Project 7: Difference-in-Differences and Synthetic Control

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

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

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

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

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

# set seed
set.seed(44)

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

```
## Rows: 663 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): State
## dbl  (3): year, uninsured_rate, population
## date (1): Date_Adopted
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Introduction

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

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

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

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

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

Data

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

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

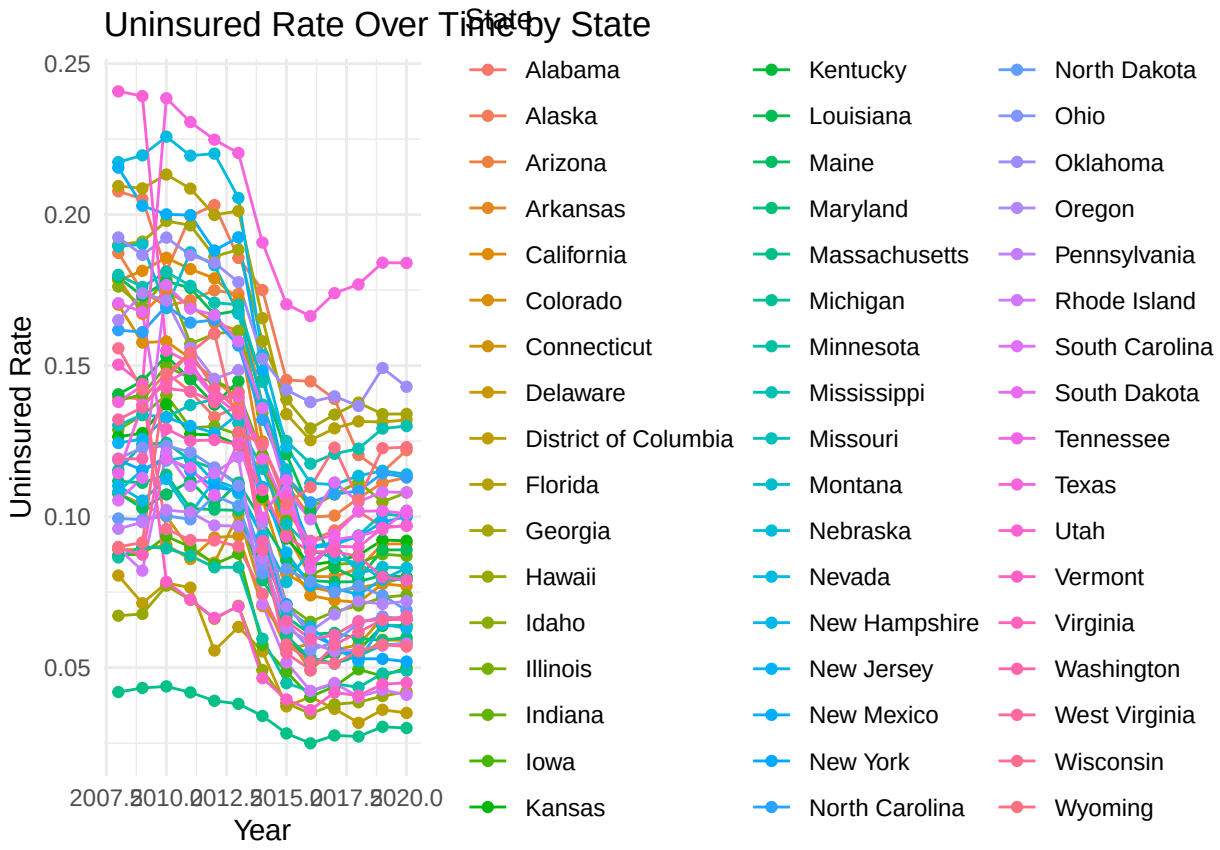
Exploratory Data Analysis

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

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

Answer: Before 2014, Massachusetts, Hawaii, and Minnesota had the lowest uninsured rate, while New Mexico, Florida, and Nevada had the highest uninsurance rates.

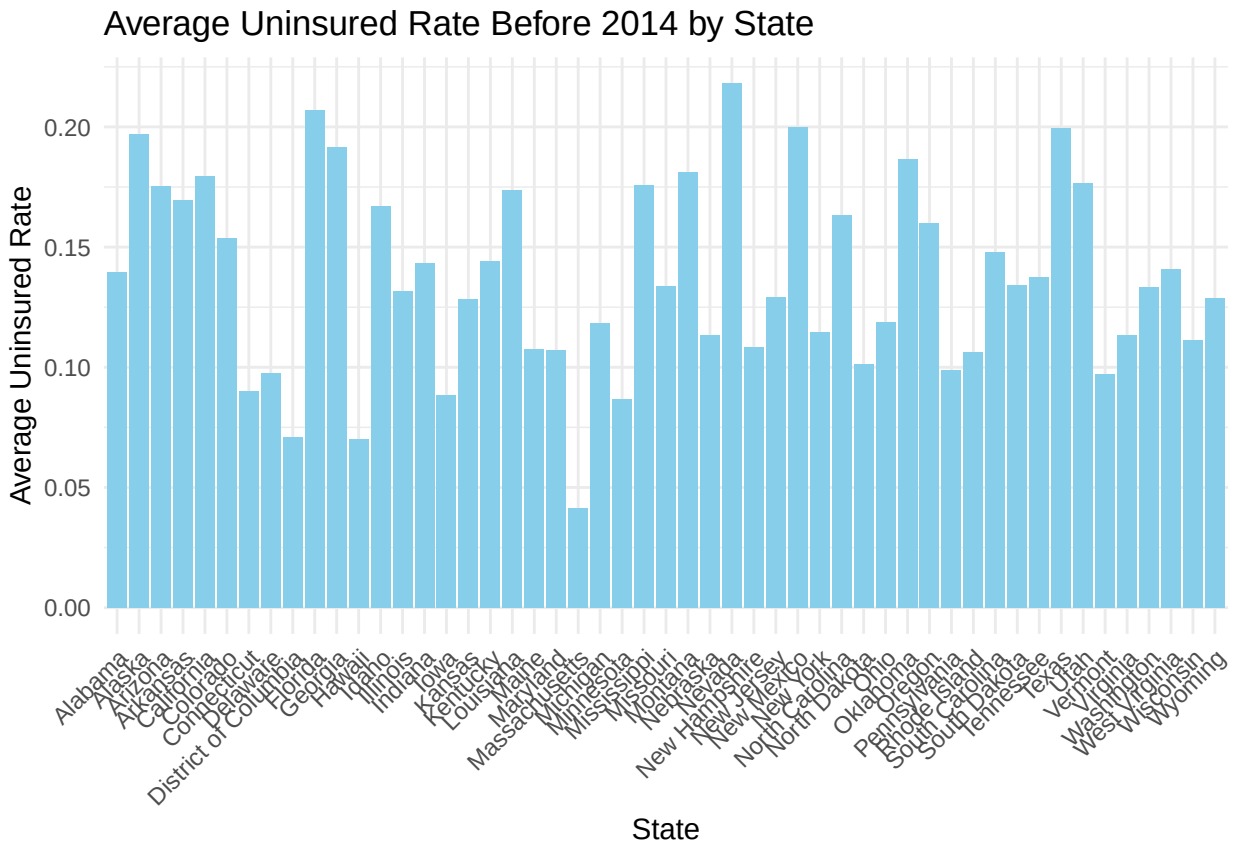
```
# highest and lowest uninsured rates
ggplot(medicaid_expansion, aes(x = year, y = uninsured_rate, group = State, color = State)) +
  geom_line() +
  geom_point() +
  labs(title = "Uninsured Rate Over Time by State",
       x = "Year",
       y = "Uninsured Rate") +
  theme_minimal()
```




