Project 7: Difference-in-Differences and Synthetic Control

Sofia Guo with assistance from Stacy Chen

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

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

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

# set seed
set.seed(44)

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

```
## 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:

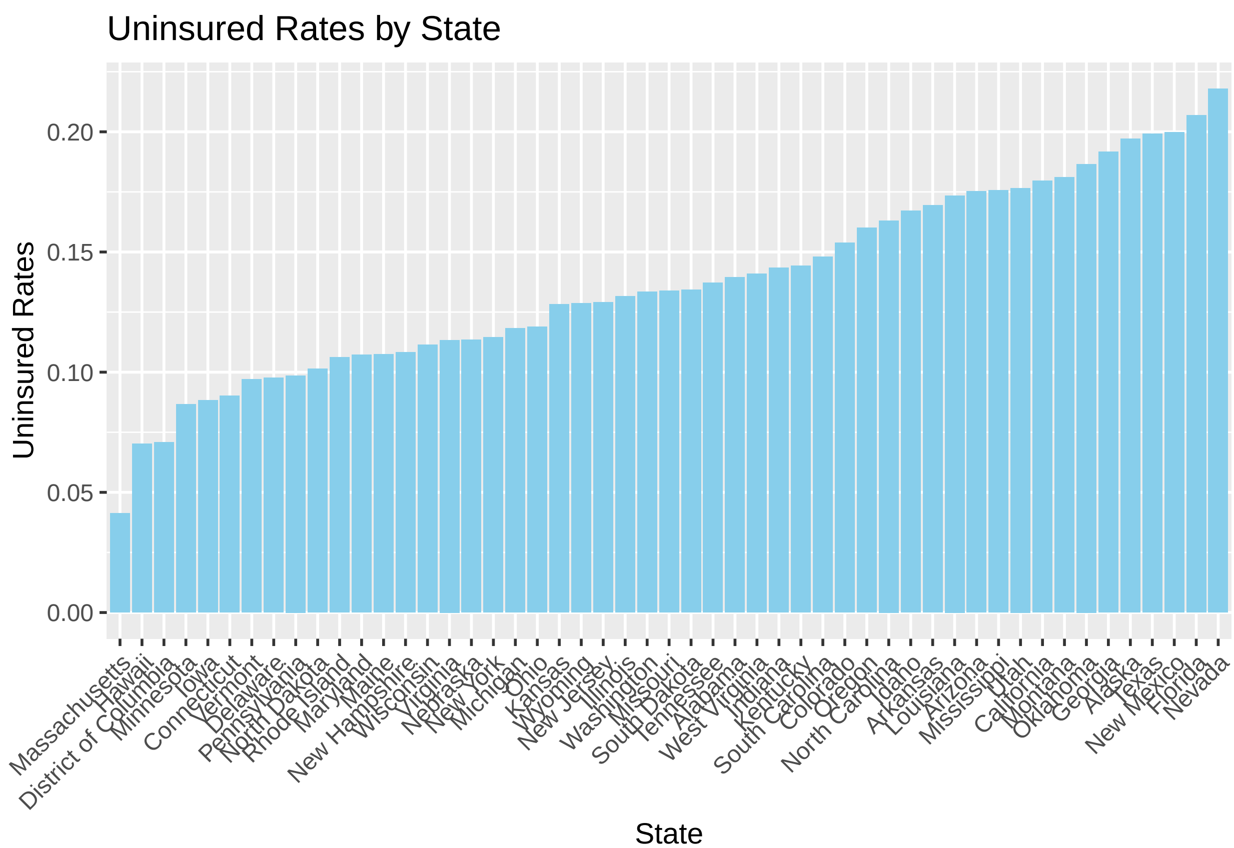
- Which states had the highest uninsured rates prior to 2014? The lowest?
  - **Answer**: Some states who had the highest uninsured rates prior to 2014 were Nevada, Texas, Utah, New Mexico, Alaska, Florida, Georgia, Montana. States who had the lowest uninsured rates prior to 2014 were Connecticut, Delaware, DC, Hawaii, and Massachusetts (below the grand mean).
- Which states were home to most uninsured Americans prior to 2014? How about in the last year in the data set?
  - **Answer**: Prior to 2014, California, Florida, and Texas were home to the most uninsured Americans. In the last year of the dataset, the same three states were also home to the most uninsured Americans, although Texas superceded California in 2020. **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
df_avg <- medicaid_expansion %>%
  filter(year >= 2008 & year <= 2013) %>%
  group_by(State) %>%
  summarise(avg_uninsured_rate = mean(uninsured_rate, na.rm = TRUE))

# highest and lowest uninsured rates
df_sorted <- df_avg[order(df_avg$avg_uninsured_rate), ]

# Create a bar plot using ggplot
```

```r
ggplot(df_sorted, aes(x = reorder(State, avg_uninsured_rate), y = avg_uninsured_rate)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Uninsured Rates by State",
       x = "State",
       y = "Uninsured Rates") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

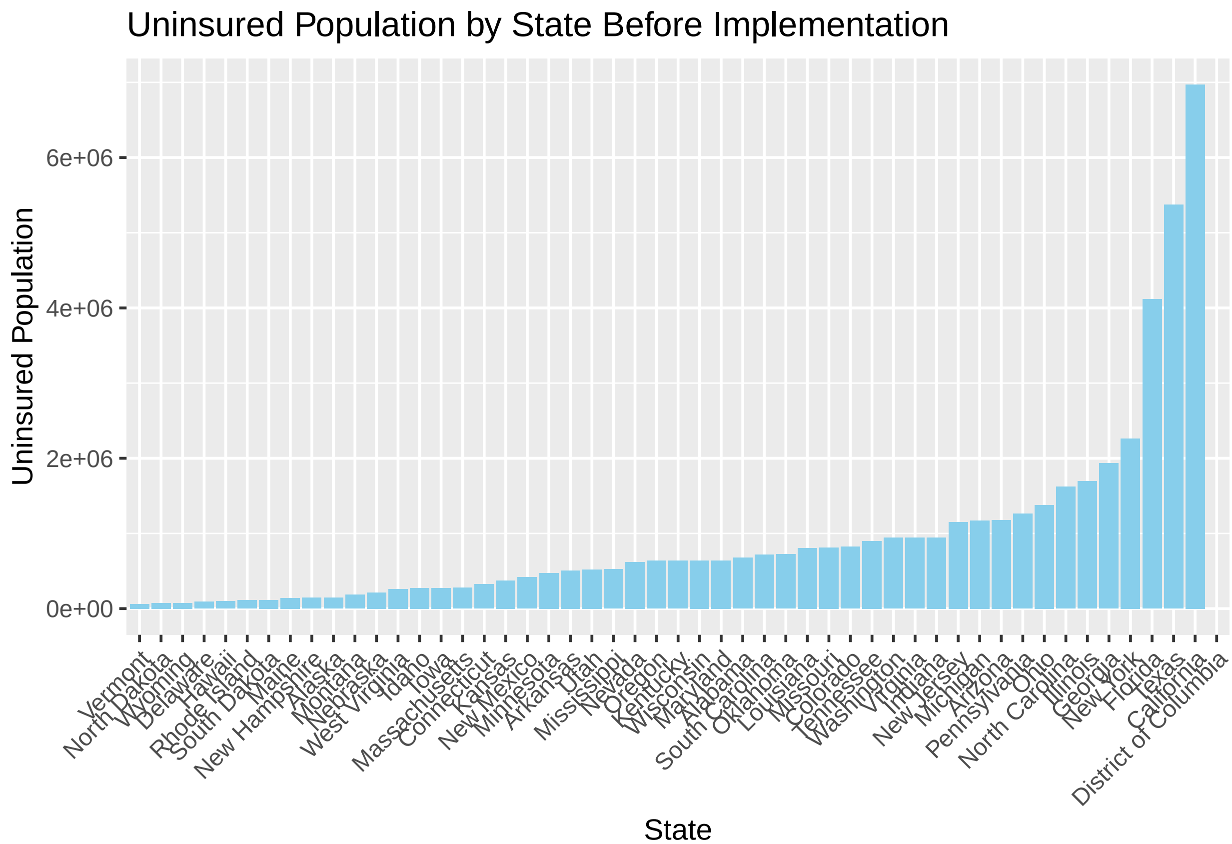



