Project 7: Difference-in-Differences and Synthetic Control

Cindy Alvarez

```
# 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 chan
ged since last install.
##   Use `force = TRUE` to force installation
```

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

```
## Warning: package 'multiSynth' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```
## Warning: 'BiocManager' not available.  Could not check Bioconductor.
##
## Please use `install.packages('BiocManager')` and then retry.
```

```
## Warning in p_install(package, character.only = TRUE, ...):
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'multiSynth'
## Warning in pacman::p_load(tidyverse, ggthemes, augsynth, gsynth, ggplot2, : Failed to
install/load:
## multiSynth
```

```
#
# chunk options
# ----------------------------------------------------------------
knitr::opts_chunk$set(
  warning = FALSE            # prevents warning from appearing after code chunk
)

# 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

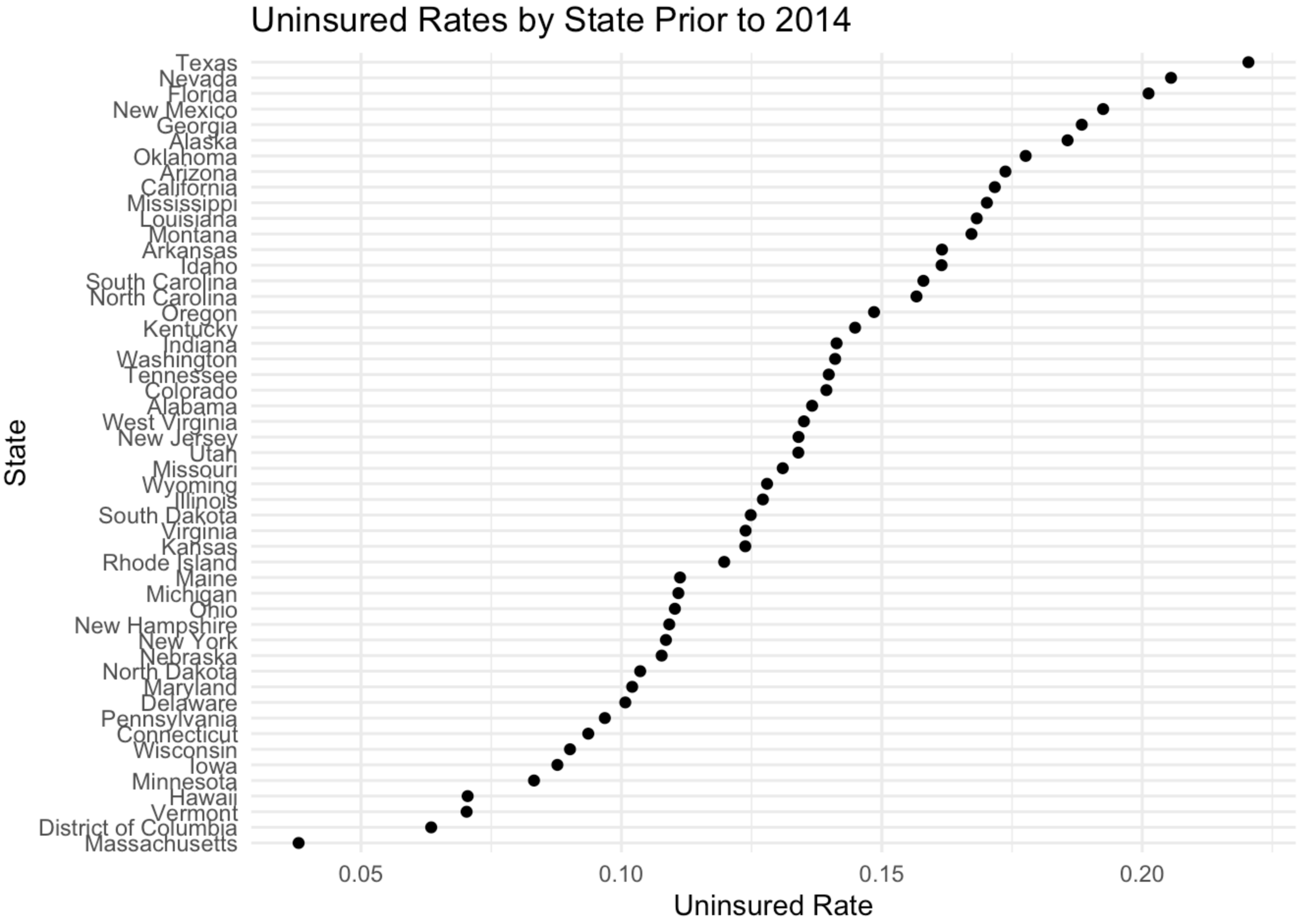
Create plots and provide 1-2 sentence analyses to answer the following questions:

- 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.

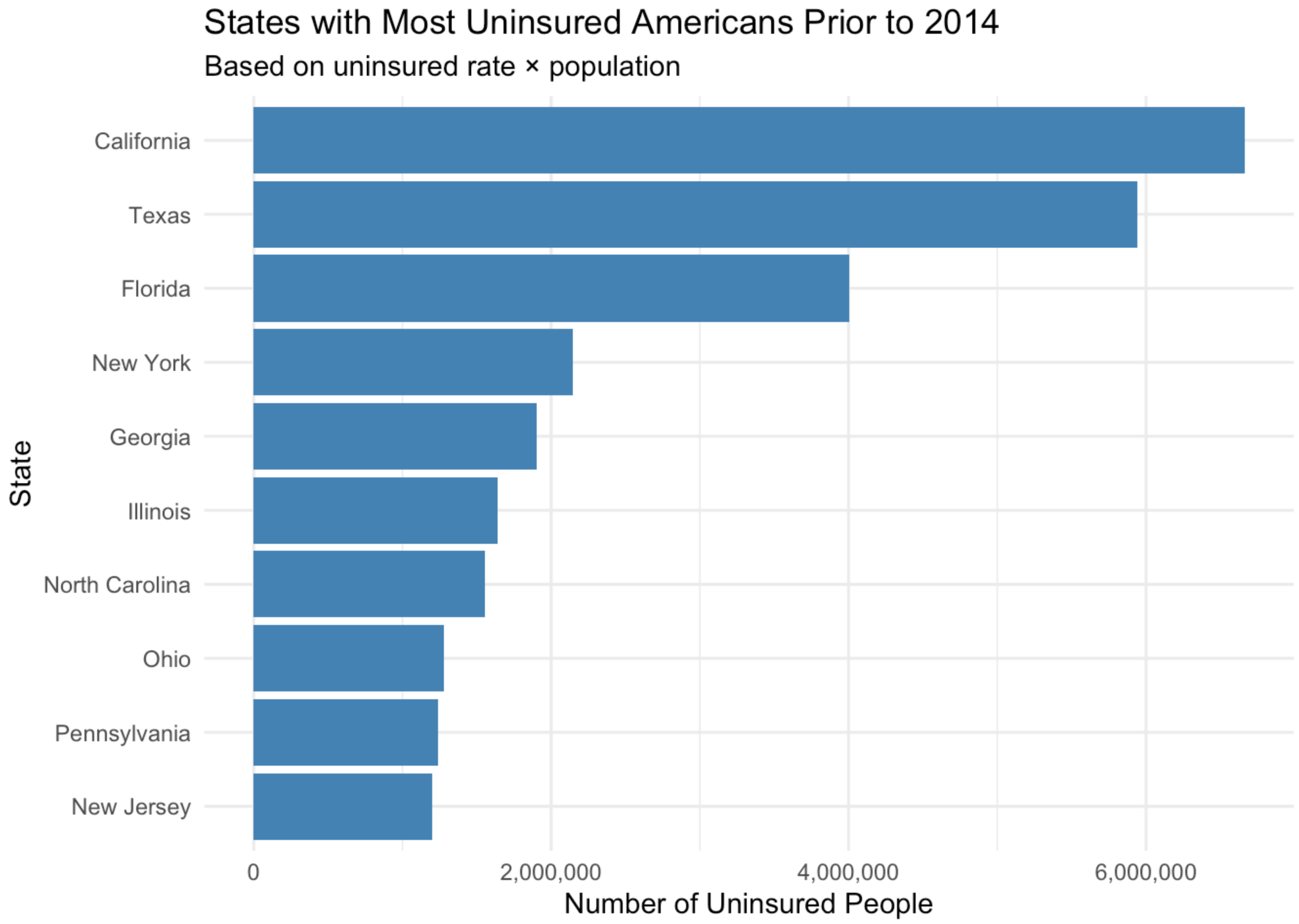
```
# highest and lowest uninsured rates

