Project 7: Difference-in-Differences and Synthetic Control

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

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

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

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




files.sources = list.files("R")
sapply(paste("R/", files.sources, sep=""), source)
```

```
##         R/augsynth.R R/augsynth_pre.R R/cv.R R/data.R R/eligible_donors.R
## value   NULL         ?                ?      "kansas" ?
## visible TRUE         FALSE            FALSE  TRUE     FALSE
##         R/fit_synth.R R/format.R R/globalVariables.R R/highdim.R R/inference.R
## value   ?             ?          character,19        ?           ?
## visible FALSE         FALSE      TRUE                FALSE       FALSE
##         R/multi_outcomes.R R/multi_synth_qp.R R/multisynth_class.R
## value   ?                  ?                  ?
## visible FALSE              FALSE              FALSE
##         R/outcome_models.R R/outcome_multi.R R/ridge.R R/ridge_lambda.R
## value   ?                  ?                 ?         ?
## visible FALSE              FALSE             FALSE     FALSE
##         R/time_regression_multi.R
## value   ?
## visible FALSE
```

```r
# 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.

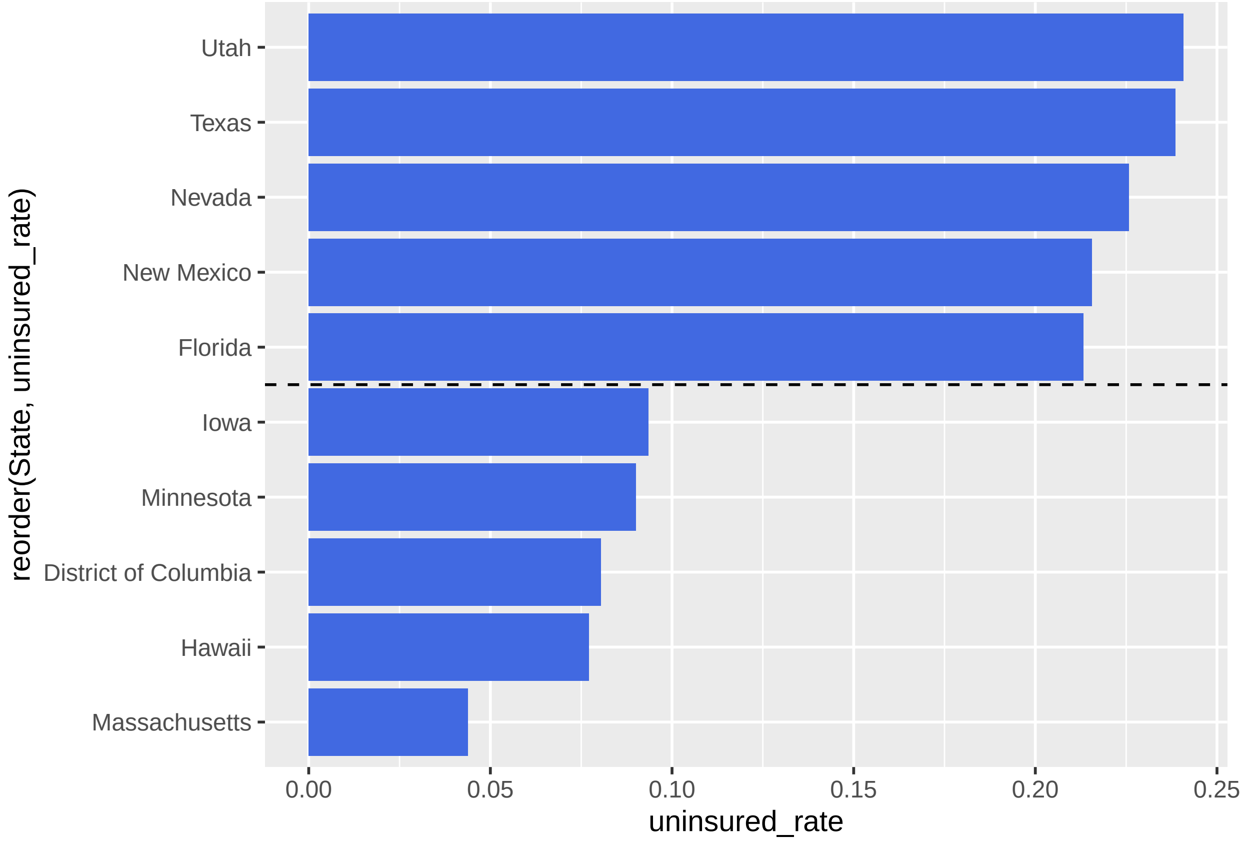
Exploratory Data Analysis

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.

