Project 7: Difference-in-Differences and Synthetic Control

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

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

```r
library(usmap)
library(ggplot2)

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

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

# set seed
set.seed(44)

# load data
medicaid_expansion <- read_csv('data/medicaid_expansion.csv', show_col_types=FALSE)
```

Introduction

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

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

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

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

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

Data

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

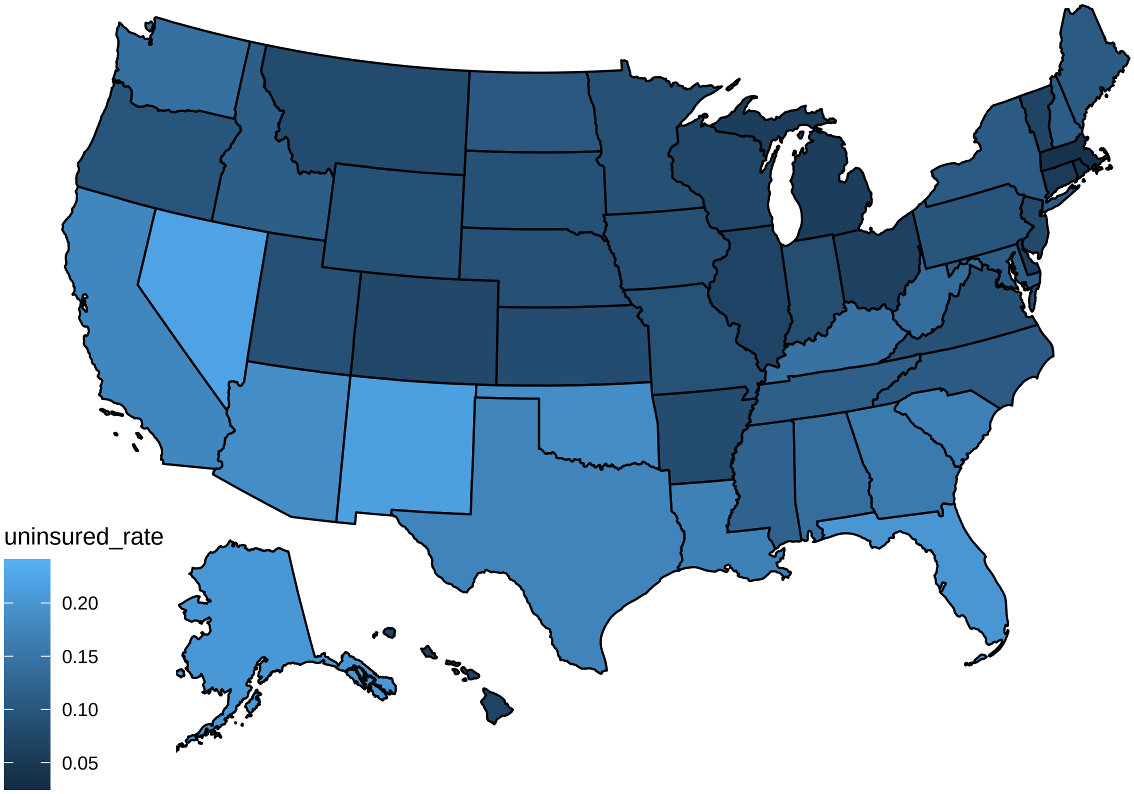
- **State**: Full name of state
- **Medicaid Expansion Adoption**: Date that the state adopted the Medicaid expansion, if it did so.
- **Year**: Year of observation.
- **Uninsured rate**: State uninsured rate in that year.

Exploratory Data Analysis

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

- Which states had the highest uninsured rates prior to 2014? The lowest?
- **Answer**: Nevada had the highest uninsured rate, and Massachusetts had the lowest uninsured rate, averaging over the years prior to 2014.
- Which states were home to most uninsured Americans prior to 2014? How about in the last year in the data set? **Note**: 2010 state population is provided as a variable to answer this question. In an actual study you would likely use population estimates over time, but to simplify you can assume these numbers stay about the same.
- **Answer**: California had the most uninsured Americans prior to 2014, but Texas had the most in 2020.

```
# highest and lowest uninsured rates

#cleaning
colnames(medicaid_expansion)[1] ="state"
#medicaid_expansion$month <- format(as.Date(medicaid_expansion$Date_Adopted, format="%Y/%m/%d"),"%m")
#medicaid_expansion$month <- month(medicaid_expansion$Date_Adopted)
#medicaid_expansion$yr_mon <- medicaid_expansion$year + ((medicaid_expansion$month - 1) / 12)

#a map for fun
plot_usmap(data = medicaid_expansion, values = "uninsured_rate")
```



```
#Q1
summ <- medicaid_expansion %>%
  subset(year < 2014) %>%
  group_by(state) %>%
  summarize(avg_rate = mean(uninsured_rate)) %>%
  as.data.frame()

summ$state[which.max(summ$avg_rate)] #Nevada is highest uninsured rate
```

```
## [1] "Nevada"
```

```
summ$state[which.min(summ$avg_rate)] #Massachusetts is lowest uninsured rate
```

```
## [1] "Massachusetts"
```

```
# most uninsured Americans
#Q2
summ <- medicaid_expansion %>%
  subset(year < 2014) %>%
  mutate(unemp_pop = uninsured_rate * population) %>%
  group_by(state) %>%
  summarize(avg_unemp_pop = mean(unemp_pop)) %>%
  as.data.frame

summ$state[which.max(summ$avg_unemp_pop)] #California
```

```
## [1] "California"
```

```
summ$state[which.min(summ$avg_unemp_pop)] #Vermont
```

```
## [1] "Vermont"
```