summary(medicaid_expansion)
```

```
##     State            Date_Adopted             year      uninsured_rate
##  Length:663         Min.   :2014-01-01   Min.   :2008   Min.   :0.02495
##  Class :character   1st Qu.:2014-01-01   1st Qu.:2011   1st Qu.:0.07702
##  Mode  :character   Median :2014-01-01   Median :2014   Median :0.10475
##                     Mean   :2014-11-30   Mean   :2014   Mean   :0.10978
##                     3rd Qu.:2014-09-18   3rd Qu.:2017   3rd Qu.:0.13888
##                     Max.   :2020-10-01   Max.   :2020   Max.   :0.24082
##                     NA's   :195
##    population
##  Min.   :  584153
##  1st Qu.: 1850326
##  Median : 4531566
##  Mean   : 6364343
##  3rd Qu.: 7061530
##  Max.   :38802500
##  NA's   :13
```

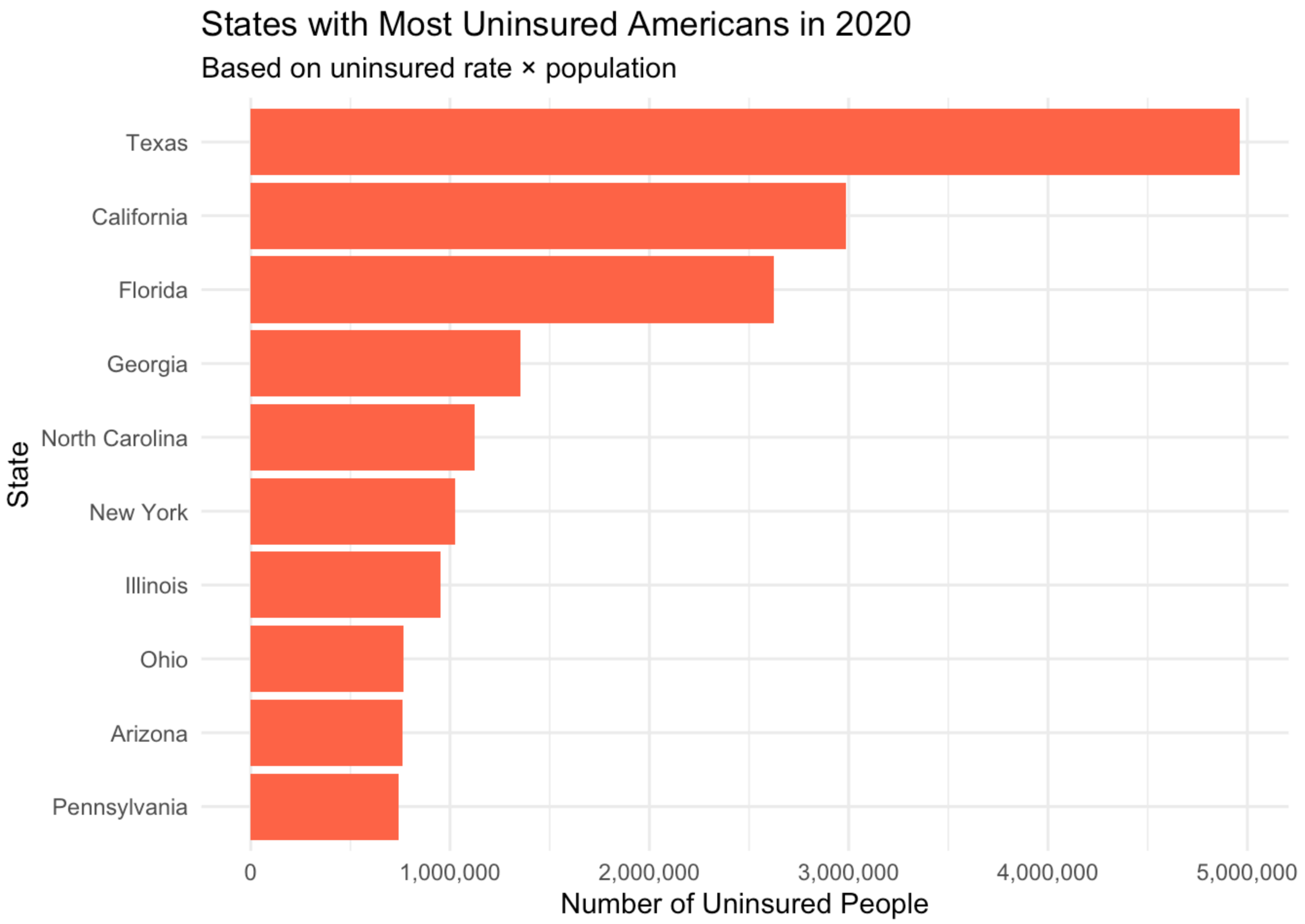
```
# Filter data for years prior to 2014 and find the latest year for each state
pre_2014 <- medicaid_expansion %>%
  filter(year < 2014) %>%
  group_by(State) %>%
  arrange(desc(year)) %>%
  slice(1) %>%
  ungroup()

# Plot states by uninsured rates (prior to 2014)
pre_2014 %>%
  arrange(uninsured_rate) %>%
  mutate(State = factor(State, levels = State)) %>%
  ggplot(aes(x = uninsured_rate, y = State)) +
  geom_point() +
  labs(title = "Uninsured Rates by State Prior to 2014",
       x = "Uninsured Rate",
       y = "State") +
  theme_minimal()
```


Uninsured Rates by State Prior to 2014
State
Texas
Nevada
Florida
New Mexico
Georgia
Alaska
Oklahoma
Arizona
California
Mississippi
Louisiana
Montana
Arkansas
Idaho
South Carolina
North Carolina
Oregon
Kentucky
Indiana
Washington
Tennessee
Colorado
Alabama
West Virginia
New Jersey
Utah
Missouri
Wyoming
Illinois
South Dakota
Virginia
Kansas
Rhode Island
Maine
Michigan
Ohio
New Hampshire
New York
Nebraska
North Dakota
Maryland
Delaware
Pennsylvania
Connecticut
Wisconsin
Iowa
Minnesota
Hawaii
Vermont
District of Columbia
Massachusetts
0.05
0.10
0.15
0.20
Uninsured Rate

```r
# most uninsured Americans

# Calculate the number of uninsured people in each state
# Using the population data consistently for all calculations
medicaid_with_counts <- medicaid_expansion %>%
  mutate(uninsured_count = uninsured_rate * population)

# For pre-2014: Which states were home to most uninsured Americans?
pre_2014_counts <- medicaid_with_counts %>%
  filter(year < 2014) %>%
  group_by(State) %>%
  arrange(desc(year)) %>%
  slice(1) %>%  # Take the most recent year for each state before 2014
  ungroup() %>%
  arrange(desc(uninsured_count)) %>%
  slice_head(n = 10)  # Top 10 states

# For the last year: Which states were home to most uninsured Americans?
last_year <- max(medicaid_expansion$year)
last_year_counts <- medicaid_with_counts %>%
  filter(year == last_year) %>%
  arrange(desc(uninsured_count)) %>%
  slice_head(n = 10)  # Top 10 states

# Create a plot for pre-2014 counts
ggplot(pre_2014_counts, aes(x = reorder(State, uninsured_count), y = uninsured_count)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "States with Most Uninsured Americans Prior to 2014",
       subtitle = "Based on uninsured rate × population",
       x = "State",
       y = "Number of Uninsured People") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



```
# Create a plot for last year counts
ggplot(last_year_counts, aes(x = reorder(State, uninsured_count), y = uninsured_count))
+
  geom_col(fill = "tomato") +
  coord_flip() +
  labs(title = paste("States with Most Uninsured Americans in", last_year),
       subtitle = "Based on uninsured rate × population",
       x = "State",
       y = "Number of Uninsured People") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



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).

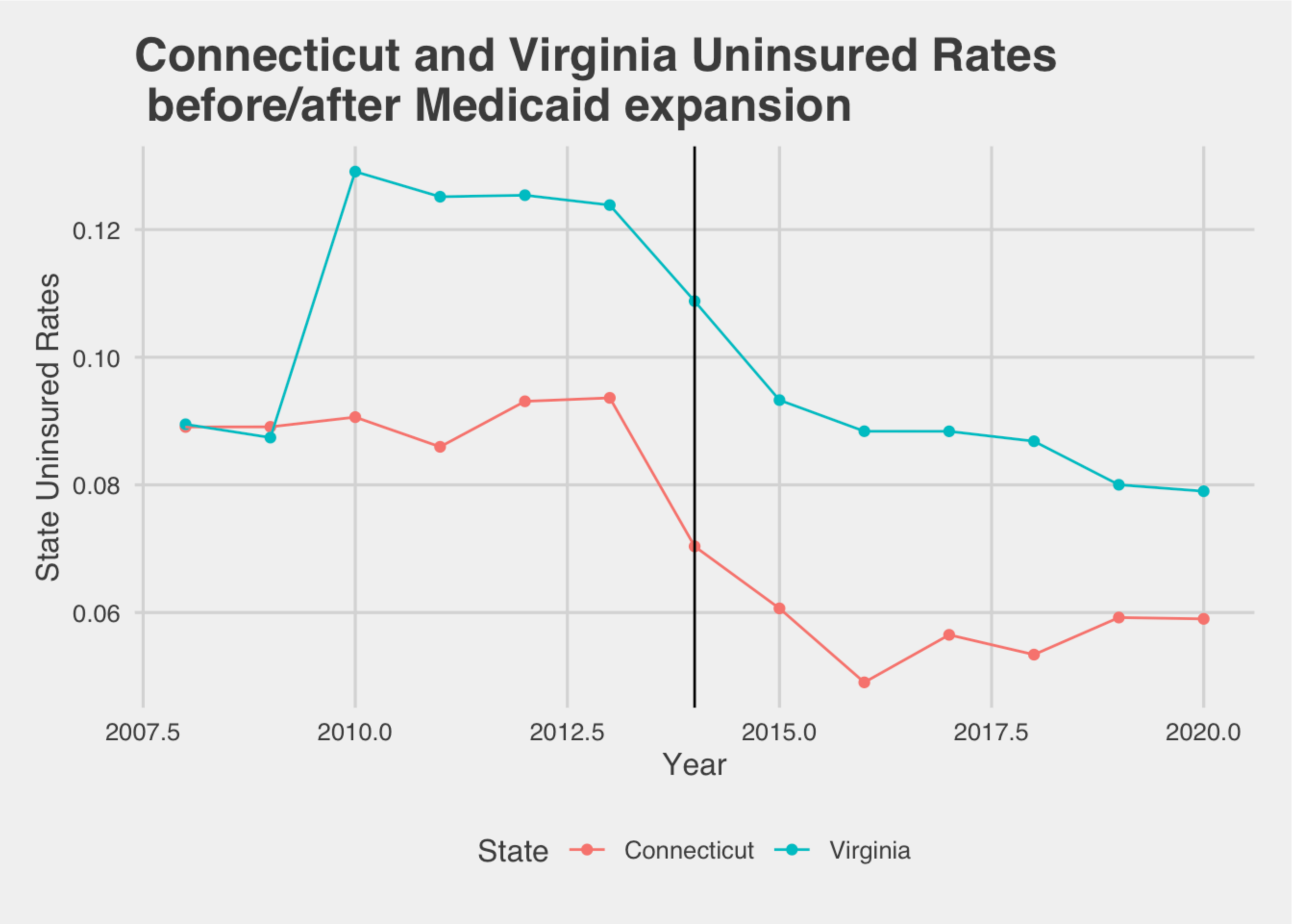
```
# Parallel Trends plot

