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')
medicaid_expansion <- read_csv('/Users/saadaamadu/git/CSS2/CSS Forked/Projects/Project 7/data/medicaid_e
```

```
## 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.
```

```
glimpse(medicaid_expansion)
```

```
## Rows: 663
## Columns: 5
## $ State          <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California~
## $ Date_Adopted   <date> NA, 2015-09-01, 2014-01-01, 2014-01-01, 2014-01-01, 20~
## $ year           <dbl> 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2~
## $ uninsured_rate <dbl> 0.139716, 0.207716, 0.187312, 0.178883, 0.178212, 0.170~
## $ population     <dbl> 4849377, 737732, 6731484, 2994079, 38802500, 5355856, 3~
```

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:

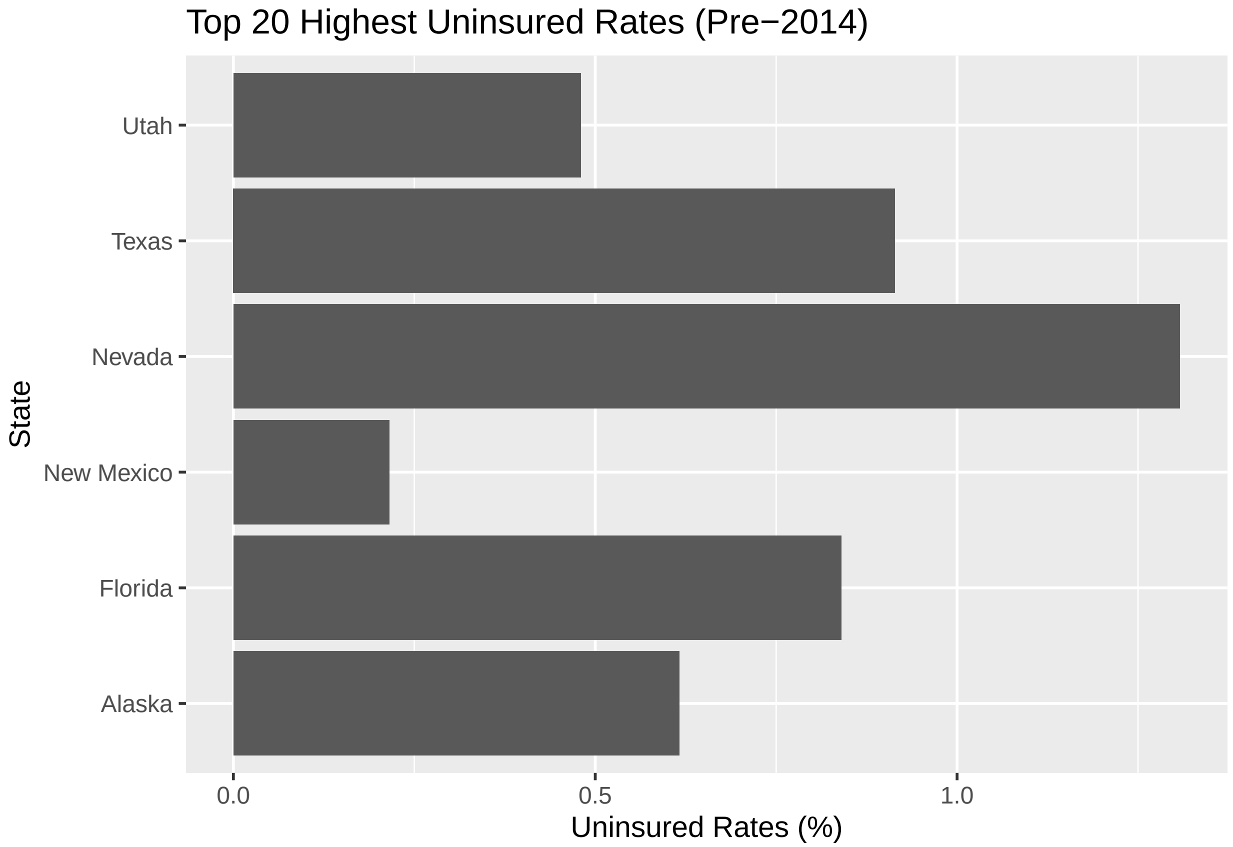
- Which states had the highest uninsured rates prior to 2014? The lowest?

- 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.

```r
pre2014_data <- medicaid_expansion %>%
  filter(year < 2014)
head(pre2014_data)
```

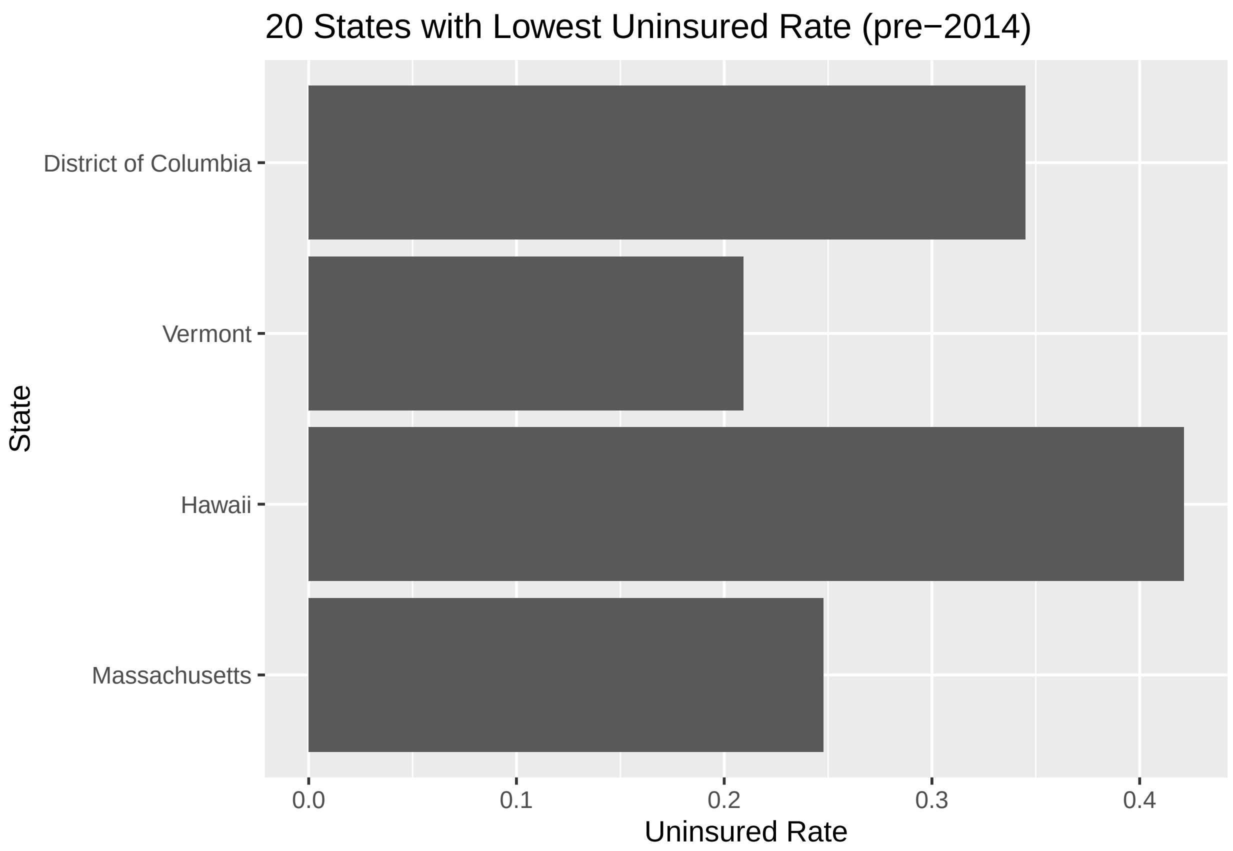
```
## # A tibble: 6 x 5
##   State      Date_Adopted  year uninsured_rate population
##   <chr>      <date>       <dbl>          <dbl>      <dbl>
## 1 Alabama    NA            2008          0.140    4849377
## 2 Alaska     2015-09-01    2008          0.208     737732
## 3 Arizona    2014-01-01    2008          0.187    6731484
## 4 Arkansas   2014-01-01    2008          0.179    2994079
## 5 California 2014-01-01    2008          0.178   38802500
## 6 Colorado   2014-01-01    2008          0.170    5355856
```

```r
library(ggplot2)
# highest and lowest uninsured rates
top20 <- pre2014_data %>%
  arrange(desc(uninsured_rate)) %>%
  head(20)

ggplot(top20, aes(x = reorder(State, uninsured_rate), y = uninsured_rate)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Top 20 Highest Uninsured Rates (Pre-2014)",
    x = "State",
    y = "Uninsured Rates (%)"
  )
```



The states with the highest uninsured rates aren't particularly surprising given what we know: health insurance coverage appears to fall along partisan lines, with Republican-led states such as Texas, Nevada, and Florida exhibiting the largest shares of uninsured residents.