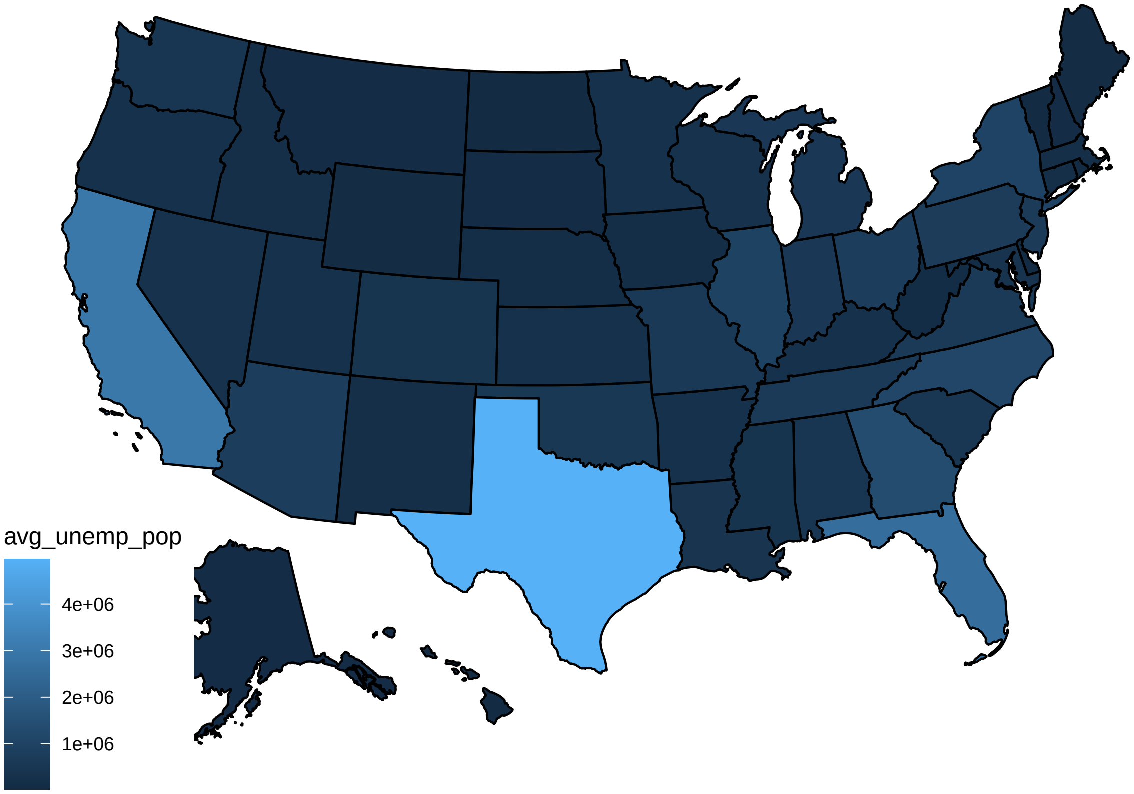
```
#in the last year = 2020
summ <- medicaid_expansion %>%
  subset(year == 2020) %>%
  mutate(unemp_pop = uninsured_rate * population) %>%
  group_by(state) %>%
  summarize(avg_unemp_pop = mean(unemp_pop)) %>%
  as.data.frame

summ$state[which.max(summ$avg_unemp_pop)] #Texas
```

```
## [1] "Texas"
```

```
summ$state[which.min(summ$avg_unemp_pop)] #Vermont
```

```
## [1] "Vermont"
```

```
#map for 2020  unemployment population
plot_usmap(data = summ, values = "avg_unemp_pop")
```