# Select state that adopted Medicaid expansion on 1/1/2014
medicaid_expansion %>%
  filter(Date_Adopted == "2014-01-01") # filter states to those that adopted on 1/1/2014
```

```
## # A tibble: 325 × 5
##    State                Date_Adopted  year uninsured_rate population
##    <chr>                <date>       <dbl>          <dbl>      <dbl>
##  1 Arizona              2014-01-01    2008         0.187     6731484
##  2 Arkansas             2014-01-01    2008         0.179     2994079
##  3 California           2014-01-01    2008         0.178    38802500
##  4 Colorado             2014-01-01    2008         0.170     5355856
##  5 Connecticut          2014-01-01    2008         0.0891    3596677
##  6 Delaware             2014-01-01    2008         0.108      935614
##  7 District of Columbia 2014-01-01    2008         0.0805         NA
##  8 Hawaii               2014-01-01    2008         0.0672    1419561
##  9 Illinois             2014-01-01    2008         0.129    12880580
## 10 Iowa                 2014-01-01    2008         0.0873    3107126
## # ℹ 315 more rows
```

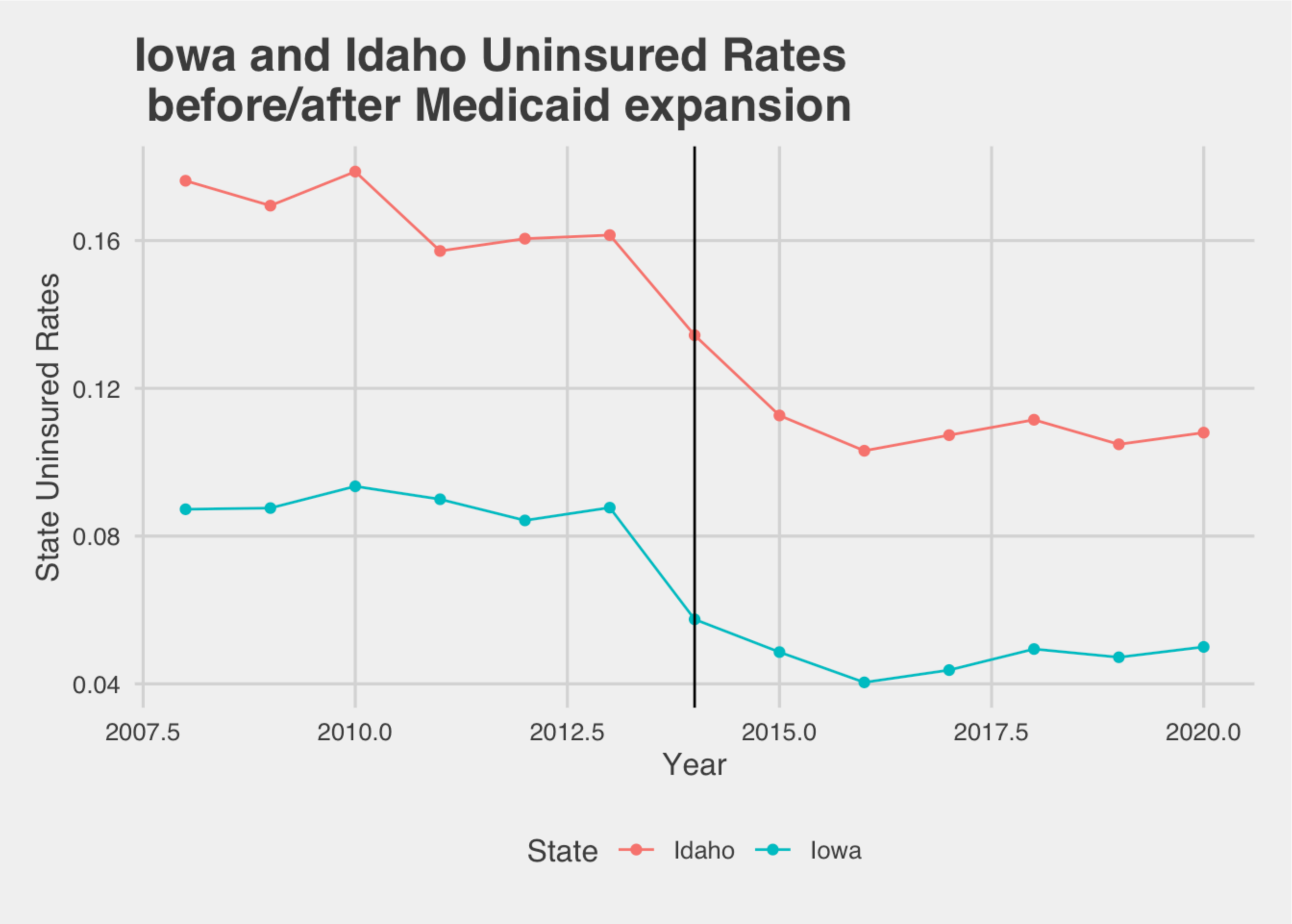
```
# Connecticut will be my first treatment state, adopting Medicaid on 1/1/2014

# Select state that did not adopt Medicaid expansion on 1/1/2014 (or between 2014 and 20
15)

medicaid_expansion %>%
  filter(Date_Adopted != "2014-01-01")
```

```
## # A tibble: 143 × 5
##    State         Date_Adopted  year uninsured_rate population
##    <chr>         <date>       <dbl>          <dbl>      <dbl>
##  1 Alaska        2015-09-01    2008         0.208      737732
##  2 Idaho         2020-01-01    2008         0.176     1634464
##  3 Indiana       2015-02-01    2008         0.138     6596855
##  4 Louisiana     2016-07-01    2008         0.179     4649676
##  5 Michigan      2014-04-01    2008         0.115     9909877
##  6 Montana       2016-01-01    2008         0.189     1023579
##  7 Nebraska      2020-10-01    2008         0.108     1881503
##  8 New Hampshire 2014-08-15    2008         0.110     1326813
##  9 Pennsylvania  2015-01-01    2008         0.0960   12787209
## 10 Utah          2020-01-01    2008         0.241     2942902
## # ℹ 133 more rows
```

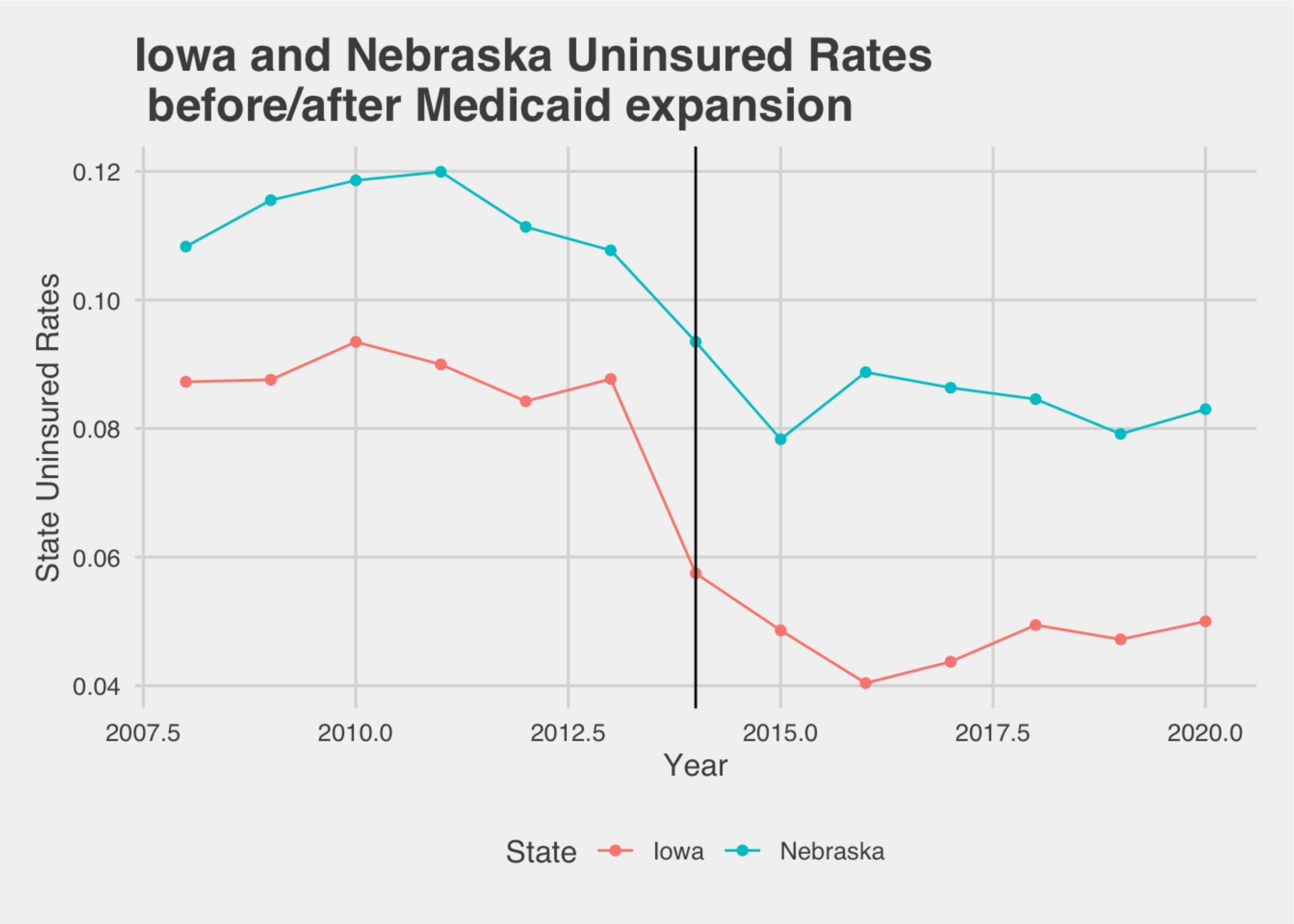
```
# Virginia will be my first control state, it adopted on 1/1/2019