```r
# most uninsured Americans
df_avg <- medicaid_expansion %>%
  filter(year >= 2008 & year <= 2013) %>%
  mutate(uninsured_pop = uninsured_rate * population) %>%
  group_by(State) %>%
  summarise(avg_uninsured_pop = mean(uninsured_pop, na.rm = TRUE))

# highest and lowest uninsured rates
df_sorted <- df_avg[order(df_avg$avg_uninsured_pop), ]

# Create a bar plot using ggplot
ggplot(df_sorted, aes(x = reorder(State, avg_uninsured_pop), y = avg_uninsured_pop)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Uninsured Population by State Before Implementation",
       x = "State",
       y = "Uninsured Population") +
  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()`).
```

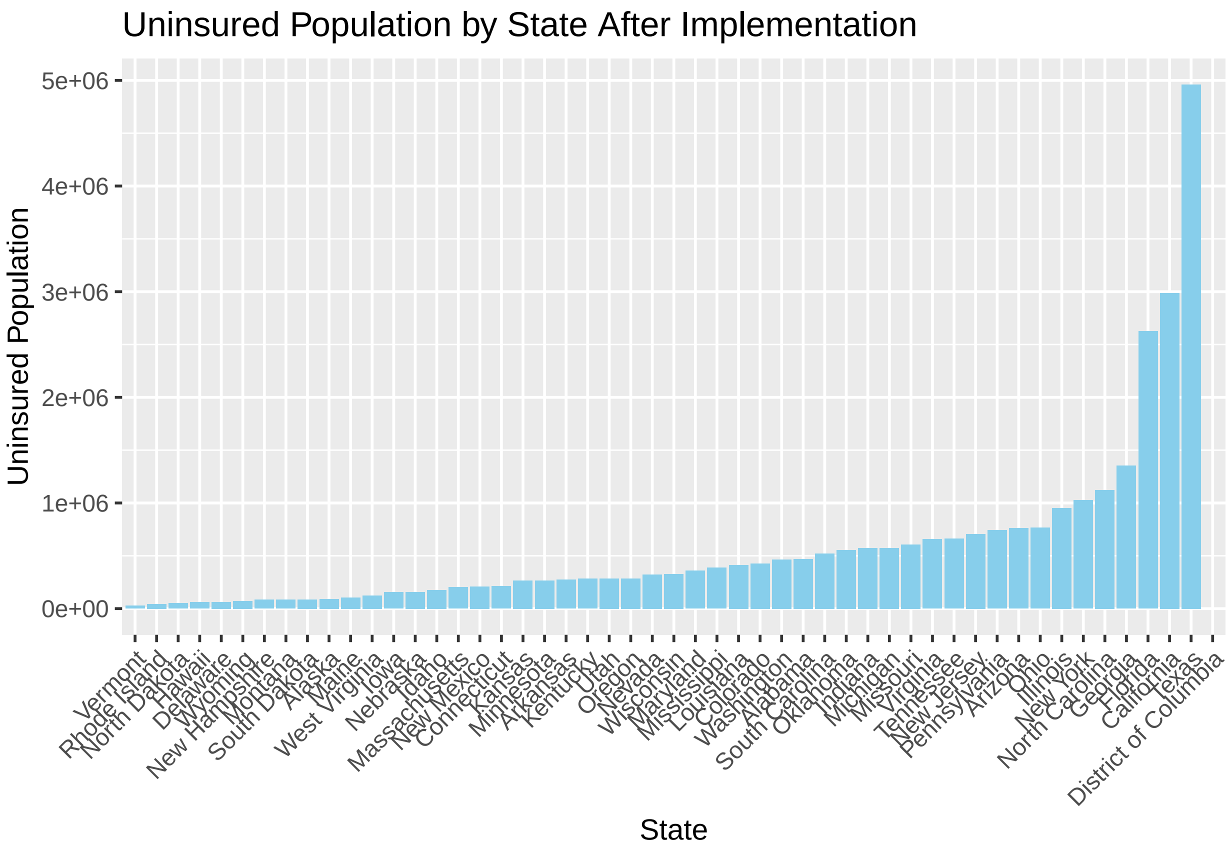



