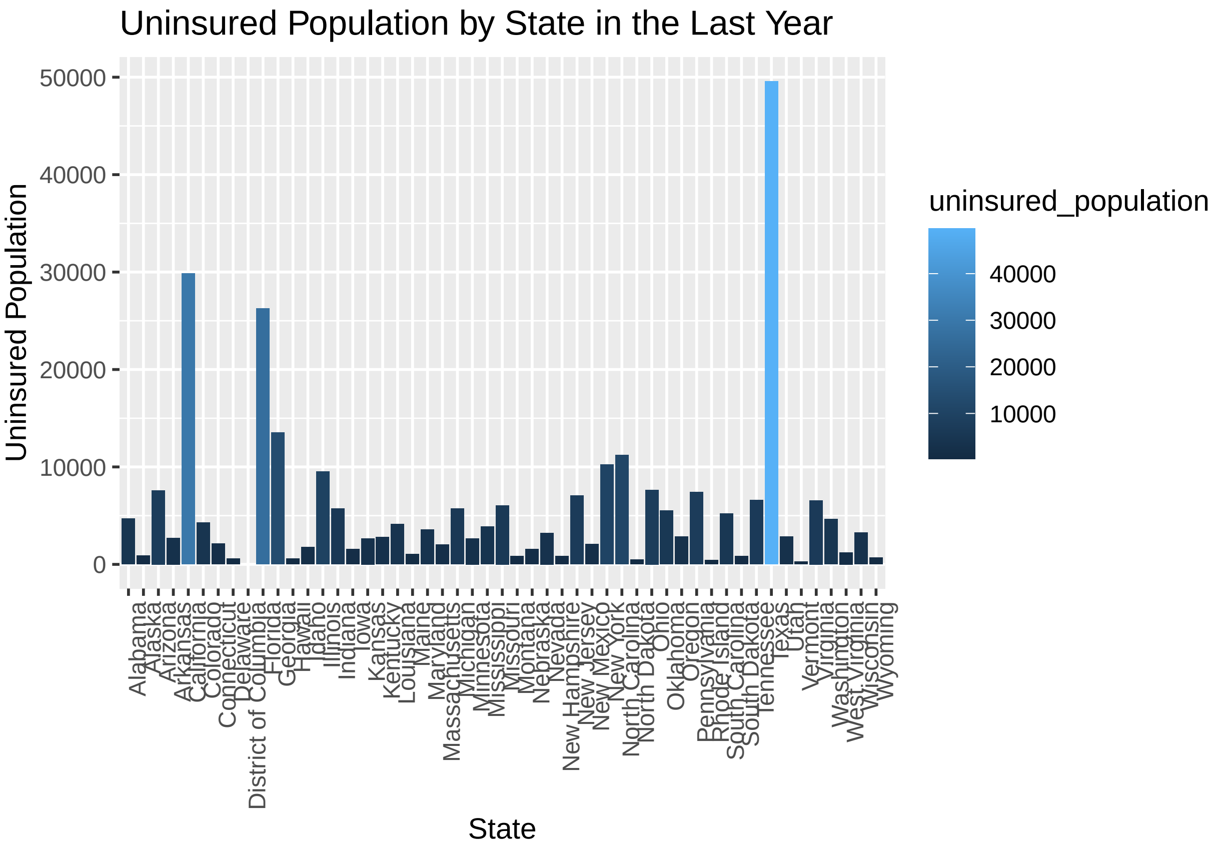
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_col()`).
```



```
# Lowest: Vermont
# Highest: Arkansas
```

```
# Last year
data_last_year = medicaid_expansion %>%
  filter(year == max(year))

# Plot of uninsured populations for each state in the last year
ggplot(data_last_year, aes(x = State, y = uninsured_population, fill = uninsured_population)) +
  geom_col() +
  labs(title = "Uninsured Population by State in the Last Year", x = "State", y = "Uninsured Population
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_col()`).
```