```
bottom20 <- pre2014_data %>%
  slice_min(order_by = uninsured_rate, n = 20)

ggplot(bottom20, aes(
    x = fct_reorder(State, uninsured_rate),
    y = uninsured_rate
  )) +
  geom_col() +
  coord_flip() +
  labs(
    title = "20 States with Lowest Uninsured Rate (pre-2014)",
    x     = "State",
    y     = "Uninsured Rate"
  )
```



The states with the lowest uninsured rates prior to 2014 are hardly surprising, especially given that Massachusetts passed its universal healthcare law in 2006. Reflecting the same partisan patterns, DC, Vermont, and Hawaii also had the smallest shares of uninsured residents.

```
pop2010 <- medicaid_expansion %>%
  filter(year == 2010) %>%
  select(State, pop2010 = population)
```

```
pre2014_data <- pre2014_data %>%
  left_join(pop2010, by = "State")
```

```
# most uninsured Americans
pre2014_data <- pre2014_data %>%
  mutate(uninsured_count = uninsured_rate * pop2010)
head(pre2014_data)
```

```
## # A tibble: 6 x 7
##   State     Date_Adopted  year uninsured_rate population pop2010 uninsured_count
##   <chr>     <date>       <dbl>          <dbl>      <dbl>   <dbl>           <dbl>
## 1 Alabama   NA            2008          0.140    4849377  4.85e6         677536.
## 2 Alaska    2015-09-01    2008          0.208     737732  7.38e5         153239.
## 3 Arizona   2014-01-01    2008          0.187    6731484  6.73e6        1260888.
## 4 Arkansas  2014-01-01    2008          0.179    2994079  2.99e6         535590.
## 5 Californ~ 2014-01-01    2008          0.178   38802500  3.88e7        6915071.
## 6 Colorado  2014-01-01    2008          0.170    5355856  5.36e6         911476.
```

```r
state_totals_pre2014 <- pre2014_data %>%
  group_by(State) %>%
  summarize(total_uninsured = sum(uninsured_count, na.rm = TRUE)) %>%
  arrange(desc(total_uninsured))

head(state_totals_pre2014, 10)
```

```
## # A tibble: 10 x 2
##    State          total_uninsured
##    <chr>                    <dbl>
##  1 California           41824710.
##  2 Texas                32232489.
##  3 Florida              24690566.
##  4 New York             13572826.
##  5 Georgia              11611631.
##  6 Illinois             10163254.
##  7 North Carolina        9732267.
##  8 Ohio                  8271125.
##  9 Pennsylvania          7565335.
## 10 Arizona               7080801.
```

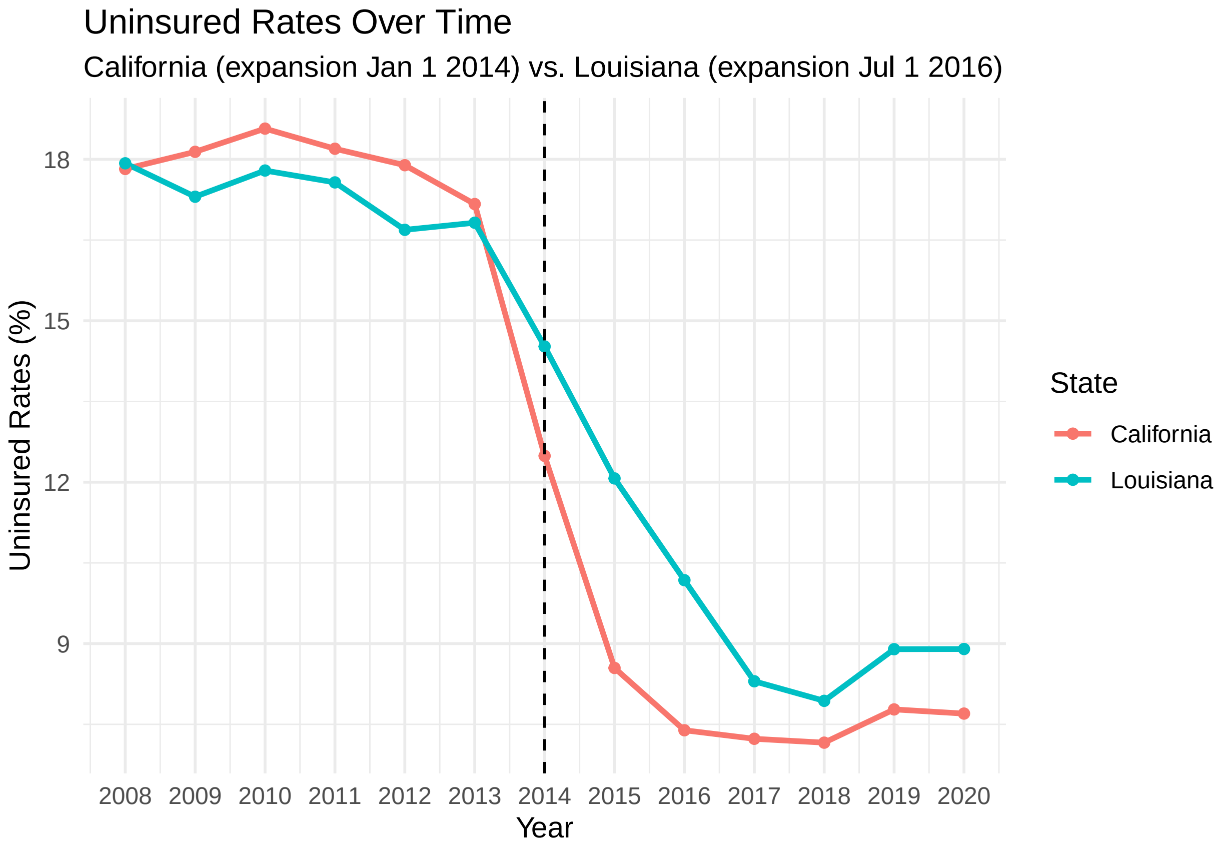
Difference-in-Differences Estimation

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).

```r
# Parallel Trends plot for California and Louisiana
plot_data <- medicaid_expansion %>%
  filter(State %in% c("California", "Louisiana"))

ggplot(plot_data, aes(x = year, y = uninsured_rate * 100, color = State)) +
  geom_line(linewidth = 1) +
  geom_point() +
  geom_vline(xintercept = 2014, linetype = "dashed") +
  scale_x_continuous(
    breaks = 2008:2020,
    limits = c(2008, 2020)
  ) +
  labs(
    title = "Uninsured Rates Over Time",
    subtitle = "California (expansion Jan 1 2014) vs. Louisiana (expansion Jul 1 2016)",
    x = "Year",
    y = "Uninsured Rates (%)"
  ) +
  theme_minimal()
```



- Estimates a difference-in-differences estimate of the effect of the Medicaid expansion on the uninsured share of the population. You may follow the lab example where we estimate the differences in one pre-treatment and one post-treatment period, or take an average of the pre-treatment and post-treatment outcomes

```
# Difference-in-Differences estimation
did_df <- medicaid_expansion %>%
  mutate(year = as.integer(substr(as.character(year), 1, 4))) %>%
  filter(State %in% c("California", "Louisiana"),
      year %in% c(2013, 2014, 2015)) %>%
  mutate(
    treated = if_else(State =="California", 1L, 0L),
    post = if_else(year >= 2014, 1L, 0L)
  )
```

```
# extracting rate
overall_diffs <- did_df %>%
  group_by(State) %>%
  summarize(
    change = last(uninsured_rate) - first(uninsured_rate),
    .groups = "drop"
  )

did_estimate_1315 <- overall_diffs %>%
  summarize(
    did = change[State=="California"]
        - change[State=="Louisiana"]
  ) %>%
  pull(did)

did_estimate_1315
```

```
## [1] -0.03869
```

The DiD estimate is -0.03869 meaning that California's uninsured rate fell about 3.87 percentage points more than Louisiana's did between 2013 and 2015.

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**: Card/Krueger's research targeted towns where little else changed. With the medicaid data, the states and Washington DC differ on demographics, policy environments, timing etc so it makes it hard to target random border. While it is possible to use proxies in the form of let's say bordering states with similar trends before treatment, it's still using entire states which different immensely compared to near-identical towns.
- What are the strengths and weaknesses of using the parallel trends assumption in difference-in-differences estimates?
- **Answer**: Easy to implement and easy to interpret visual check. Possible to control for *time-invariant* differences across units. However, it doesn't address time-varying confounders such as policies and trends. And it relies on pre-treatment constant gap which may not hold.

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

synth_data <- medicaid_expansion %>%

# the adoption date into a Date object
mutate(
    expansion_date = ymd(Date_Adopted),
    year           = as.integer(substr(as.character(year), 1, 4))
  ) %>%

filter(
    is.na(expansion_date) |
    expansion_date > ymd("2016-01-01")
  ) %>%

mutate(treated = if_else(State == "Louisiana", 1L, 0L))
```