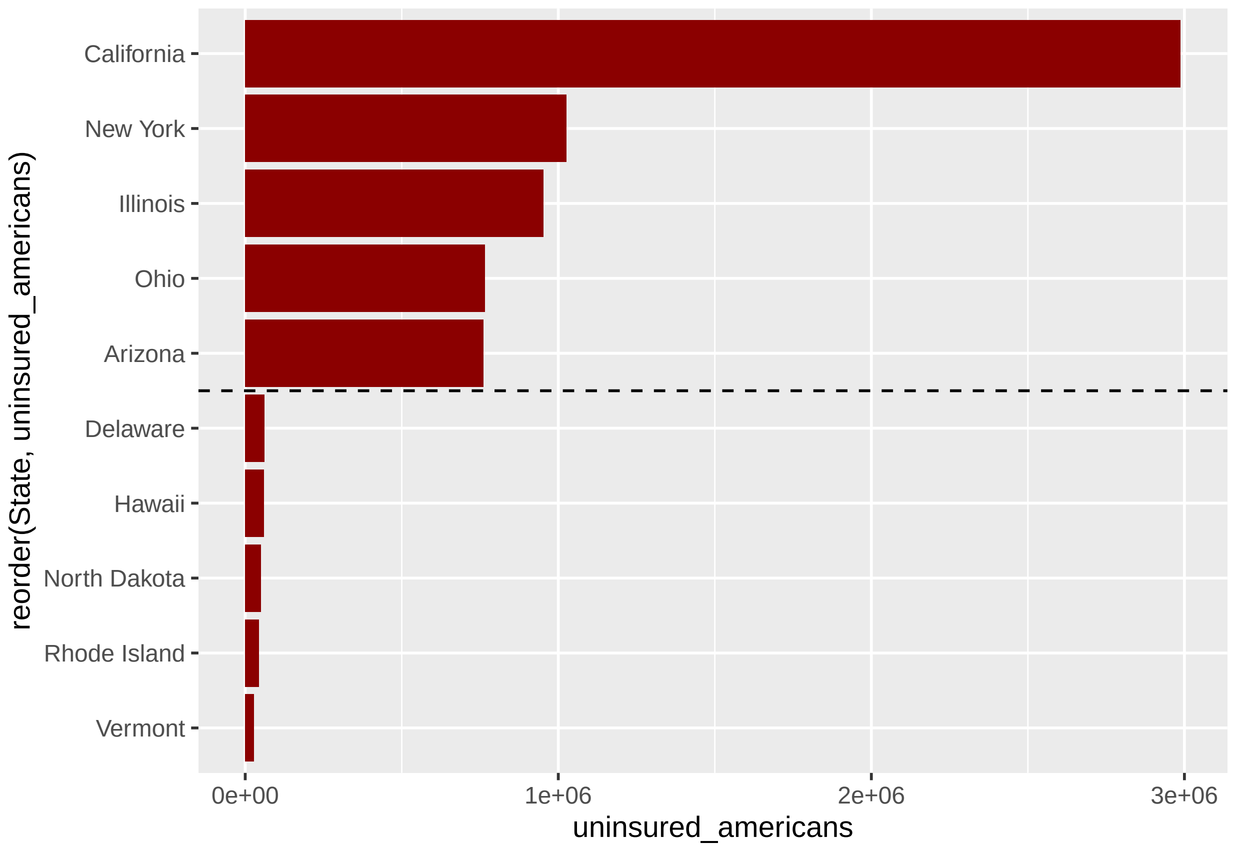
```
# highest and lowest uninsured rates
medicaid_expansion %>%
  filter(year < 2014) %>%
  arrange(desc(uninsured_rate)) %>%
  distinct(State, .keep_all = TRUE) %>%
  slice(c(1:5, (n()-4):n())) %>%
  ggplot(aes(x=reorder(State, uninsured_rate),
             y=uninsured_rate)) +
  geom_bar(stat = "identity", fill="royalblue") +
  geom_vline(xintercept = 5.5, linetype = "dashed", color = "black") +
  coord_flip() +
  theme(legend.position="none")
```



```
# most uninsured Americans before 2014
medicaid_expansion %>%
  filter(year < 2014) %>%
  mutate(uninsured_americans = uninsured_rate*population) %>%
  drop_na() %>%
  arrange(desc(uninsured_americans)) %>%
  distinct(State, .keep_all = TRUE) %>%
  slice(c(1:5, (n()-4):n())) %>% select(State, uninsured_americans)
```

```
## # A tibble: 10 x 2
##    State        uninsured_americans
##    <chr>                      <dbl>
##  1 California              7205973.
##  2 New York                2364907.
##  3 Illinois                1806179.
##  4 Ohio                    1431647.
##  5 Pennsylvania            1306814.
##  6 Rhode Island             127901.
##  7 Hawaii                   109501.
##  8 Delaware                 100659.
##  9 Vermont                   94105.
## 10 North Dakota              79058.
```

```
# most uninsured Americans in the last year (2020)
medicaid_expansion %>%
  filter(year == max(year)) %>%
  mutate(uninsured_americans = uninsured_rate*population) %>%
  drop_na() %>%
  arrange(desc(uninsured_americans)) %>%
  slice(c(1:5, (n()-4):n())) %>%
  ggplot(aes(x=reorder(State, uninsured_americans),
             y=uninsured_americans)) +
  geom_bar(stat = "identity", fill="darkred") +
  geom_vline(xintercept = 5.5, linetype = "dashed", color = "black") +
  coord_flip() +
  theme(legend.position="none")
```