```
# Lowest: Vermont
# Highest: Texas
```

Difference-in-Differences Estimation

Estimate Model

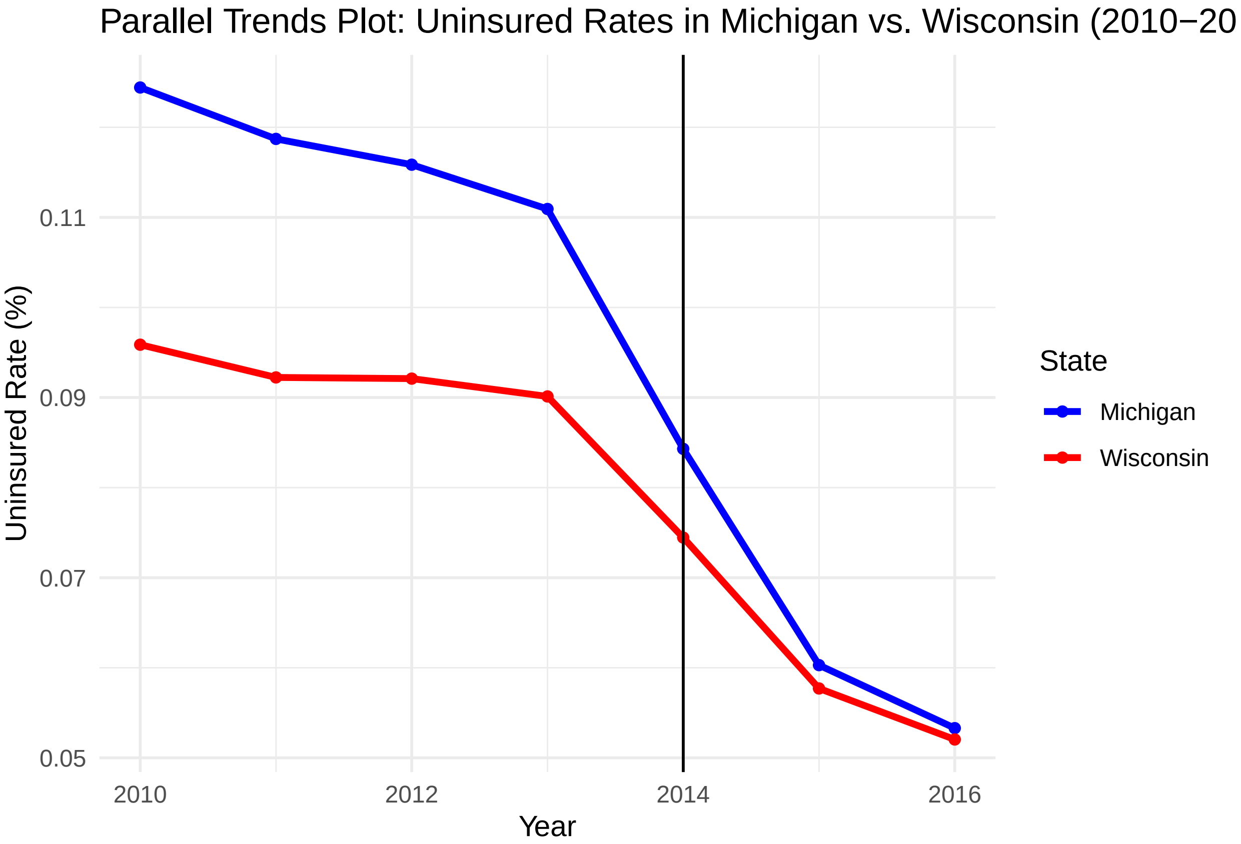
Do the following:

- 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

analysis_data <- medicaid_expansion %>%
  filter(State %in% c("Michigan", "Wisconsin"), year >= 2010 & year <= 2016) %>%
  mutate(treatment = ifelse(State == "Michigan" & year >= 2014, 1, 0),
         post = ifelse(year >= 2014, 1, 0))

ggplot(analysis_data, aes(x = year, y = uninsured_rate, color = State, group = State)) +
  geom_line(linewidth = 1.2) +
  geom_point(linewidth  = 2) +
  labs(title = "Parallel Trends Plot: Uninsured Rates in Michigan vs. Wisconsin (2010-2016)",
       x = "Year",
       y = "Uninsured Rate (%)") +
```

```r
  theme_minimal() +
  scale_color_manual(values = c("Michigan" = "blue", "Wisconsin" = "red")) +
  geom_vline(aes(xintercept = 2014))
```

```
## Warning in geom_point(linewidth = 2): Ignoring unknown parameters: `linewidth`
```



