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")
```

```
## Using GitHub PAT from the git credential store.
```

```
## Skipping install of 'augsynth' from a github remote, the SHA1 (982f650b) 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('./data/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.

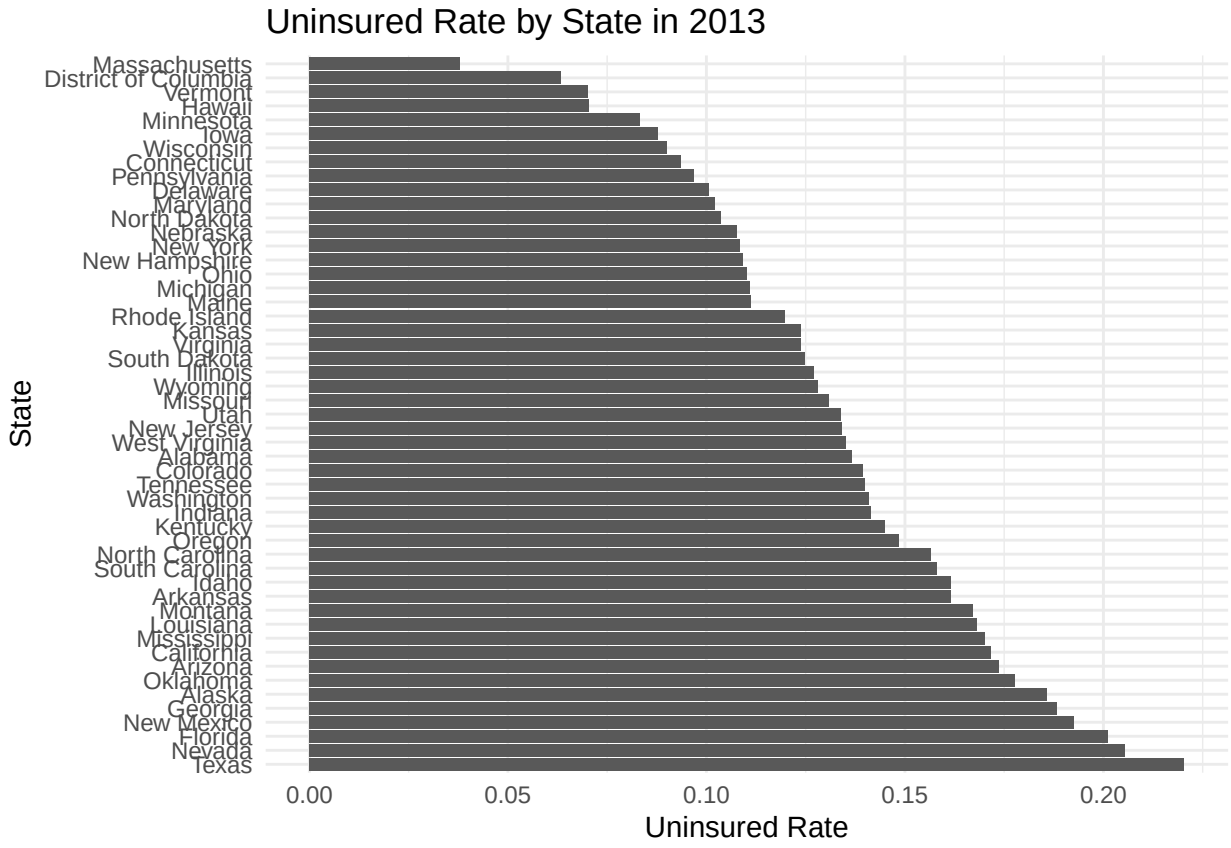
Exploratory Data Analysis

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?
- 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.

```
# highest and lowest uninsured rates

#Filter to 2013
df_2013 <- medicaid_expansion %>%
  filter(year == 2013) %>%
  select(State, uninsured_rate)

df_2013 <- df_2013 %>%
  arrange(desc(uninsured_rate))

# plot
ggplot(df_2013, aes(x = reorder(State, -uninsured_rate), y = uninsured_rate)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Uninsured Rate by State in 2013",
```

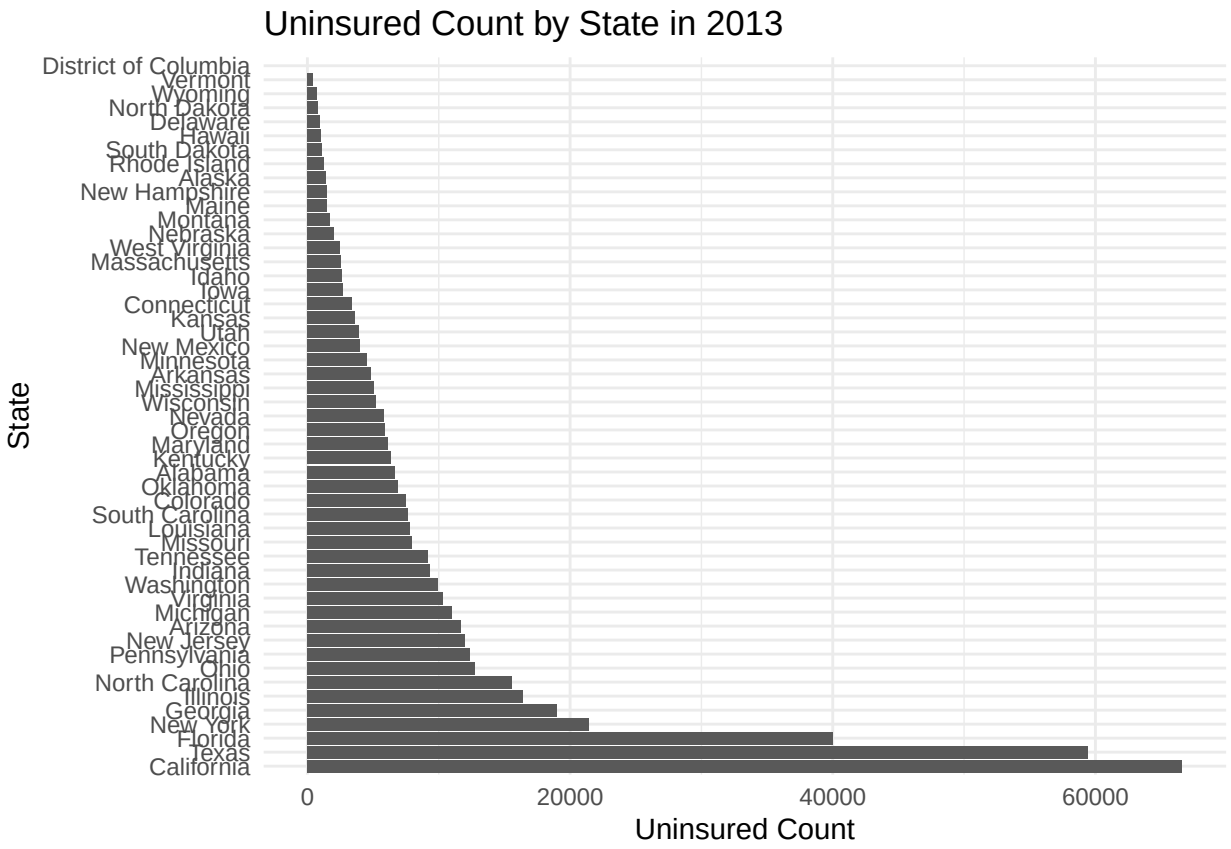
```
       x = "State",
       y = "Uninsured Rate") +
  theme_minimal()
```