```r
# Filter dataframe to include only years before 2014
filtered_data <- medicaid_expansion %>%
  filter(year < 2014)

# Calculate average uninsured rate grouped by state
average_uninsured_rate <- filtered_data %>%
  group_by(State) %>%
  summarise(average_uninsured_rate = mean(uninsured_rate))%>%
  arrange(average_uninsured_rate)  # Sort by average uninsured rate

# Plotting
ggplot(average_uninsured_rate, aes(x = State, y = average_uninsured_rate)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Average Uninsured Rate Before 2014 by State",
       x = "State",
       y = "Average Uninsured Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



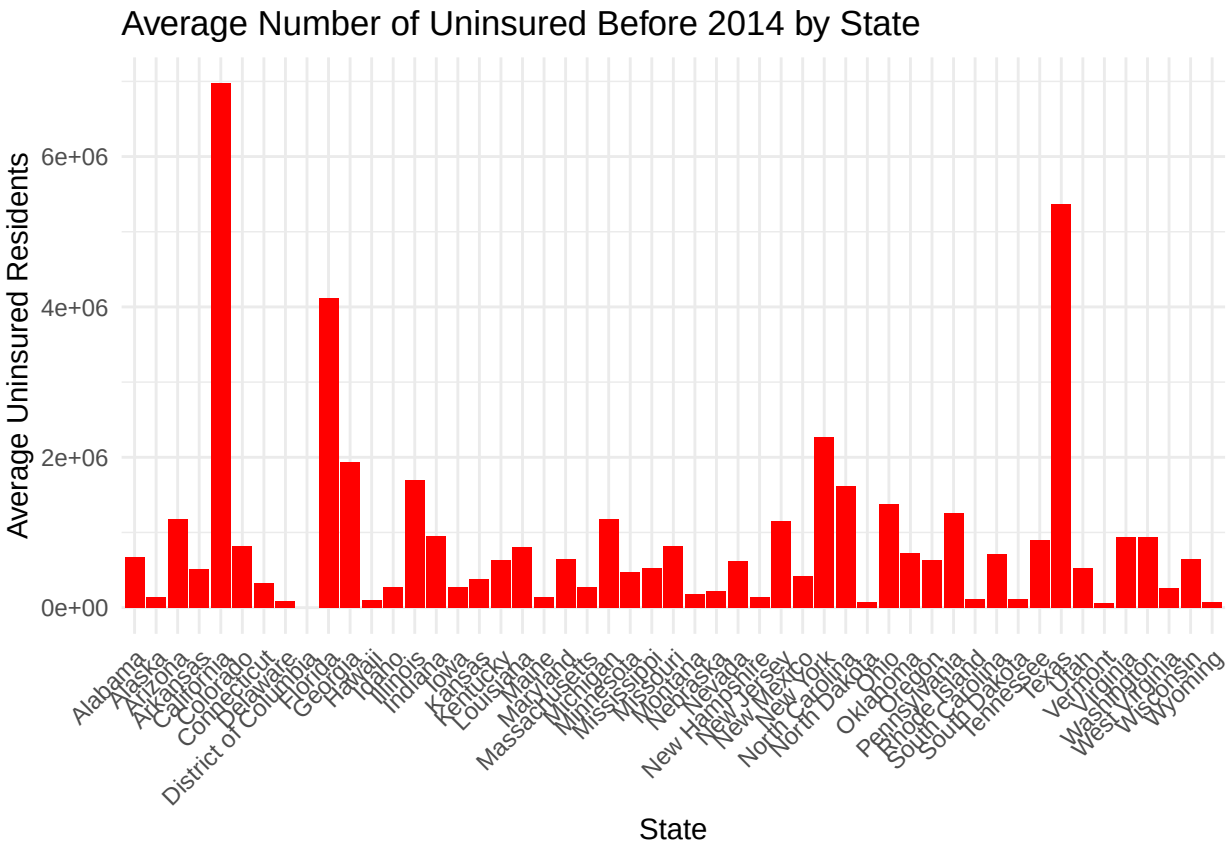
Answer: Before expansion and averaging the uninsured rate in all pre-expansion years, the states with the highest number of uninsured residents are California, Texas, and Florida.

```
# most uninsured Americans pre-2014
n_population<-medicaid_expansion%>%
  group_by(State)%>%summarise(population=mean(population))

n_uninsured <- average_uninsured_rate %>%
  arrange(State)%>%mutate(n_pop=n_population$population)%>%mutate(n_unins=n_pop*average_uninsured_rate)%

#plot
ggplot(n_uninsured, aes(x = State, y = n_unins)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Average Number of Uninsured Before 2014 by State",
       x = "State",
       y = "Average Uninsured Residents") +
  theme_minimal() +
  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()`).
```