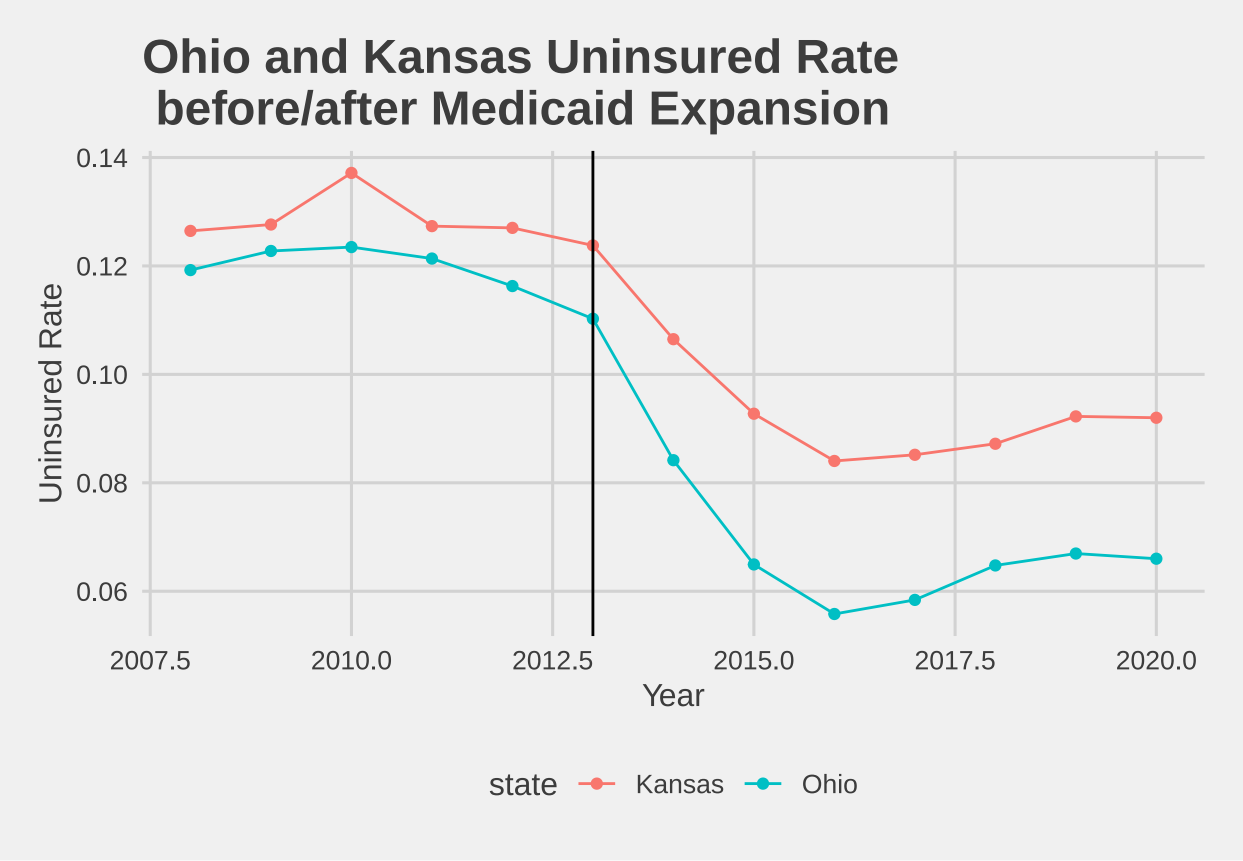
Difference-in-Differences Estimation

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

#let's do Ohio and Kansas

medicaid_expansion %>%
  filter(state %in% c("Ohio","Kansas")) %>%
  ggplot() +
  geom_point(aes(x = year,
                 y = uninsured_rate,
                 color = state)) +
  geom_line(aes(x = year,
                y = uninsured_rate,
                color = state)) +
  geom_vline(aes(xintercept = 2013)) +
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  ggtitle('Ohio and Kansas Uninsured Rate \n before/after Medicaid Expansion') +
  xlab('Year') +
  ylab('Uninsured Rate')
```



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

```
# Difference-in-Differences estimation
# ohio-kansas
```

```
ok <- medicaid_expansion %>%
  filter(state %in% c("Ohio","Kansas")) %>%
  filter(year >= 2013 & year<= 2014)

# pre-treatment difference

pre_diff <- ok %>%
  filter(year == 2013) %>%
  select(state,
         uninsured_rate) %>%
  spread(state,
         uninsured_rate) %>%
  summarise(Kansas - Ohio)

# post-treatment difference

post_diff <- ok %>%
  filter(year == 2014) %>%
  select(state,
         uninsured_rate) %>%
  spread(state,
         uninsured_rate) %>%
  summarise(Kansas - Ohio)

# diff-in-diffs

diff_in_diffs <- post_diff - pre_diff
diff_in_diffs
```

```
##   Kansas - Ohio
## 1        0.0088
```

Looks like our treatment effect is about .009 (as a percent of population uninsured).

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**: This data is at the state level, which aggregates features such as demographics and economics, making the comparison more broad than the Card/Krueger example.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: The strengths of the parallel trends assumption in DiD estimates are that it allows for causal inference and is relatively more efficient than other techniques, such as RDs, at large sample size. Some weaknesses of the parallel trends assumption in DiD estimates are that it is relatively easy to break the assumption, requires group similarity that is often difficult to convincingly prove, and confounders such as time-varying factors can easily break the assumption.

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.

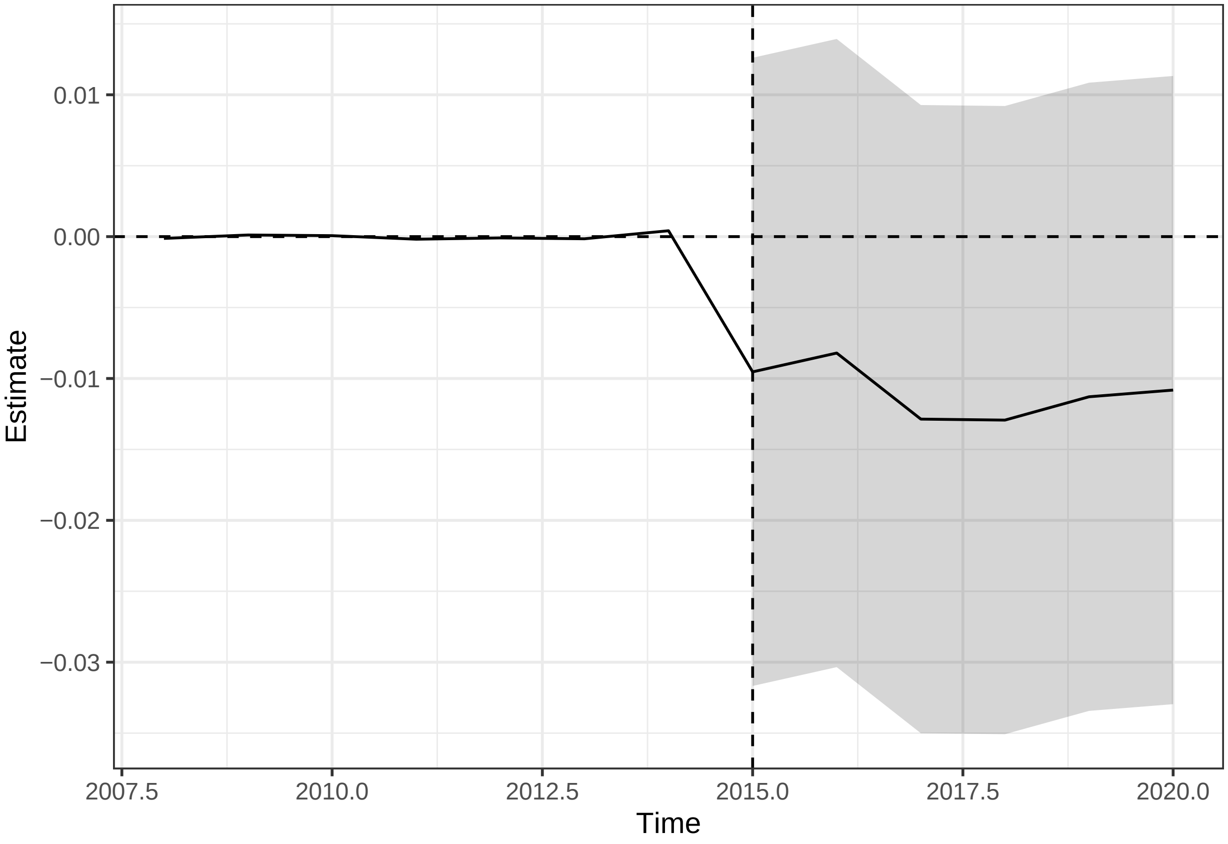
Do the following:

- 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
# look at Pennsylvania
# construct non-augmented synthetic control

pa <- medicaid_expansion %>%
    mutate(treatment = ifelse(state == "Pennsylvania" & year >= 2015,
                            1,
                            0))

syn <- augsynth(uninsured_rate ~ treatment, state, year, pa,
                progfunc = "None", scm = T)
```