```r
df_avg <- medicaid_expansion %>%
  filter(year == 2020) %>%
  mutate(uninsured_pop = uninsured_rate * population) %>%
  group_by(State) %>%
  summarise(avg_uninsured_pop = mean(uninsured_pop, na.rm = TRUE))

# highest and lowest uninsured rates
df_sorted <- df_avg[order(df_avg$avg_uninsured_pop), ]

# Create a bar plot using ggplot
ggplot(df_sorted, aes(x = reorder(State, avg_uninsured_pop), y = avg_uninsured_pop)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Uninsured Population by State After Implementation",
       x = "State",
       y = "Uninsured Population") +
  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:

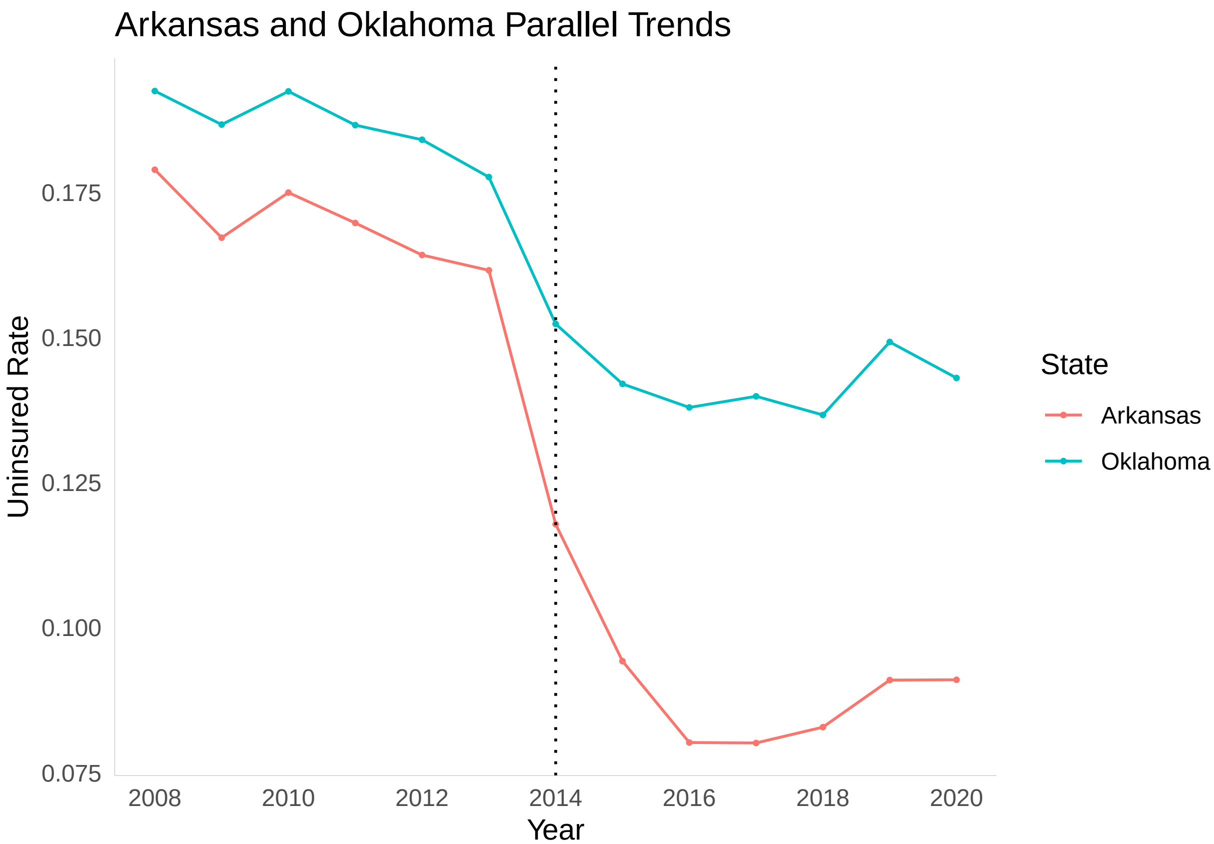
- Choose a state that adopted the Medicaid expansion on January 1, 2014 and a state that did not. **Hint**: Do not pick Massachusetts as it passed a universal healthcare law in 2006, and also avoid picking a state that adopted the Medicaid expansion between 2014 and 2015.
- Assess the parallel trends 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). I tried CA/FL, ND/SD, LA/MS (which had over laying trends but they weren't identical enough). None of these combinations had convincing parallel trends.

```
#find states that did not adopt between 2014 and 2015
qualified_states <- subset(medicaid_expansion, Date_Adopted == "2014-01-01" | Date_Adopted >= "2015-01-0
```

```
# Parallel Trends plot
# filter out Massachusetts and states that adopted prior to 2014 and after 2015
# ---------------------------------------------------------------
medicaid_expansion %>%
  # process
  # ---------
  filter(State %in% c("Arkansas","Oklahoma")) %>%
  # plotting all of the time periods -- not filtering out any of them
  # plot
  # ---------
  ggplot() +
```

```r
# add in point layer
geom_point(aes(x = year,
               y = uninsured_rate,
               color = State), size = 0.5) +
# add in line layer
geom_line(aes(x = year,
              y = uninsured_rate,
              color = State)) +
# add a horizontal line
geom_vline(aes(xintercept = 2014), linetype = "dotted") +
# themes
theme_minimal() +
theme(plot.caption = element_text(hjust = 0),
      legend.position = "right",
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      axis.line = element_line(colour = "grey", size = .1),
      panel.spacing = unit(3, "lines"))+
scale_x_continuous(breaks = seq(2008,2020,2)) +
# labels
ggtitle('Arkansas and Oklahoma Parallel Trends') +
xlab('Year') +
ylab('Uninsured Rate')
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



- 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
medicaid_expansion_states <-
  medicaid_expansion %>%
  mutate(treatment = case_when(State == "Arkansas" & year >= 2014 ~ 1, TRUE ~ 0)) %>%
  filter(State %in% c("Arkansas", "Oklahoma"))

medicaid_expansion <-
  medicaid_expansion %>%
  mutate(treatment = case_when(State == "Arkansas" & year >= 2014 ~ 1, TRUE ~ 0))

# multisynth model states
multi <- medicaid_expansion %>%
    filter(State != "District of Columbia") %>%
    mutate(treated = ifelse(year >= year(Date_Adopted), 1, 0))

                # pre-treatment difference
pre_diff <- medicaid_expansion %>%
  # filter out only the quarter we want
  filter(year == 2013) %>%
  # subset to select only vars we want
  select(State,
         uninsured_rate) %>%
  # make the data wide
```

```r
  pivot_wider(names_from = State,
              values_from = uninsured_rate) %>%
  # subtract to make calculation
  summarise(Oklahoma - Arkansas)

# post-treatment difference
# ----------
post_diff <-
  medicaid_expansion %>%
  # filter out only the quarter we want
  filter(year == 2015) %>%
  # subset to select only vars we want
  select(State,
         uninsured_rate) %>%
  # make the data wide
  pivot_wider(names_from = State,
              values_from = uninsured_rate) %>%
  # subtract to make calculation
  summarise(Oklahoma - Arkansas)

# diff-in-diffs
# ----------
diff_in_diffs <- post_diff - pre_diff
diff_in_diffs
```

```
##   Oklahoma - Arkansas
## 1              0.0317
```