- 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
synth_data %>%
  group_by(State) %>%
  summarize(
    n_years = n(),
    min_year = min(year),
    max_year = max(year),
    treated  = first(treated)
  )
```

```
## # A tibble: 20 x 5
##    State          n_years min_year max_year treated
##    <chr>            <int>    <int>    <int>   <int>
##  1 Alabama             13     2008     2020       0
##  2 Florida             13     2008     2020       0
##  3 Georgia             13     2008     2020       0
##  4 Idaho               13     2008     2020       0
##  5 Kansas              13     2008     2020       0
##  6 Louisiana           13     2008     2020       1
##  7 Maine               13     2008     2020       0
##  8 Mississippi         13     2008     2020       0
##  9 Missouri            13     2008     2020       0
## 10 Nebraska            13     2008     2020       0
## 11 North Carolina      13     2008     2020       0
## 12 Oklahoma            13     2008     2020       0
## 13 South Carolina      13     2008     2020       0
## 14 South Dakota        13     2008     2020       0
## 15 Tennessee           13     2008     2020       0
## 16 Texas               13     2008     2020       0
## 17 Utah                13     2008     2020       0
## 18 Virginia            13     2008     2020       0
## 19 Wisconsin           13     2008     2020       0
## 20 Wyoming             13     2008     2020       0
```

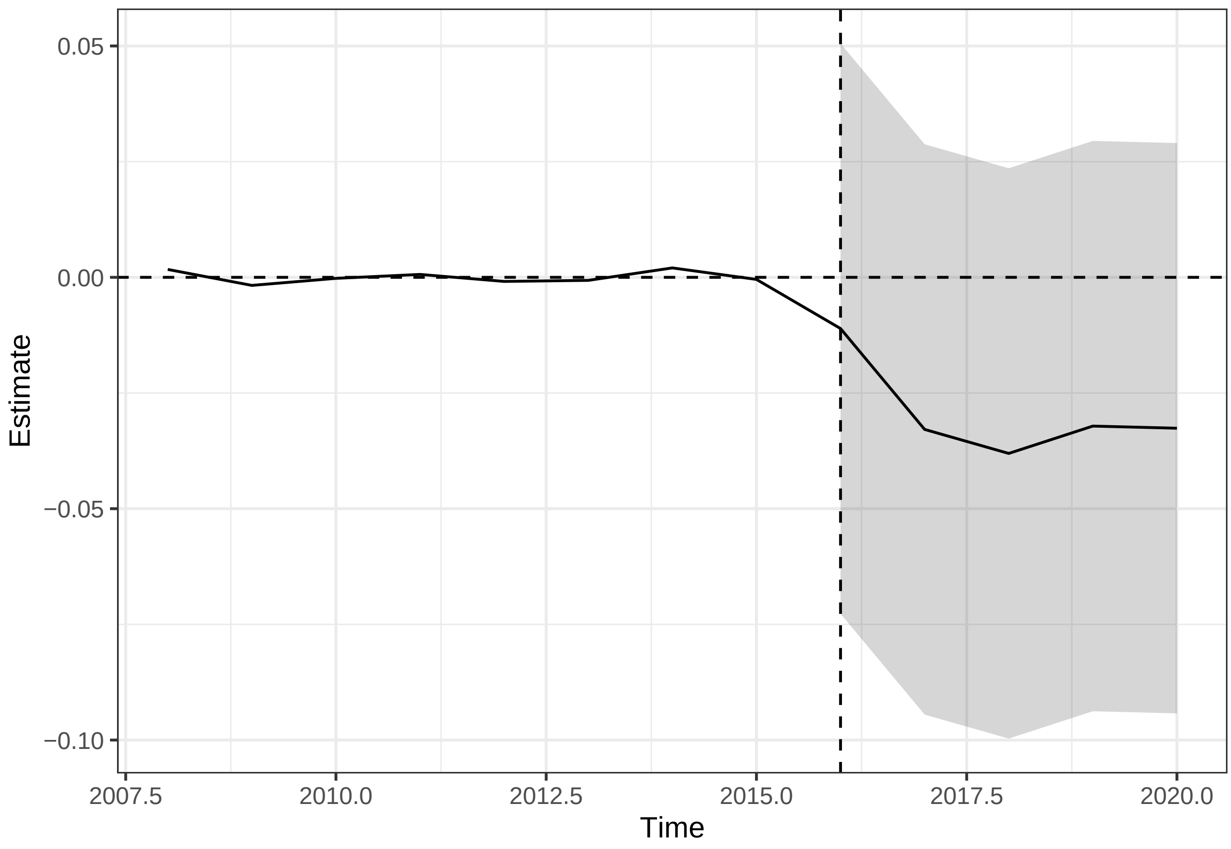
```r
sc_naug <- augsynth(
  uninsured_rate ~ treated,
  State,
  year,
  t_int   = 2016,
  data    = synth_data,
  progfunc = "none",
  scm     = TRUE
)
```

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

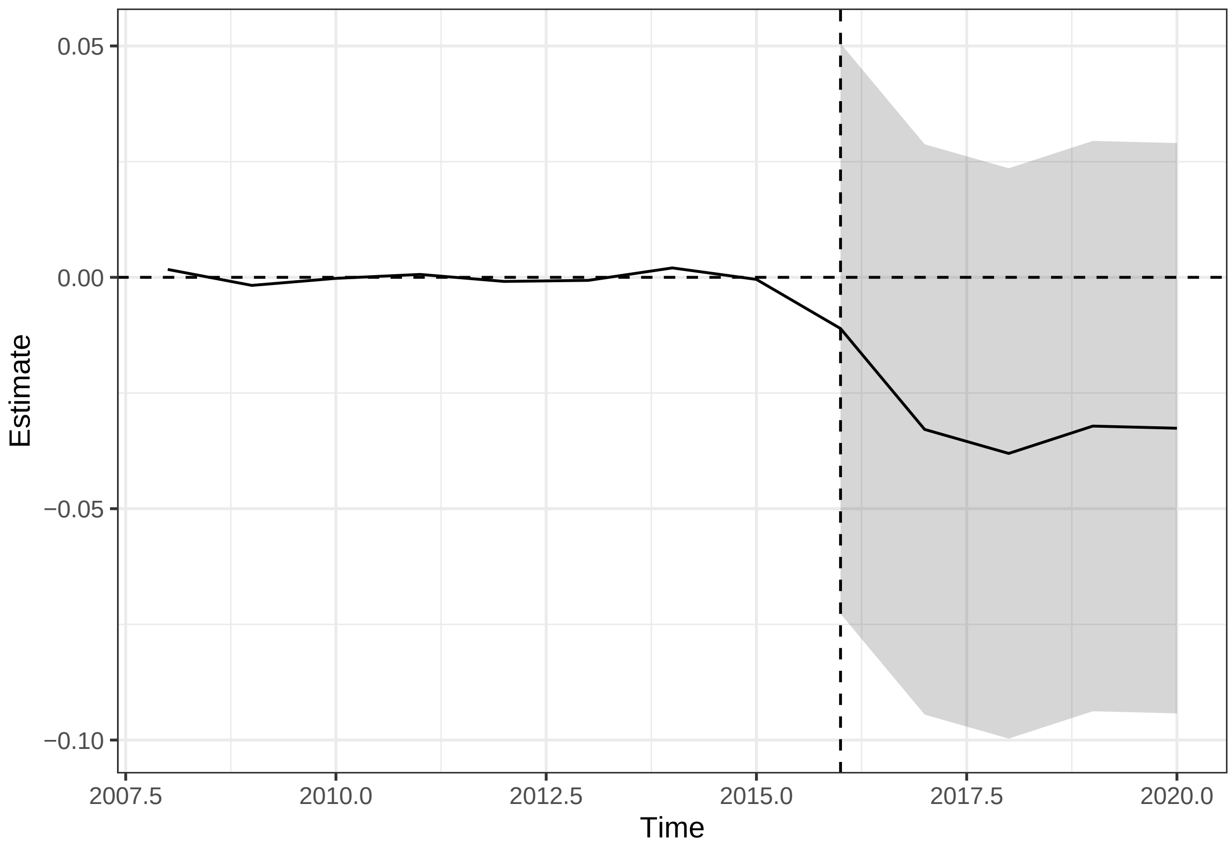
```r
# report average ATT & L2
summary(sc_naug)
```