In 2013, prior to Medicaid expansion, the states with the highest uninsured rates were Texas, Nevada and Florida. The states with the lowest uninsured rates were Massachusetts, DC and Vermont. Uninsured rate ranged from less than 5% to close to 25%.

```
# most uninsured Americans

#2013 plot
df_2013 <- medicaid_expansion %>%
  filter(year == 2013) %>%
  select(State, uninsured_rate, population)

df_2013 <- df_2013 %>%
  mutate(uninsured_count = (uninsured_rate / 100) * population) %>%
  arrange(desc(uninsured_count))

# plot
ggplot(df_2013, aes(x = reorder(State, -uninsured_count), y = uninsured_count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Uninsured Count by State in 2013",
       x = "State",
       y = "Uninsured Count") +
  theme_minimal()
```

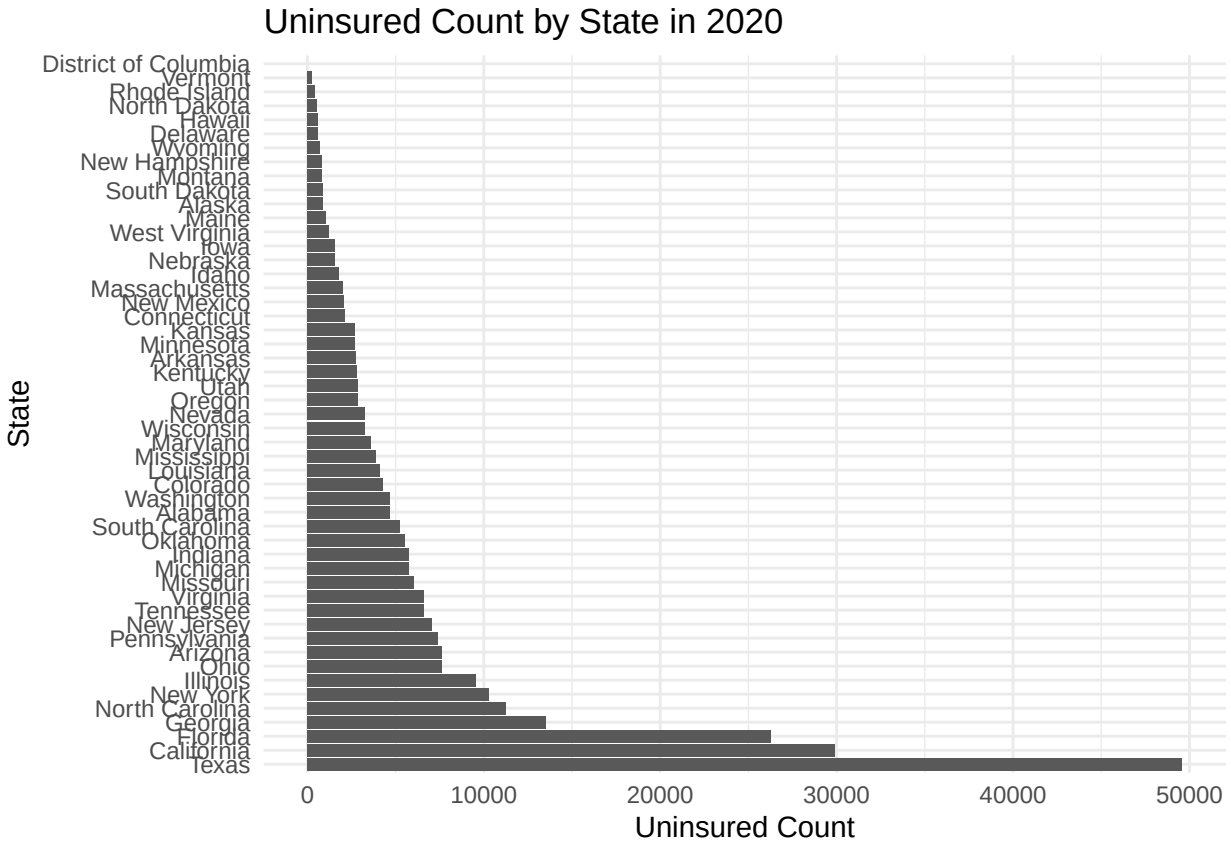
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



In 2013, prior to Medicaid expansion, the states with the most uninsured Americans were California, Texas and Florida. The states with the least uninsured Americans were DC, Vermont and Wyoming.

```
#2020 plot
df_2020 <- medicaid_expansion %>%
  filter(year == 2020) %>%
  select(State, uninsured_rate, population)

df_2020 <- df_2020 %>%
  mutate(uninsured_count = (uninsured_rate / 100) * population) %>%
  arrange(desc(uninsured_count))

# plot

ggplot(df_2020, aes(x = reorder(State, -uninsured_count), y = uninsured_count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Uninsured Count by State in 2020",
       x = "State",
       y = "Uninsured Count") +
  theme_minimal()
```

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



In 2020, the states with the most uninsured Americans were Texas, California and Florida. The states with the least uninsured Americans were DC, Vermont and Rhode Island. Absolute numbers of uninsured Americans appear to have decreased in the period post the ACA implementation.