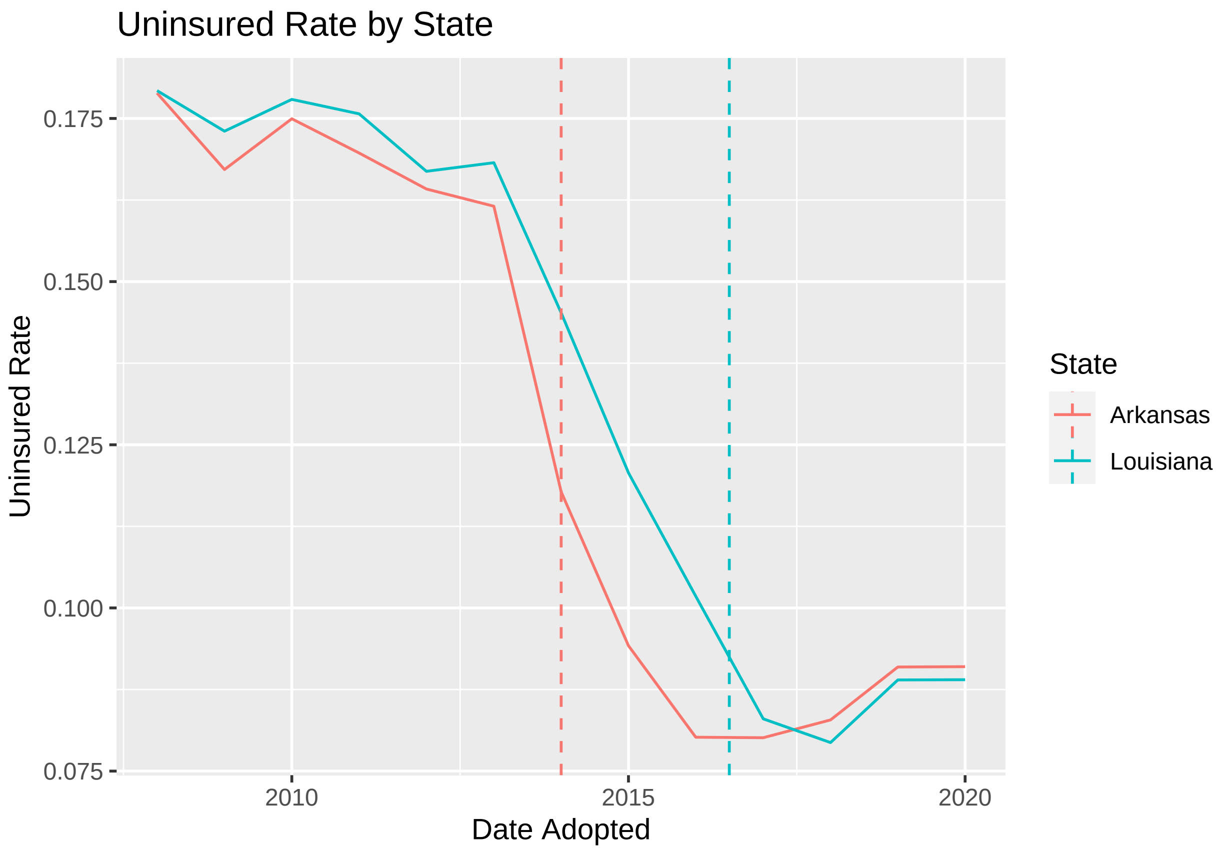
Difference-in-Differences Estimation

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
selection <- medicaid_expansion %>% filter(State %in% c("Arkansas",
                                                        "Louisiana"))

selection <- selection %>%
  mutate(Date_Adopted = as.Date(Date_Adopted),
         year = as.Date(paste0(year, "-01-01")))

adoption_table <- selection %>%
  group_by(State) %>%
  reframe(adopted = Date_Adopted) %>%
  distinct()

ggplot() +
  geom_line(data=selection, aes(x = year, y = uninsured_rate, group=State, color=State)) +
  labs(title = "Uninsured Rate by State", x = "Date Adopted", y = "Uninsured Rate") +
  geom_vline(data=adoption_table, aes(xintercept=adopted, color=State), linetype="dashed")
```



- 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

```
selection %>% mutate(post = as.numeric(year >= Date_Adopted),
                     treat = as.numeric(State == "Arkansas")) %>%
  lm(uninsured_rate ~ post*treat, data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = uninsured_rate ~ post * treat, data = .)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.054742 -0.007323  0.000143  0.010146  0.026777
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.156530   0.006187  25.299  < 2e-16 ***
## post        -0.071445   0.011154  -6.405 1.91e-06 ***
## treat        0.012878   0.009783   1.316    0.202
## post:treat  -0.006950   0.015200  -0.457    0.652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01856 on 22 degrees of freedom
## Multiple R-squared:  0.8192, Adjusted R-squared:  0.7945
## F-statistic: 33.23 on 3 and 22 DF,  p-value: 2.385e-08
```

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**: That intuition is very difficult to replicate here because these states are massive geographic units, with large amounts of within-state variation in insurance rates, whereas Card and Krueger used smaller units – cities. With the aggregated units, we lose much of the value that promiximity might otherwise bring us.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: The parallel trends assumption is only possible in the past – we can show they two groups were parallel in the past, but the counterfactual – that they would have been the same in the future, isn't something we can prove, and is a liability for this identification strategy.

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.

```r
# non-augmented synthetic control

medicaid_expansion <- medicaid_expansion %>%
  mutate(Date_Adopted = as.Date(Date_Adopted),
         year = as.Date(paste0(year, "-01-01")))


state_sel <- "Arkansas"

medicaid_expansion <- medicaid_expansion %>%
  mutate(treated = 1 * ( (year>=Date_Adopted) & (State==state_sel) ))

medicaid_expansion%>% group_by(State) %>% count()
```

```
## # A tibble: 51 x 2
## # Groups:   State [51]
##    State                    n
##    <chr>                <int>
##  1 Alabama                 13
##  2 Alaska                  13
##  3 Arizona                 13
##  4 Arkansas                13
##  5 California              13
##  6 Colorado                13
##  7 Connecticut             13
##  8 Delaware                13
##  9 District of Columbia    13
## 10 Florida                 13
## # i 41 more rows
```