- 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

```r
# Difference-in-Differences estimation

mw <- medicaid_expansion %>%
  filter(State %in% c("Michigan", "Wisconsin"))

pre_diff <- mw %>%
  filter(year < 2014) %>%
  group_by(State) %>%
  summarise(avg_uninsured_rate = mean(uninsured_rate, na.rm = TRUE)) %>%
  pivot_wider(names_from = State, values_from = avg_uninsured_rate) %>%
  summarise(pre_diff = Michigan - Wisconsin)

post_diff <- mw %>%
  filter(year >= 2014) %>%
  group_by(State) %>%
  summarise(avg_uninsured_rate = mean(uninsured_rate, na.rm = TRUE)) %>%
  pivot_wider(names_from = State, values_from = avg_uninsured_rate) %>%
  summarise(post_diff = Michigan - Wisconsin)

diff_in_diffs <- post_diff$post_diff - pre_diff$pre_diff
diff_in_diffs
```

```
## [1] -0.005002733
```

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**: Replicating their approach with this data is challenging because Medicaid expansion decisions were made at the state level, rather than at the local level. While towns on either side of the Delaware river share similar characteristics due to geographical proximity, states have diverse populations, economies, and healthcare systems. Therefore, finding suitable control states that closely resemble the treatment states in all relevant aspects is more difficult. Additionally, Medicaid expansion implementation may have varied effects depending on factors such as state policies, demographics, and healthcare infrastructure, further complicating the identification of comparable control states.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: The main strength of relying on this assumption is its simplicity, as it only requires observing trends in the outcome variable before and after treatment for both treated and control groups. However, parallel trends prior to the treatment time does not guarantee parallel trends after it even if treatment is not adopted for both groups, so we need to carefully select the controls to make sure there are no other variables affecting the trend at the same time when treatment is in place.

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.

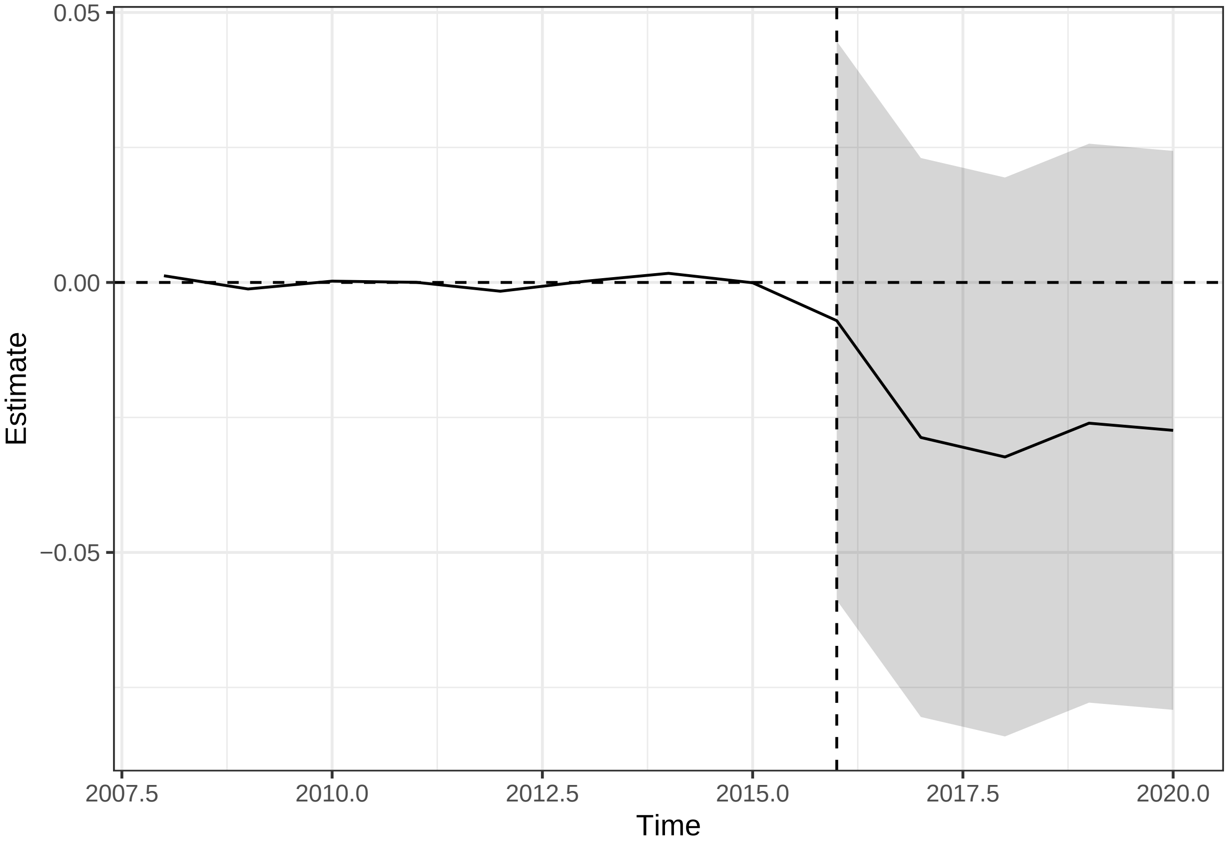
Do the following:

- 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

data <- medicaid_expansion %>%
  select(year, State, uninsured_rate) %>%
  mutate(treatment = case_when(State == "Louisiana" & year >= 2016 ~ 1,
                               TRUE ~ 0))

syn <- augsynth(uninsured_rate ~ treatment,
                unit = State,
                time = year,
                data = data,
                progfunc = "None",
                scm = TRUE)
```

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

```
plot(syn)
```



```
summary(syn)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "None", scm = TRUE)
##
## Average ATT Estimate (p Value for Joint Null):  -0.0243   ( 0.17 )
## L2 Imbalance: 0.003
## Percent improvement from uniform weights: 97.2%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.007             -0.059              0.045   0.225
##  2017   -0.029             -0.080              0.023   0.123
##  2018   -0.032             -0.084              0.019   0.121
##   2019   -0.026              -0.078              0.026   0.121
##   2020   -0.027              -0.079              0.024   0.106
```

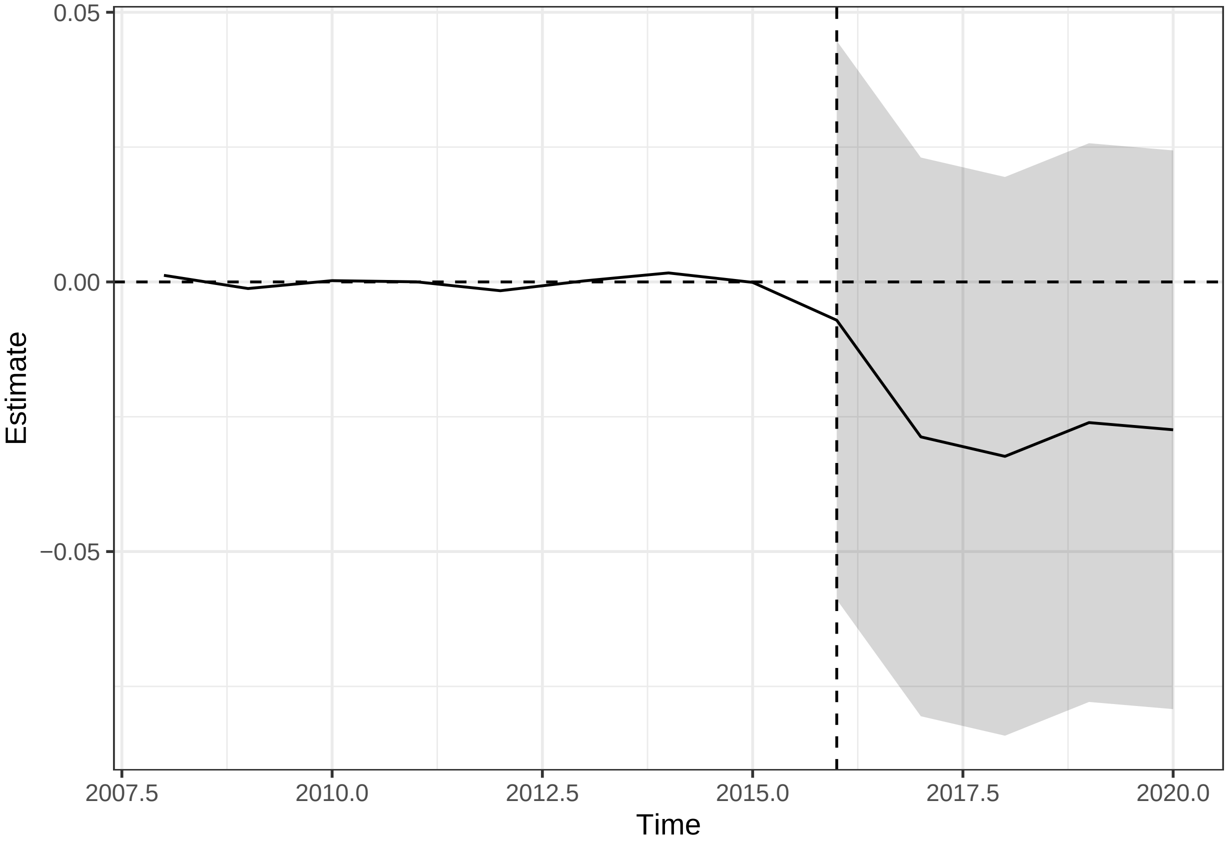
```
# Average ATT Estimate (p Value for Joint Null):  -0.0243   ( 0.17 )
# L2 Imbalance: 0.003
```

- 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

syn_augmented <- augsynth(uninsured_rate ~ treatment,
                          unit = State,
                          time = year,
                          data = data,
                          progfunc = "Ridge",  # Ridge augmentation
                          scm = TRUE)
```