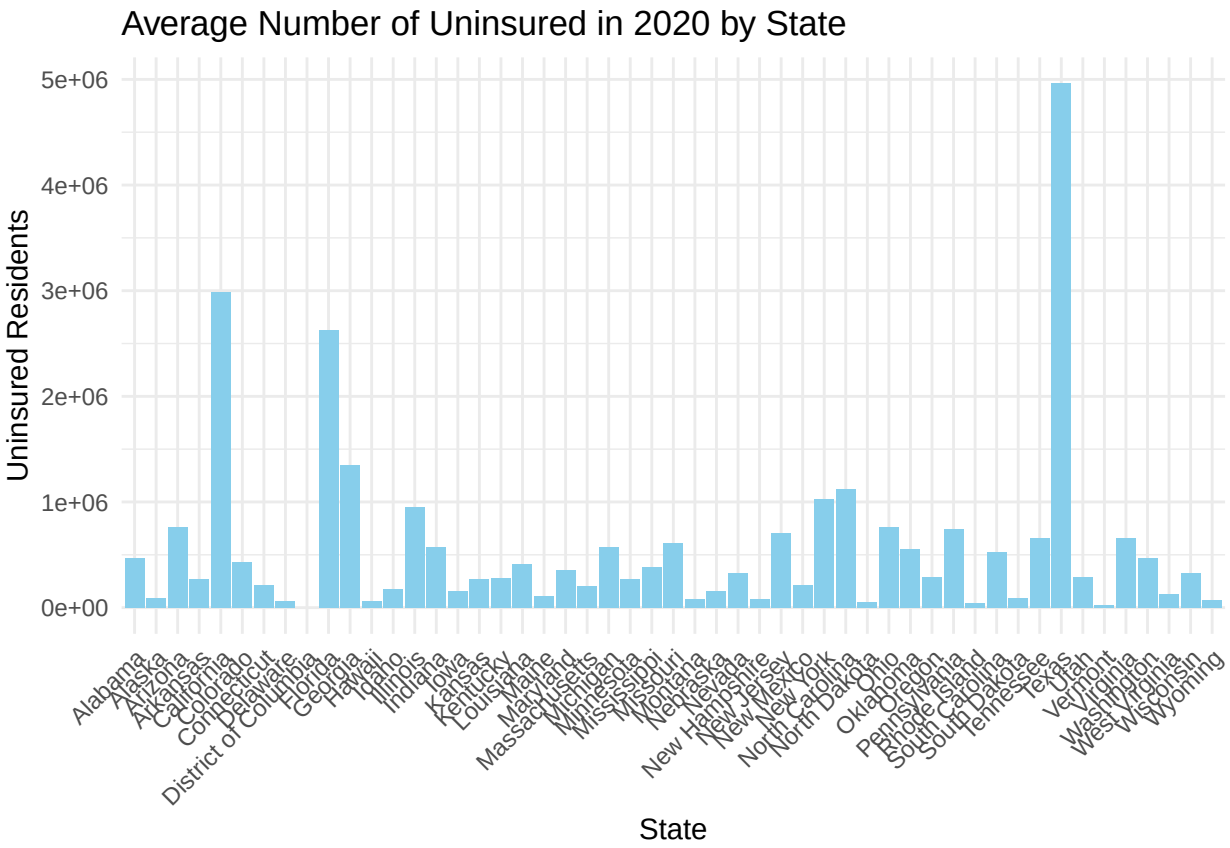


A**Answer**: In 2020, the states with the highest number of uninsured residents were Texas, California, and Florida.

```
# check in the last year of the dataset
last_year_data<-medicaid_expansion%>%filter(year==2020)
uninsured_2020 <- last_year_data %>%
  arrange(State)%>%mutate(n_pop=n_population$population)%>%mutate(n_unins=n_pop*uninsured_rate)%>%arran

#plot
ggplot(uninsured_2020, aes(x = State, y = n_unins)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Average Number of Uninsured in 2020 by State",
       x = "State",
       y ="Uninsured Residents") +
  theme_minimal() +
  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()`).
```



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).

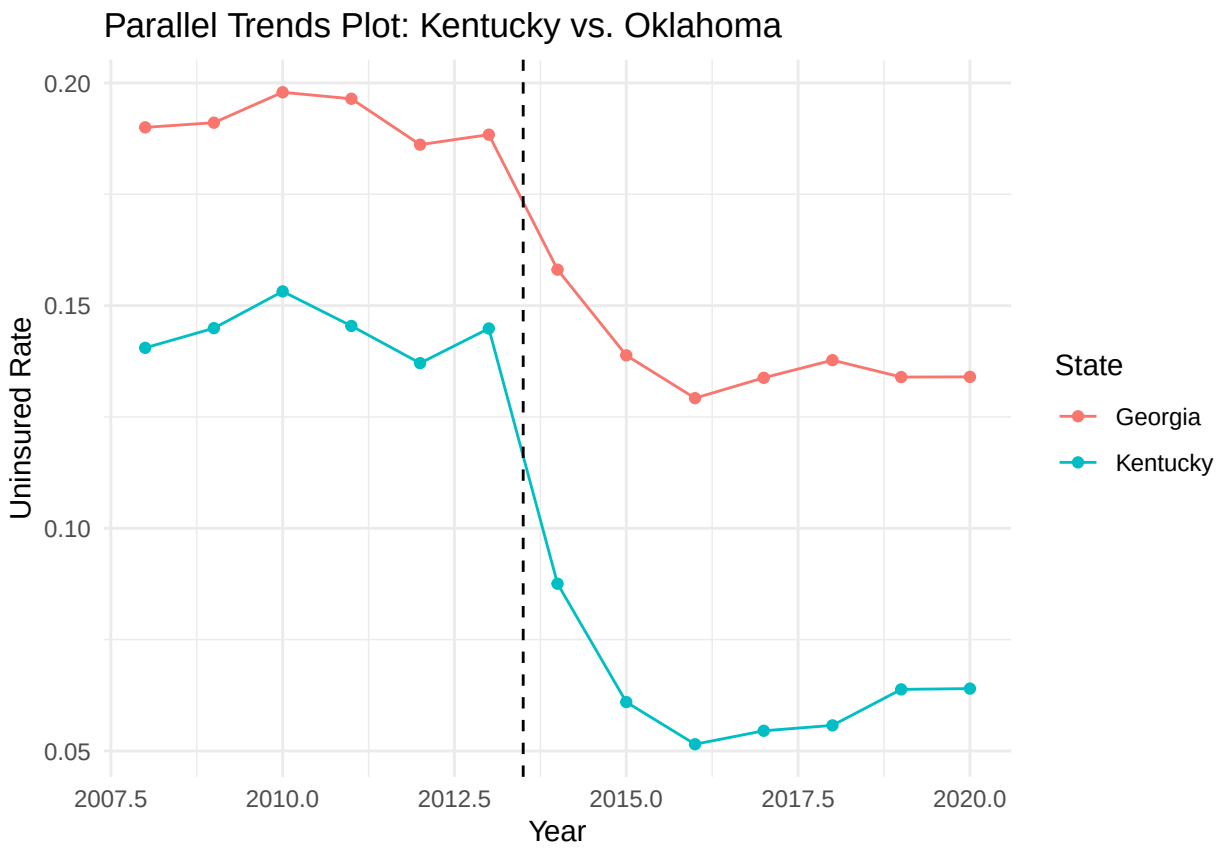
Answer: I picked Kentucky as my treated state because I was a little surprised Mitch McConnell's state was an early adopter. I tried Oklahoma and didn't love the fit, but Georgia looks a bit better.

```
# Parallel Trends plot

#treated state=kentucky
did_data<-medicaid_expansion%>%filter(State=="Kentucky"|State=="Georgia")

#plot parallel trends
ggplot(did_data, aes(x = year, y = uninsured_rate, group = State, color = State)) +
  geom_line() +
  geom_point() +
  geom_vline(xintercept = 2013.5, linetype = "dashed") +  # Vertical line at 2014
  labs(title = "Parallel Trends Plot: Kentucky vs. Oklahoma",
       x = "Year",
       y = "Uninsured Rate",
```

```
       color = "State",
       linetype = "Medicaid Expansion") +
  theme_minimal()
```