Difference-in-Differences Estimation

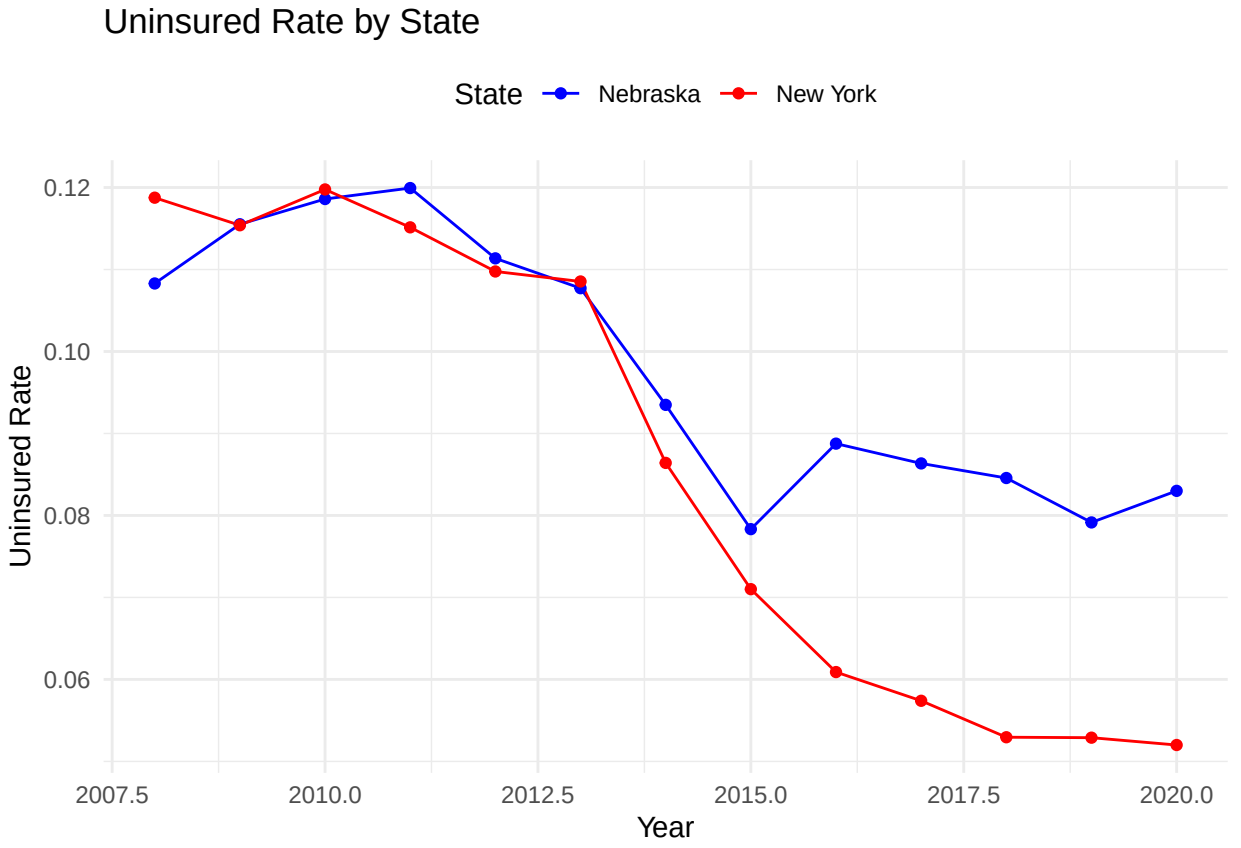
Estimate Model

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
#Early Medicaid expansion adopter: NY, Late Medicaid expansion adopter: Virginia (first attempt, has a
df <- medicaid_expansion %>%
  filter(State %in% c("New York", "Nebraska")) %>%
  select(State, year, uninsured_rate)

ggplot(df, aes(x = year, y = uninsured_rate, color = State)) +
  geom_line() +
  geom_point() +
  labs(title = "Uninsured Rate by State",
       x = "Year",
       y = "Uninsured Rate") +
```

```
  theme_minimal() +
  scale_color_manual(values = c("blue", "red")) +
  theme(legend.position = "top")
```



- 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

#Indicator for treatment (1 if state adopted Medicaid expansion and for all years after)
library(lubridate)

medicaid_expansion <- medicaid_expansion %>%
  mutate(Date_Adopted = ymd(Date_Adopted)) %>%
           mutate(expansion_year = year(Date_Adopted)) %>%
           mutate(treatment = ifelse(!is.na(expansion_year) & year >= expansion_year, 1, 0))


#Estimate DiD for NY and Nebraska

nn <- medicaid_expansion %>%
  filter(State %in% c("New York", "Nebraska")) %>%
  select(State, year, uninsured_rate, treatment)

did_model <- lm(uninsured_rate ~ year + treatment + year:treatment, data = nn)
```

```r
summary(did_model)
```

```
##
## Call:
## lm(formula = uninsured_rate ~ year + treatment + year:treatment,
##     data = nn)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0163368 -0.0056405 -0.0002242  0.0049685  0.0251712
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.9502058  1.3325812   5.966 5.28e-06 ***
## year           -0.0038985  0.0006622  -5.888 6.34e-06 ***
## treatment      -2.7055125  3.3984507  -0.796    0.434
## year:treatment  0.0013308  0.0016852   0.790    0.438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009282 on 22 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8452
## F-statistic: 46.49 on 3 and 22 DF,  p-value: 1.087e-09
```