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

```
plot(syn)
```



```
summary(syn)
```

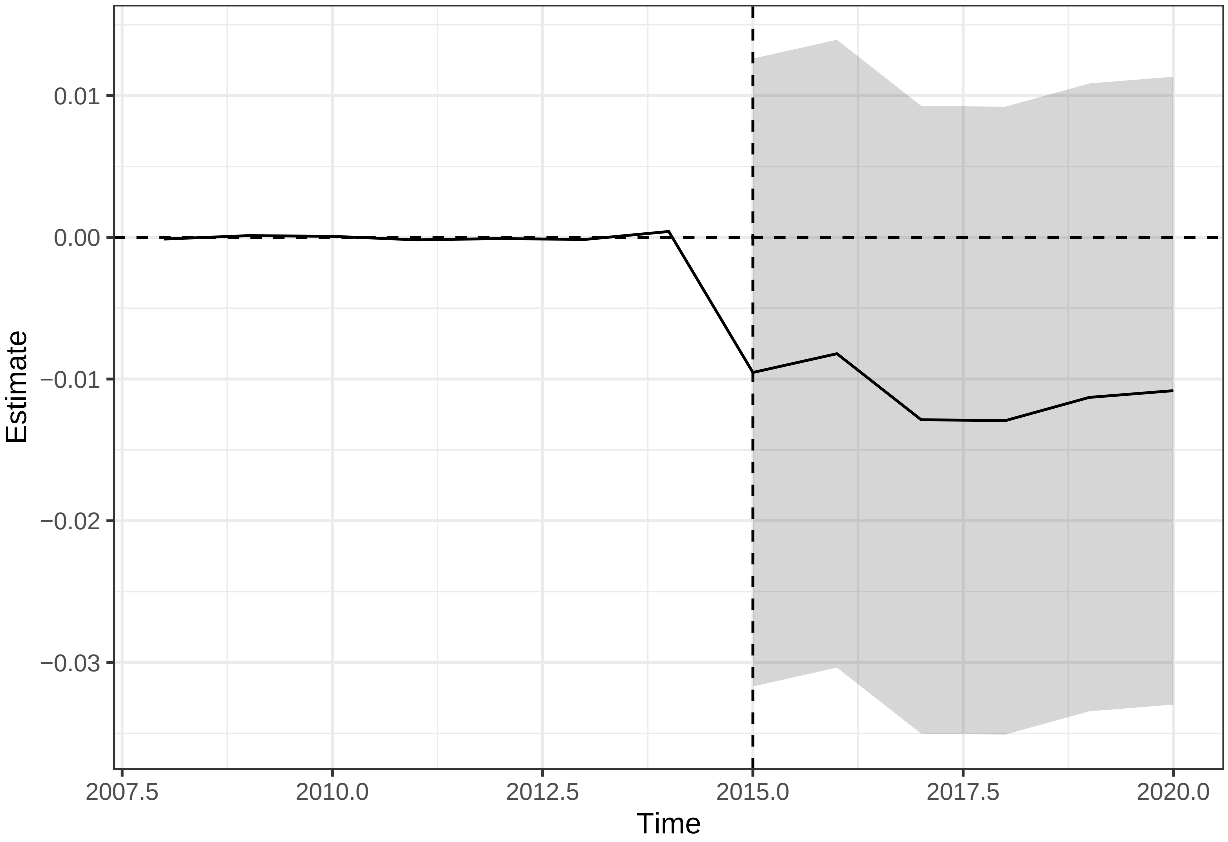
```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "None", scm = ..2)
##
## Average ATT Estimate (p Value for Joint Null):  -0.011   ( 0.609 )
## L2 Imbalance: 0.001
## Percent improvement from uniform weights: 99.5%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2015   -0.010             -0.032              0.013   0.250
##  2016   -0.008             -0.030              0.014   0.250
##  2017   -0.013             -0.035              0.009   0.250
##  2018   -0.013             -0.035              0.009   0.375
##  2019   -0.011             -0.033              0.011   0.250
##  2020   -0.011             -0.033              0.011   0.250
```

- 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_aug <- augsynth(uninsured_rate ~ treatment, state, year, pa,
                progfunc = "ridge", scm = T)
```

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