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

```
plot(syn_augmented)
```



```
summary(syn_augmented)
```

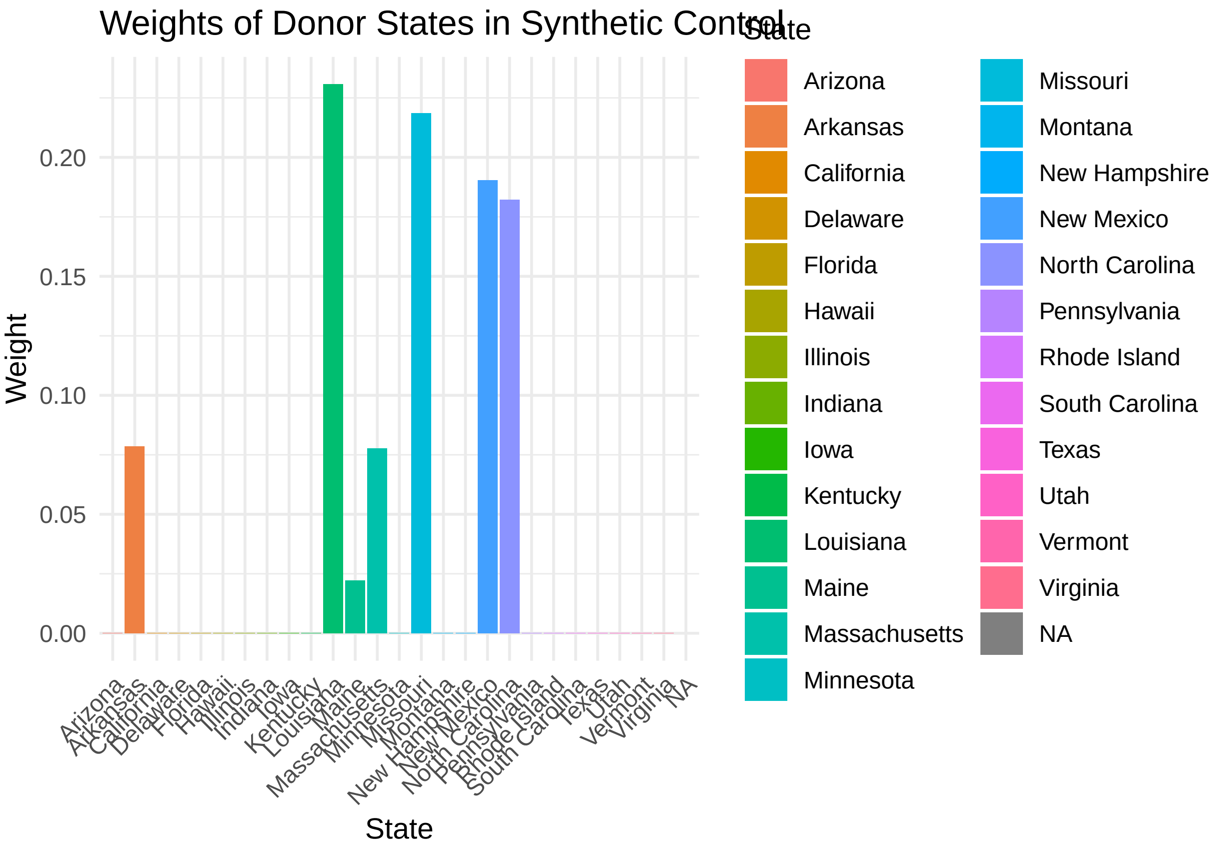
```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "Ridge", scm = TRUE)
##
## Average ATT Estimate (p Value for Joint Null):  -0.0243   ( 0.16 )
## L2 Imbalance: 0.003
## Percent improvement from uniform weights: 97.2%
##
## Avg Estimated Bias: 0.000
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.007             -0.059              0.045   0.200
##  2017   -0.029             -0.081              0.023   0.098
##  2018   -0.032             -0.084              0.019   0.096
##  2019   -0.026             -0.078              0.026   0.117
##  2020   -0.027             -0.079              0.024   0.101
```

```
# Output:
# Average ATT Estimate (p Value for Joint Null):  -0.0243   ( 0.19 )
# L2 Imbalance: 0.003
```