- 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
pre_treatment <- did_data %>% filter(year < 2014)
post_treatment <- did_data %>% filter(year >= 2014)

# Calculate the average uninsured rate in the pre-treatment and post-treatment periods for each state
pre_avg <- pre_treatment %>% group_by(State) %>% summarise(avg_uninsured_rate_pre = mean(uninsured_rate
post_avg <- post_treatment %>% group_by(State) %>% summarise(avg_uninsured_rate_post = mean(uninsured_ra

# Merge the pre-treatment and post-treatment averages
did_data_results<- merge(pre_avg, post_avg, by = "State")

# Calculate the DiD estimate
did_data_results <- did_data_results %>%
  mutate(difference = avg_uninsured_rate_post - avg_uninsured_rate_pre)

did_estimate <- diff(did_data_results$difference)
did_estimate
```

```
## [1] -0.0280306
```

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**: Different states are likely to have different healthcare environments, including regulations, providers, payers, and public health and social services. For example, insurance providers are locked within state lines, which is why we have a fractured set of Blue Cross/Blue Shield plans, for example. In addition, different states can have vastly different demographics, economies and employment opportunities, and other characteristics that could all influence the uninsurance rate.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?

Answer: It's helpful as a visual check, but it's easy for folks to hand wave that this looks good when it really shouldn't. It's also hard to statistically test (although there is movement here with the Roth pre-trends test). Determining how many pre-periods is required for valid testing is also unclear.

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.

Answer: Here the average ATT is -.0275 with an L2 imbalance of .021.

```
# non-augmented synthetic control - picking Montana and then only picking the states that were never tr
synth1_data<-medicaid_expansion%>%filter(State=="Montana"| is.na(Date_Adopted))%>%mutate(treated=case_wh

nonaugsynth<-augsynth(uninsured_rate ~ treated, State, year, synth1_data,
                progfunc = "None", scm = T)
```

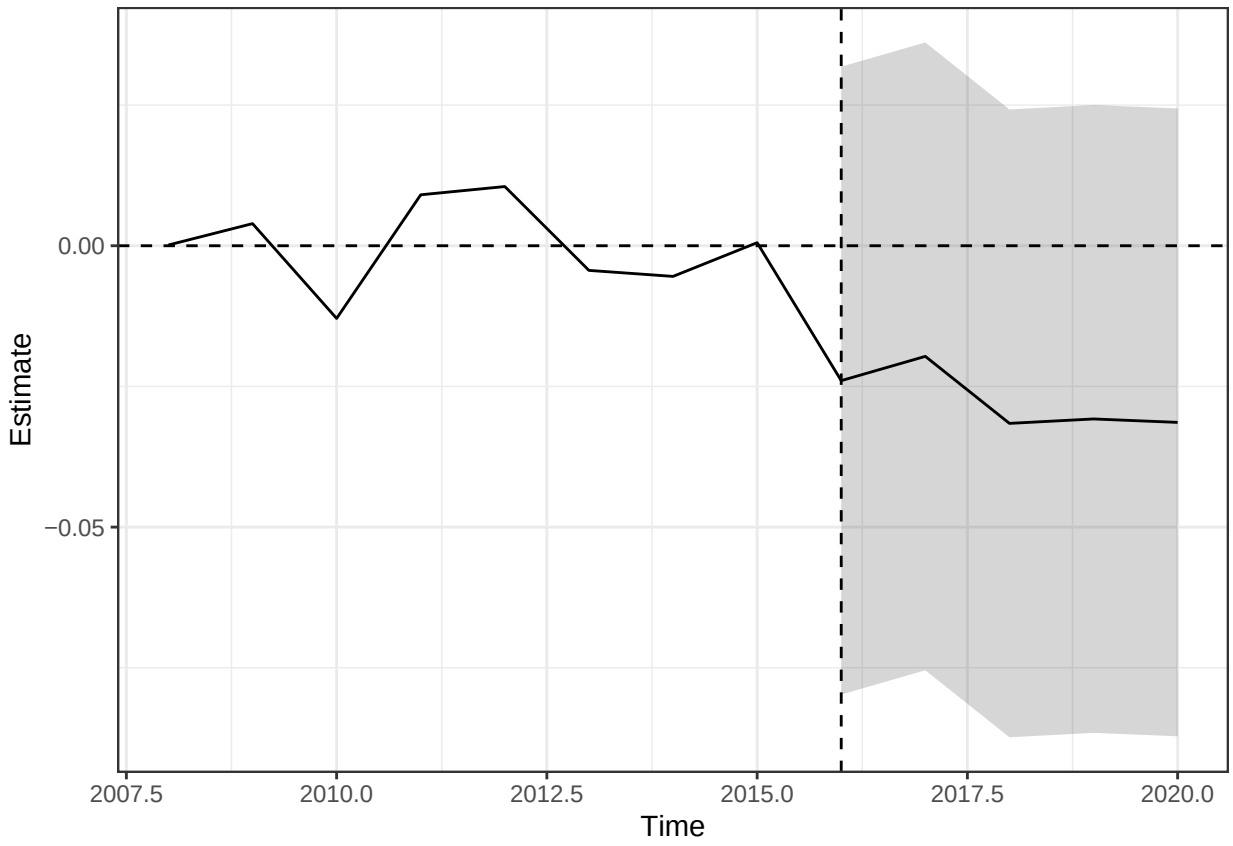
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
nonaug_sum<-summary(nonaugsynth)
nonaug_sum
```

```
##
## 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   ( 0.018 )
## 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   0.110
##  2017   -0.020             -0.075              0.036   0.101
##  2018   -0.032             -0.087              0.024   0.109
##  2019   -0.031             -0.087              0.025   0.113
##  2020   -0.031             -0.087              0.024   0.118
```

```
plot(nonaug_sum)
```



- 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.

