its_synthetic

Bogdan Vasilescu

2024-04-04

Load the smoking data.

```
data(smoking)
head(smoking)
```

```
## # A tibble: 6 x 7
##   state         year cigsale lnincome  beer age15to24 retprice
##   <chr>        <dbl>   <dbl>    <dbl> <dbl>     <dbl>    <dbl>
## 1 Rhode Island  1970   124.        NA    NA     0.183     39.3
## 2 Tennessee     1970    99.8       NA    NA     0.178     39.9
## 3 Indiana       1970   135.        NA    NA     0.177     30.6
## 4 Nevada        1970   190.        NA    NA     0.162     38.9
## 5 Louisiana     1970   116.        NA    NA     0.185     34.3
## 6 Oklahoma      1970   108.        NA    NA     0.175     38.4
```

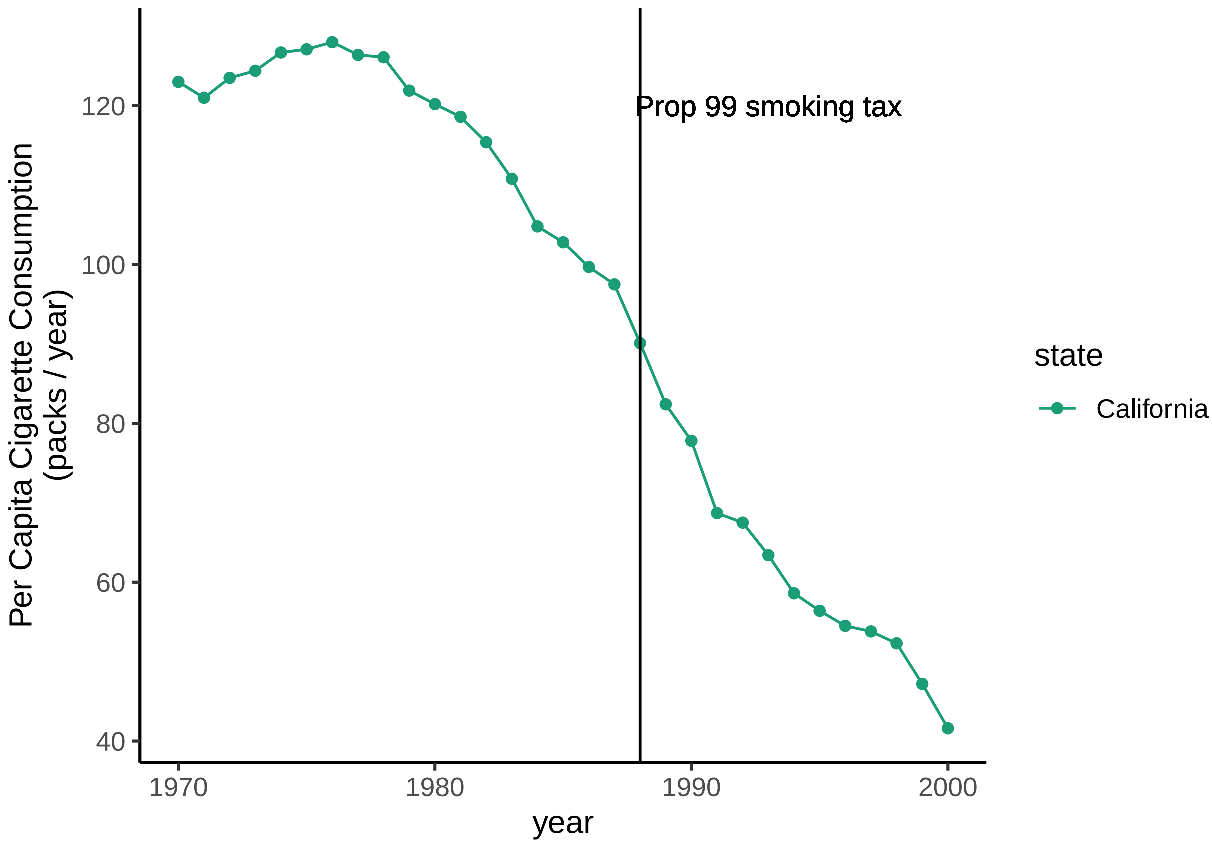
As part of the method's assumptions and requirements, it is important to exclude from the donor pool (this is how the collection of control units is traditionally called) any unit that may not be a true control - i.e. any unit that has implemented a similar intervention. In our case, some states also introduced anti-smoking programs or substantially increased the tax for cigarettes. These have already been excluded:

```
length(unique(smoking$state))
```

```
## [1] 39
```