medicaid_expansion %>%

  # process
  # ---------
  filter(State %in% c("Connecticut","Virginia")) %>%
  # plotting all of the time periods -- not filtering out any of them

  # plot
  # ---------
  ggplot() +
  # add in point layer
  geom_point(aes(x = year,
               y = uninsured_rate,
               color = State)) +
  # add in line layer
  geom_line(aes(x = year,
              y = uninsured_rate,
              color = State)) +
  # add a horizontal line
  geom_vline(aes(xintercept = 2014)) +

  # themes
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +

  # labels
  ggtitle('Connecticut and Virginia Uninsured Rates \n before/after Medicaid expansion')
+
  xlab('Year') +
  ylab('State Uninsured Rates')
```


Connecticut and Virginia Uninsured Rates
before/after Medicaid expansion
State Uninsured Rates
0.12
0.10
0.08
0.06
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
Year
State
Connecticut
Virginia

```
# Parallel trends between Connecticut and Virginia were not satisfied. Selecting a new p
air of states to try.

# Iowa and Idaho:

medicaid_expansion %>%

  # process
  # ---------
  filter(State %in% c("Iowa","Idaho")) %>%
  # plotting all of the time periods -- not filtering out any of them

  # plot
  # ---------
  ggplot() +
  # add in point layer
  geom_point(aes(x = year,
               y = uninsured_rate,
               color = State)) +
  # add in line layer
  geom_line(aes(x = year,
              y = uninsured_rate,
              color = State)) +
  # add a horizontal line
  geom_vline(aes(xintercept = 2014)) +

  # themes
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +

  # labels
  ggtitle('Iowa and Idaho Uninsured Rates \n before/after Medicaid expansion') +
  xlab('Year') +
  ylab('State Uninsured Rates')
```


Iowa and Idaho Uninsured Rates
before/after Medicaid expansion
State Uninsured Rates
0.16
0.12
0.08
0.04
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
Year
State
Idaho
Iowa

```
# Parallel trends between Iowa and Idaho also don't seem to be satisfied, because VA uni
nsured rates increased a lot in 2010. I am also concerned (as we were in our lab) by the
decrease in both state's trends before treatment. This would potentially violate the no
treatment anticipation assumption, so I will try another pair.

# Iowa and Nebraska:

medicaid_expansion %>%

  # process
  # ---------
  filter(State %in% c("Iowa","Nebraska")) %>%

  # plot
  # ---------
  ggplot() +
  # add in point layer
  geom_point(aes(x = year,
                 y = uninsured_rate,
                 color = State)) +
  # add in line layer
  geom_line(aes(x = year,
                y = uninsured_rate,
                color = State)) +
  # add a horizontal line
  geom_vline(aes(xintercept = 2014)) +

  # themes
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +

  # labels
  ggtitle('Iowa and Nebraska Uninsured Rates \n before/after Medicaid expansion') +
  xlab('Year') +
  ylab('State Uninsured Rates')
```



```
# Iowa and Nebraska look pretty good in terms of parallel trends but as in our lab, I am
a bit worried about the drop in uninsured rates that I see leading up to Medicaid expans
ion (treatment) because this suggests some potential anticipation.
```

- 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

# Filter data for Iowa (treatment) and Nebraska (control)
did_data <- medicaid_expansion %>%
  filter(State %in% c("Iowa", "Nebraska"))

# Create variables for DiD analysis
did_data <- did_data %>%
  mutate(
    # Treatment indicator (1 for Iowa, 0 for Nebraska)
    treatment_i = ifelse(State == "Iowa", 1, 0),

    # Post-treatment period indicator (1 for years >= 2014, 0 otherwise)
    post = ifelse(year >= 2014, 1, 0),

    # Interaction term for DiD (treatment × post)
    treatment_i_post = treatment_i * post
  )

# Run the DiD regression model
did_model <- lm(uninsured_rate ~ treatment_i + post + treatment_i_post, data = did_data)

# View the regression results
summary(did_model)
```

```
##
## Call:
## lm(formula = uninsured_rate ~ treatment_i + post + treatment_i_post,
##     data = did_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.007719 -0.003651 -0.000455  0.001933  0.009353
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.113572   0.002000  56.796  < 2e-16 ***
## treatment_i      -0.025201   0.002828  -8.911 9.41e-09 ***
## post             -0.028766   0.002725 -10.556 4.47e-10 ***
## treatment_i_post -0.011488   0.003854  -2.981  0.00689 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004898 on 22 degrees of freedom
## Multiple R-squared:  0.9644, Adjusted R-squared:  0.9596
## F-statistic: 198.8 on 3 and 22 DF,  p-value: 4.385e-16
```

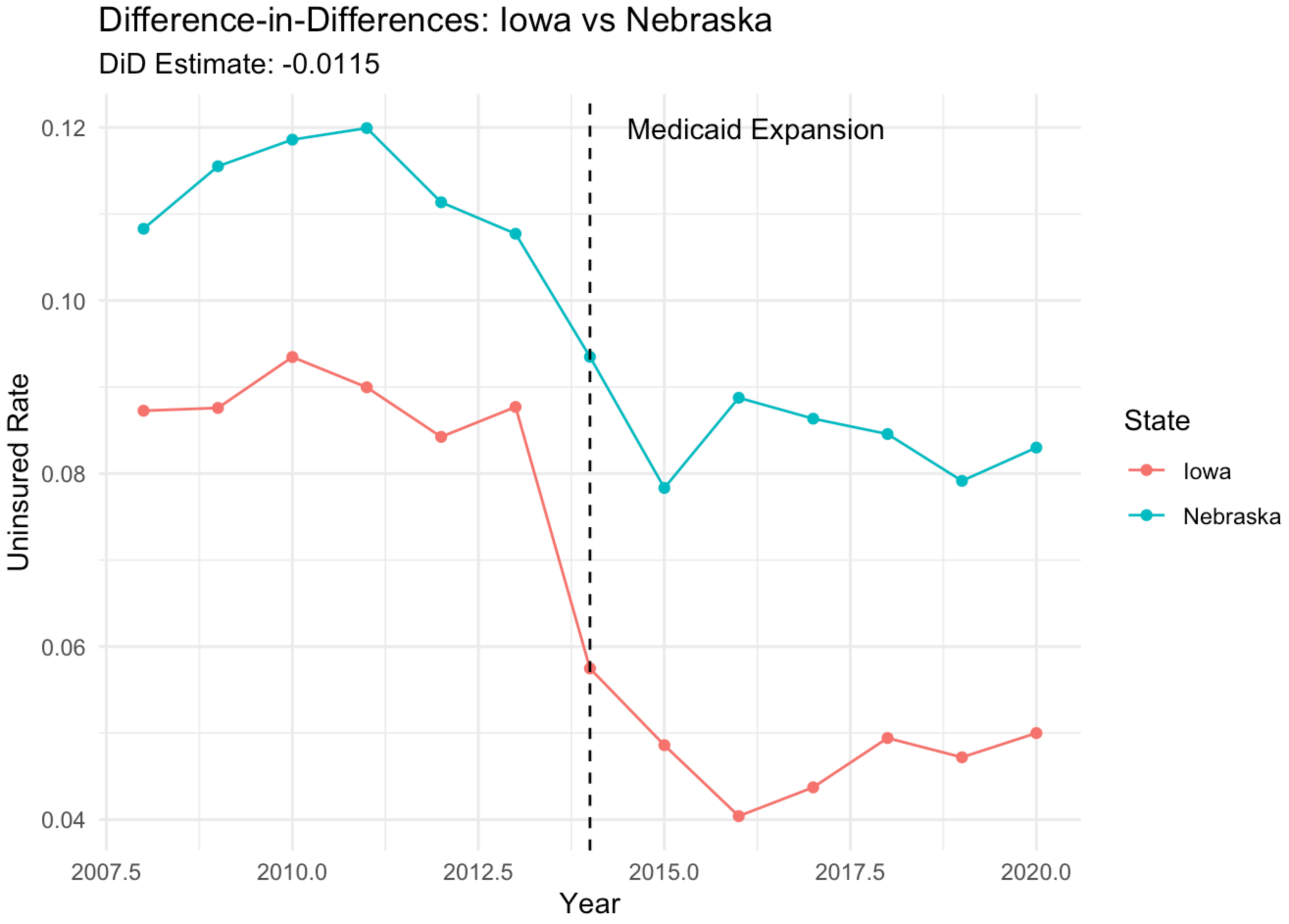
```
# Calculate means for each group and time period
means <- did_data %>%
  group_by(treatment_i, post) %>%
  summarize(mean_uninsured = mean(uninsured_rate, na.rm = TRUE), .groups = "drop")

# Print means table
print(means)
```

```
## # A tibble: 4 × 3
##   treatment_i  post mean_uninsured
##         <dbl> <dbl>          <dbl>
## 1           0     0         0.114
## 2           0     1         0.0848
## 3           1     0         0.0884
## 4           1     1         0.0481
```

```
# DiD calculation to verify regression results
# (Treatment post - Treatment pre) - (Control post - Control pre)
treated_pre <- means %>% filter(treatment_i == 1, post == 0) %>% pull(mean_uninsured)
treated_post <- means %>% filter(treatment_i == 1, post == 1) %>% pull(mean_uninsured)
control_pre <- means %>% filter(treatment_i == 0, post == 0) %>% pull(mean_uninsured)
control_post <- means %>% filter(treatment_i == 0, post == 1) %>% pull(mean_uninsured)

# Calculate DiD estimate manually (like our lab)
did_estimate <- (treated_post - treated_pre) - (control_post - control_pre)
print(paste("DiD Estimate:", round(did_estimate, 4)))
```

```
## [1] "DiD Estimate: -0.0115"
```

```
# Visualize the DiD result
ggplot(did_data, aes(x = year, y = uninsured_rate, color = State)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 2014, linetype = "dashed") +
  annotate("text", x = 2014.5, y = max(did_data$uninsured_rate),
           label = "Medicaid Expansion", hjust = 0) +
  labs(title = "Difference-in-Differences: Iowa vs Nebraska",
       subtitle = paste("DiD Estimate:", round(did_estimate, 4)),
       x = "Year",
       y = "Uninsured Rate") +
  theme_minimal()
```