```
##
## 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):  -0.0293   ( 0.017 )
## L2 Imbalance: 0.003
## Percent improvement from uniform weights: 94.3%
##
## Avg Estimated Bias: NA
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.011             -0.073              0.051   0.110
##  2017   -0.033             -0.094              0.029   0.101
##  2018   -0.038             -0.100              0.024   0.109
##  2019   -0.032             -0.094              0.029   0.113
##  2020   -0.033             -0.094              0.029   0.118
```

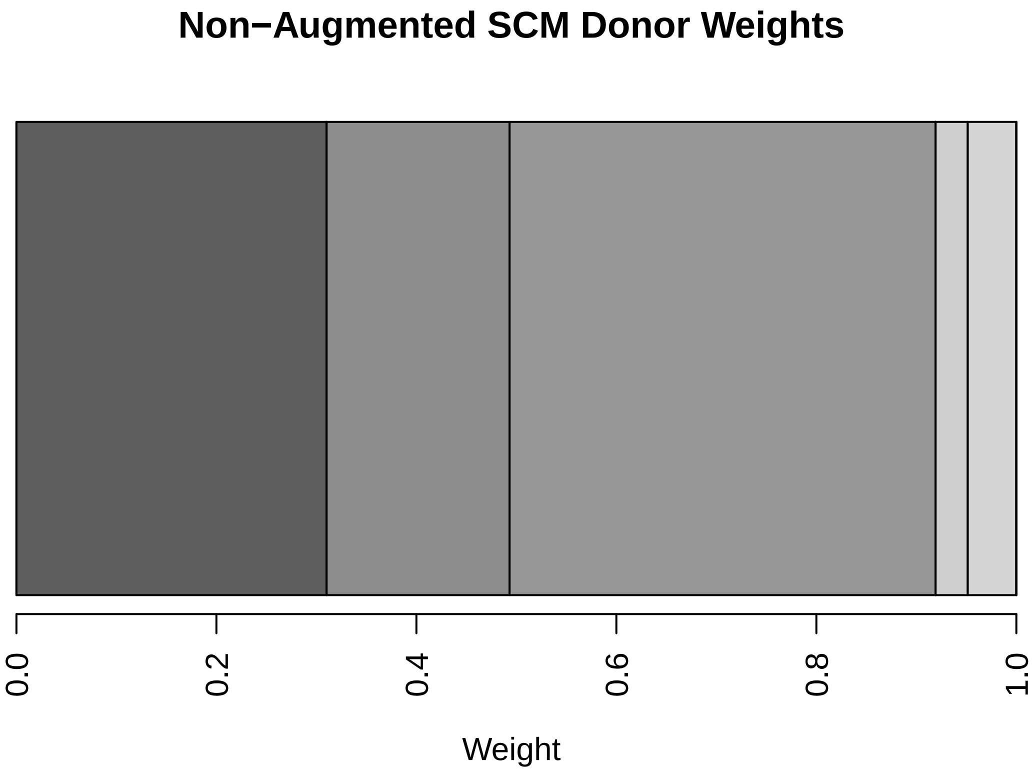
```r
# pre/post fit
plot(sc_naug, type = "path")
```



```r
# gap series
plot(sc_naug, type = "gap")
```



```
# donor weights
barplot(
  sc_naug$weights,
  las   = 2,
  horiz = TRUE,
  xlab  = "Weight",
  main  = "Non-Augmented SCM Donor Weights"
)
```



```
# augmented synthetic control

library(dplyr)
library(augsynth)


# estimating non_augmented for LA
sc_aug <- augsynth(
  uninsured_rate ~ treated,  # outcome
  State,                     # unit column
  year,                      # time column
  2016,                      # intervention year
  data     = synth_data,
  # progfunc defaults to "ridge", and scm=TRUE by default
)
```

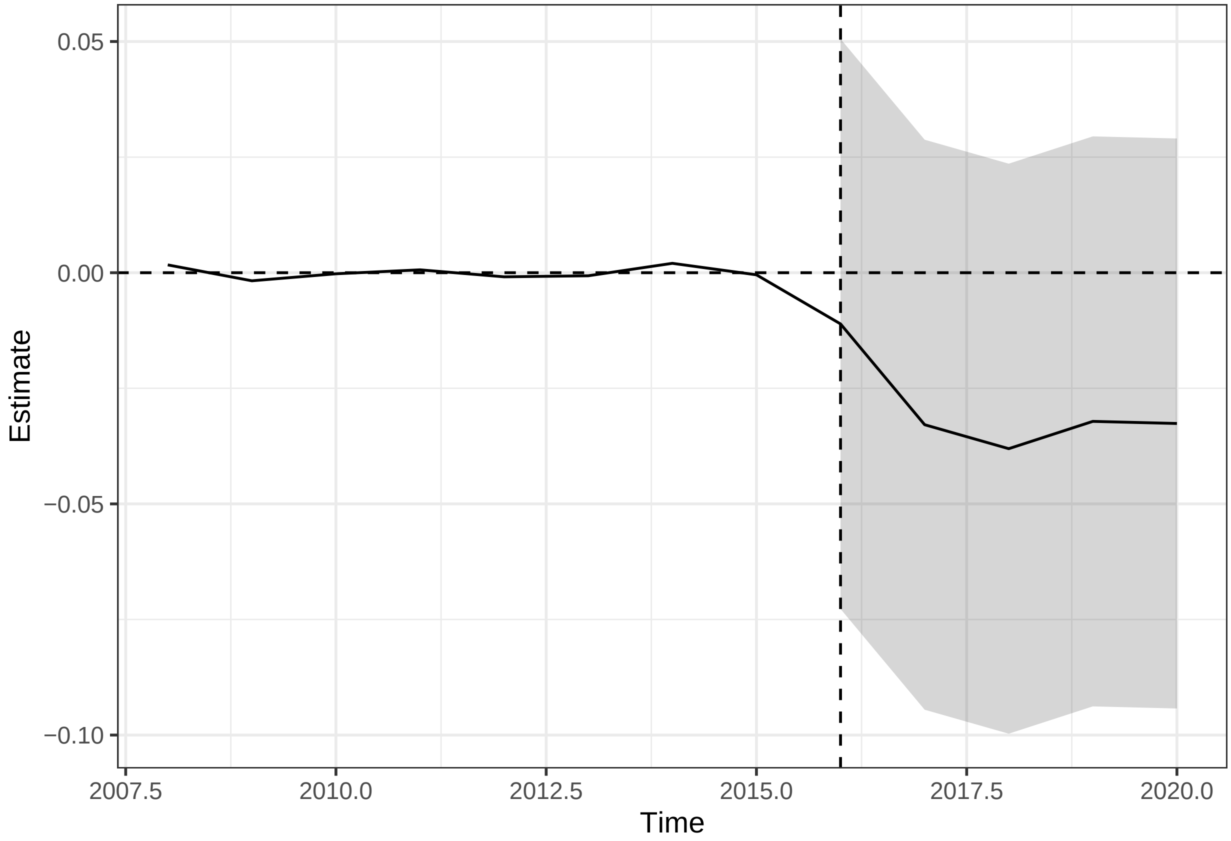
```
## One outcome and one treatment time found. Running single_augsynth.
```