```
# restricting to the treated state and other untreated states
filtered_data <- medicaid_expansion%>% filter( (is.na(Date_Adopted)) | (State==state_sel) )

filtered_data %>% group_by(State) %>% count()
```

```
## # A tibble: 16 x 2
## # Groups:   State [16]
##    State              n
##    <chr>          <int>
##  1 Alabama           13
##  2 Arkansas          13
##  3 Florida           13
##  4 Georgia           13
##  5 Kansas            13
##  6 Maine             13
##  7 Mississippi       13
##  8 Missouri          13
##  9 North Carolina    13
## 10 Oklahoma          13
## 11 South Carolina    13
## 12 South Dakota      13
## 13 Tennessee         13
## 14 Texas             13
## 15 Wisconsin         13
## 16 Wyoming           13
```

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

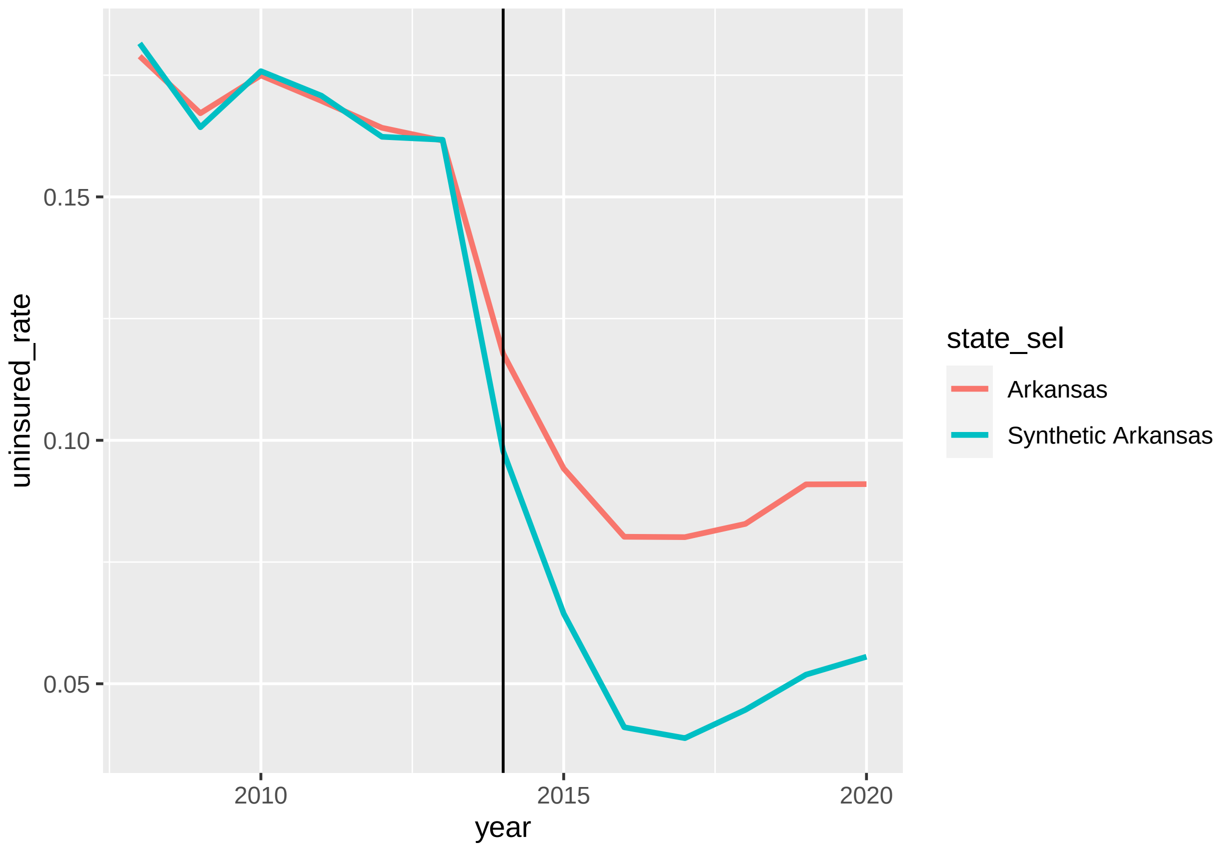
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
syn_sum <- summary(syn)


synth_state <- filtered_data %>%
filter(State == state_sel) %>%
bind_cols(difference = syn_sum$att$Estimate) %>%
mutate(synthetic_state = uninsured_rate + difference)

synth_state
```

```
## # A tibble: 13 x 8
##    State    Date_Adopted year       uninsured_rate population treated difference
##    <chr>    <date>       <date>              <dbl>      <dbl>   <dbl>      <dbl>
##  1 Arkansas 2014-01-01   2008-01-01         0.179     2994079       0   0.00265
##  2 Arkansas 2014-01-01   2009-01-01         0.167     2994079       0  -0.00287
##  3 Arkansas 2014-01-01   2010-01-01         0.175     2994079       0   0.000866
##  4 Arkansas 2014-01-01   2011-01-01         0.170     2994079       0   0.00107
##  5 Arkansas 2014-01-01   2012-01-01         0.164     2994079       0  -0.00185
##  6 Arkansas 2014-01-01   2013-01-01         0.162     2994079       0   0.000186
##  7 Arkansas 2014-01-01   2014-01-01         0.118     2994079       1  -0.0200
##  8 Arkansas 2014-01-01   2015-01-01         0.0942    2994079       1  -0.0298
##  9 Arkansas 2014-01-01   2016-01-01         0.0802    2994079       1  -0.0391
## 10 Arkansas 2014-01-01   2017-01-01         0.0801    2994079       1  -0.0413
## 11 Arkansas 2014-01-01   2018-01-01         0.0828    2994079       1  -0.0382
## 12 Arkansas 2014-01-01   2019-01-01         0.0910    2994079       1  -0.0391
## 13 Arkansas 2014-01-01   2020-01-01         0.091     2994079       1  -0.0354
## # i 1 more variable: synthetic_state <dbl>
```

```r
synth_state %>% ggplot() +
  geom_line(aes(x=year, y=uninsured_rate, color=state_sel), size=1) +
  geom_line(aes(x=year, y=synthetic_state, color=paste("Synthetic", state_sel)), size=1) +
  geom_vline(aes(xintercept = Date_Adopted[1]))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```r
print("L2 imbalance")
```

```
## [1] "L2 imbalance"
```

```r
print(syn$l2_imbalance)
```

```
## [1] 0.004540479
```

```r
print("ATT")
```

```
## [1] "ATT"
```

```r
print(syn_sum$average_att[1,1])
```

```
## [1] -0.03470768
```

```r
print("ATT p-value")
```

```
## [1] "ATT p-value"
```

```r
print(syn_sum$average_att[1,5])
```

```
## [1] 0.376
```