The trend in California

```
ggplot(data = subset(smoking, state == "California"),
       aes(x = year,
           y = cigsale,
           color = state,
           group = state)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1988) +
  labs(y = "Per Capita Cigarette Consumption\n(packs / year)") +
  geom_text(aes(x=1993,
                label="Prop 99 smoking tax",
                y=120),
            # angle=90,
            color="black") +
  scale_color_brewer(palette = 'Dark2') +
  theme_classic(base_size = 12)
```



Adding the average of the other 38 states as a control

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.4.1
```

```
##
## Attaching package: 'data.table'
```

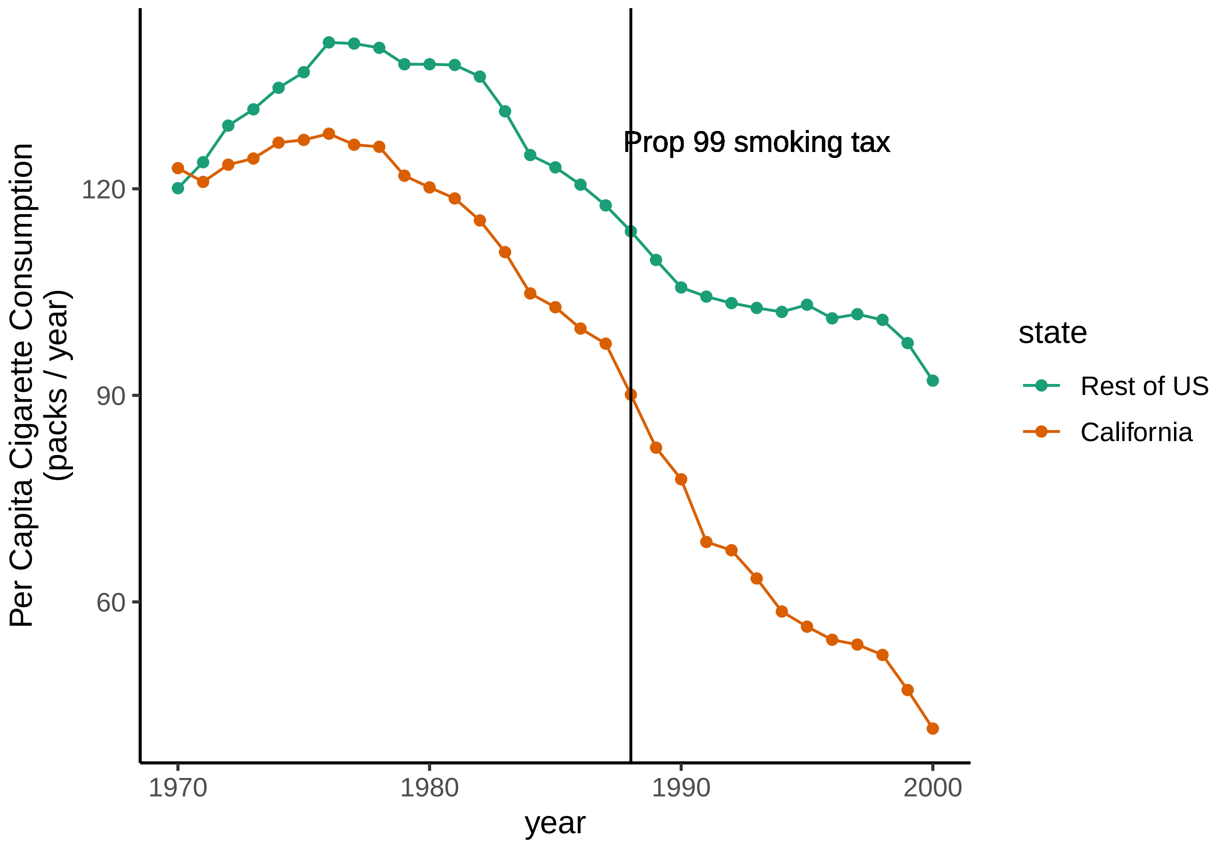
```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
smoking.s <- as.data.table(
  smoking[smoking$state != "California", c("state","year","cigsale")])