```
summary(sc_aug)
```

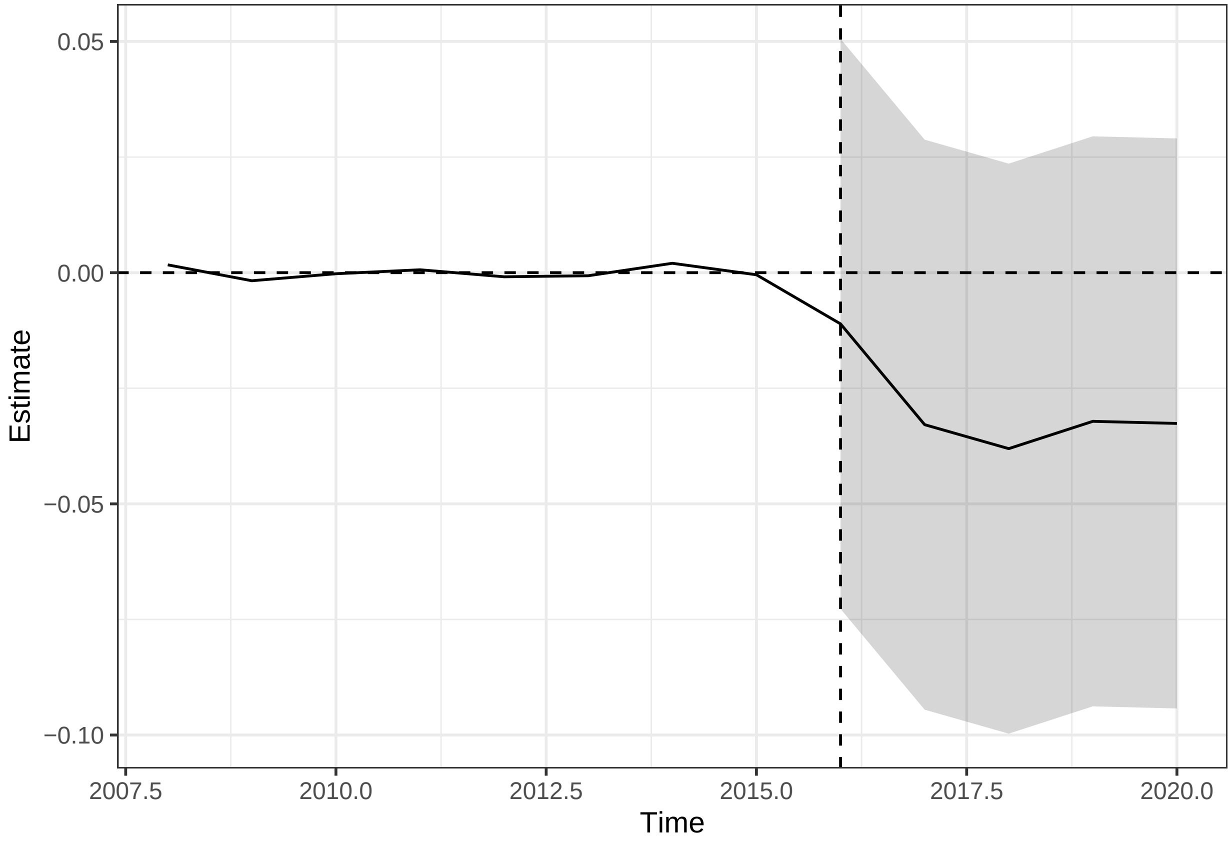
```
##
## Call:
## single_augsynth(form = form, unit = !!enquo(unit), time = !!enquo(time),
##     t_int = t_int, data = data, progfunc = ..1)
##
```

```
## Average ATT Estimate (p Value for Joint Null):  -0.0294   ( 0.027 )
## L2 Imbalance: 0.003
## Percent improvement from uniform weights: 94.3%
##
## Avg Estimated Bias: 0.000
##
## Inference type: Conformal inference
##
##  Time Estimate 95% CI Lower Bound 95% CI Upper Bound p Value
##  2016   -0.011             -0.073              0.051   0.121
##  2017   -0.033             -0.095              0.029   0.123
##  2018   -0.038             -0.100              0.024   0.121
##  2019   -0.032             -0.094              0.029   0.121
##  2020   -0.033             -0.094              0.029   0.106
```

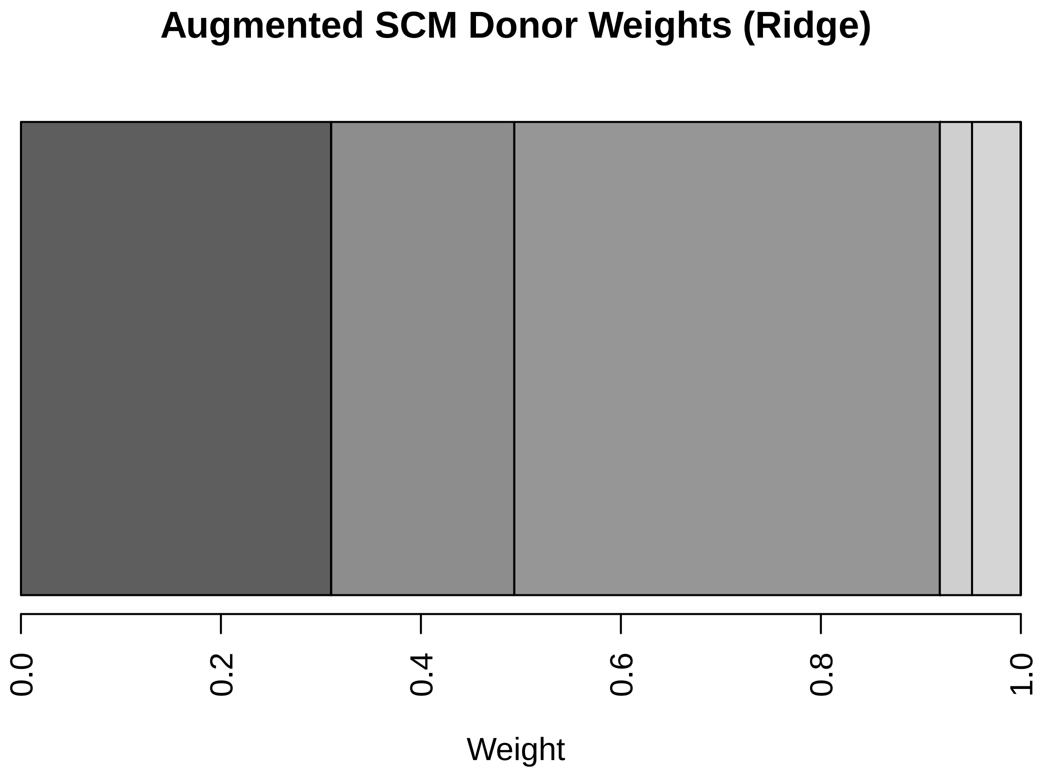
```
# pre/post fit over time
plot(sc_aug, type = "path")
```



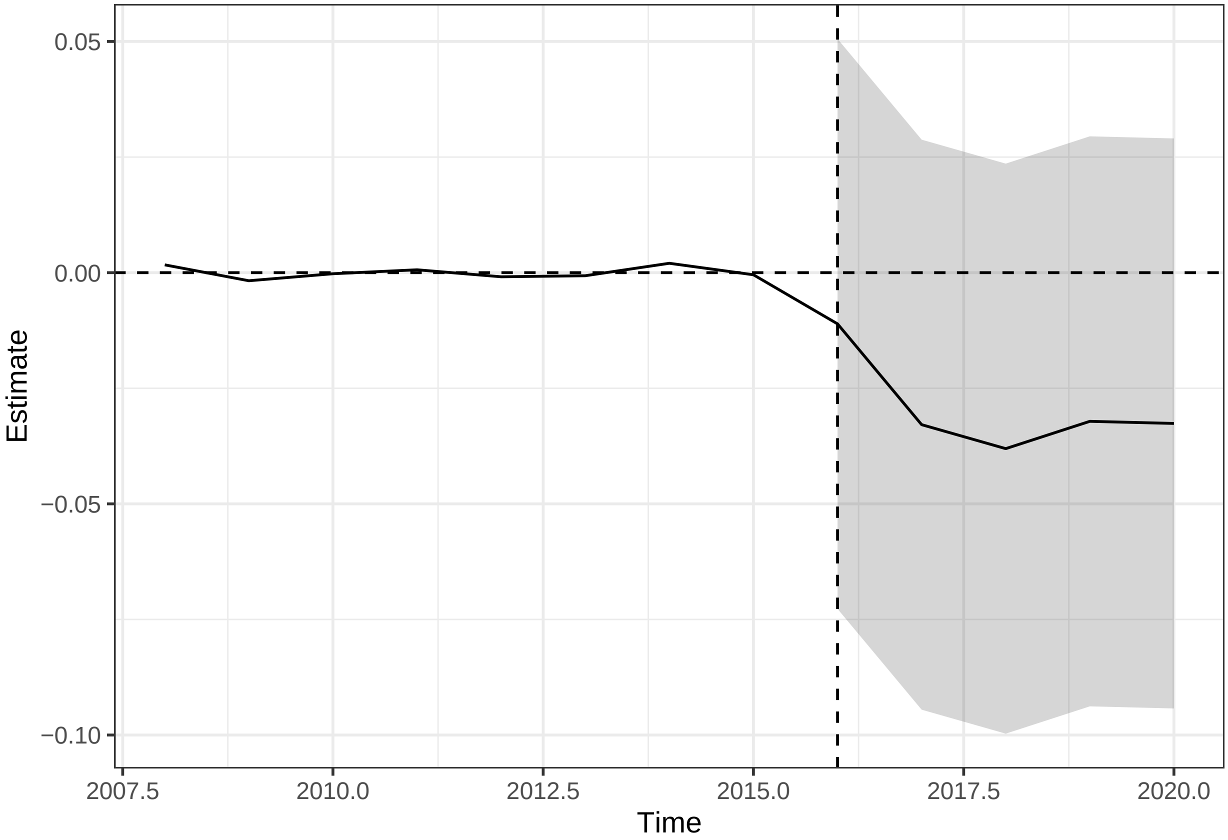
```
# gap series
plot(sc_aug, type = "gap")
```



```r
# donor weights

barplot(
  sc_aug$weights,
  las   = 2,
  horiz = TRUE,
  xlab  = "Weight",
  main  = "Augmented SCM Donor Weights (Ridge)"
)
```