- 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.

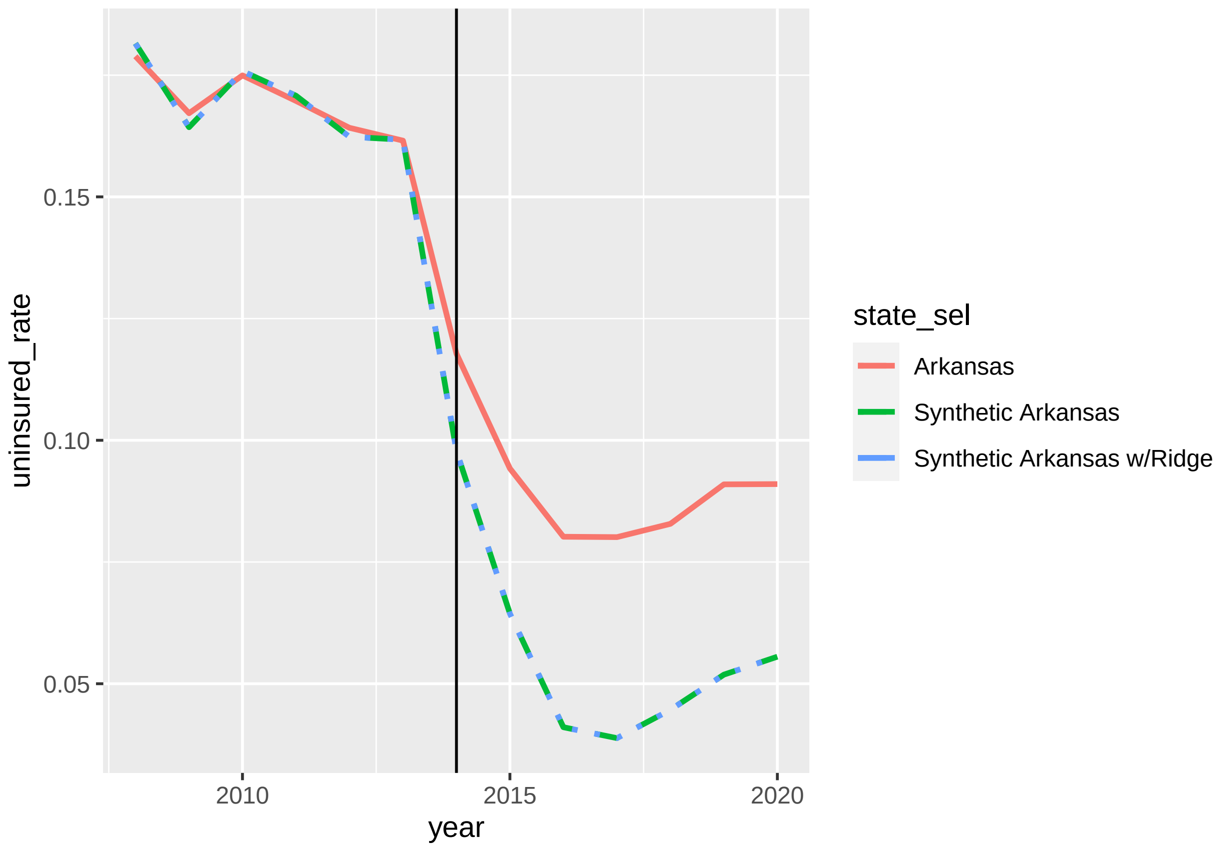
```r
# augmented synthetic control
syn_ridge <- augsynth(uninsured_rate ~ treated, State, year, filtered_data, progfunc = "Ridge", scm = T
```