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**: In this situation, we are interested in differences that arise across states, and the states are not as similar to one another as towns on either side of a river. In the example I chose, New York and Nebraska, it is likely that these states are different for many reasons. So, we rely on the parallel trends assumption to help us understand the differences in trends between the two states, which accounts for the fact that these states are not similar to one another, but rather similar to themselves over time.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: Parallel trends allows us to state more confidently whether there are potentially unobserved confounders that would make the estimate less reliable or not. However, it is a strong assumption that may not hold in practice. Still, one strength is that it is a testable assumption, compared to many other assumptions in econometrics which cannot be tested directly. On the other hand it can require a less systematic approach than desireable (and better achieved by synthetic control).

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.

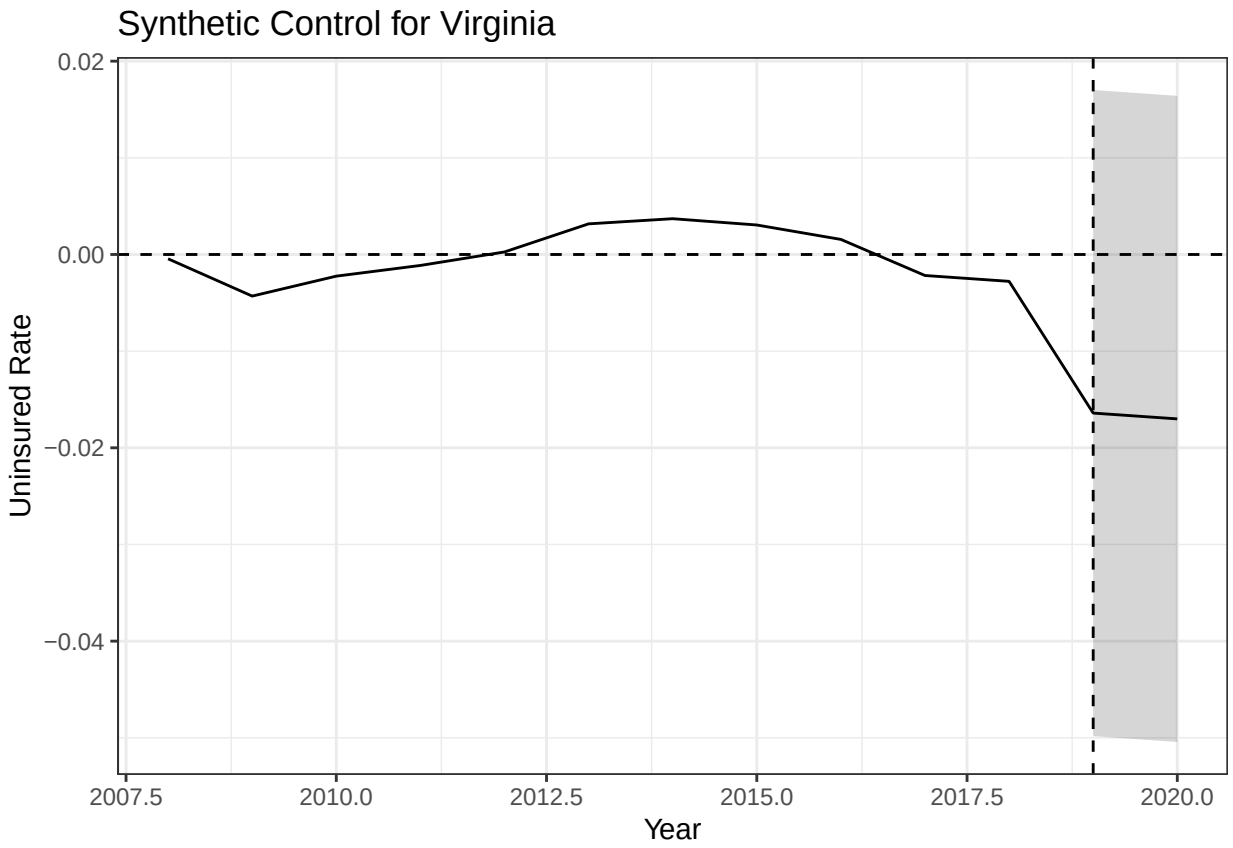
```
# non-augmented synthetic control (Virginia)
treated_state <- "Virginia"
sc_df <- medicaid_expansion %>%
  mutate(Date_Adopted = lubridate::ymd(Date_Adopted),
         expansion_year = lubridate::year(Date_Adopted),
         treated = ifelse(State == treated_state, 1, 0),
         treat_year = ifelse(State == treated_state, expansion_year, 0))



syn_model <- augsynth(
  uninsured_rate ~ treated,
  unit = State,
  time = year,
  t_int = unique(sc_df$treat_year[sc_df$State == treated_state]),
  data = sc_df,
  progfunc = "none",
  scm = T)
```