smoking.control <- smoking.s[,
                             lapply(.SD,
                                    mean,
                                    na.rm=TRUE),
                             by=year,
                             .SDcols=c("cigsale")]

smoking.control[,
                state := rep("Rest of US", nrow(smoking.control))]

smoking.control <- rbind(
```

```r
  smoking.control,
  as.data.table(smoking[smoking$state == "California", c("state","year","cigsale")])
  )

smoking.control[, ":="(
  treated = year >= 1988,
  years_after_intervention = ifelse(year < 1988, 0, year - 1988),
  year0 = year - 1988
)]

smoking.control$state <- factor(smoking.control$state,
                                 levels = c("Rest of US", "California"))

smoking.means <- smoking.control[,
  lapply(.SD,
         mean,
         na.rm=TRUE),
  by=c("state","treated"),
  .SDcols=c("cigsale")]
smoking.means
```

```
##         state treated   cigsale
##        <fctr>  <lgcl>     <num>
## 1: Rest of US   FALSE 131.49985
## 2: Rest of US    TRUE 102.96316
## 3: California   FALSE 117.66111
## 4: California    TRUE  62.63846
```

```r
ggplot(data = smoking.control,
       aes(x = year,
           y = cigsale,
           color = state,
           group = state)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1988) +
  labs(y = "Per Capita Cigarette Consumption\n(packs / year)") +
  geom_text(aes(x=1993,
                label="\nProp 99 smoking tax",
                y=130),
            # angle=90,
            color="black") +
  scale_color_brewer(palette = 'Dark2') +
  theme_classic(base_size = 12)
```



Simple DID model

Comparing means before and after, by group.

```
m1 <- lm(cigsale ~ state * treated,
         data = smoking.control)
summary(m1)
```

```
##
## Call:
## lm(formula = cigsale ~ state * treated, data = smoking.control)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0385  -6.7951   0.5966   6.5620  27.4615
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  131.500      2.309  56.944  < 2e-16 ***
## stateCalifornia              -13.839      3.266  -4.237 8.20e-05 ***
## treatedTRUE                  -28.537      3.566  -8.002 6.07e-11 ***
## stateCalifornia:treatedTRUE  -26.486      5.043  -5.252 2.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.797 on 58 degrees of freedom
## Multiple R-squared:  0.8741, Adjusted R-squared:  0.8676
```