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

```
syn_sum_ridge <- summary(syn_ridge)

synth_state_ridge <- synth_state %>%
bind_cols(ridge_difference = syn_sum_ridge$att$Estimate) %>%
mutate(ridge_synthetic_state = uninsured_rate + ridge_difference)


synth_state_ridge
```

```
## # A tibble: 13 x 10
##    State    Date_Adopted year       uninsured_rate population treated difference
##    <chr>    <date>       <date>              <dbl>      <dbl>   <dbl>      <dbl>
##  1 Arkansas 2014-01-01   2008-01-01         0.179     2994079       0   0.00265
##  2 Arkansas 2014-01-01   2009-01-01         0.167     2994079       0  -0.00287
##  3 Arkansas 2014-01-01   2010-01-01         0.175     2994079       0   0.000866
##  4 Arkansas 2014-01-01   2011-01-01         0.170     2994079       0   0.00107
##  5 Arkansas 2014-01-01   2012-01-01         0.164     2994079       0  -0.00185
##  6 Arkansas 2014-01-01   2013-01-01         0.162     2994079       0   0.000186
##  7 Arkansas 2014-01-01   2014-01-01         0.118     2994079       1  -0.0200
##  8 Arkansas 2014-01-01   2015-01-01         0.0942    2994079       1  -0.0298
##  9 Arkansas 2014-01-01   2016-01-01         0.0802    2994079       1  -0.0391
## 10 Arkansas 2014-01-01   2017-01-01         0.0801    2994079       1  -0.0413
## 11 Arkansas 2014-01-01   2018-01-01         0.0828    2994079       1  -0.0382
## 12 Arkansas 2014-01-01   2019-01-01         0.0910    2994079       1  -0.0391
## 13 Arkansas 2014-01-01   2020-01-01         0.091     2994079       1  -0.0354
## # i 3 more variables: synthetic_state <dbl>, ridge_difference <dbl>,
## #   ridge_synthetic_state <dbl>
```

```
synth_state_ridge %>% ggplot() +
  geom_line(aes(x=year, y=uninsured_rate, color=state_sel), size=1) +
  geom_line(aes(x=year, y=synthetic_state, color=paste("Synthetic", state_sel)), size=1, linetype="dashe
  geom_line(aes(x=year, y=ridge_synthetic_state, color=paste("Synthetic", state_sel, "w/Ridge")), size=
  geom_vline(aes(xintercept = Date_Adopted[1]))
```



```
print("L2 imbalance")
```

```
## [1] "L2 imbalance"
```

```
print(syn_ridge$l2_imbalance)
```

```
## [1] 0.004527312
```

```
print("ATT")
```

```
## [1] "ATT"
```

```
print(syn_sum_ridge$average_att[1,1])
```

```
## [1] -0.03471667
```

```
print("ATT p-value")
```

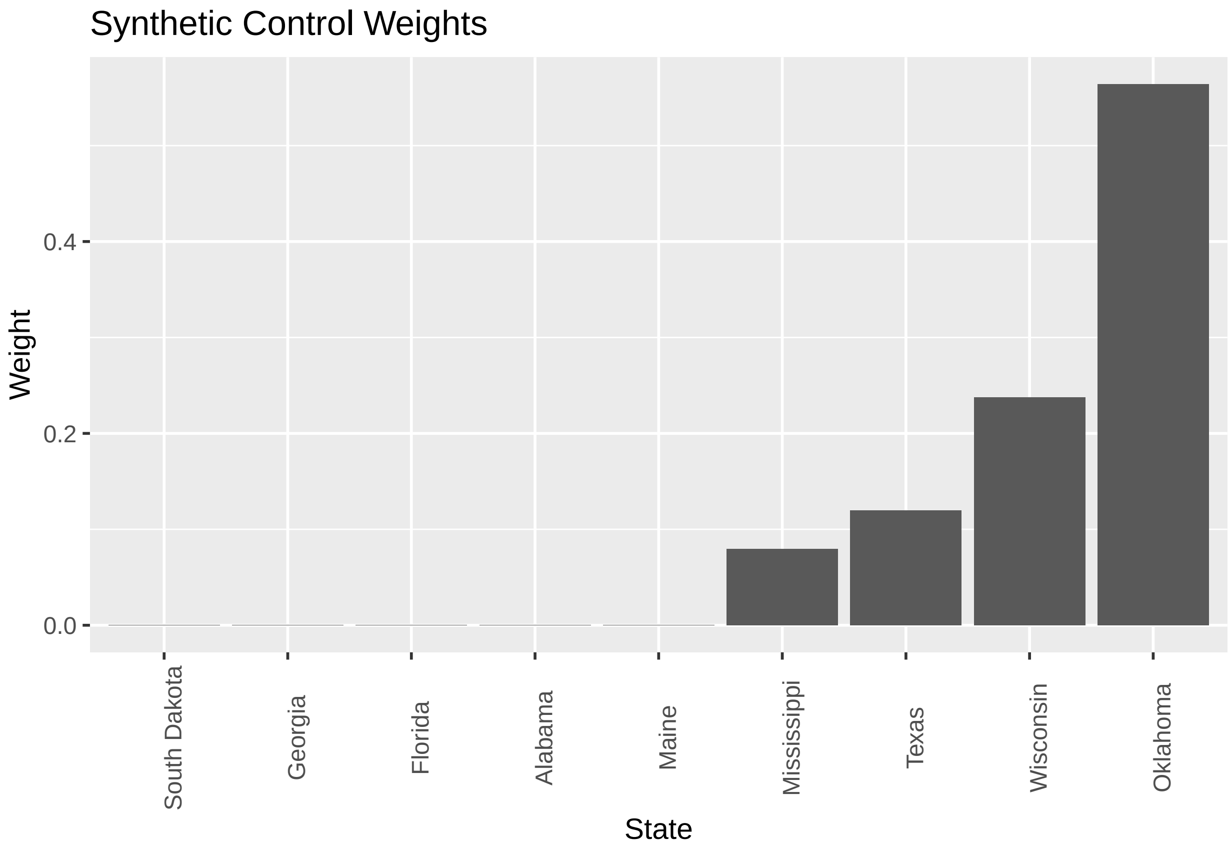
```
## [1] "ATT p-value"
```

```
print(syn_sum_ridge$average_att[1,5])
```

```
## [1] 0.511
```