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

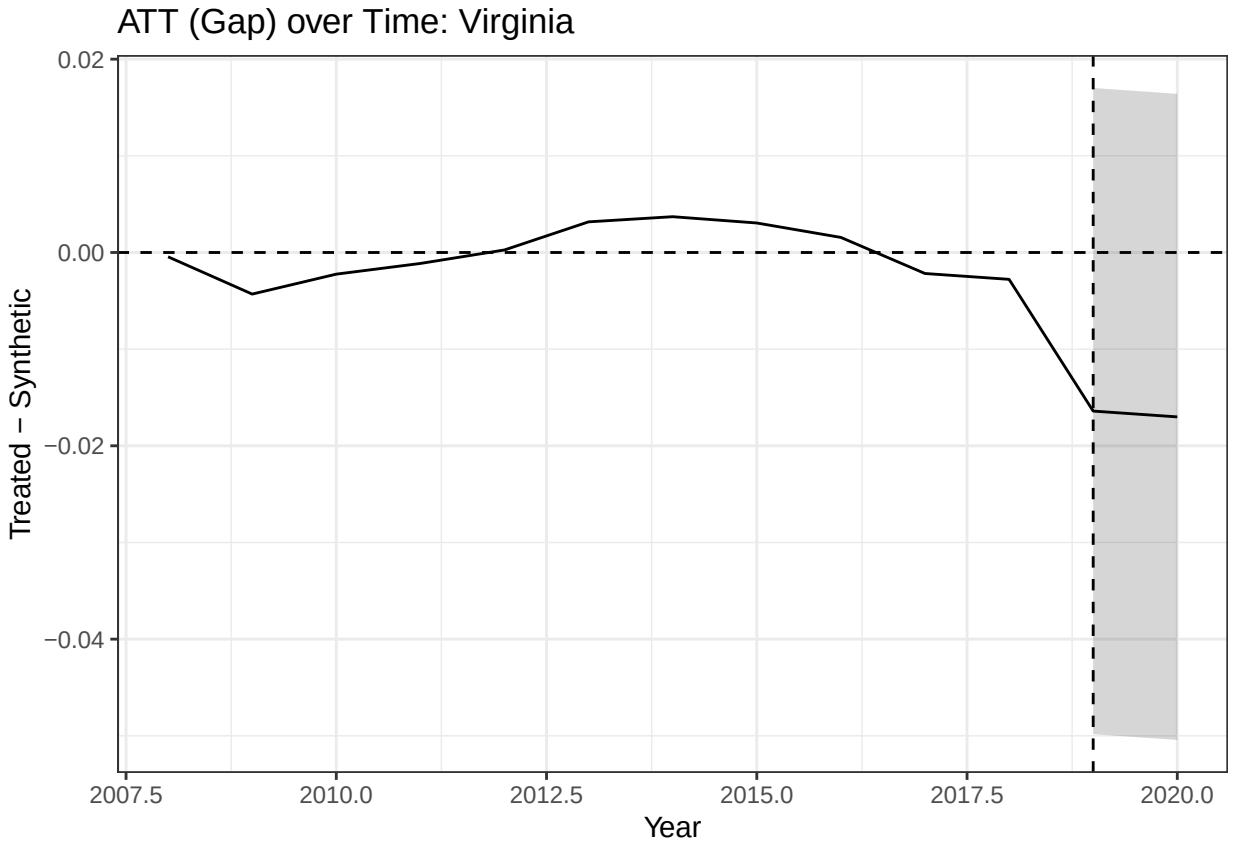
```
#report average ATT and L2 imbalance
summary(syn_model)
```

```
##
## 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):  -0.0167   ( 0.095 )
## L2 Imbalance: 0.009
## Percent improvement from uniform weights: 88.9%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2019   -0.016              -0.05               0.017   0.079
##  2020   -0.017              -0.05               0.016   0.081
```

```
# Plot treated vs synthetic
plot(syn_model, type = "unit") +
  ggtitle(paste("Synthetic Control for", treated_state)) +
  ylab("Uninsured Rate") +
  xlab("Year")
```




```
# Plot the ATT over time
plot(syn_model, type = "gap") +
  ggtitle(paste("ATT (Gap) over Time:", treated_state)) +
  ylab("Treated - Synthetic") +
  xlab("Year")
```



- 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

ridge_sc <- augsynth(
  uninsured_rate ~ treated,
  unit = State,
  time = year,
  t_int = unique(sc_df$treat_year[sc_df$State == treated_state]),
  data = sc_df,
  progfunc = "ridge",
  scm = T)
```

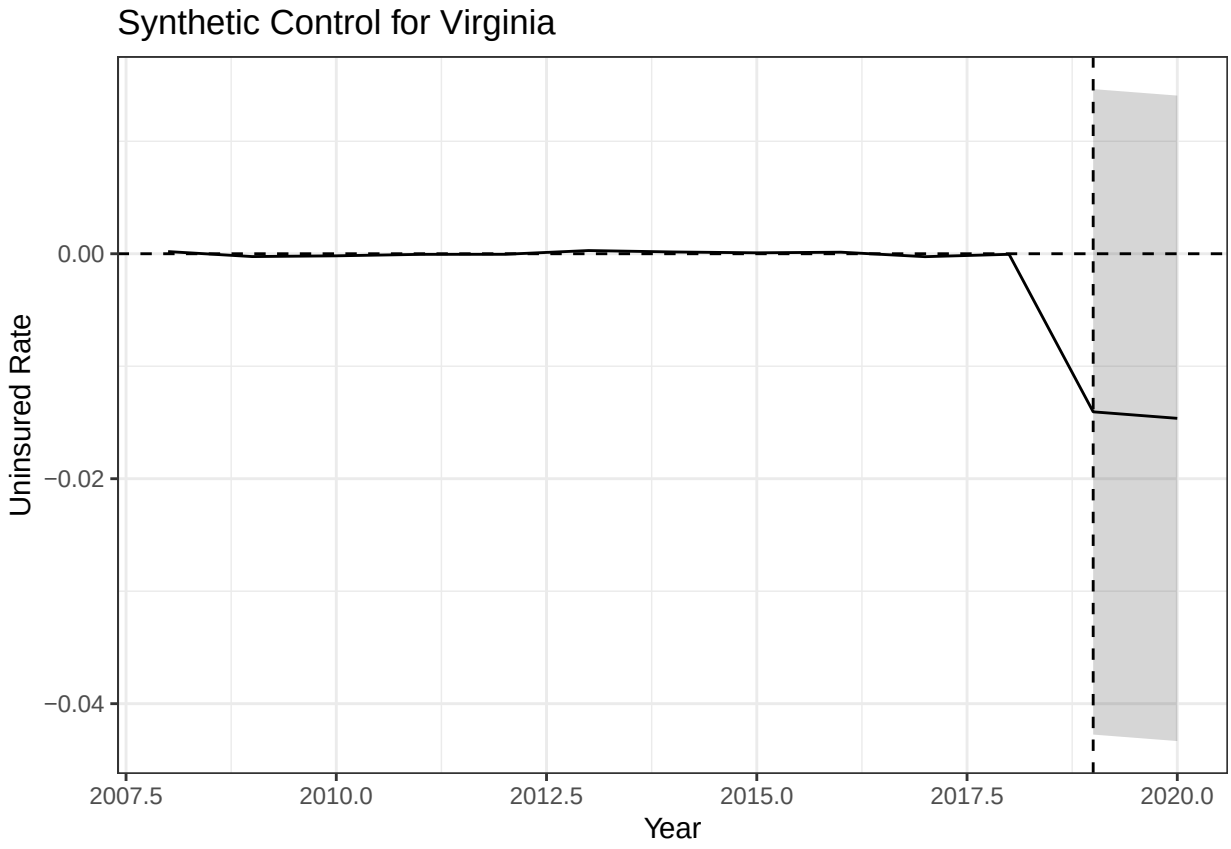
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
#report average ATT and L2 imbalance
summary(ridge_sc)
```

```
##
## 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):  -0.0143   ( 0.06 )
## L2 Imbalance: 0.001
## Percent improvement from uniform weights: 99.2%
##
## Avg Estimated Bias: -0.002
```