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**: These data encompass differences in implementation of the same federal policy that allows states to make it more/less difficult to enroll, so the estimated effects might not be accounting for differences that are not explained by being in similar situations/proximity. Furthermore, state-level regulations for fast food labor versus medicaid eligibility criteria are likely quite different in their variance (due to the variability in population demographics, state resources, political influences), so Card/Krueger's assumptions are more difficult to uphold in this policy context.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: The strength of using the parallel trends assumption is that it does not require controlling for every measurable confounder, only that there is a constant difference in the observed outcome between the entities; a weakness of the diff-in-diff approach is that it assumes the counterfactual of the treated unit would have maintainted that constant distance. However, it is impossible to know whether that assumption is actually true since the counterfactual is never observed.

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.

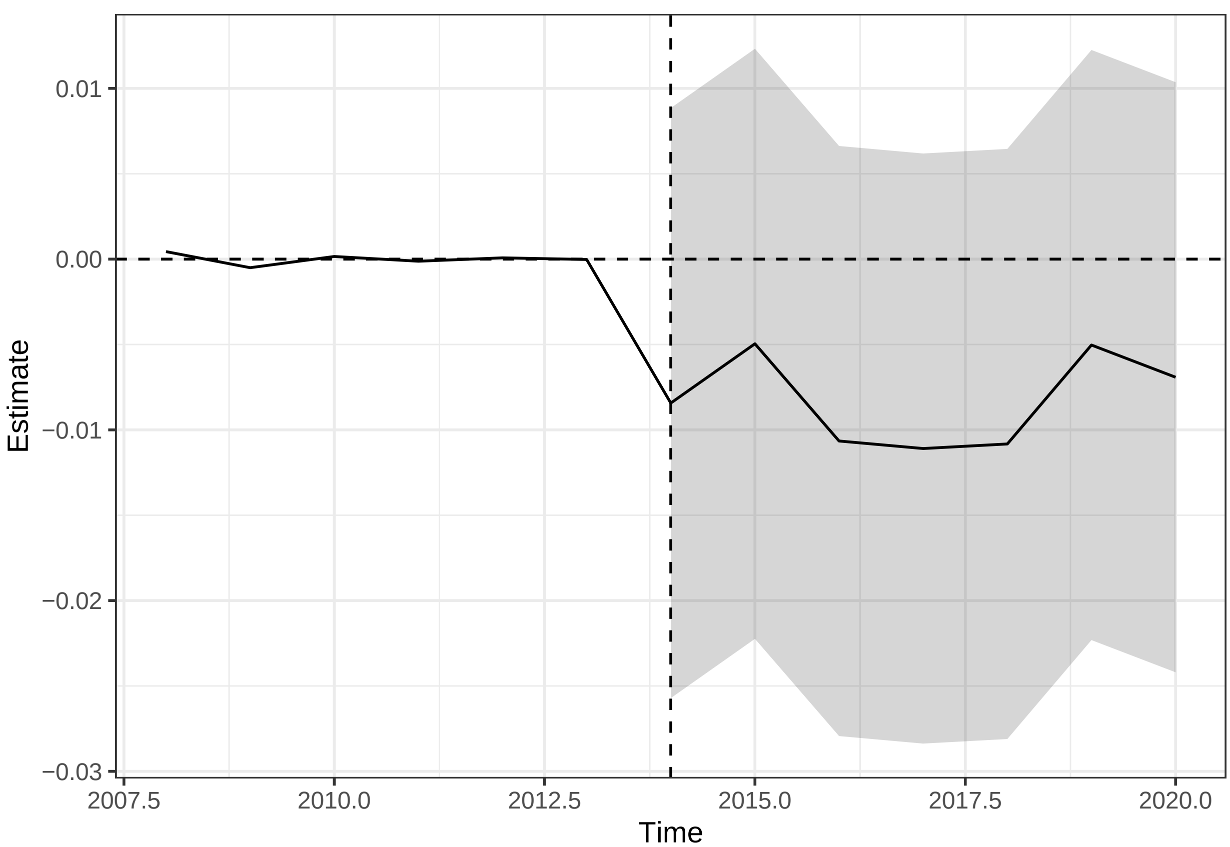
Do the following:

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

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

```
# Plot results
plot(syn)
```




```
#Also report the average ATT and L2 imbalance
summary_stats <- summary(syn)
summary_stats
```

```
##
## 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):  -0.00827   ( 0.06 )
## L2 Imbalance: 0.001
## Percent improvement from uniform weights: 99.1%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2014   -0.008             -0.026              0.009   0.154
##  2015   -0.005             -0.022              0.012   1.000
##  2016   -0.011             -0.028              0.007   0.860
##  2017   -0.011             -0.028              0.006   0.844
##  2018   -0.011             -0.028              0.006   1.000
##  2019   -0.005             -0.022              0.012   1.000
##  2020   -0.007             -0.024              0.010   1.000
```

```
# Average ATT Estimate (p Value for Joint Null):  -0.00827   ( 0.068 )
# L2 Imbalance: 0.001
```

- 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
# recalculate with Ridge function that penalizes really high weights
# ----------------------------------------------------------------
ridge_syn <-
  augsynth(uninsured_rate ~ treatment,
                         State,     # unit
                         year,  # time
                         medicaid_expansion,
          progfunc = "ridge",  # specify
          scm = T)
```

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

```
summary(ridge_syn)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "ridge", scm = ..2)
##
## Average ATT Estimate (p Value for Joint Null):  -0.00828   ( 0.052 )
## L2 Imbalance: 0.001
## Percent improvement from uniform weights: 99.1%
##
## Avg Estimated Bias: 0.000
##
## Inference type: Conformal inference
##
```

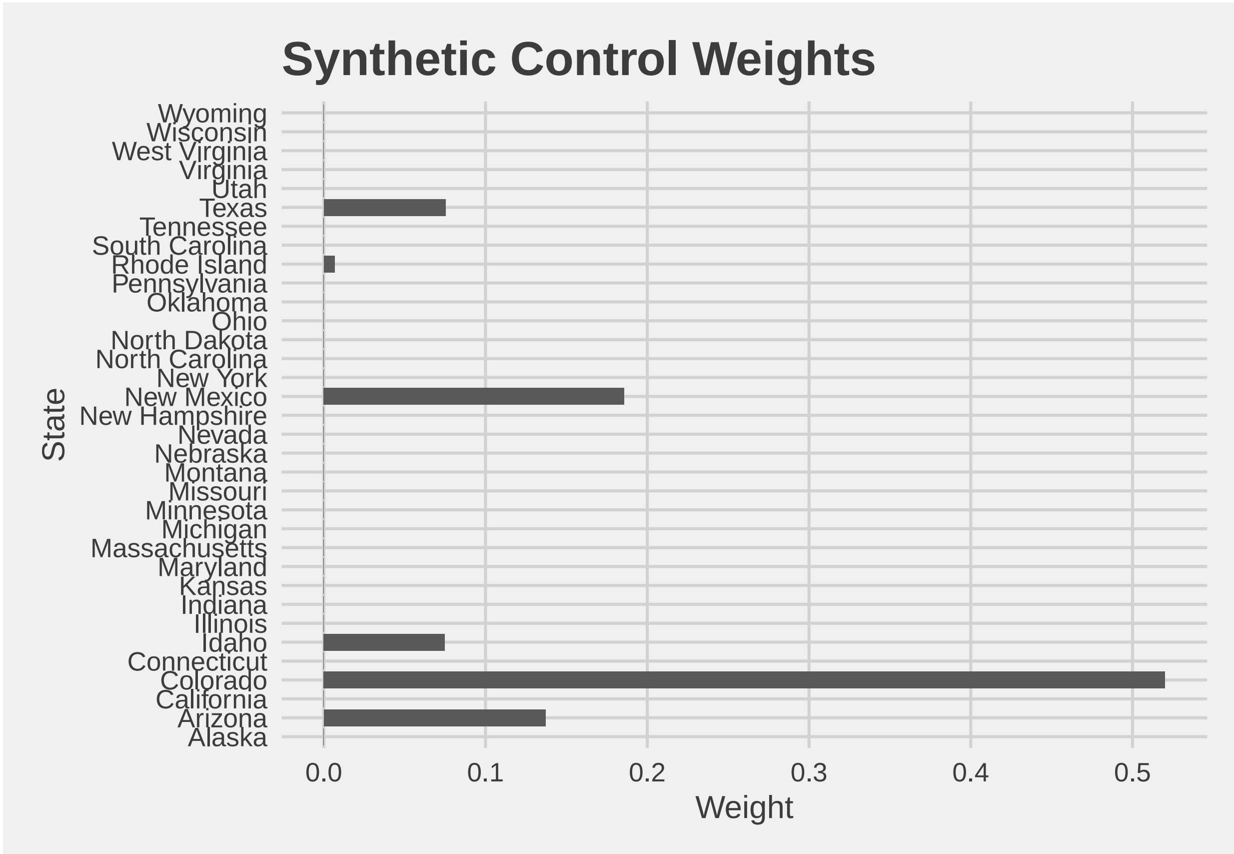
```
##   Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##   2014   -0.008             -0.026              0.009   0.133
##   2015   -0.005             -0.022              0.012   1.000
##   2016   -0.011             -0.028              0.007   0.846
##   2017   -0.011             -0.028              0.006   0.846
##   2018   -0.011             -0.028              0.006   1.000
##   2019   -0.005             -0.022              0.012   1.000
##   2020   -0.007             -0.024              0.010   1.000
```

```r
# Average ATT Estimate (p Value for Joint Null):  -0.00828   ( 0.052 )
# L2 Imbalance: 0.001
```

- Plot barplots to visualize the weights of the donors.