- Plot barplots to visualize the weights of the donors.

```
# barplots of weights
data.frame(syn_ridge$weights) %>%
# change index to a column
tibble::rownames_to_column('State') %>%
  filter(syn_ridge.weights > 0) %>%
ggplot() +
# stat = identity to take the literal value instead of a count for geom_bar()
geom_bar(aes(x = reorder(State, syn_ridge.weights),
y = syn_ridge.weights),
stat = 'identity') +
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?

Discussion Questions

- What are the advantages and disadvantages of synthetic control compared to difference-in-differences estimators?
- **Answer**: Synthetic control allows us to create a much more similar comparison group – a weighted combination of all the candidates. This creates very parellel trends in the pre period, and allows us to more easily assume parallel trends in the post.
- 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**: The negative weights in the augmented SCM reflect the fact that some control units may have a negative association with the treated unit in the pre-treatment period. This can happen, for example, if the control units have a negative correlation with the treated unit on some unobserved confounding variable. The negative weights allow the synthetic control to down-weight these control units and improve the balance in the pre-treatment period.

In terms of balancing the considerations of negative weights versus improvements in pre-treatment balance, we can evaluate the robustness of their findings to different weighting schemes. We can also assess the plausibility of negative weights and whether they make sense given the context of a given analysis.

Staggered Adoption Synthetic Control

Estimate Multisynth

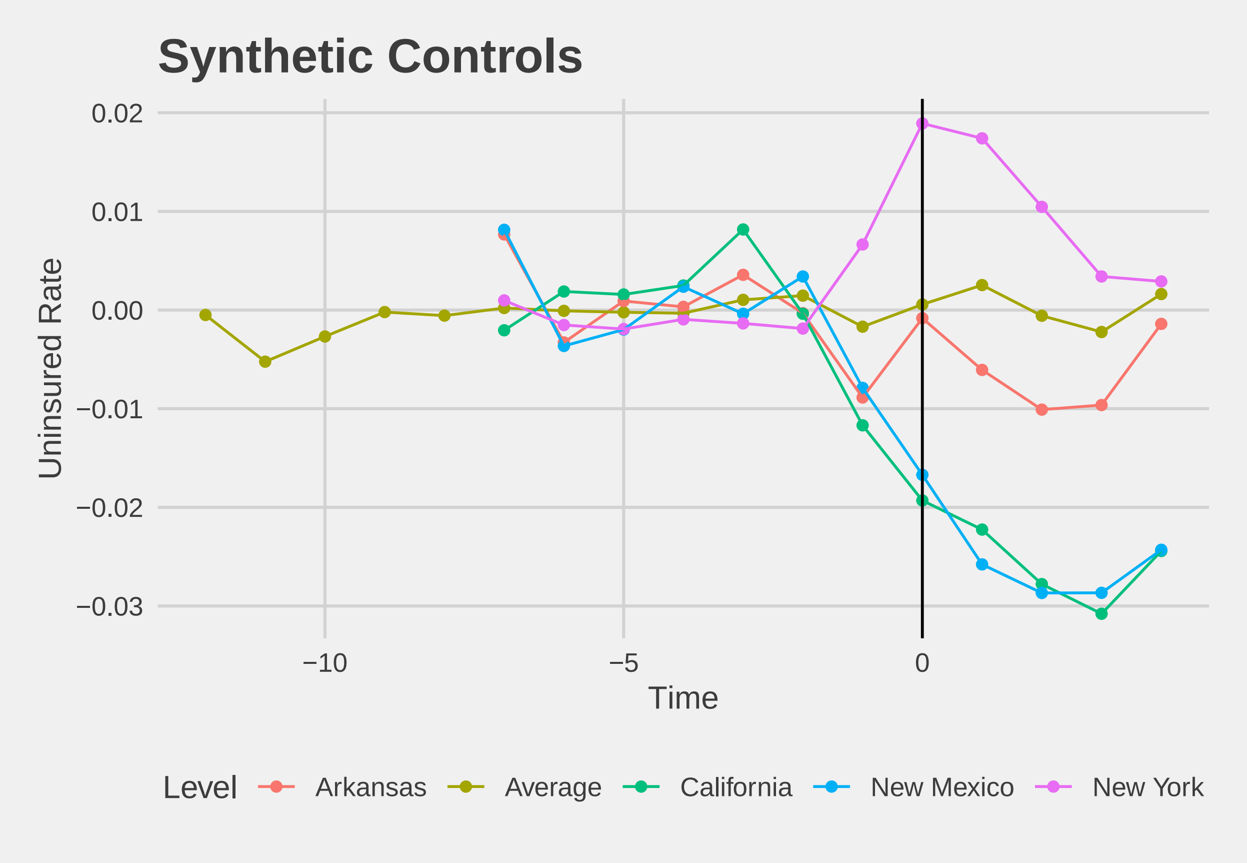
Do the following:

- 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
medicaid_expansion_clean <- medicaid_expansion %>%
  mutate(expanded = 1*(!is.na(Date_Adopted) & (year > Date_Adopted)))

syn_multi <- multisynth(uninsured_rate ~ expanded, State, year, medicaid_expansion_clean,
                        progfunc = "None", scm = T, n_leads = 5)


summary(syn_multi)$att %>%
  filter(Level %in% c("California", "New York", "Arkansas", "New Mexico", "Average")) %>%
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') +
xlab('Time') +
ylab('Uninsured Rate')
```

```
## Warning: Removed 25 rows containing missing values (`geom_point()`).
```