Discussion Questions

- Card/Krueger's original piece utilized the fact that towns on either side of the Delaware river are likely to be quite similar to one another in terms of demographics, economics, etc. Why is that intuition harder to replicate with this data?

- **Answer**:This application is different from the minimum wage study because states are the unit of analysis, and states are likely less homogenous than neighboring towns (for example, in policies, demographics, economics, etc.). In this example, states also chose whether to expand Medicaid (often based on political leanings as described in the background information provided for this project), creating selection bias. In the minimum wage study, on the other hand, the researchers benefited from having an exogenous policy change that affected only one side of the river. States also adopted expansion at different times rather than simultaneously.

- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?

- **Answer**:In terms of strengths, teh parallel trends assumption helps us compare what actually happened to what would have happened without a policy change by looking at similar groups. It's simple to explain with visual graphs showing before-and-after comparisons. We can check if the groups were moving in similar directions before the policy to build confidence in our findings. However, the challenge is we can never truly know if the comparison group perfectly represents what would have happened without the policy change. So, the method is less believable if other important events or policy changes happen at the same time as the intervention. Results can also change depending on which comparison group is chosen, making findings feel less reliable.

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.

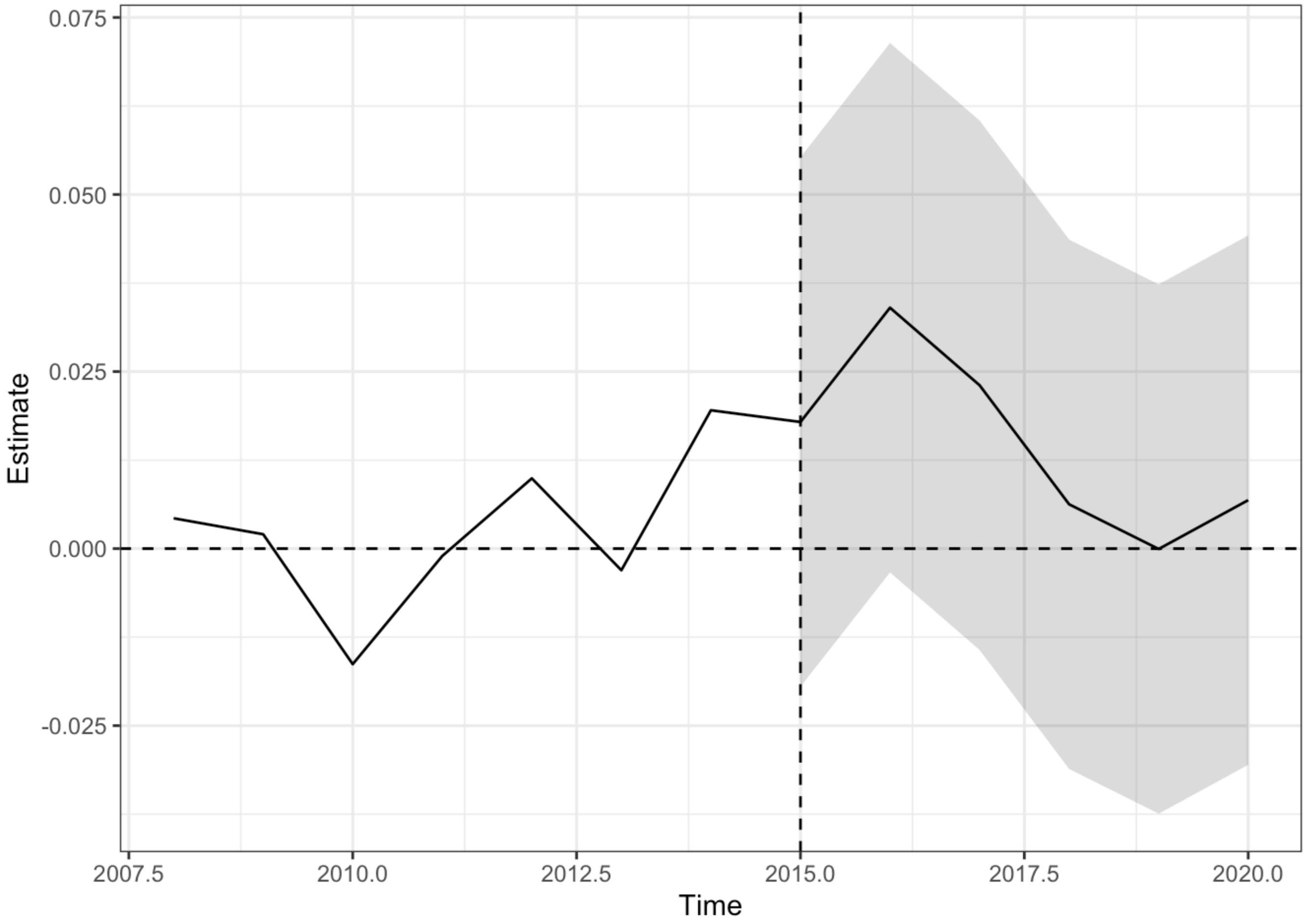
```
# non-augmented synthetic control

# Alaska expanded in Sept 2015

# Preprocessing for synthetic control analysis
data_for_synth <- medicaid_expansion %>%
  # Create treatment and time variables
  mutate(
    # For Alaska treatment
    treat = ifelse(State == "Alaska", 1, 0),
    # Create treatment time variable (Alaska expanded in 2015)
    post = ifelse(year >= 2015, 1, 0)
  ) %>%
  # Filter out states that expanded on Jan 1, 2014 from donor pool
  filter(is.na(Date_Adopted) | year(Date_Adopted) > 2014 | State == "Alaska")

# Run synthetic control model
syn_control <- augsynth(
  form = uninsured_rate ~ treat,
  unit = State,
  time = year,
  t_int = 2015,
  data = data_for_synth,
  progfunc = "None", # Plain synthetic control/non-augmented
  scm = TRUE
)
```

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

```
# View summary statistics - ATT and L2 imbalance
summary(syn_control)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "None", scm = TRUE)
##
## Average ATT Estimate (p Value for Joint Null):  0.0147   ( 0.18 )
## L2 Imbalance: 0.028
## Percent improvement from uniform weights: 78.1%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2015    0.018             -0.019              0.055   0.489
##  2016    0.034             -0.003              0.071   0.381
##  2017    0.023             -0.014              0.060   0.501
##  2018    0.006             -0.031              0.044   0.741
##  2019    0.000             -0.037              0.037   1.000
##  2020    0.007             -0.031              0.044   0.744
```

```
# Plot pre-treatment fit and post-treatment differences
plot(syn_control)
```



Above, we see that the Average ATT Estimate is 0.0147 and the L2 Imbalance is: 0.028.

- 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.