```r
# barplots of weights

data.frame(syn$weights) %>% # coerce to data frame since it's in vector form
  # process
  # ---------
  # change index to a column
  tibble::rownames_to_column('State') %>% # move index from row to column (similar to index in row as i
  filter(syn.weights > 0) %>% # filter out weights less than 0
  # ---------
  ggplot() +
  # stat = identity to take the literal value instead of a count for geom_bar()
  geom_bar(aes(x = State,
               y = syn.weights),
           stat = 'identity') +  # override count() which is default of geom_bar(), could use geom_col(
  coord_flip() +   # flip to make it more readable
  # themes
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  # labels
  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?

Discussion Questions

- What are the advantages and disadvantages of synthetic control compared to difference-in-differences estimators?
- **Answer**: Synthetic control is helpful because it combines the weights of multiple control units instead of relying on a single control unit (therefore accounting for heterogeneity). This approach may be better than DiD in small samples where averages may be more sensitive to which units are controls or not. One disadvantage is that it is more complicated than DiD and thus not as straightforward to interpret, and its results are more model-dependent (prone to model specification errors).
- 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**: It would create interpretation problems because negative weights would mean there are nonzero differences between treatment and control units. Balancing the consideration includes contductin sensitivity tests and thinking critically about the results in context and using qualitative/theoretical knowledge.

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
# ppool_syn <- multisynth(uninsured_rate ~ treated,
#                         State,                         # unit
#                         year,                          # time
#                         nu = 0.5,                      # varying degree of pooling
#                         multi, # data
#                         n_leads = 10)                  # post-treatment periods to estimate
# with default nu
# ---------
ppool_syn <- multisynth(uninsured_rate ~ treated,
                        State,                         # unit
                        year,                          # time
                        multi,  # data
                        n_leads = 10)                  # post-treatment periods to estimate

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

```
## [1] 0.2998142
```

```
ppool_syn
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
##     data = multi, n_leads = 10)
##
## Average ATT Estimate: -0.017
```