```r
# using the augsynth helper
plot(sc_aug, type = "weight")
```



- Plot barplots to visualize the weights of the donors.

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 lets you construct a convex combination of many donor units to closely match the treated unit's pre-treatment trajectory, rather than relying on a single control. The non-negative weights that sum to one make it transparent which donors drive the counterfactual, and year-by-year gap plots provide clear diagnostics of fit. In contrast, difference-in-differences can easily handle multiple treated units and leverages standard regression inference, whereas classical synthetic control was designed for one treated unit (or a small cohort) at a time. Moreover, synthetic control requires a long pre-treatment panel and a sufficiently similar donor pool and if those are lacking, the method can overfit or produce implausible weights.
- 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**: Negative weights imply "subtracting" a donor's trajectory from the synthetic control, which undermines the intuitive notion of the counterfactual as a weighted average of real units. However, ridge

or matrix-completion augmentation often achieves substantially better pre-treatment balance (lower mean squared error). A sensible compromise is to present the non-augmented convex combination as the primary specification, highlighting its interpretability, and then show the augmented fit alongside it, explicitly reporting any negative weights. If those weights are small, the gain in bias reduction may justify their use; if they're large, it's best to stick with the simpler, non-negative formulation.

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

ms_data <- medicaid_expansion %>%
  mutate(
    cal_year       = as.integer(substr(as.character(year), 1, 4)),
    expansion_year = year(ymd(Date_Adopted)),
    treated_tv     = if_else(!is.na(expansion_year) & cal_year >= expansion_year, 1L, 0L)
  )
```

```
ms_states_units <- multisynth(
  uninsured_rate ~ treated_tv, #form
  State,                       #unit
  cal_year,                    #time
  ms_data,                     #data
  time_cohort = FALSE
)
```

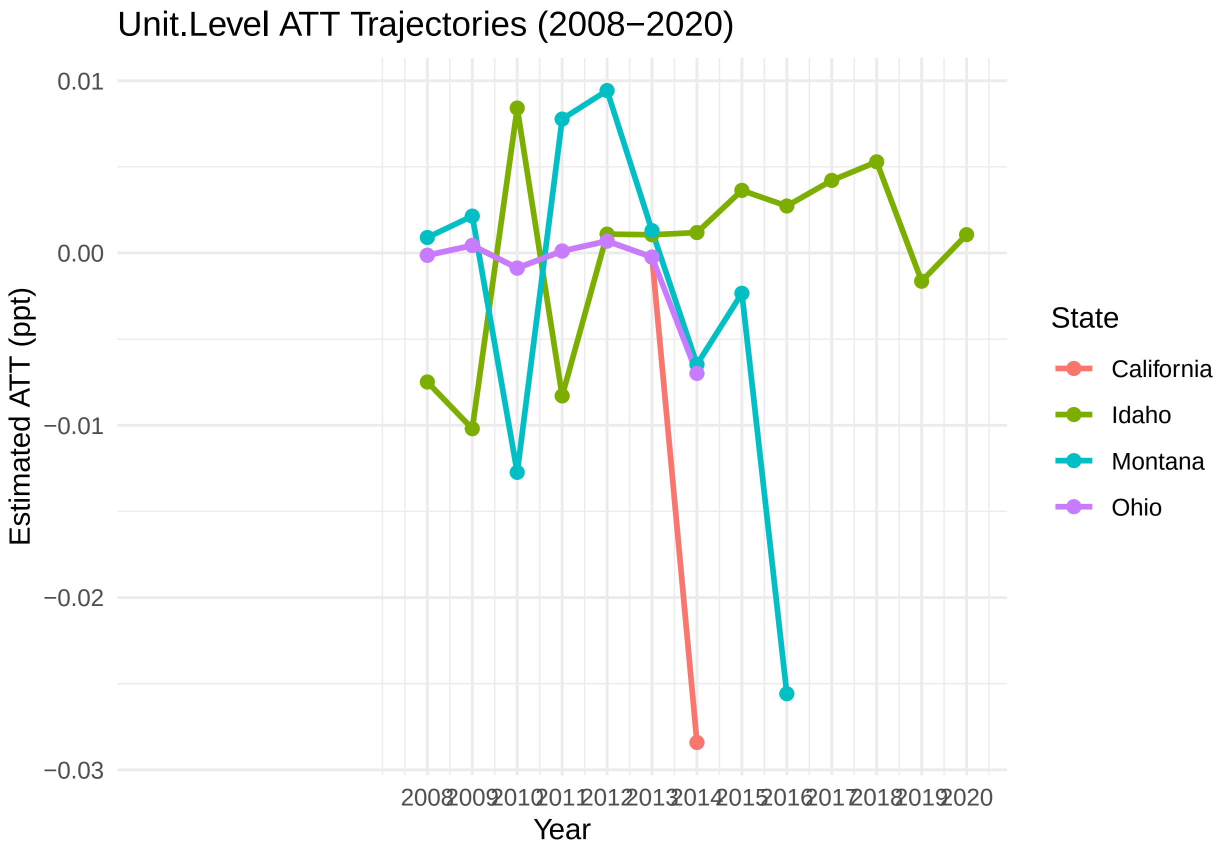
```
# extracting state event-time ATT
ms_att_unit <- summary(ms_states_units, by = "unit", inf_type = "none")$att %>%
  as_tibble() %>%
  rename(
    unit = Level,
    time = Time,
    att  = Estimate
  ) %>%
  left_join(
    ms_data %>% select(State, expansion_year) %>% distinct(),
    by = c("unit" = "State")
  ) %>%
  mutate(calendar_year = expansion_year + time)
```

```
# filter out four states
ms_cal_sub <- ms_att_unit %>%
  filter(unit %in% c("California", "Idaho", "Montana", "Ohio"))
```

```r
ggplot(ms_cal_sub, aes(x = calendar_year, y = att, color = unit)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  scale_x_continuous(breaks = 2008:2020) +
  labs(
    title = "Unit-Level ATT Trajectories (2008-2020)",
    x     = "Year",
    y     = "Estimated ATT (ppt)",
    color = "State"
  ) +
  theme_minimal()
```

```
## Warning: Removed 20 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```

```
## Warning: Removed 20 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



```r
#cohort version of multisynth
ms_states_cohort <- multisynth(
  uninsured_rate ~ treated_tv,
  State,
  cal_year,
  ms_data,
  time_cohort = TRUE
)
```

```
#extracting cohort ATT
ms_att_cohort <- summary(
  ms_states_cohort,
  by       = "cohort",
  inf_type = "none"
)$att %>%
  as_tibble()
```

- 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.

```
#which states adopted in 2016?
states_2016 <- ms_data %>%
  filter(expansion_year == 2016) %>%
  pull(State) %>%
  unique()
```

```
ms_2016_units <- ms_att_unit %>%
  filter(unit %in% states_2016)
```

```
print(states_2016)
```

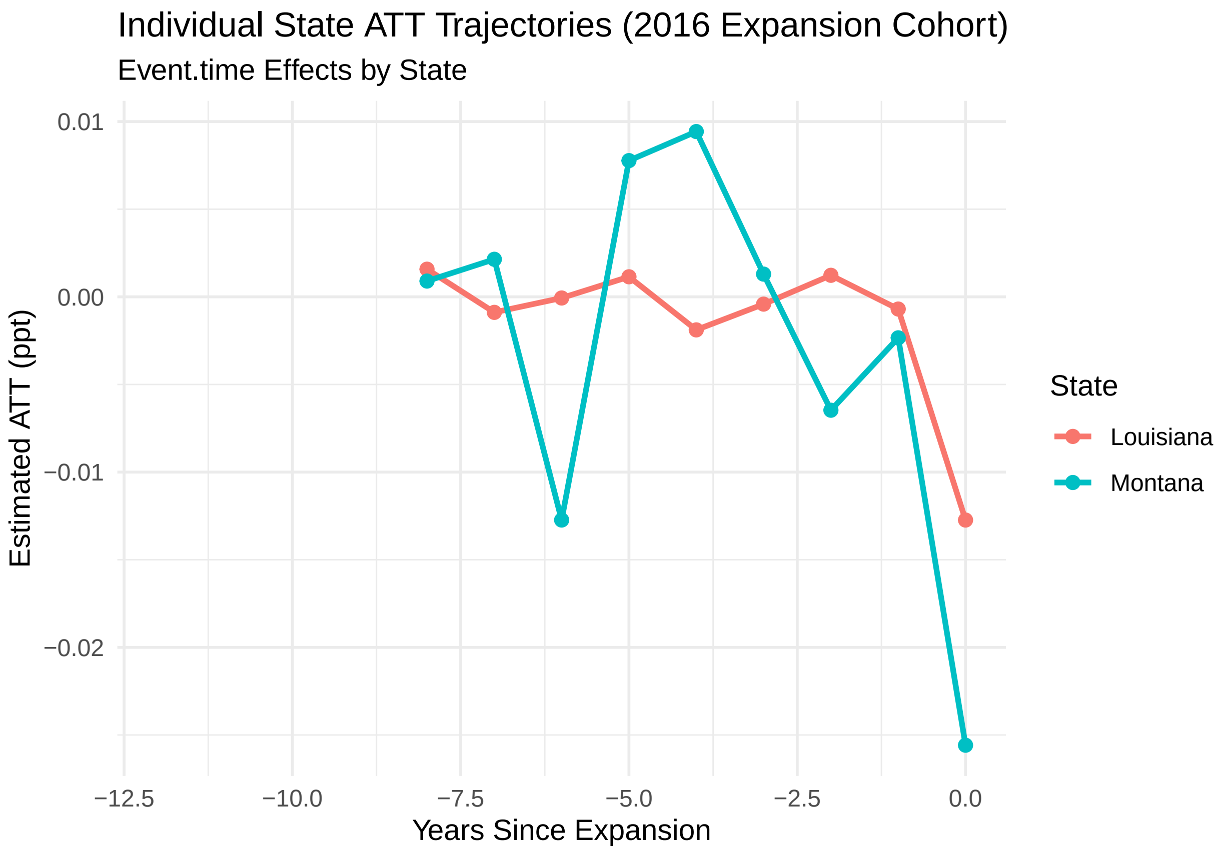
```
## [1] "Louisiana" "Montana"
```

```
print(table(ms_2016_units$unit))
```

```
##
## Louisiana   Montana
##        14        14
```