- Plot barplots to visualize the weights of the donors.

```
# barplots of weights

donor_states <- data$State[data$treatment == 0 & !duplicated(data$State)]

weights <- syn_augmented$weights

weights_data <- data.frame(State = donor_states,
                           Weight = weights[1:length(donor_states)])

# Filter out zero weights
weights_data <- weights_data[weights_data$Weight > 0,]

ggplot(weights_data, aes(x = State, y = Weight, fill = State)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Weights of Donor States in Synthetic Control",
       x = "State",
       y = "Weight") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



Discussion Questions

- What are the advantages and disadvantages of synthetic control compared to difference-in-differences estimators?
- **Answer**: Synthetic control enables the selection of control units based on specific characteristics relevant to the treatment effect. Also, synthetic control can capture treatment effects in settings where there is significant heterogeneity across units.
- 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**:Negative weights imply that some control units have a counterfactual effect that opposes the treatment effect. This presents a conceptual hurdle as it suggests that certain control units are associated with a decrease in the outcome of interest when the treatment is applied. Such counterintuitive interpretations can be difficult to explain and justify. We can conduct sensitivity analyses to assess the impact of negative weights on the estimated treatment effect and then decide whether to adopt this approach or not.

Staggered Adoption Synthetic Control

Estimate Multisynth

Do the following:

- Estimate a multisynth model that treats each state individually.

```r
# multisynth model states

medicaid_expansion$treatment <- ifelse(medicaid_expansion$year >= medicaid_expansion$Adopted_Year, 1, 0

multi_syn <- multisynth(uninsured_rate ~ treatment,
                        State,
                        year,
                        medicaid_expansion,
                        n_leads = 5)