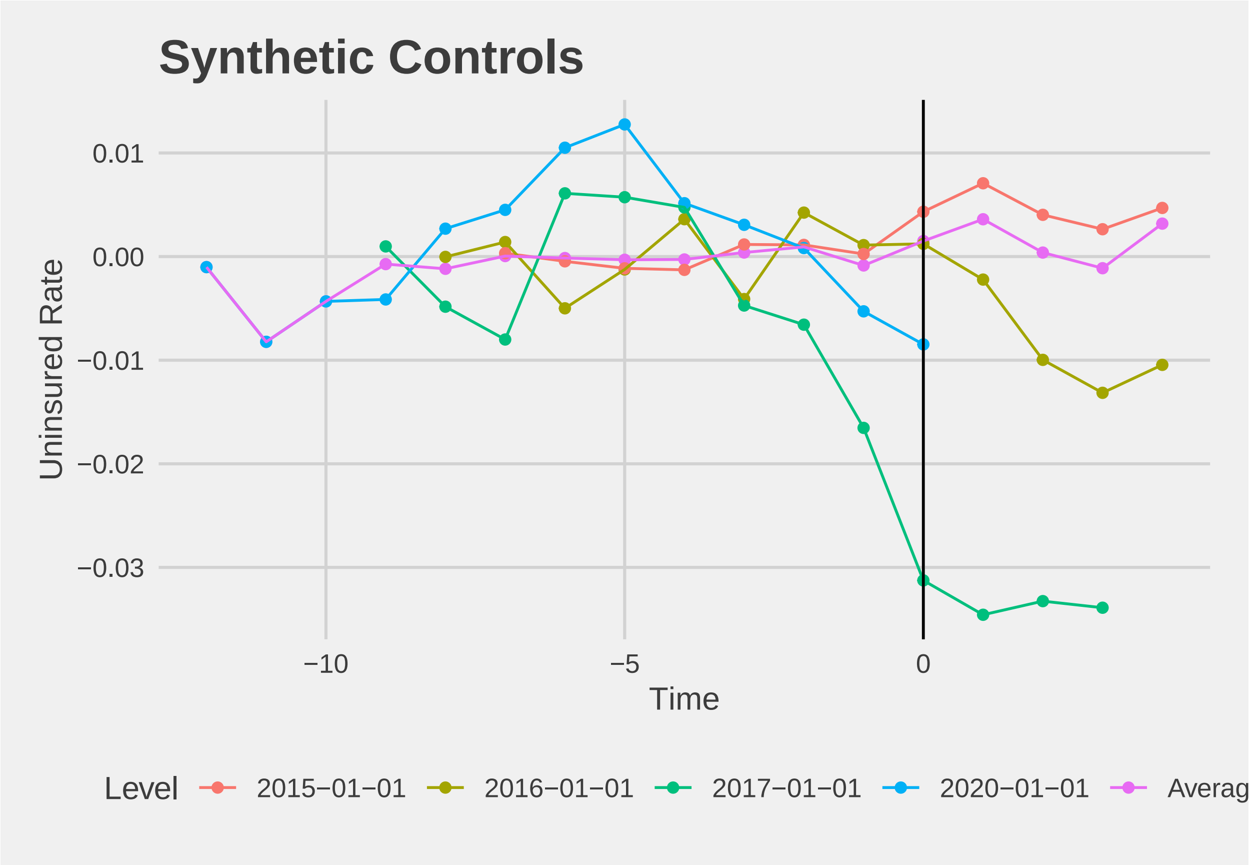
```
## Warning: Removed 25 rows containing missing values (`geom_line()`).
```



- 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
syn_multi_cohort <- multisynth(uninsured_rate ~ expanded, State, year, medicaid_expansion_clean,
                        progfunc = "None", scm = T, n_leads = 5, time_cohort = TRUE)


summary(syn_multi_cohort)$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') +
xlab('Time') +
ylab('Uninsured Rate')
```

```
## Warning: Removed 22 rows containing missing values (`geom_point()`).

## Warning: Removed 22 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**: I see considerable evidence that there was heterogeneity in treatment effects – indeed, I see effects of different signs, with New York reporting an increase in the uninsured whereas California and New Mexico report decreases in uninsurance rates.
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: Unfortunately I did not see that evidence, I think I must have mis-specified my model or else incorrectly restricted control candidates.

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?
- **Answer**: DiD and synthetic control methods are well suited to studies of aggregated units because we can much more easily assume homogeneity in treatment effects (important for DiD) in the context of

states than in the context of individuals. Likewise, cities, states, and other aggregated units having parallel trends is an easier argument that arguing that many individuals have parallel trends.

- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?
- **Answer**: In RD designs, the assumption is that individuals or groups on either side of a pre-determined threshold are similar in all relevant aspects except for their proximity to the threshold, which determines selection into treatment. Therefore, selection into treatment is explicitly taken into account in the design, and the estimates are based on the assumption that treatment status is as good as random around the threshold. In DiD/SCM, the assumption is that the treatment and control groups are similar in all relevant aspects except for the treatment itself. Therefore, selection into treatment is a potential confounder that isn't accounted for formally.