```
## F-statistic: 134.2 on 3 and 58 DF,  p-value: < 2.2e-16
```

Add trends

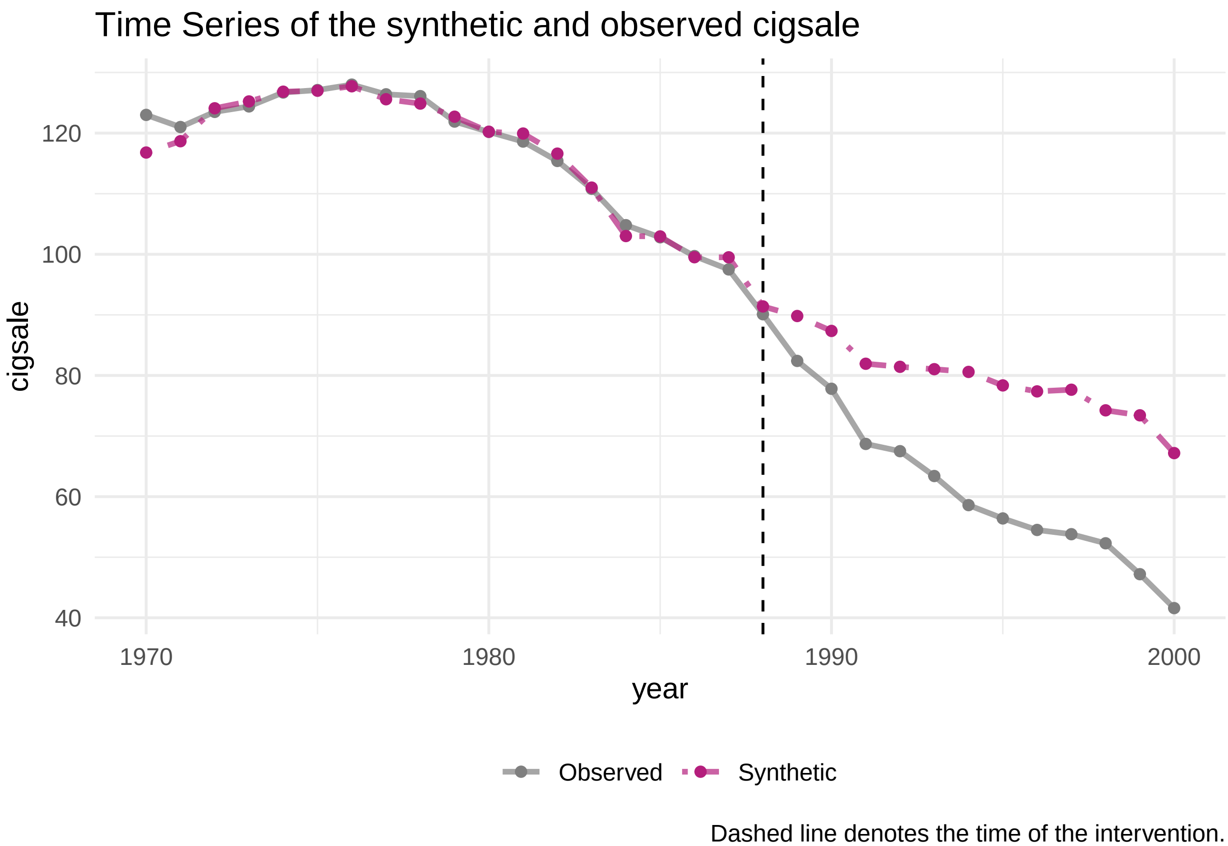
```
m2 <- lm(cigsale ~
           year0
         + treated
         + years_after_intervention
         + state
         + year0:state
         + treated:state
         + years_after_intervention:state
         , data = smoking.control)
summary(m2)
```