Answer: Here the L2 imbalance is .013 and the ATT estimate is -.0301.

```
# augmented synthetic control
ridge_syn <- augsynth(uninsured_rate ~ treated, State, year, synth1_data,
                progfunc = "ridge", scm = T)
```

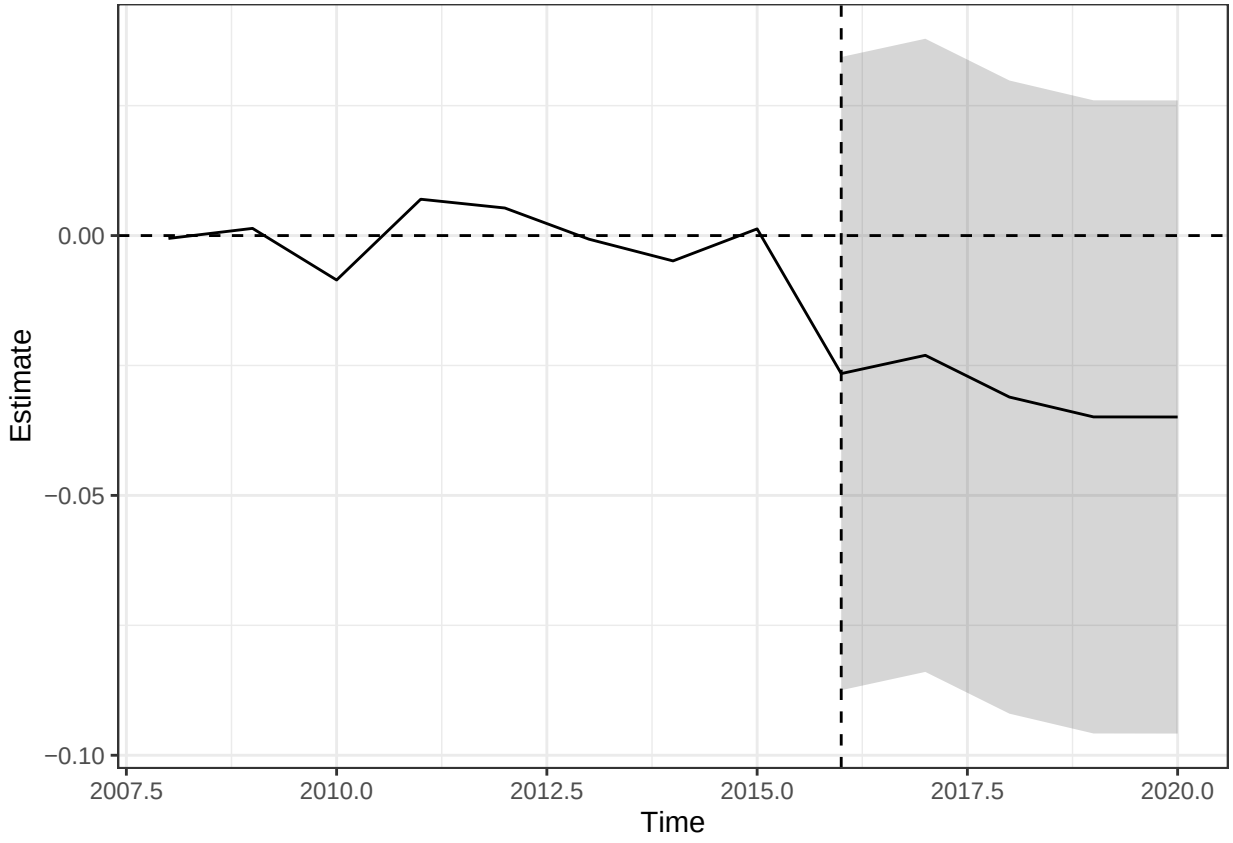
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
ridge_syn_sum<-summary(ridge_syn)
ridge_syn_sum
```

```
##
## 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   ( 0.58 )
## 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   0.121
##  2017   -0.023             -0.084              0.038   0.123
##  2018   -0.031             -0.092              0.030   0.121
##  2019   -0.035             -0.096              0.026   0.121
##  2020   -0.035             -0.096              0.026   0.106
```

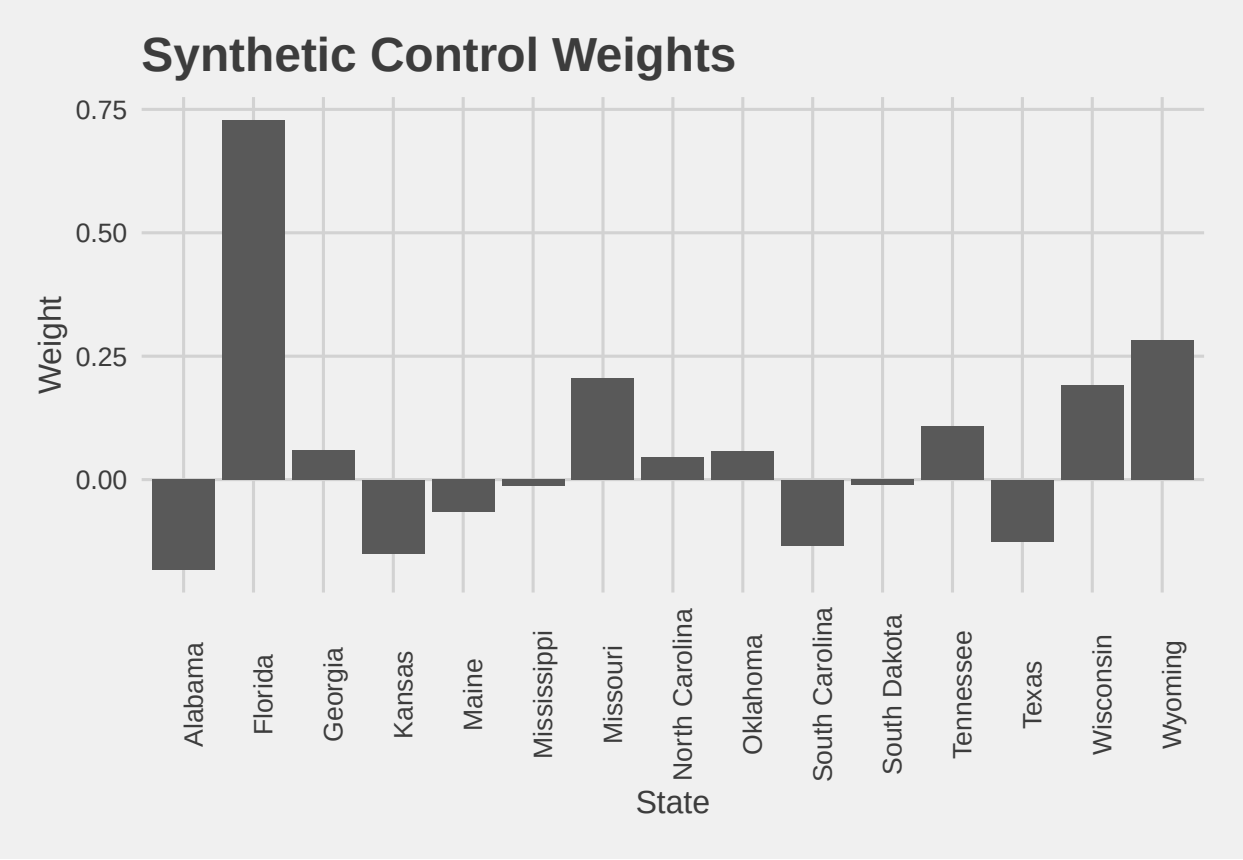
```
plot(ridge_syn_sum)
```



- Plot barplots to visualize the weights of the donors.

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