```
plot(syn_aug)
```



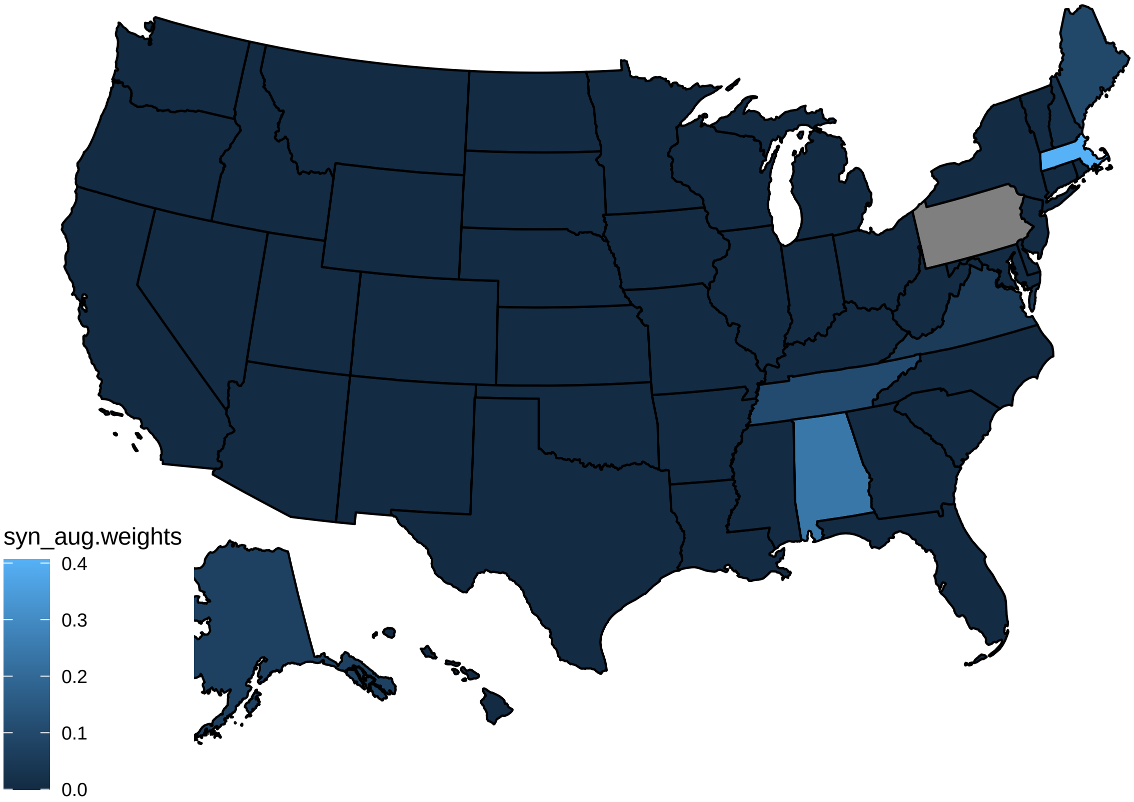
```
summary(syn_aug)
```

```
##
## 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):  -0.011   ( 0.618 )
## L2 Imbalance: 0.001
## Percent improvement from uniform weights: 99.5%
##
## Avg Estimated Bias: 0.000
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2015   -0.010             -0.032              0.013   0.250
##  2016   -0.008             -0.030              0.014   0.250
##  2017   -0.013             -0.035              0.009   0.250
##  2018   -0.013             -0.035              0.009   0.375
##  2019   -0.011             -0.033              0.011   0.250
##  2020   -0.011             -0.033              0.011   0.250
```

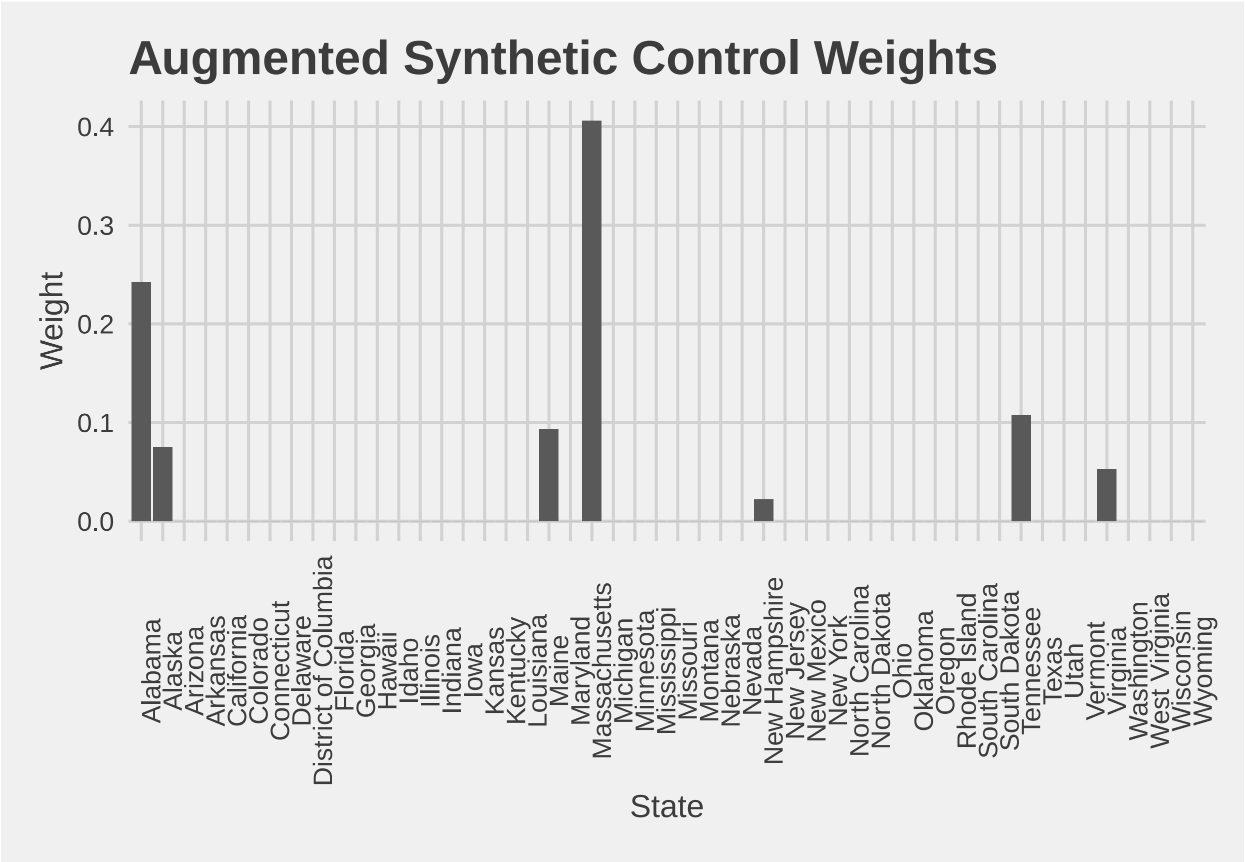
- Plot barplots to visualize the weights of the donors.