```
##
## Call:
## lm(formula = cigsale ~ year0 + treated + years_after_intervention +
##     state + year0:state + treated:state + years_after_intervention:state,
##     data = smoking.control)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1024  -3.9153   0.6397   3.8963   9.1155
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                128.4970     2.7923  46.018  < 2e-16
## year0                                       -0.3161     0.2580  -1.225  0.22579
## treatedTRUE                                -18.2506     4.0811  -4.472 4.02e-05
## years_after_intervention                    -0.8978     0.4937  -1.819  0.07451
## stateCalifornia                            -25.8604     3.9490  -6.549 2.22e-08
## year0:stateCalifornia                       -1.2654     0.3648  -3.469  0.00104
## treatedTRUE:stateCalifornia                 -0.4278     5.7715  -0.074  0.94119
## years_after_intervention:stateCalifornia  -1.0740     0.6981  -1.538  0.12981
##
## (Intercept)                                ***
## year0
## treatedTRUE                                ***
## years_after_intervention                   .
## stateCalifornia                            ***
## year0:stateCalifornia                      **
## treatedTRUE:stateCalifornia
## years_after_intervention:stateCalifornia
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.678 on 54 degrees of freedom
## Multiple R-squared:  0.9606, Adjusted R-squared:  0.9555
## F-statistic: 188.2 on 7 and 54 DF,  p-value: < 2.2e-16
```

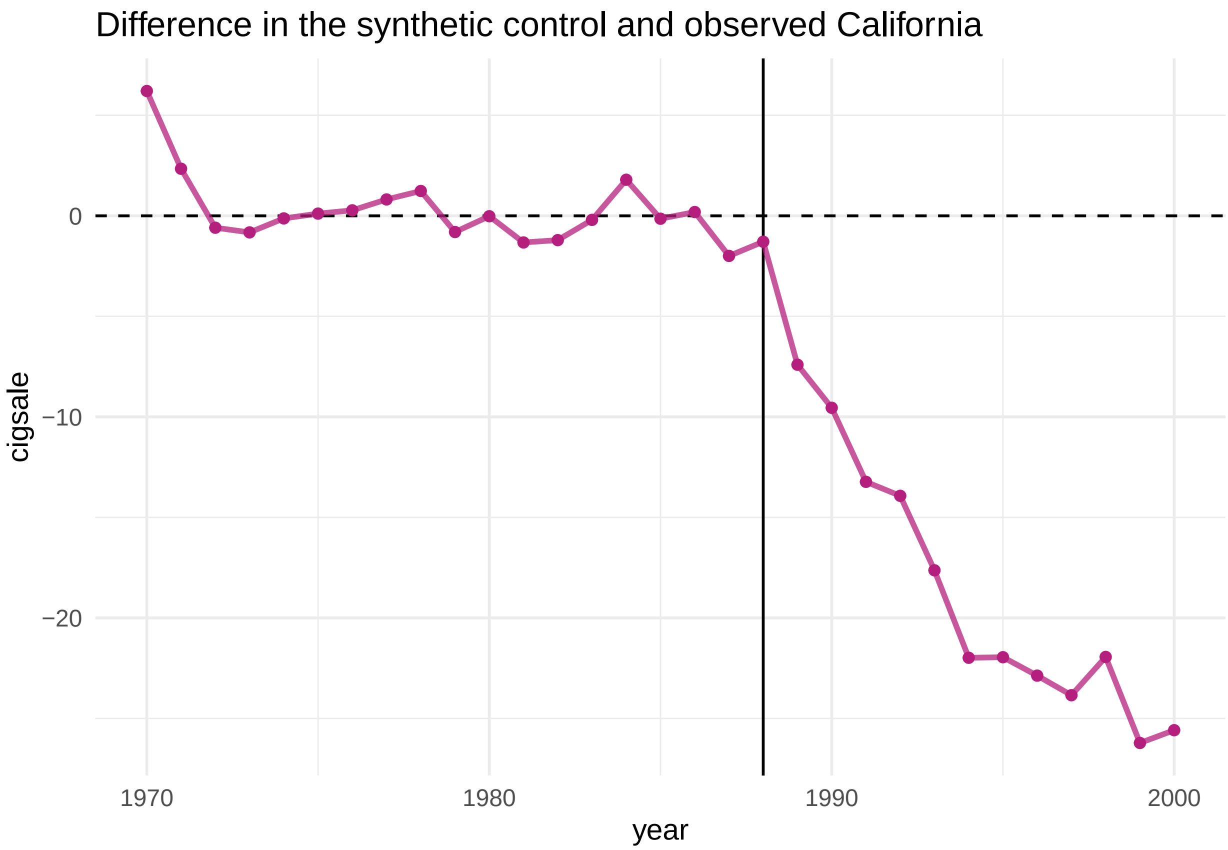
Synthetic control

```r
smoking_out <-

  smoking %>%

  # initial the synthetic control object
  synthetic_control(outcome = cigsale, # outcome
                    unit = state, # unit index in the panel data
                    time = year, # time index in the panel data
                    i_unit = "California", # unit where the intervention occurred
                    i_time = 1988, # time period when the intervention occurred
                    generate_placebos=T # generate placebo synthetic controls (for inference)
                    ) %>%

  # Generate the aggregate predictors used to fit the weights

  # average log income, retail price of cigarettes, and proportion of the
  # population between 15 and 24 years of age from 1980 - 1988
  generate_predictor(time_window = 1980:1988,
                     ln_income = mean(lnincome, na.rm = T),
                     ret_price = mean(retprice, na.rm = T),
                     youth = mean(age15to24, na.rm = T)) %>%