print(multi_syn)
```

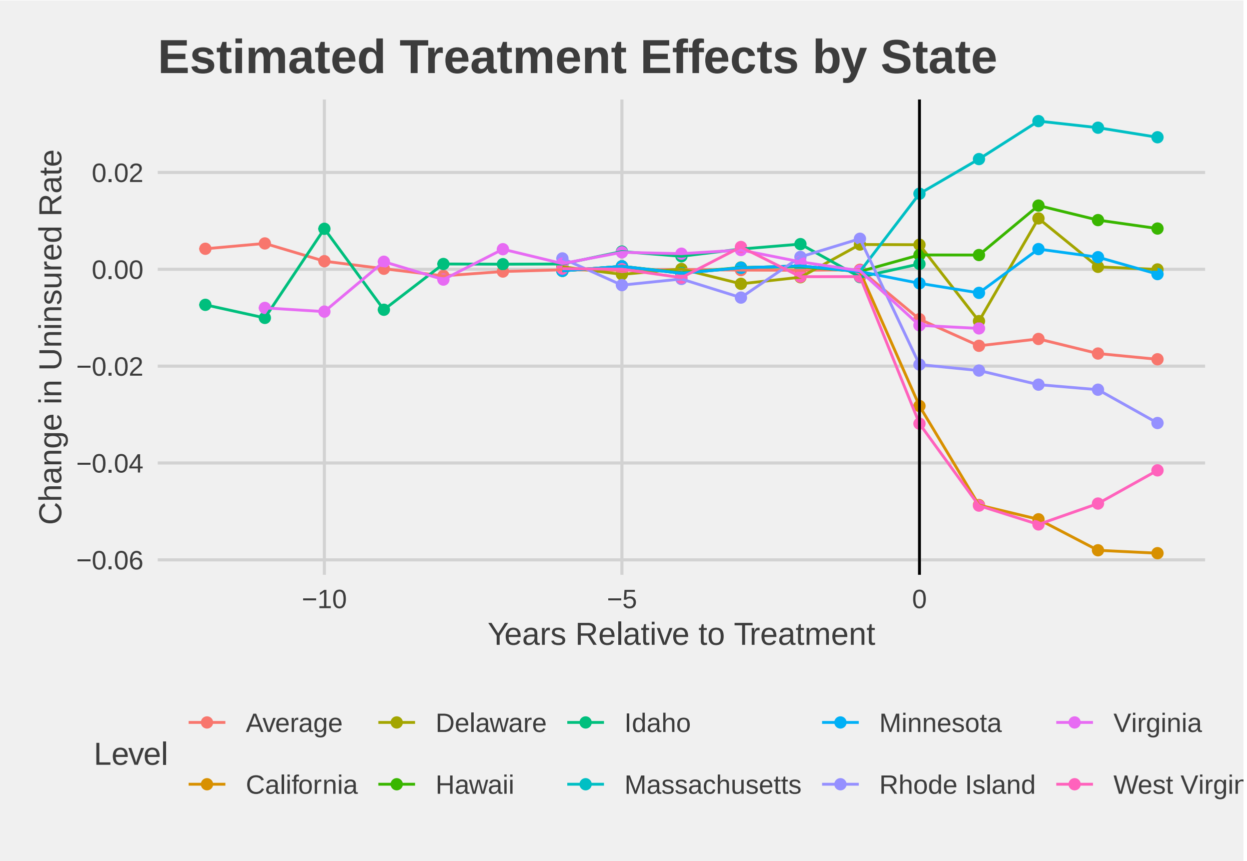
```
##
## Call:
## multisynth(form = uninsured_rate ~ treatment, unit = State, time = year,
##     data = medicaid_expansion, n_leads = 5)
##
## Average ATT Estimate: -0.014
```

- Choose a fraction of states that you can fit on a plot and examine their treatment effects.

```r
# Choose 10 states that can fit on a plot and examine their treatment effects.

multi_syn_summ <- summary(multi_syn)

states_to_plot <- sample(unique(multi_syn_summ$att$Level), 10)

filtered_data <- multi_syn_summ$att %>%
  filter(Level %in% states_to_plot)

filtered_data %>%
  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('Estimated Treatment Effects by State') +
  xlab('Years Relative to Treatment') +
  ylab('Change in Uninsured Rate')
```

```
## Warning: Removed 60 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 60 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.

```
# multisynth model time cohorts

multi_syn_time <- multisynth(uninsured_rate ~ treatment,
                             State,
                             year,
                             medicaid_expansion,
                             n_leads = 5,
                             time_cohort = TRUE)


multi_syn_time_summ <- summary(multi_syn_time)
multi_syn_time_summ
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treatment, unit = State, time = year,
##     data = medicaid_expansion, n_leads = 5, time_cohort = TRUE)
##
## Average ATT Estimate (Std. Error): -0.014  (0.006)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.007
## Percent improvement from uniform global weights: 99.3
```