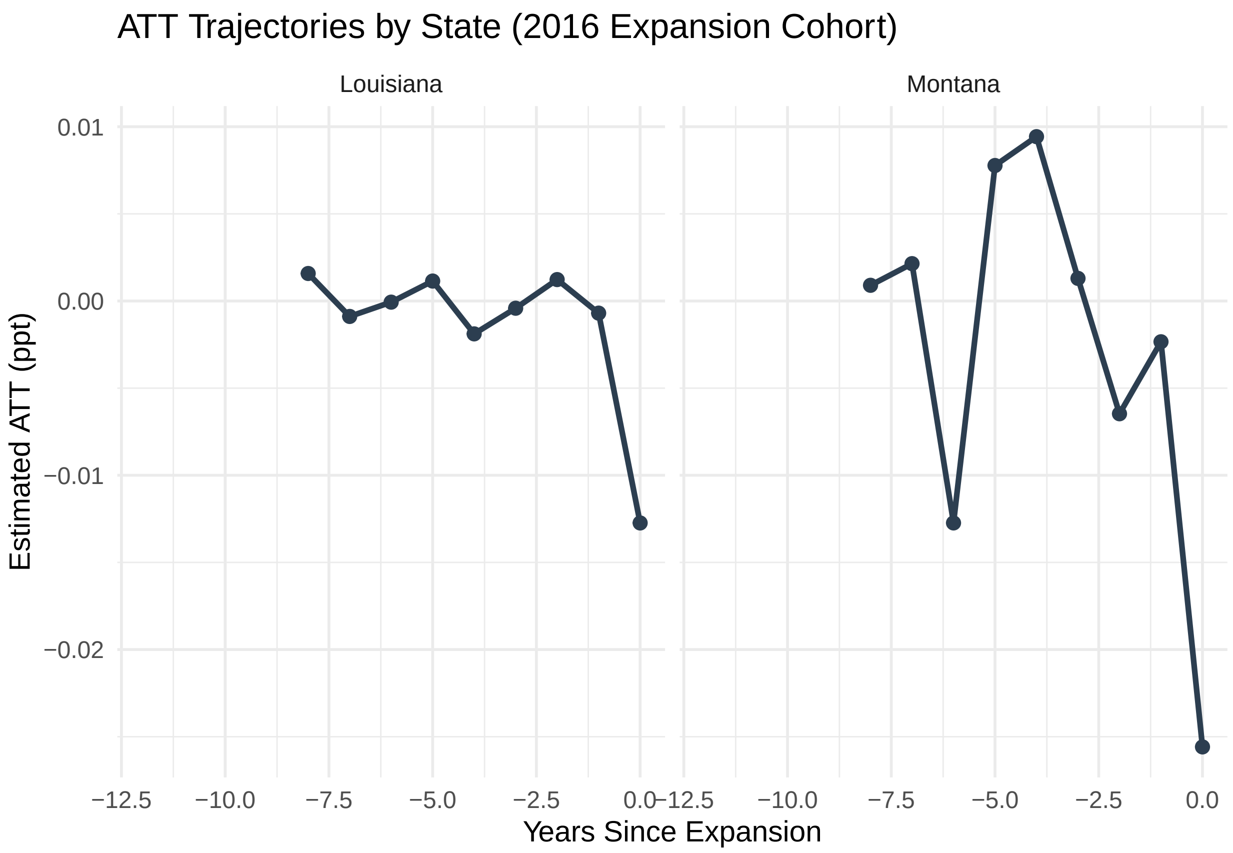
```
ggplot(ms_2016_units, aes(x = time, y = att, color = unit)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title    = "Individual State ATT Trajectories (2016 Expansion Cohort)",
    subtitle = "Event-time Effects by State",
    x        = "Years Since Expansion",
    y        = "Estimated ATT (ppt)",
    color    = "State"
  ) +
  theme_minimal()
```

```
## Warning: Removed 10 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
## Warning: Removed 10 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



```r
# separate panel per state
ggplot(ms_2016_units, aes(x = time, y = att)) +
  geom_line(linewidth = 1, color = "#2c3e50") +
  geom_point(size = 2, color = "#2c3e50") +
  facet_wrap(~ unit, ncol = 2) +
  labs(
    title = "ATT Trajectories by State (2016 Expansion Cohort)",
    x     = "Years Since Expansion",
    y     = "Estimated ATT (ppt)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

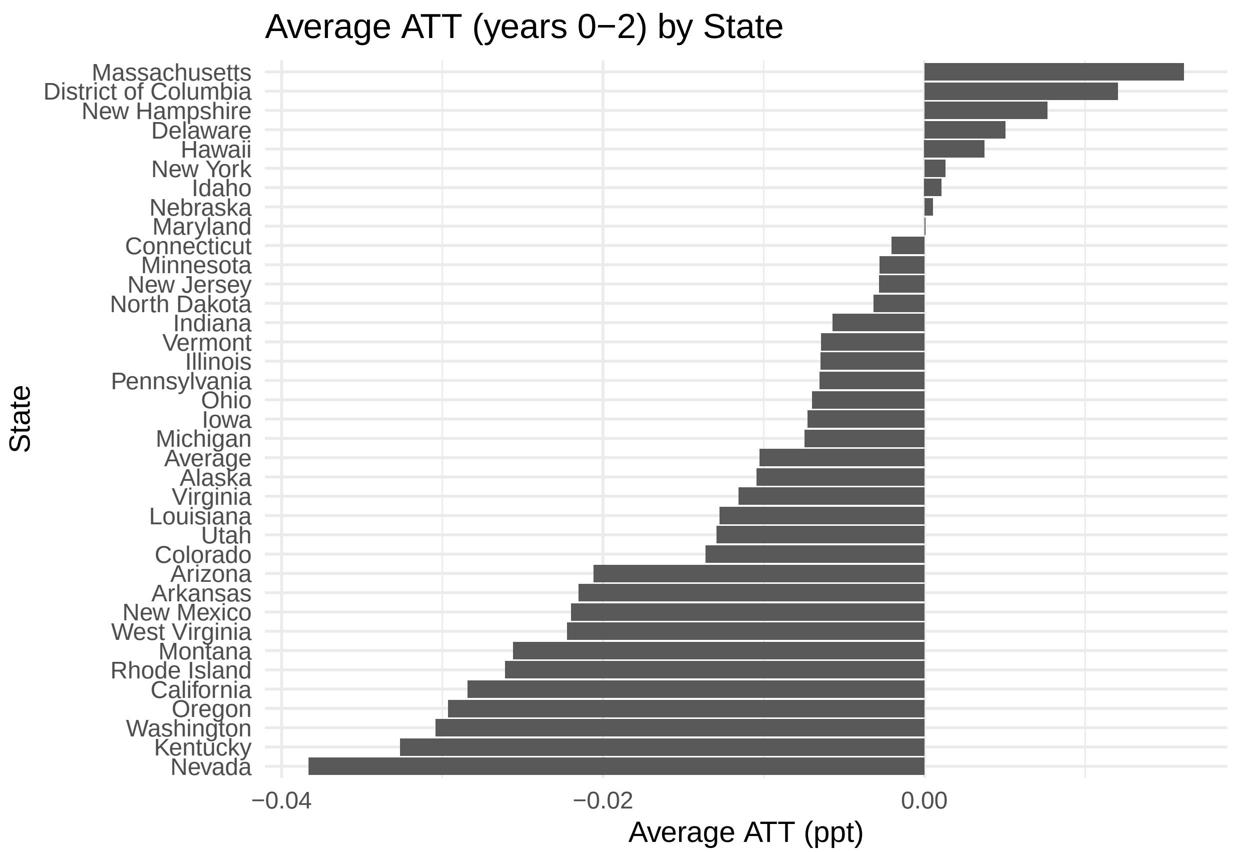
```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (‘geom_line()‘).

## Warning: Removed 10 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



I'm not sure if the I did the staggered adoption synthetic control section right.