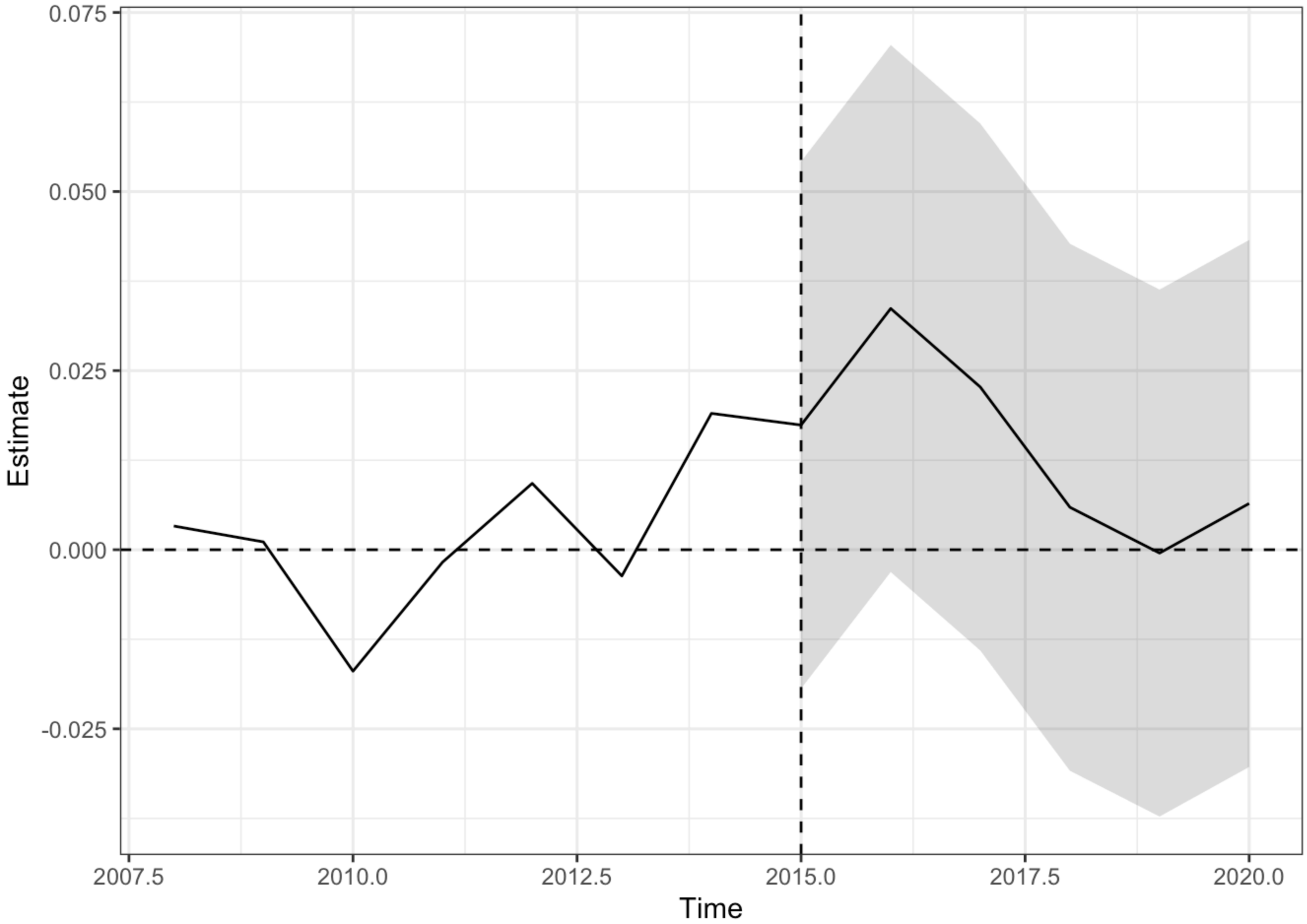
```
# augmented synthetic control

# Using the same Alaska example but with Ridge augmentation

aug_syn_control <- augsynth(
  form = uninsured_rate ~ treat,
  unit = State,
  time = year,
  t_int = 2015,
  data = data_for_synth,
  progfunc = "Ridge", # Ridge augmentation
  scm = TRUE
)
```

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

```
# View summary statistics - ATT and L2 imbalance
summary(aug_syn_control)
```

```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = "Ridge", scm = TRUE)
##
## Average ATT Estimate (p Value for Joint Null):  0.0143   ( 0.19 )
## L2 Imbalance: 0.028
## Percent improvement from uniform weights: 78.3%
##
## Avg Estimated Bias: 0.000
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2015    0.017             -0.019              0.054   0.469
##  2016    0.034             -0.003              0.070   0.388
##  2017    0.023             -0.014              0.059   0.517
##  2018    0.006             -0.031              0.043   0.730
##  2019    0.000             -0.037              0.036   1.000
##  2020    0.006             -0.030              0.043   0.753
```

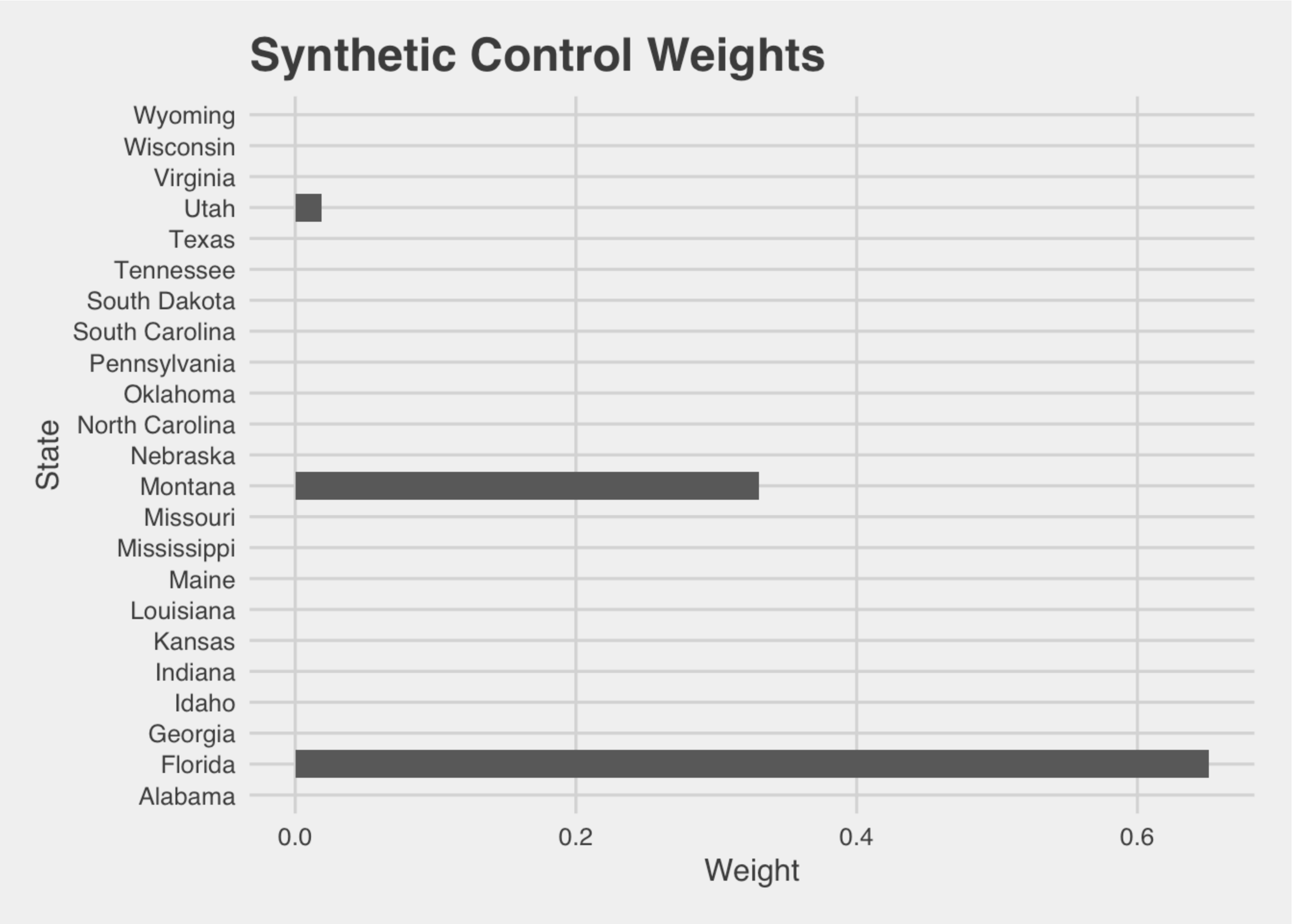
```
# Plot pre-treatment fit and post-treatment differences
plot(aug_syn_control)
```