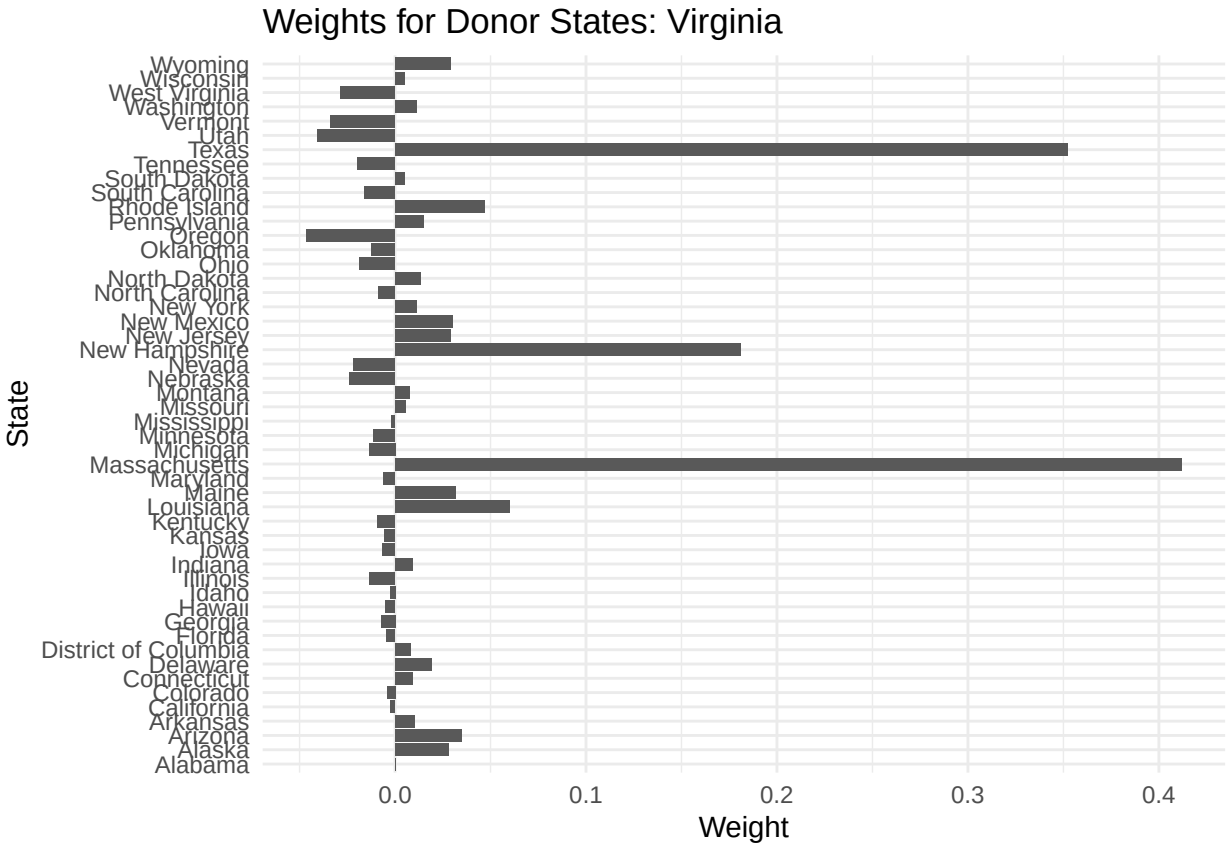
```
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2019   -0.014             -0.043              0.015   0.095
##  2020   -0.015             -0.043              0.014   0.095
```

```r
#plot
plot(ridge_sc, type = "unit") +
  ggtitle(paste("Synthetic Control for", treated_state)) +
  ylab("Uninsured Rate") +
  xlab("Year")
```



- Plot barplots to visualize the weights of the donors.

```r
# barplots of weights

data.frame(ridge_sc$weights) %>%
  tibble::rownames_to_column('State') %>%
  ggplot() +
  geom_bar(aes(x = State, y=ridge_sc.weights),
           stat = 'identity')+
  coord_flip() +
  labs(title = paste("Weights for Donor States:", treated_state),
       x = "State",
       y = "Weight") +
  theme_minimal()
```



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**: DiD relies on the parallel trends assumption. In cases where we think this holds, it can be a clear method for estimating policy effects. Still, the main advantage of synthetic control is that it allows for a more flexible approach to estimating the counterfactual which can be helpful when we think there might be violations of PTA. It is also particularly useful when there are a small number of treated units. However, it is also more complex and requires more data than a simple DiD estimator. In addition, given that the control is now created from a set of donor states, it is less transparent and potentially more difficult to interpret/communicate.
- 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**: In the standard synthetic control, given the way weights are restricted, we can interpret the synthetic control as a combination or blend of the donor states. In augmented synthetic control, we may get negative weights which are difficult to interpret and it can look less like what a real control unit might look like. However, the tradeoff is that we are able to get a better fit in the pre-treatment period, which is important for the validity of the method, so augmented SC can give us better estimates at the expense of worse interpretability. The decision of which to use should be informed by the particular goals of the project at hand.