```
      axis.text.x = element_text(angle = 90)) +
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?

Answer: To make units comparable, it is best practice to rescale each individual time-series to ensure comparability across the aggregate of time-series in selecting weights for donor states, normally based on mean/variance of the treated units. Failing to do so can lead to incorrect weighting.

Discussion Questions

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

Answer: Synthetic control can allow a researcher to conduct research where parallel trends may not necessarily hold. Rather than hand wave that some PTA chart shows what the researcher wants, the SC admits that we can't guarantee pretrends and adds weights to make the study happen. At the same time, SC can be far more challenging to interpret. With several treated units, SC also constructs several different individual controls for each treated unit. This only adds to the complexity and the discomfort people have in interpreting these methods as causal. We conjure synthetic Delawares or North Carolina out of thin air; this is the hubris of the empirical researcher, bending the curve of what we observe in pursuit of facilitating an analysis. Is this folly? It depends on whom you ask and their discipline.

- 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: I have taken this class and I struggle with the concept of the negative weight and how a control unit may "negatively" factor into the estimate; folks with far less training on these methods specifically will be even more confused. As such, this feature is likely to make interpretation more difficult and believability even more farfetched; this step strays further from the real world. That being said, augmented synthetic control does improve the ability to match treated units to the control units, especially with limited control units. Ultimately, I would say it probably makes sense to not do this unless you are absolutely forced to - you then just have to live with the consequences of a reviewer not fully getting it.

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
#probably should kick dc - sorry lil buddy (you do deserve statehood)
data_no_dc<-medicaid_expansion%>%filter(State!="District of Columbia")

#need treatment indicator whereby it's 1 if the year of expansion and 0 if not
data_no_dc<-data_no_dc%>%
  mutate(year_expanded = str_extract(Date_Adopted, "\\d{4}"))%>% #takes the year since it's the first f
  mutate(year_expanded=ifelse(is.na(year_expanded), 9999, year_expanded)) %>%
  mutate(year_expanded=as.integer(year_expanded), treated=ifelse(year>=year_expanded, 1, 0))

ppool_syn <- multisynth(uninsured_rate ~ treated, State, year,
                        nu = 0.5, data_no_dc, n_leads = 5)

ppool_syn_summ <- summary(ppool_syn)
ppool_syn_summ
```

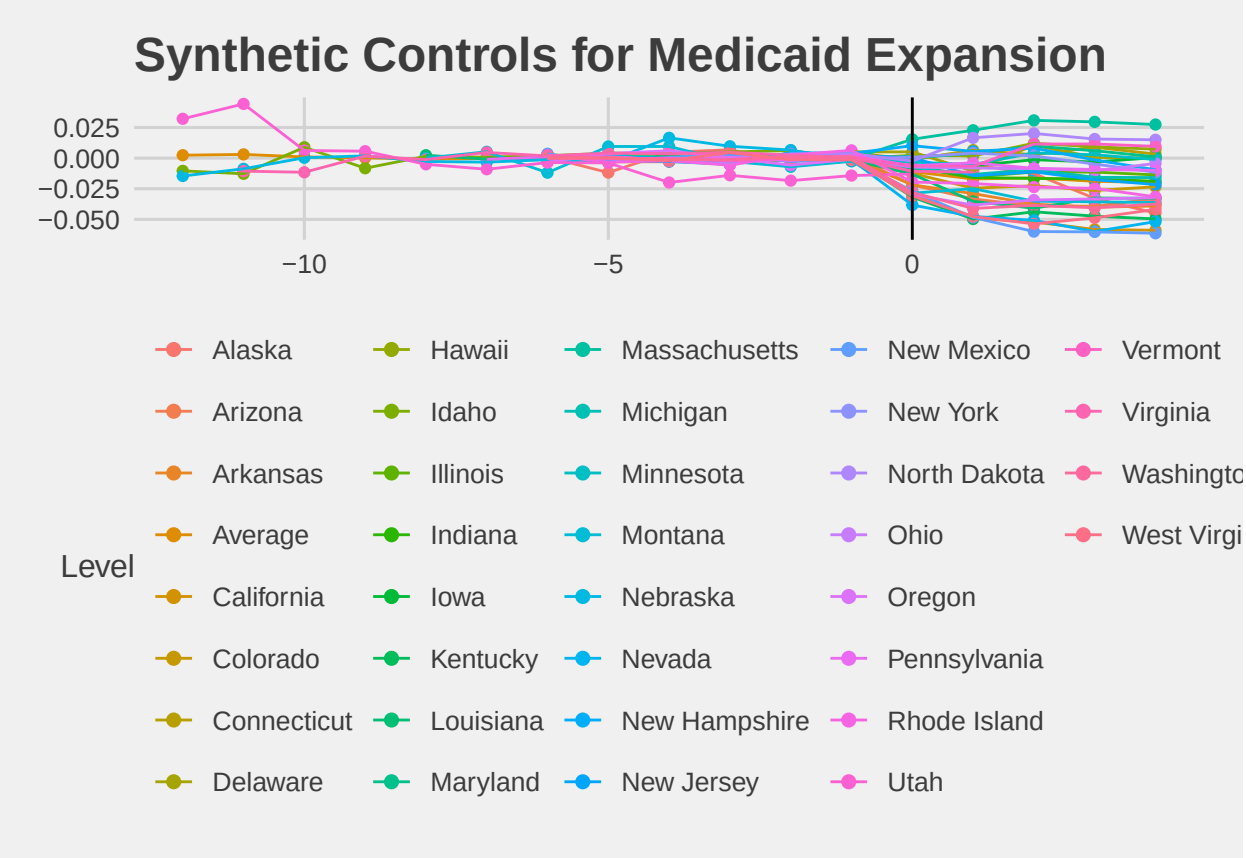
```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
##     data = data_no_dc, n_leads = 5, nu = 0.5)
##
## Average ATT Estimate (Std. Error): -0.015  (0.005)
##
## Global L2 Imbalance: 0.000
## Scaled Global L2 Imbalance: 0.009
## Percent improvement from uniform global weights: 99.1
##
## Individual L2 Imbalance: 0.004
## Scaled Individual L2 Imbalance: 0.098
## Percent improvement from uniform individual weights: 90.2
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01101702 0.004283977 -0.01941738 -0.002424649
##                     1 Average -0.01703809 0.006226894 -0.02903123 -0.005286680
##                     2 Average -0.01611506 0.006810711 -0.02952920 -0.002685157
```