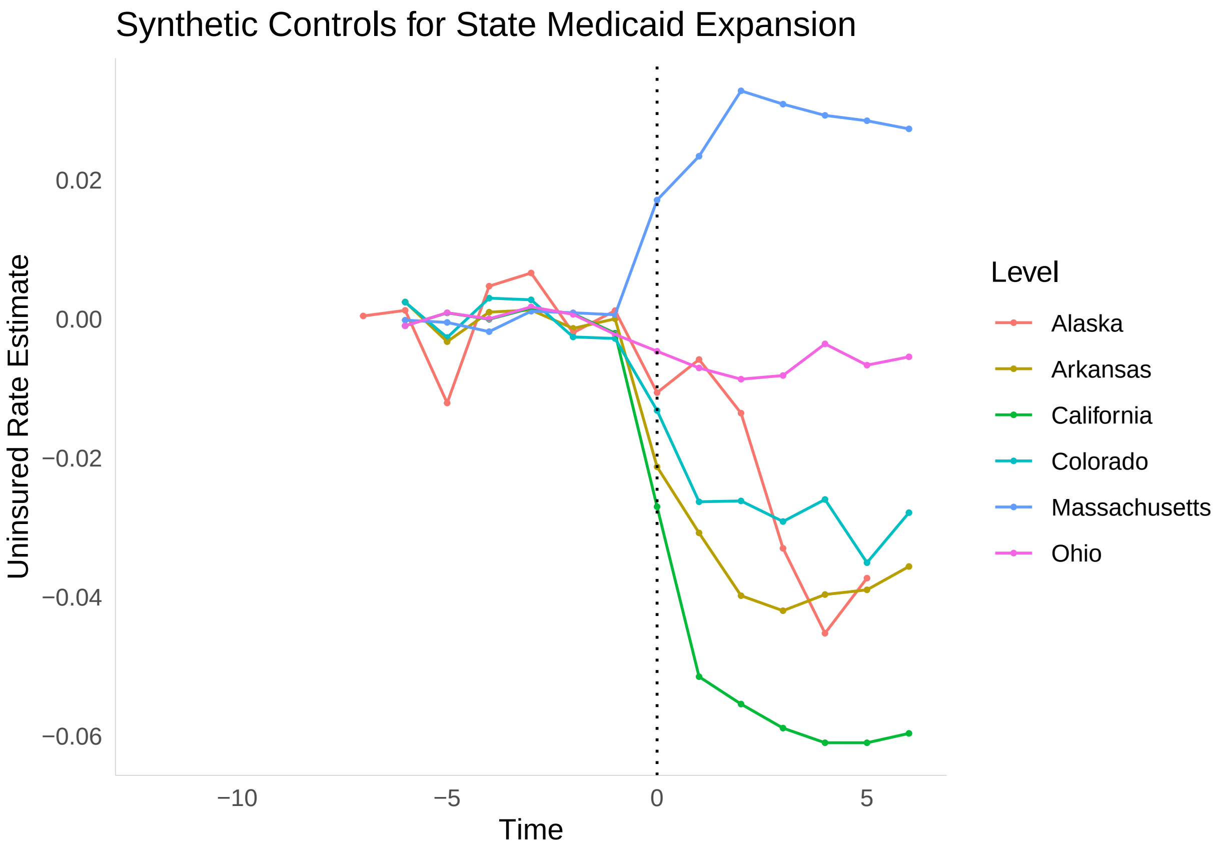
```
ppool_syn_summ <- summary(ppool_syn)
```

```
# Specify small subset
desired_levels <- c("Alaska", "Arkansas","Alabama", "California", "Colorado", "Ohio", "Massachusetts",
# Filter data for the desired states
filtered_data <- ppool_syn_summ$att %>%
  filter(Level %in% desired_levels)
# Plot
filtered_data %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point(size = 0.5) +
  geom_line() +
  geom_vline(xintercept = 0, linetype="dotted") +
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
```

```
        panel.spacing = unit(3, "lines"))+
  ggtitle('Synthetic Controls for State Medicaid Expansion') +
  xlab('Time') +
  ylab('Uninsured Rate Estimate')
```

```
## Warning: Removed 42 rows containing missing values or values outside the scale range
## (`geom_point()`).

## Warning: Removed 42 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

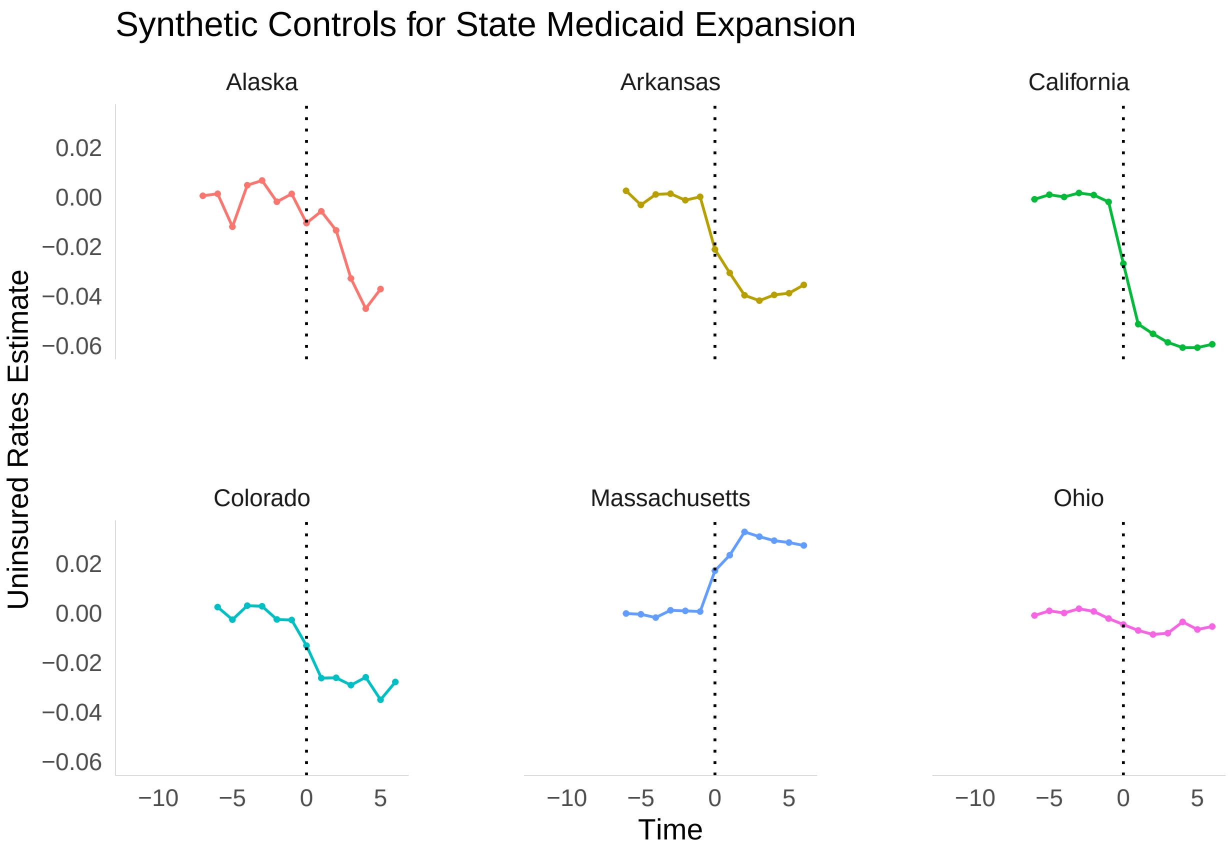



```
# plot actual estimates not values of synthetic controls - use a facet_wrap for readability
# ---------
filtered_data %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point(size = 0.5) +
  geom_line() +
  geom_vline(xintercept = 0, linetype = "dotted") +
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        panel.spacing = unit(3, "lines"),
        legend.position = 'None') +
  ggtitle('Synthetic Controls for State Medicaid Expansion') +
  xlab('Time') +
  ylab('Uninsured Rates Estimate') +
```

```r
  facet_wrap(~Level)   # facet-wrap by level (state in this case) for clearer presentation
```

```
## Warning: Removed 42 rows containing missing values or values outside the scale range
## (`geom_point()`).

## Warning: Removed 42 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.

```r
# multisynth model time cohorts
ppool_syn_time <- multisynth(uninsured_rate ~ treated,
                             State,
                             year,
                             multi,
                             n_leads = 10,
                             time_cohort = TRUE)           # time cohort set to TRUE

# save summary
ppool_syn_time_summ <- summary(ppool_syn_time)

# view
ppool_syn_time_summ
```