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

stag_syn <- multisynth(
  uninsured_rate ~ treatment,
  unit = State,
  time = year,
  nu = 0,
  medicaid_expansion,
  n_leads = 3)

summary(stag_syn)
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treatment, unit = State, time = year,
##     data = medicaid_expansion, n_leads = 3, nu = 0)
##
## Average ATT Estimate (Std. Error): -0.014  (0.005)
##
## Global L2 Imbalance: 0.001
## Scaled Global L2 Imbalance: 0.039
## Percent improvement from uniform global weights: 96.1
##
## Individual L2 Imbalance: 0.004
## Scaled Individual L2 Imbalance: 0.089
## Percent improvement from uniform individual weights: 91.1
##
##  Time Since Treatment   Level    Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.01099755 0.004619381 -0.01958718 -0.002111696
##                     1 Average -0.01695027 0.006264771 -0.02888033 -0.005101834
##                     2 Average -0.01564573 0.007151582 -0.02983788 -0.002156679
```

```
subset_states <- c("New York", "Virginia",  "Nebraska", "California")

plot(stag_syn, "unit", units = subset_states) +
  ggtitle("Multisynth Model for Selected States") +
  ylab("Uninsured Rate") +
  xlab("Year")
```

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

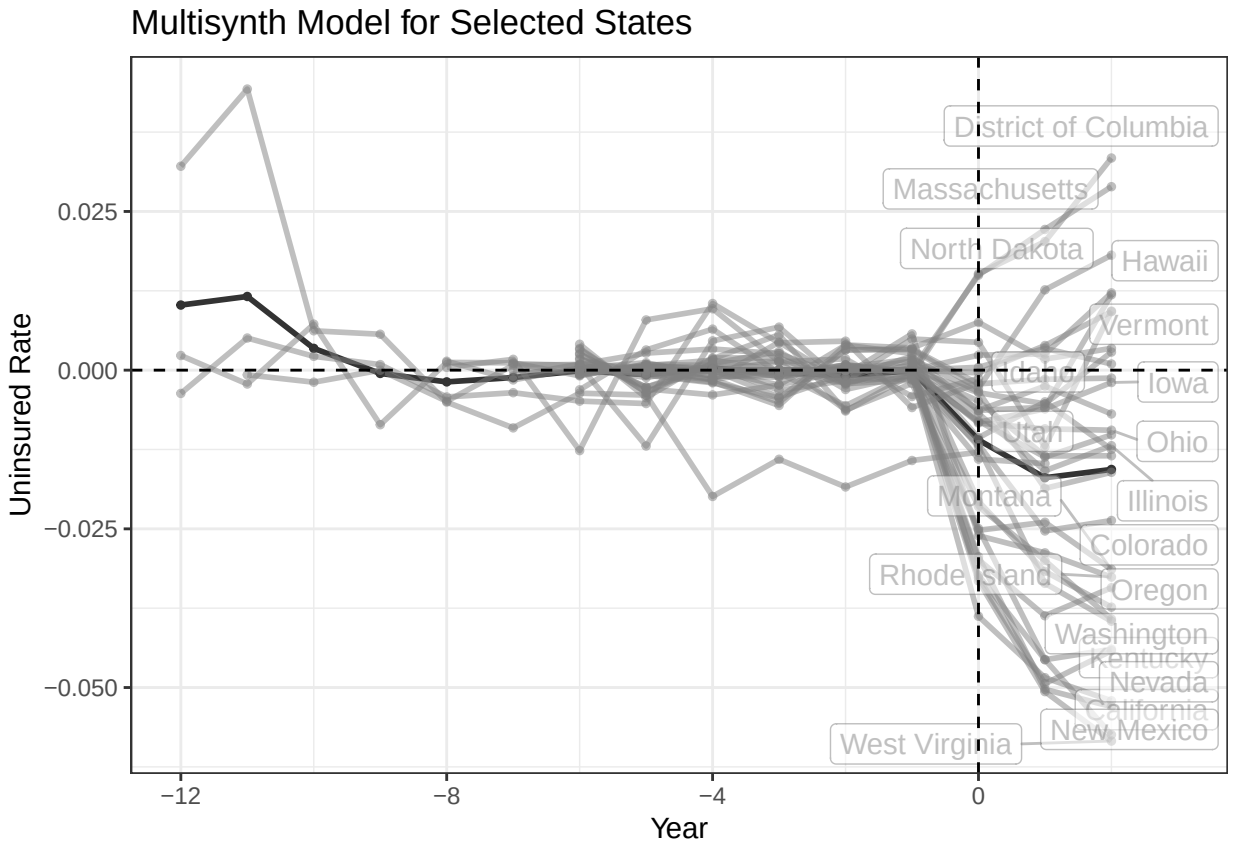
```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the augsynth package.
##   Please report the issue to the authors.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 230 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 230 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
## Warning: ggrepel: 17 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