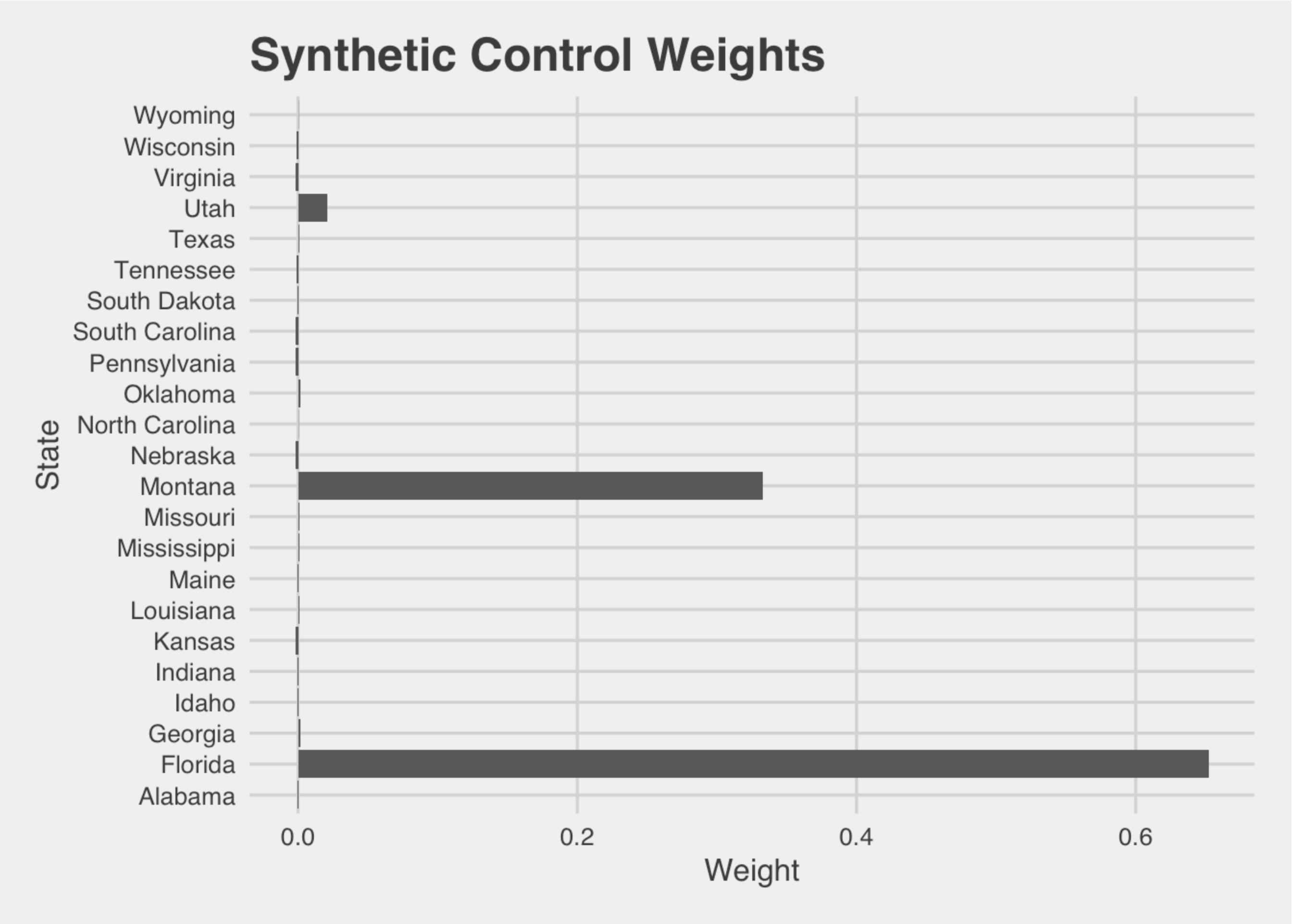
Above, we see that the Average ATT Estimate is 0.0143 and the L2 Imbalance is: 0.028.

- Plot barplots to visualize the weights of the donors.

```
# barplots of weights

# Non-augmented weights

# Extract donor weights for the non-augmented model
data.frame(syn_control$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 in Python)
  # plot
  # ---------
  ggplot() +
  # stat = identity to take the literal value instead of a count for geom_bar()
  geom_bar(aes(x = State,
               y = syn_control.weights),
           stat = 'identity') +  # override count() which is default of geom_bar(), could use geom_col() instead
  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')
```



```
# Augmented weights
data.frame(aug_syn_control$weights) %>%
  tibble::rownames_to_column('State') %>%
  ggplot() +
  geom_bar(aes(x = State, y = aug_syn_control.weights),
           stat = 'identity') +
  coord_flip() + # coord flip
  theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  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 advantageous over a diff-in-diff approach when we only have one treated unit or when the parallel trends assumption seems questionable because it essentially creates a weighted combination of control units to better match the pre-treatment trajectory of the treated unit. However, it might not make sense to use synthetic control if you have few good matches or poor pre-treatment fit. It also requires that we have a long pre-treatment period so we can create a good synthetic control. DiD is simpler and better when we have multiple treated and control units and clear parallel trends.
- 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**: Yes, augmentation introduces some interpretation challenges because the fact that control units can contribute negatively makes it harder to conceptualize the “counterfactual.” I would use only augmentation if I find that traditional synthetic control leads to poor pre-treatment fit. After doing so, I would also check teh magnitude and distribution of negative weights because having small or few

negatives might be more acceptable than if I had many large negative weights. If I were trying to publish a paper using augmented syntehtic control methods, I would report both augmented and traditional estimates to show how sensitive my findings are.

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.

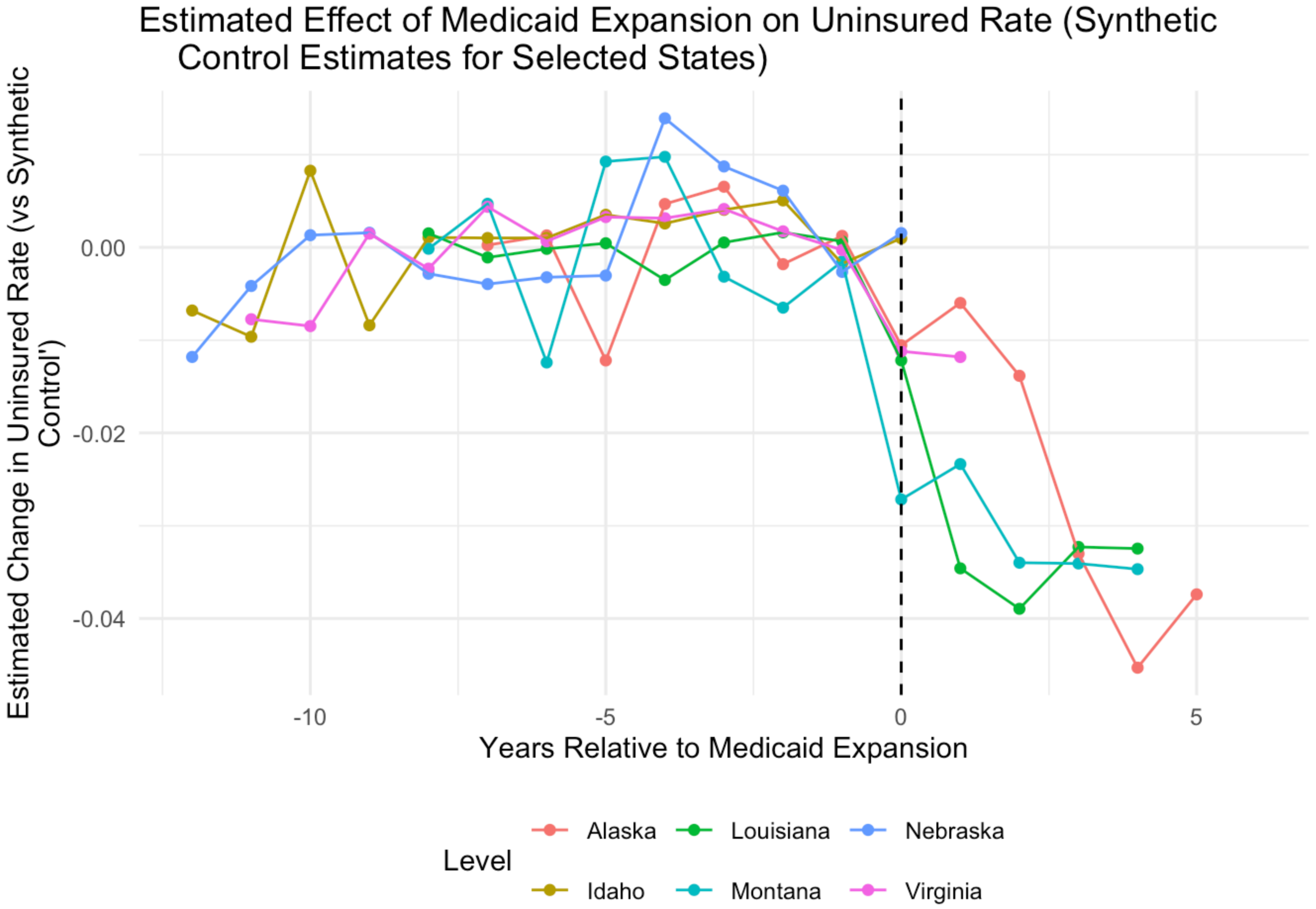
```
# multisynth model states

# Create a 'treated' indicator by state-year
medicaid_expansion <- medicaid_expansion %>%
  group_by(State) %>%
  mutate(treated = ifelse(!is.na(Date_Adopted) & year >= year(Date_Adopted), 1, 0)) %>%
  ungroup()

# Drop always-treated units (no pre-treatment observations)
pre_treat_counts <- medicaid_expansion %>%
  group_by(State) %>%
  summarize(pre_treat_obs = sum(treated == 0))

valid_states <- pre_treat_counts %>%
  filter(pre_treat_obs >= 1) %>%
  pull(State)

cleaned_data <- medicaid_expansion %>%
  filter(State %in% valid_states)