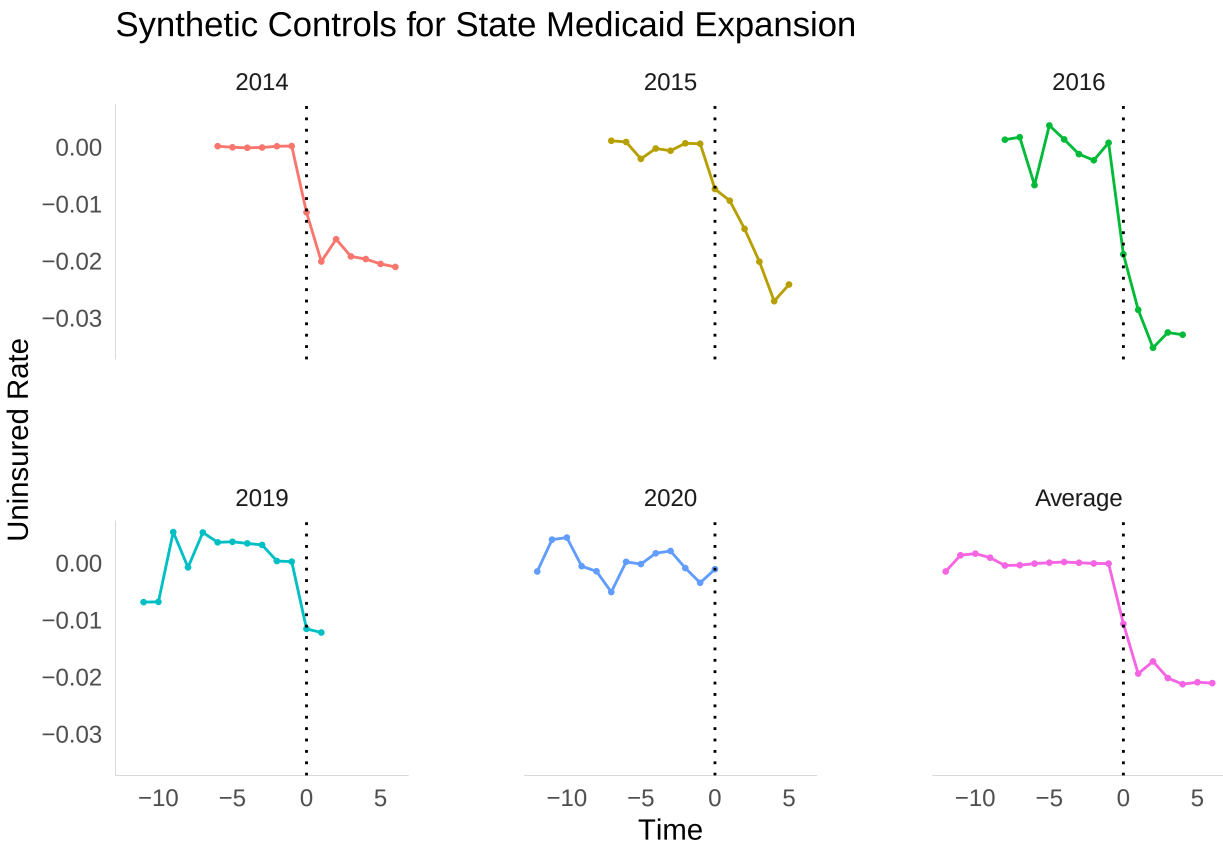
```
##
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year,
```

```
##     data = multi, n_leads = 10, time_cohort = TRUE)
##
## Average ATT Estimate (Std. Error): -0.017  (0.006)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.008
## Percent improvement from uniform global weights: 99.2
##
## Individual L2 Imbalance: 0.005
## Scaled Individual L2 Imbalance: 0.017
## Percent improvement from uniform individual weights: 98.3
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01068255 0.004828774 -0.02104673 -0.001961552
##                     1 Average -0.01942100 0.006029684 -0.03142681 -0.008607062
##                     2 Average -0.01727609 0.005878019 -0.02894542 -0.006475934
##                     3 Average -0.02019425 0.006222753 -0.03282509 -0.009002417
##                     4 Average -0.02127762 0.006103968 -0.03374563 -0.010438341
##                     5 Average -0.02093346 0.005767378 -0.03206622 -0.010318718
##                     6 Average -0.02108744 0.006277225 -0.03346102 -0.009229644
```

```
ppool_syn_time_summ$att %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point(size = 0.5) +
  geom_line() +
  geom_vline(xintercept = 0, linetype = "dotted") +
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        panel.spacing = unit(3, "lines"),
        legend.position = 'None') +
  ggtitle('Synthetic Controls for State Medicaid Expansion') +
  xlab('Time') +
  ylab('Uninsured Rate') +
  facet_wrap(~Level)
```

```
## Warning: Removed 36 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 36 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**: I do see evidence that states have different treatment effect sizes due to differences in implementation. For example, a more liberal state like California seems to have a larger treatment effect (bigger decline) than a more conservative state like Arkansas or Ohio which have smaller declines in the uninsured rates post treatment.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: Multisynth modeling suggests that the 2015 and 2016 adopters saw the largest declines in their uninsured rate (whereas the 2014 cohort was a more modest decline). Meanwhile, late adopters after 2016 seemed to show little declines in their uninsured rate. So it would seem that any benefit from "early" adoption would land in the 2015-16 time, not necessarily the very first time cohort of 2014 (but yes, generally earlier adopters saw greater benefits).

General Discussion Questions

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

- **Answer**: DiD helps control for time varying confounders which are likely to be present when comparing groups such as states/cities/countries, and being able to use parallel trends allows for causal interpretation of what otherwise (without DiD) would be maybe difficult to prove and explain. Large aggregated groups tend to have similarities which lend themselves well to the parallel trends assumption but they also may have enough differences within a large enough sample to choose properly.
- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?
- **Answer**: RD is typically used when there is substantial disaggregated (e.g. individual test scores, health measures) cross sectional data while synthetic control is used for comparing aggregated units over time. In terms of selection into treatment, RD is able to use an arbitrary cutoff to rule out selection (since it may be difficult to really differentiate between individuals just above or just below) while synthetic control does not (parallel trends assumption does not consider whether the selected units are appropriated matched based on how biased or more/less likely to take the treatment).