```
test <- data.frame(syn_aug$weights) %>%
  tibble::rownames_to_column('state')
```

```
plot_usmap(data = test, values = "syn_aug.weights")
```



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



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

Discussion Questions

- What are the advantages and disadvantages of synthetic control compared to difference-in-differences estimators?
- **Answer**: Synthetic controls can handle n's of treated units and manufactures its own control while DiD usually needs multiple treated and controlled observations. Synthetic controls also can use more informationf from pre-treatment controls to create the weights whereas DiD relies solely on differences over time in treatment and control groups. There are disadvantages to synthetic controls though. In particular, synthetic controls require strong assumptions of the validity of the weights. It is also less clear to interpret compared to DiD models. The control group is synthetic and so may not be directly interpretable.
- 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 may make interpretability difficult but may be useful in capturing complex treatment effects. One way to balance these considerations is to conduct sensitivity analyses that compare the results of synthetic control with positive-only weights to the results of augmentation with negative weights, and to assess the robustness of the treatment effect estimates to different specifications of the control group and the weighting scheme. Another way is to use a priori information or qualitative reasoning to justify the use of negative weights in specific contexts, such as when the control units exhibit negative spillover effects that need to be captured in the synthetic control.

Staggered Adoption Synthetic Control

Estimate Multisynth

Do the following:

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

```
# multisynth model states
# let's look at the midatlantic states, excluding DC
# Delaware, Maryland, New Jersey, New York, Pennsylvania, Virginia, West Virginia, and Washington, D.C.

# do some cleaning
multisynth_grp <- medicaid_expansion %>%
  #filter(state %in% c("Delaware", "Maryland", "New Jersey", "New York", "Pennsylvania", "Virginia", "W
  mutate(year_adopted = as.numeric(format(Date_Adopted,'%Y'))) %>%
  mutate(treated = ifelse(year >= year_adopted, 1, 0))

# with default nu
ppool_syn <- multisynth(uninsured_rate ~ treated, state, year,
                        multisynth_grp, n_leads = 10)

print(ppool_syn$nu)
```

```
## [1] 0.2933713
```

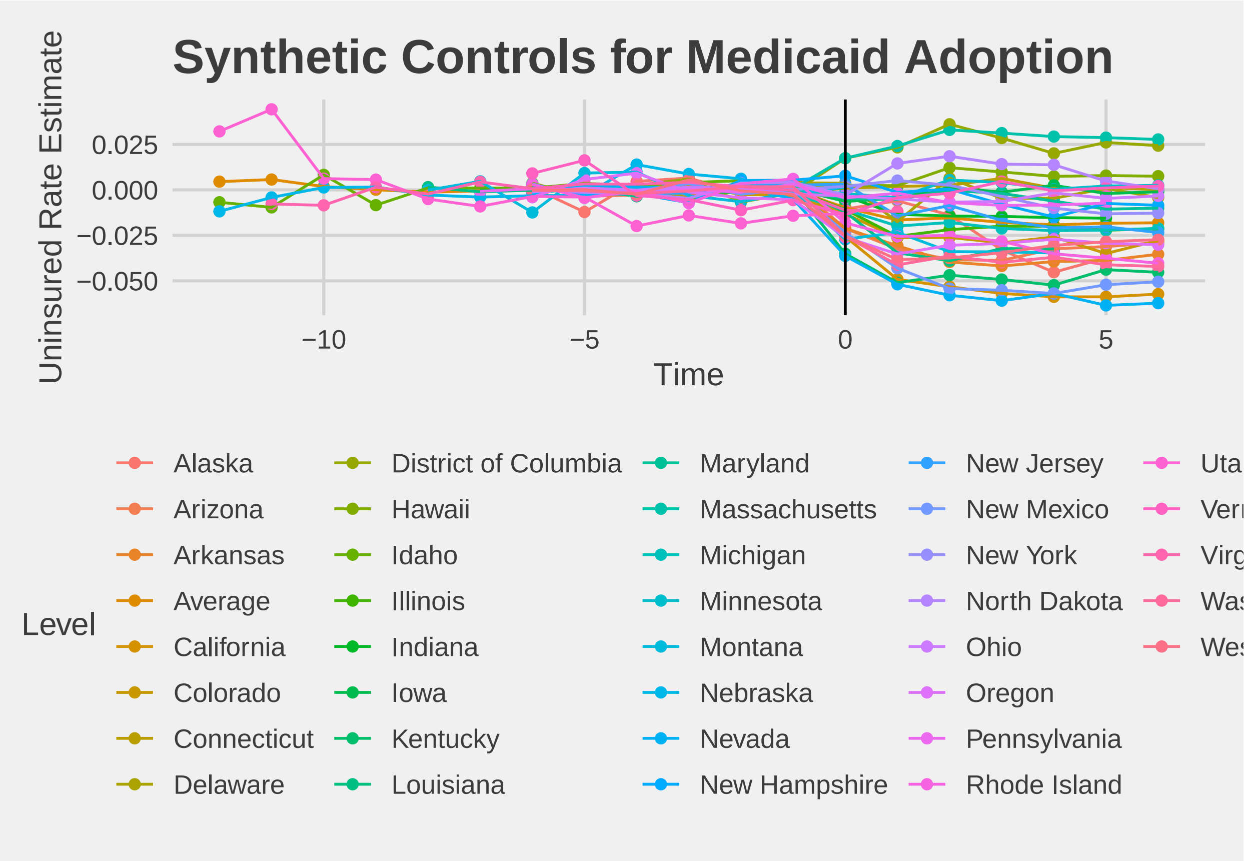
```
ppool_syn
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = state, time = year,
##     data = multisynth_grp, n_leads = 10)
##
## Average ATT Estimate: -0.015
```