- 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

cohort <- multisynth(
  uninsured_rate ~ treatment,
  unit = State,
  time = year,
  t_int = 'expansion_year',
  data = medicaid_expansion,
  n_leads = 3,
  progfunc = 'Ridge',
  scm = T,
  time_cohort = TRUE)

summary(cohort)
```

```
##
## Call:
## multisynth(form = uninsured_rate ~ treatment, unit = State, time = year,
##     data = medicaid_expansion, n_leads = 3, scm = T, time_cohort = TRUE,
##     t_int = "expansion_year", progfunc = "Ridge")
##
## Average ATT Estimate (Std. Error): -0.013  (0.005)
##
## 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: 0.016
## Percent improvement from uniform individual weights: 98.4
##
##  Time Since Treatment   Level     Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.009831319 0.004179071 -0.01827401 -0.002370055
##                     1 Average -0.015536949 0.005841387 -0.02665153 -0.004557869
##                     2 Average -0.014968769 0.006370219 -0.02764028 -0.003367269
```

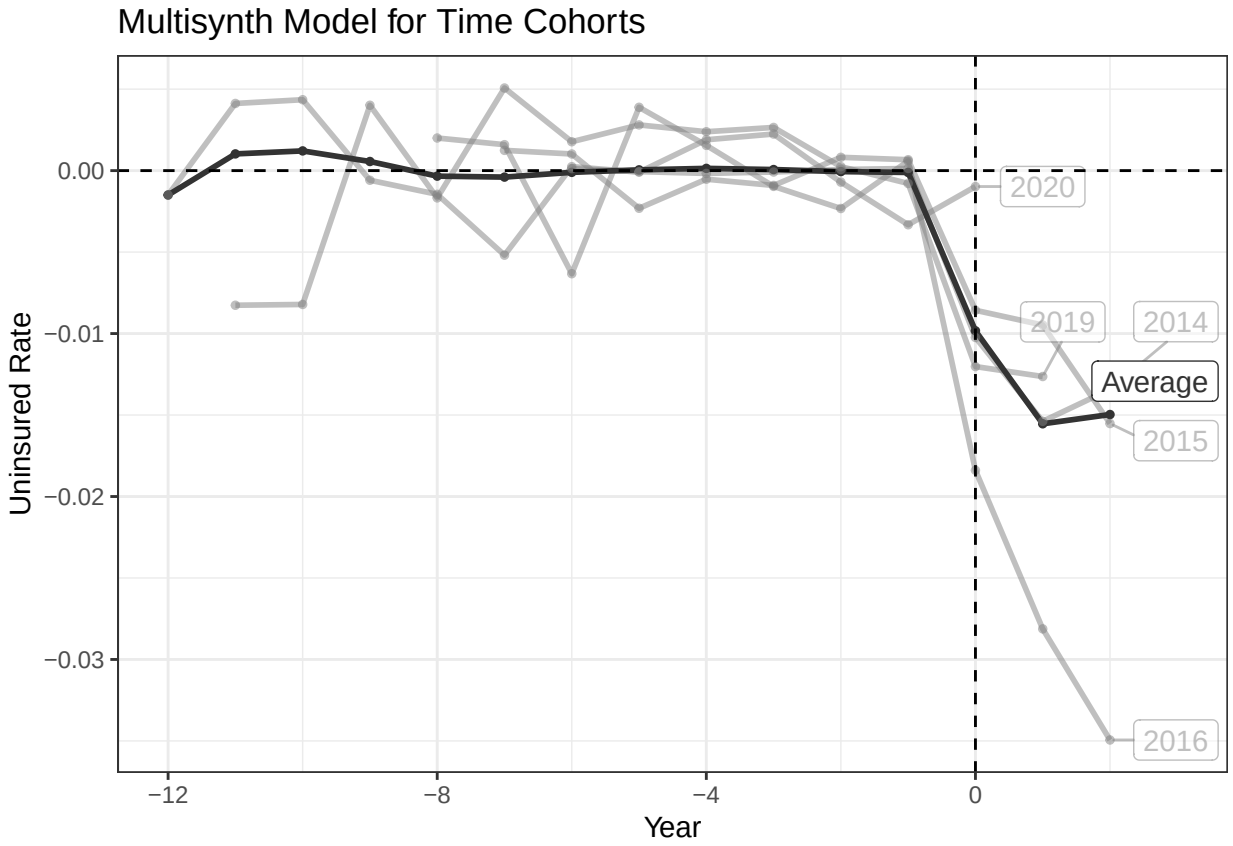
```
plot(cohort, "unit") +
  ggtitle("Multisynth Model for Time Cohorts") +
  ylab("Uninsured Rate") +
  xlab("Year")
```

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

```
## Warning: Removed 25 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 25 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



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. The cohort synthetic control model shows that states that expanded Medicaid in 2016 had a larger treatment effect than states that expanded in 2015, demonstrating this heterogeneity. In addition, we see in the model that looks at individual states that there is a lot of variation in the treatment effect sizes, with some states even having increases in the uninsured rate after expansion, lending more evidence to the fact that Medicaid expansion is in fact a coarse measure of the implementation of a complex set of policies.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: To some extent. The cohort which seems to have had the largest decrease in the uninsured rate was the 2016 cohort. So, those who expanded in 2016 have had better outcomes than those who expanded later (i.e. 2019), but also better outcomes than those that expanded in 2014 and 2015. I believe this is likely due to the fact that the 2016 cohort was made up of states where the expansion of Medicaid was more impactful due to other factors in that state's policy environment.

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?
- **Answer**: These methods rely on access to panel data. These are much less common for individual level observations, but much more common to have repeated cross section data at the state or country level. By leveraging multiple years of data over the same units of observation, these methods are able to account for unobserved confounders that are constant over time or constant over the unit of measurement.
- 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, we formalize the issue of selection into treatment using the parallel trends assumption. This assumption would require that there is not differential selection into treatment in a way that systematically relates to the outcome. Still, this does not account for time-varying confounders. So, it makes sense to use these methods when we have panel data and a strong belief (or evidence) that the parallel trends assumption holds. Regression discontinuity designs think about selection into treatment via assumptions as well. In this case, we are concerned about manipulation around the running variable. We can test for this using various methods to identify whether there is bunching around the cutoff and therefore should be concerned that individuals are selecting into treatment (values right above the cutoff of the running variable). RDD makes sense to use when we have a clear cutoff and don't have violation of the no manipulation assumption.