```
avg_att_by_state <- ms_att_unit %>%
  filter(time >= 0, time <= 2) %>%
  group_by(unit) %>%
  summarize(avg_att = mean(att, na.rm = TRUE)) %>%
  arrange(avg_att)
```

```
ggplot(avg_att_by_state, aes(x = reorder(unit, avg_att), y = avg_att)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Average ATT (years 0-2) by State",
    x     = "State",
    y     = "Average ATT (ppt)"
  ) +
  theme_minimal()
```



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?

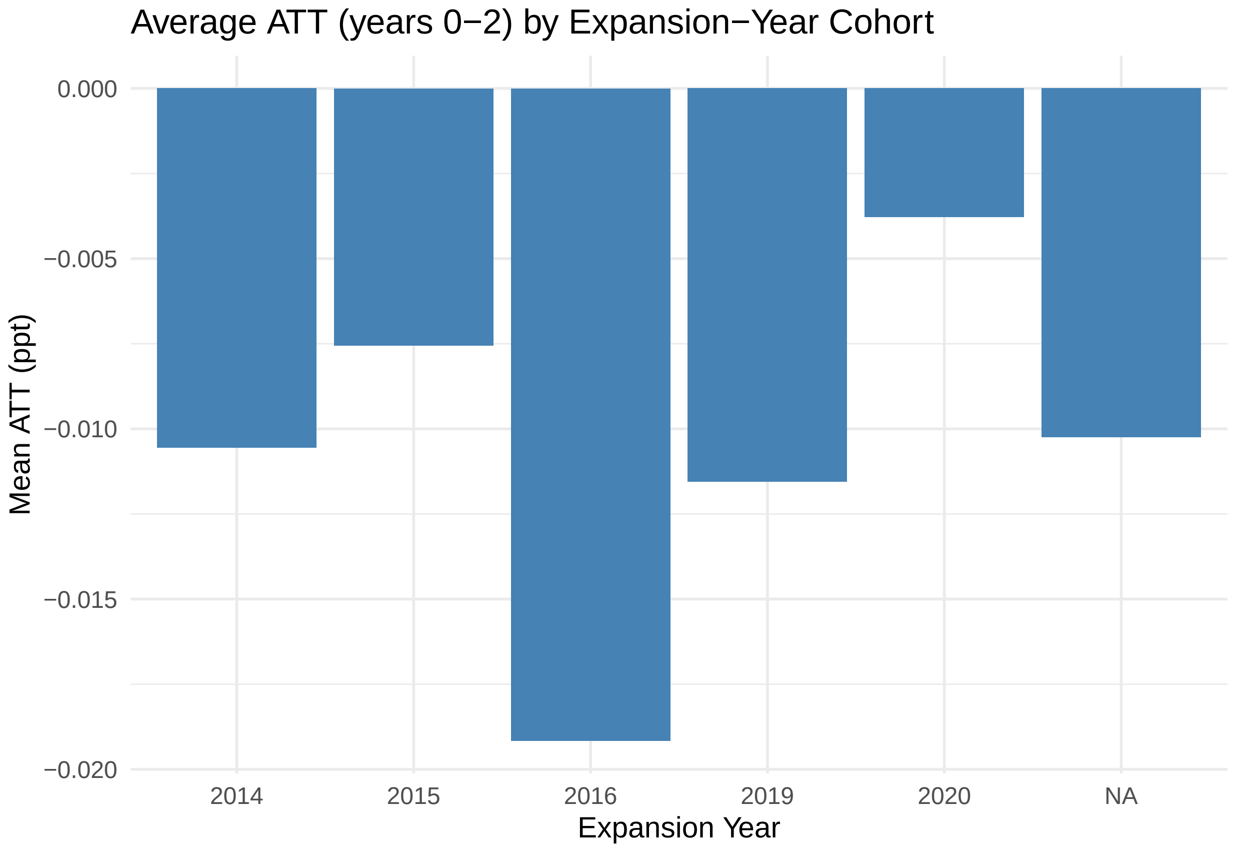
- **Answer**: The average ATT over the first two post-expansion years varies substantially across states. For example, Nevada experienced one of the largest declines (around –0.04 ppt), while Connecticut's change was nearly zero, and states like New Hampshire and Hawaii actually saw slight increases in uninsured rates over that window. This wide spread—ranging from roughly –0.04 ppt to +0.02 ppt—suggests that state-level implementation details (outreach efforts, administrative capacity, supplemental eligibility rules, etc.) materially shaped the magnitude of the Medicaid expansion's impact. Even without a specific "work-requirement" flag, these differences show that the "same" federal policy can play out quite differently across states depending on local context and execution.

```
avg_att_by_state <- avg_att_by_state %>%
  left_join(
    ms_data %>% select(State, expansion_year) %>% distinct(),
    by = c("unit" = "State")
  )
```

```r
cohort_summary <- avg_att_by_state %>%
  group_by(expansion_year) %>%
  summarize(
    cohort_avg_att = mean(avg_att, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(expansion_year)

print(cohort_summary)
```

```
## # A tibble: 6 x 2
##   expansion_year cohort_avg_att
##            <dbl>          <dbl>
## 1           2014       -0.0106
## 2           2015       -0.00755
## 3           2016       -0.0192
## 4           2019       -0.0116
## 5           2020       -0.00378
## 6             NA       -0.0102
```

```r
ggplot(cohort_summary, aes(x = factor(expansion_year), y = cohort_avg_att)) +
  geom_col(fill = "steelblue") +
  labs(
    title    = "Average ATT (years 0-2) by Expansion-Year Cohort",
    x        = "Expansion Year",
    y        = "Mean ATT (ppt)"
  ) +
  theme_minimal()
```



- Do you see evidence for the idea that early adopters of Medicaid expansion enjoyed a larger decrease in the uninsured population?
- **Answer**: Comparing the average two-year ATT across expansion-year cohorts, there doesn't seem to be an advantage for states with early adoption date pattern. In fact:the 2016 cohort experienced the largest average decline (around –0.02 ppt). The 2014 cohort (the earliest adopters) had an average ATT of roughly –0.01 ppt, not that larger than the 2019 cohort and only slightly more than the 2020 cohort. This all suggest that timing alone; being early to adopt, did not guarantee the largest coverage gains

General Discussion Questions

- Why are DiD and synthetic control estimates well suited to studies of aggregated units like cities, states, countries, etc?
- **Answer**: Both methods require only a single summary measure (e.g. state-level uninsured rate) per period. There is no need for individual-level observations, making them ideal when only aggregate statistics (cities, states, countries) are available. In DiD, its is easy to see the regression specification, interaction term, and additional covariates if included. In synthetic control, the donor weights make clear which units drive the counterfactual, and one can inspect how fit improves with augmentation. DiD can accommodate varying treatment times with appropriate modifications (event-study designs). Synthetic control generalizes via multisynth to handle multiple treated units entering at different times.
- What role does selection into treatment play in DiD/synthetic control versus regression discontinuity? When would we want to use either method?

- **Answer**: DiD assumes that, absent treatment, treated and control units would follow parallel trends. Synthetic control assumes it is possible to build combination of donor units whose weighted pre-treatment path matches the treated unit, thereby capturing time-invariant and slowly-varying confounders. Whereas in regression discontinuity treatment is assigned by a cutoff. Near the cutoff, units are as good as randomly assigned, so the only systematic difference at the margin is treatment status. In DiD/synthetic control,selection on unobservables is allowed so long as it does not violate parallel trends or can be absorbed by the synthetic weights. It is possible to test the plausibility of these assumptions via pre-treatment diagnostics such as trend tests, gap-plots. In regression discontinuity, you exploit a design where selection is discontinuous at the cutoff but smooth around it. Units just above and below the threshold are assumed to differ only in treatment status, not in other characteristics, therefore not parallel-trends assumption is needed. DiD/synthetic control are ideal when treatment is rolled out at known times without a clean discontinuity rule, especially for aggregate units and when you have a reasonable pre-treatment panel to test and match on trends. Regression discontinuity is best when assignment is determined by a precise, observable cutoff, and you have enough observations close to that threshold.