# Run model by state
multi_state_model <- multisynth(
  form = uninsured_rate ~ treated,
  unit = State,
  time = year,
  data = cleaned_data,
  n_leads = 10
)

# print results

print(multi_state_model$nu)
```

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

```
multi_state_model
```

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

```
# Extract the state-specific ATTs
multi_state_model_summ <- summary(multi_state_model)

# extract dataframe
att_df <- multi_state_model_summ$att

# Choose a subset of states for visibility
subset_states <- c("Alaska", "Montana", "Louisiana", "Virginia", "Idaho", "Nebraska")

# Filter the ATT dataframe to just those states
att_subset_df <- att_df %>%
  filter(Level %in% subset_states)

att_subset_df %>%
  ggplot(aes(x = Time, y = Estimate, color = Level)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
  theme_minimal() +
  theme(
    legend.position = "bottom"
  ) +
  labs(
    title = "Estimated Effect of Medicaid Expansion on Uninsured Rate (Synthetic
    Control Estimates for Selected States)",
  x = "Years Relative to Medicaid Expansion",
  y = "Estimated Change in Uninsured Rate (vs Synthetic\nControl')"
  )
```

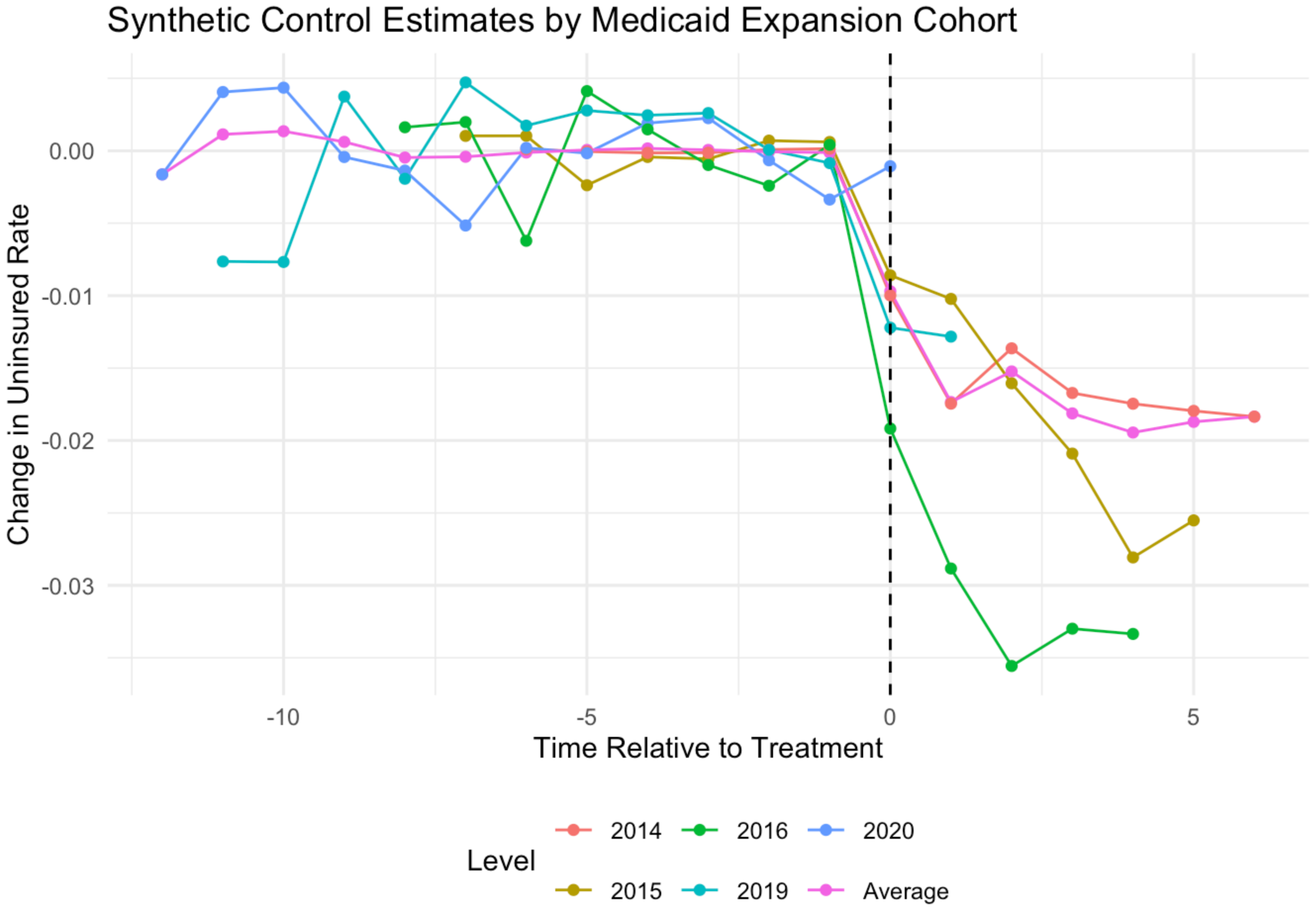


- Estimate a multisynth model using time cohorts. For the purpose of this exercise, you can simplify the treatment time so that states that adopted Medicaid expansion within the same year (i.e. all states that adopted epxansion in 2016) count for the same cohort. Plot the treatment effects for these time cohorts.

```r
# multisynth model time cohorts

# Create cohort and treatment indicator
medicaid_expansion <- medicaid_expansion %>%
  mutate(
    cohort = year(Date_Adopted),
    treated = ifelse(!is.na(Date_Adopted) & year >= year(Date_Adopted), 1, 0)
  )

# Drop always-treated (no pre-treatment)
valid_states <- medicaid_expansion %>%
  group_by(State) %>%
  summarize(pre_treat_obs = sum(treated == 0)) %>%
  filter(pre_treat_obs > 0) %>%
  pull(State)

cohort_data <- medicaid_expansion %>%
  filter(State %in% valid_states)

# Estimate model with time cohorts
cohort_model <- multisynth(uninsured_rate ~ treated,
  State,
  year,
  cohort_data,
  n_leads = 10,
  time_cohort = TRUE
)

# Summarize results
cohort_model_summ <- summary(cohort_model)
print(cohort_model_summ)
```

```
## 
## Call:
## multisynth(form = uninsured_rate ~ treated, unit = State, time = year, 
##     data = cohort_data, n_leads = 10, time_cohort = TRUE)
## 
## Average ATT Estimate (Std. Error): -0.016  (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: 0.015
## Percent improvement from uniform individual weights: 98.5
## 
##  Time Since Treatment   Level     Estimate   Std.Error lower_bound  upper_bound
##                     0 Average -0.009699016 0.004314594 -0.01872530 -0.001948973
##                     1 Average -0.017346723 0.005819906 -0.02894032 -0.006084722
##                     2 Average -0.015234130 0.005871422 -0.02696368 -0.003760760
##                     3 Average -0.018128760 0.006177458 -0.03077496 -0.006252357
##                     4 Average -0.019448849 0.006007999 -0.03152522 -0.008130039
##                     5 Average -0.018711810 0.005729399 -0.02956803 -0.007915411
##                     6 Average -0.018354004 0.006233672 -0.03055281 -0.006306519
```

```
# Plot
# Extract ATT data
att_df <- cohort_model_summ$att

# Plot
ggplot(att_df, aes(x = Time, y = Estimate, color = Level)) +
  geom_line() +
  geom_point() +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_minimal() +
  labs(
    title = "Synthetic Control Estimates by Medicaid Expansion Cohort",
    x = "Time Relative to Treatment",
    y = "Change in Uninsured Rate"
  ) +
  theme(legend.position = "bottom")
```



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 first graph above suggests that different states experienced different treatment effect sizes. After the treatment (post-year 0), the magnitude of the reduction in the uninsured rate varies by state. So, some states show a steep and sustained decline, while others show a more modest drop. This variation supports the idea that states implemented Medicaid expansion differently (for example, potentially due to administrative choices like waivers for work requirements or varying outreach and enrollment efforts).
- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: The second graph above shows some evidence that earlier adopters may have experienced a larger or more sustained decrease in the uninsured rate. Some of the lines that diverge most sharply from zero post-treatment are the 2014-2016 lines, but it is a bit challenging to gauge the effect on later years (2020, for example) because the post treatment time is not available for that cohort (we only had data through 2020). I would not feel confident making this explicit claim unless I had more cohorts/years of data - especially since the 2019 cohort shows a steeper drop than the 2015 cohort.

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?
- **Answer**: They are well suited because they are meant to estimate causal effects when RCTs are not feasible. Aggregated units like states or countries often have limited sample sizes, complex policies, and diverse characteristics. DiD allows us to use variation over time and between treated and untreated groups to control for confounders. Synthetic control methods go further by constructing a weighted combination of untreated units that more closely approximates the treated unit's pre-treatment trajectory, offering better counterfactual estimates when units are heterogeneous.
- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?
- **Answer**: Selection into treatment is a concern in DiD and synthetic control because units are not randomly assigned. These methods assume that, in the absence of treatment, treated and control units would have followed similar trends (parallel trends assumption in DiD; similar pre-treatment trajectories in synthetic control). Regression discontinuity (RD), on the other hand, relies on a cutoff (like an income threshold or age limit) where treatment assignment is as good as random around the threshold. If treatment is self-selected or influenced by unobserved factors, RD might be better/ more credible—but only near the cutoff. I would choose RD when there's a clear rule-based assignment, and DiD/synthetic control when there is variation across time and units without a sharp cutoff.