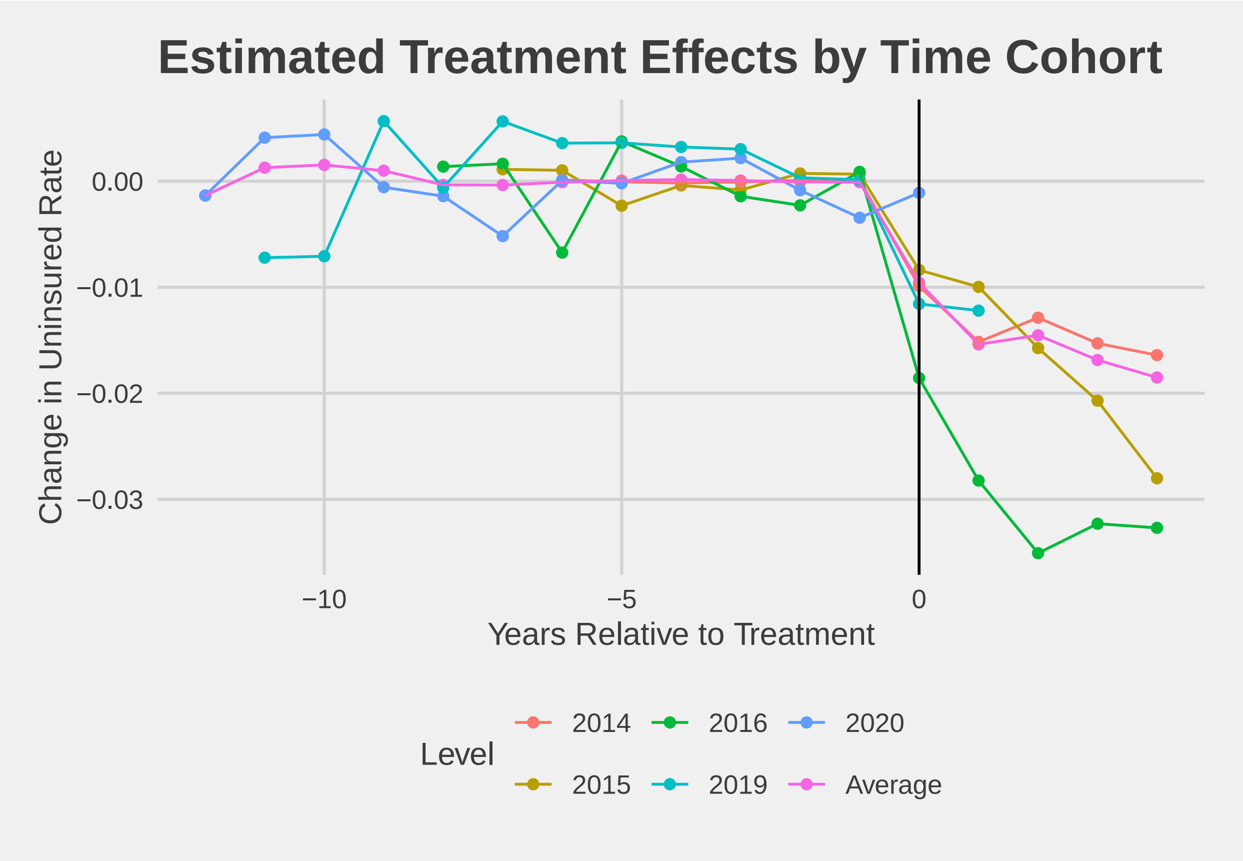
```
##
## Individual L2 Imbalance: 0.005
## Scaled Individual L2 Imbalance: 0.015
## Percent improvement from uniform individual weights: 98.5
##
##  Time Since Treatment   Level     Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.009569598 0.004770268 -0.01959715 -0.001095949
##                     1 Average -0.015386484 0.006121986 -0.02778910 -0.003660139
##                     2 Average -0.014526643 0.006516000 -0.02807591 -0.002391281
##                     3 Average -0.016859429 0.006703827 -0.03051070 -0.004395855
##                     4 Average -0.018509238 0.006481106 -0.03186880 -0.006604689
```

Plot the treatment effects for these time cohorts.

```
# Plot
multi_syn_time_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level, label = Level)) +
  geom_point() +
  geom_line(show.legend = TRUE) +  # Add show.legend = TRUE to include legend
  geom_vline(xintercept = 0) +
  theme_fivethirtyeight() +
  theme(axis.title = element_text(),
        legend.position = 'bottom') +  # Position legend at the bottom
  ggtitle('Estimated Treatment Effects by Time Cohort') +
  xlab('Years Relative to Treatment') +
  ylab('Change in Uninsured Rate')
```

```
## Warning: Removed 29 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 29 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
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: No

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 synthetic control control for time-varying factors that affect treatment and control units similarly over time. This is especially important in studies of aggregated units where there may be common trends or unobserved factors influencing outcomes.
- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?
- **Answer**: In DiD and synthetic control approaches, selection bias occurs if there are systematic differences between treated and control units that influence their treatment status and outcomes. In

regression discontinuity designs, selection into treatment is based on a specific cutoff or threshold. Selection bias arises if subjects manipulate their assignment to treatment based on their characteristics around the cutoff. RD designs are appropriate when treatment assignment is based on a known cutoff or threshold, such as eligibility criteria for a program or policy. DiD and synthetic control are suitable when treatment assignment is based on exogenous factors or policy changes that affact some units but not others.