```
##                     3 Average -0.01892813 0.006995493 -0.03287521 -0.005642859
##                     4 Average -0.01990362 0.006727735 -0.03319316 -0.006694685
```

```
ppool_syn_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme_fivethirtyeight() +
  ggtitle('Synthetic Controls for Medicaid Expansion') +
  xlab('Time') +
  ylab('Uninsured Rate Estimate') +
  guides(color = guide_legend(override.aes = list(size = 2)))
```

```
## Warning: Removed 231 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 231 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

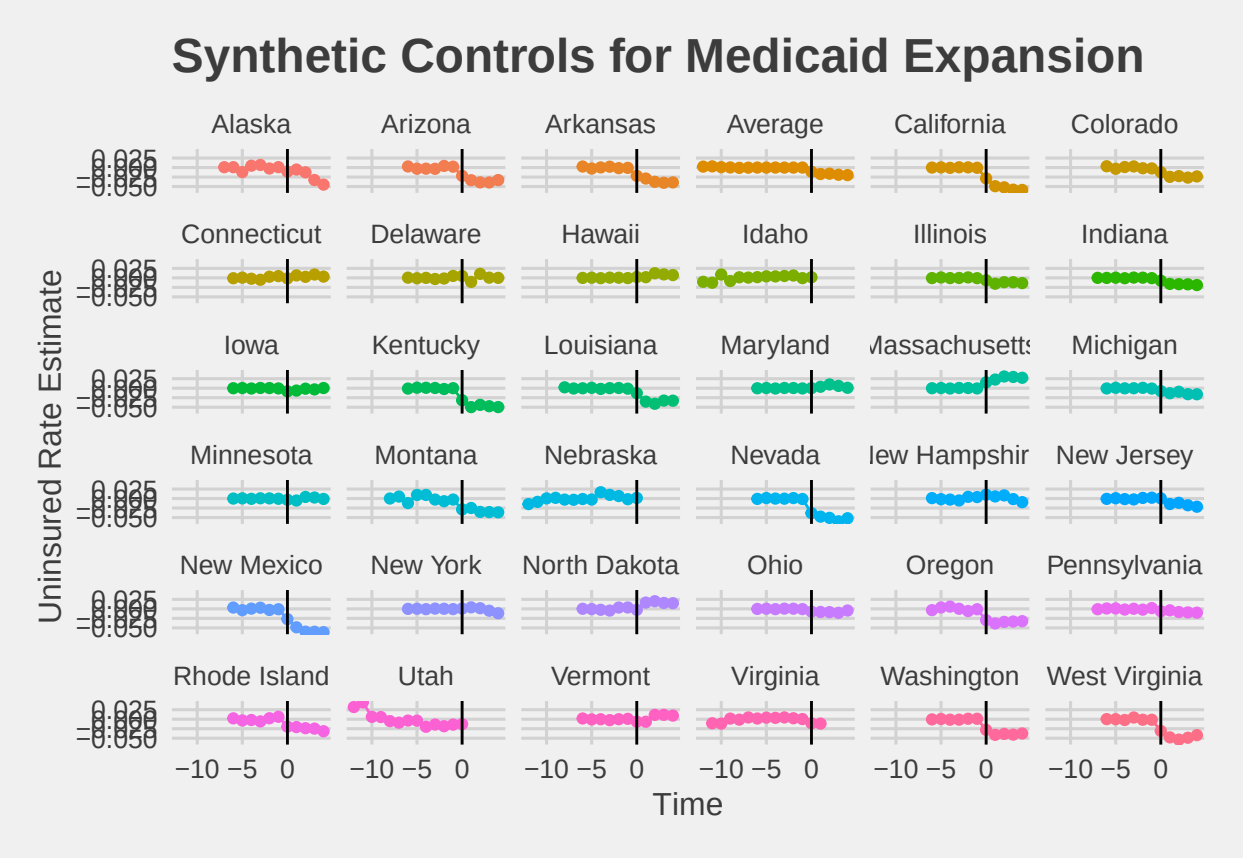



```
ppool_syn_summ$att %>%
ggplot(aes(x = Time, y = Estimate, color = Level)) +
geom_point() +
geom_line() +
geom_vline(xintercept = 0) +
theme_fivethirtyeight() +
theme(axis.title = element_text(),
legend.position = 'None') +
```

```
ggtitle('Synthetic Controls for Medicaid Expansion') +
xlab('Time') +
ylab('Uninsured Rate Estimate') +
facet_wrap(~Level)
```

```
## Warning: Removed 231 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 231 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
ppool_syn2 <- multisynth(uninsured_rate ~ treated, State, year,
                         nu = 0.5, data_no_dc, n_leads = 5, time_cohort = TRUE)

ppool_syn2_summ <- summary(ppool_syn2)
ppool_syn2_summ
```