```
ppool_syn_summ <- summary(ppool_syn)

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 Adoption') +
  xlab('Time') +
  ylab('Uninsured Rate Estimate')
```

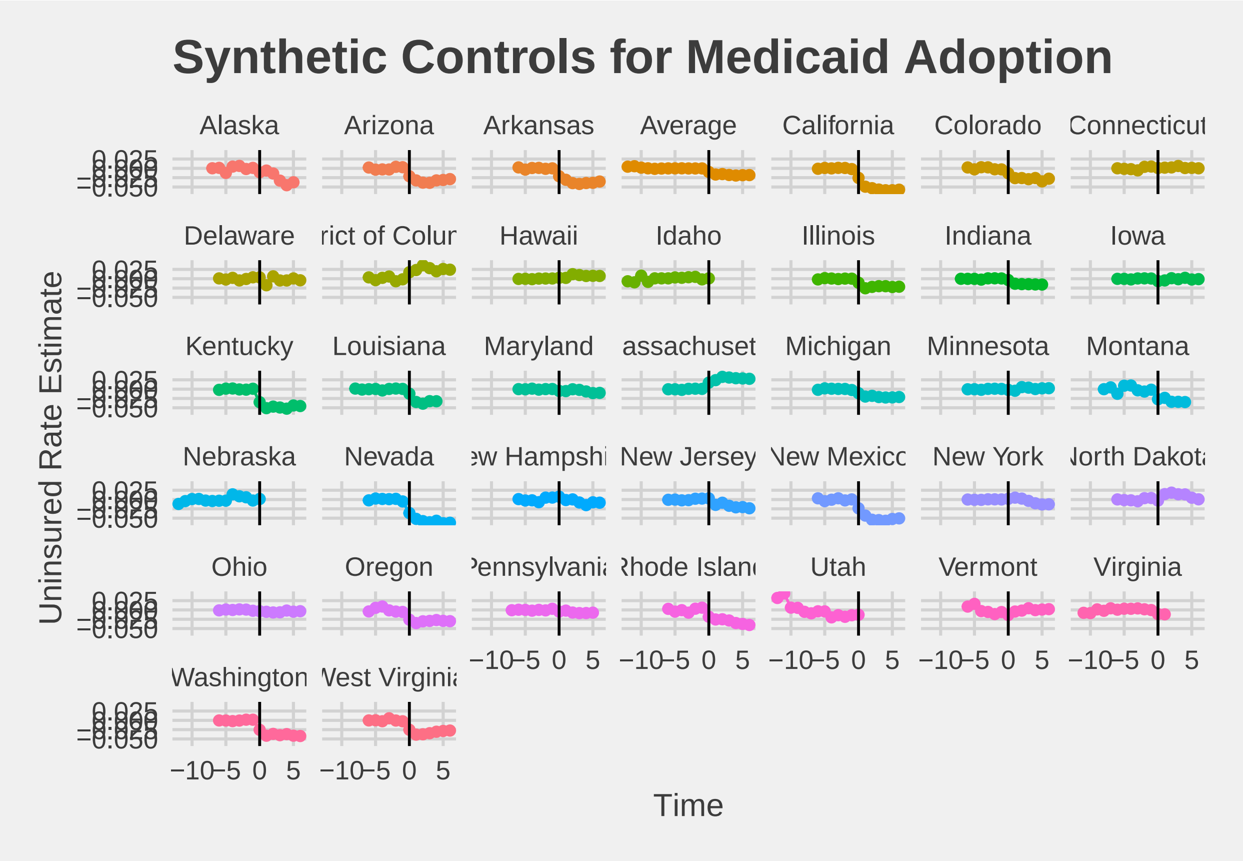
```
## Warning: Removed 253 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 253 rows containing missing values (`geom_line()`).
```



```
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 = 'None') +
  ggtitle('Synthetic Controls for Medicaid Adoption') +
  xlab('Time') +
  ylab('Uninsured Rate Estimate') +
  facet_wrap(~Level)
```

```
## Warning: Removed 253 rows containing missing values (`geom_point()`).
## Removed 253 rows containing missing values (`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
ppool_syn_time <- multisynth(uninsured_rate ~ treated, state, year,
                        multisynth_grp, n_leads = 10, time_cohort = TRUE)

ppool_syn_time_summ <- summary(ppool_syn_time)