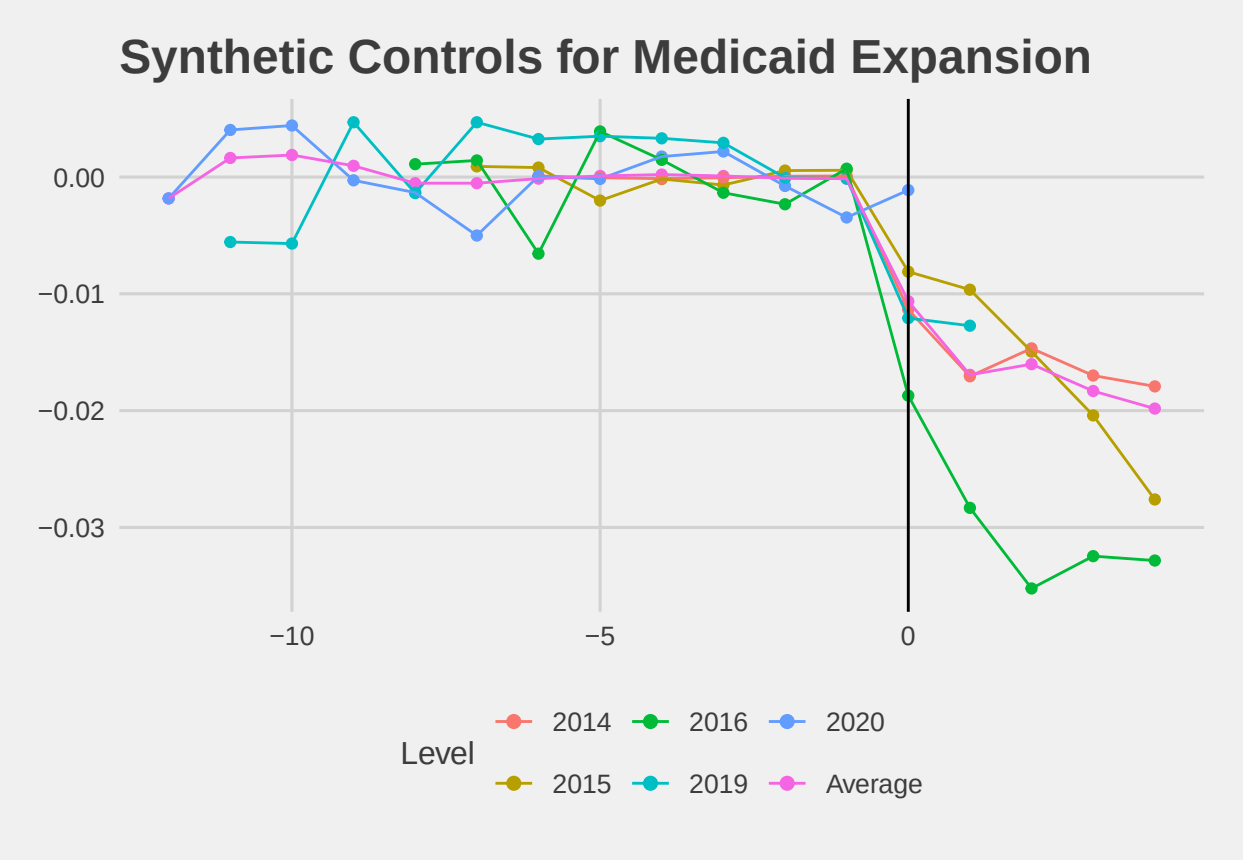
```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
##     data = data_no_dc, n_leads = 5, nu = 0.5, time_cohort = TRUE)
##
## Average ATT Estimate (Std. Error): -0.015  (0.006)
```

```
##
## 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: 0.016
## Percent improvement from uniform individual weights: 98.4
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01063406 0.004573462 -0.01960816 -0.001455500
##                     1 Average -0.01693489 0.006069817 -0.02819005 -0.005132574
##                     2 Average -0.01602318 0.006323102 -0.02769918 -0.003389289
##                     3 Average -0.01832277 0.006459247 -0.02999520 -0.005231273
##                     4 Average -0.01982241 0.006193285 -0.03181242 -0.007197888
```

```
ppool_syn2_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0) +
  theme_fivethirtyeight() +
  ggtitle('Synthetic Controls for Medicaid Expansion') +
  xlab('Time') +
  ylab('Uninsured Rate Estimate') +
  guides(color = guide_legend(override.aes = list(size = 2)))
```

```
## 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, you can see differences here when reviewing the graphs above - it's a potential concern, too, with just relying on the time-based cohorts, right? That analysis aggregates by time cohort, but ostensibly states in the same time cohort may not implement the law in the same way, yielding variance in effect that gets aggregated up. This is a reason to prefer the state-based analyses.

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

Answer: Yes, the effects appear to increase with the law being in effect for more years. This is sensible because states may capture some immediate folks who want to get insured after expansion, but there will be a subset of folks who are not covered and do not know they could get covered. States who adopt earlier will be able to reach the latter group more readily than those expanding later.

General Discussion Questions

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

Answer: One of the biggest issues with longitudinal studies at the individual level is loss to follow up. DiD and synthetic control obviously require extended years of follow up such that individuals are unlikely to have the full panel of data required to facilitate these estimations. Data on aggregate units is more widely available and complete than individuals or smaller units like towns, etc. Also, the types of interventions DiD and synthetic controls are used to analyze normally are spread out across these aggregate units vs targeting specific individuals.

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

Answer: DiD/synthetic control doesn't really care about how individual units are selected into treatment so long as PTA holds (and this doesn't actually have to hold necessarily with synthetic control). Regression discontinuity does present a concern when selection into treatment is not random around the cutoff; for example, if the running variable is manipulable by individuals around a specific cutoff, then treatment is no longer random and the discontinuity is biased. We observe this in the union elections literature where concerns about the 50% cutoff threaten the validity of several studies, as well as health services research - for example, concerns about RDD used to study 340B eligibility and its impacts.

If the treatment of interest is across a continuous variable with a known cutoff, RDD makes sense. If treatment is present in some units and not others and you have longitudinal data, consider DiD; if PTA does not necessarily hold but you need to do your study, leverage SC.