ppool_syn_time_summ
```

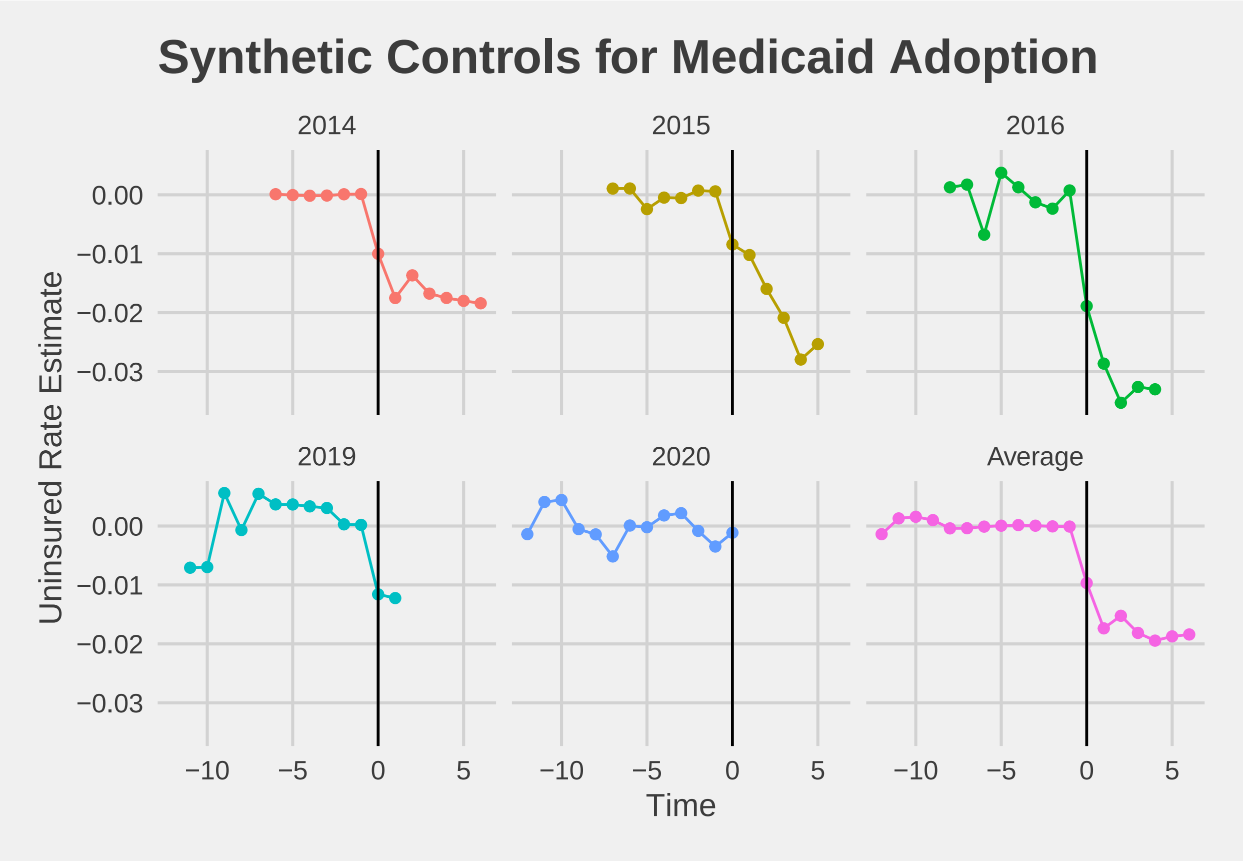
```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = state, time = year,
##     data = multisynth_grp, n_leads = 10, time_cohort = TRUE)
##
## Average ATT Estimate (Std. Error): -0.016  (0.006)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.007
## Percent improvement 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
##                      0 Average -0.009668818 0.004702434 -0.01868080
##                      1 Average -0.017359993 0.005890871 -0.02852409
##                      2 Average -0.015229414 0.005957062 -0.02692056
##                      3 Average -0.018133781 0.006157548 -0.03049711
##                      4 Average -0.019443367 0.006038734 -0.03138100
##                      5 Average -0.018723960 0.005741660 -0.02970989
##                      6 Average -0.018399231 0.006230170 -0.02982072
##    upper_bound
##  -0.0008023836
##  -0.0053845460
##  -0.0037532031
##  -0.0064041500
##  -0.0079508357
##  -0.0071205550
##  -0.0052228144
```

```
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 Adoption') +
  xlab('Time') +
  ylab('Uninsured Rate Estimate') +
  facet_wrap(~Level)
```

```
## Warning: Removed 36 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 36 rows containing missing values (`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, if adoption were the same across the board at the state level, the effects would be similar in size across states after adoption. However, this is not the case. However, it is difficult to tell how that changes across time after adoption. Changing policies after adoption may not be well modeled here since the main effect of study is the effect of adoption at a broad level, not varieties of adoption.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: It's difficult to tell a trend since there weren't any adopters in 2017 and 2018. The decrease in the uninsured rate is smaller in 2019 than previous years, which does suggest that early adopter had a larger decrease in uninsured rates. However, that trend does not hold within the early years, with an larger decreases every year in uninsured rates from 2014 to 2016

General Discussion Questions

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

- **Answer**:Difference-in-Differences (DiD) and Synthetic Control methods are well-suited for studies of aggregated units like cities, states, and countries because they are designed to address the problem of unobserved heterogeneity across units. In observational studies, the assignment of units (such as cities or states) to different treatments or policies is usually not random, which means that there may be other factors that affect the outcome of interest besides the treatment or policy of interest. DiD and Synthetic Controls attempt to control for this unobserved heterogeneity.
- 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, selection into treatment is typically addressed by assuming that the treated and control groups would have followed the same trend in the outcome of interest in the absence of the treatment. In contrast, regression discontinuity designs address selection into treatment by exploiting the fact that the treatment is assigned based on a continuous variable (the "assignment variable") that is correlated with the outcome of interest. This allows researchers to estimate the causal effect of the treatment by comparing outcomes of units just above and just below a threshold in the assignment variable. In general, DiD and synthetic control methods are best suited for situations where treatment is assigned at the group or aggregate level, and where there is a clear control group of untreated units. Regression discontinuity designs are best suited for situations where treatment is assigned based on a continuous variable, and where there is a clear cutoff or threshold in the assignment variable.