  # average beer consumption in the donor pool from 1984 - 1988
  generate_predictor(time_window = 1984:1988,
                     beer_sales = mean(beer, na.rm = T)) %>%

  # Lagged cigarette sales
  generate_predictor(time_window = 1975,
                     cigsale_1975 = cigsale) %>%
  generate_predictor(time_window = 1980,
                     cigsale_1980 = cigsale) %>%
  generate_predictor(time_window = 1988,
                     cigsale_1988 = cigsale) %>%


  # Generate the fitted weights for the synthetic control
  generate_weights(optimization_window = 1970:1988, # time to use in the optimization task
                   margin_ipop = .02,sigf_ipop = 7,bound_ipop = 6 # optimizer options
  ) %>%

  # Generate the synthetic control
  generate_control()
```

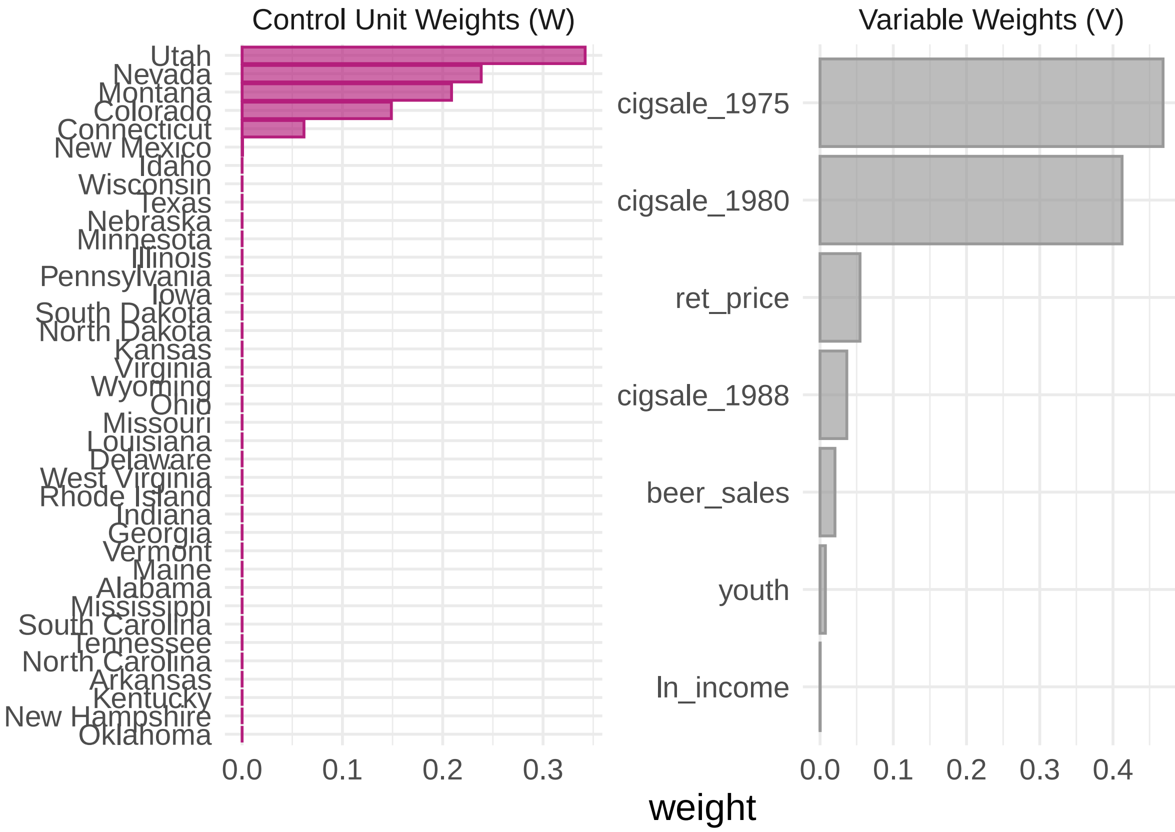
```r
plot_trends(smoking_out)
```




```r
plot_differences(smoking_out)
```




```r
plot_weights(smoking_out)
```

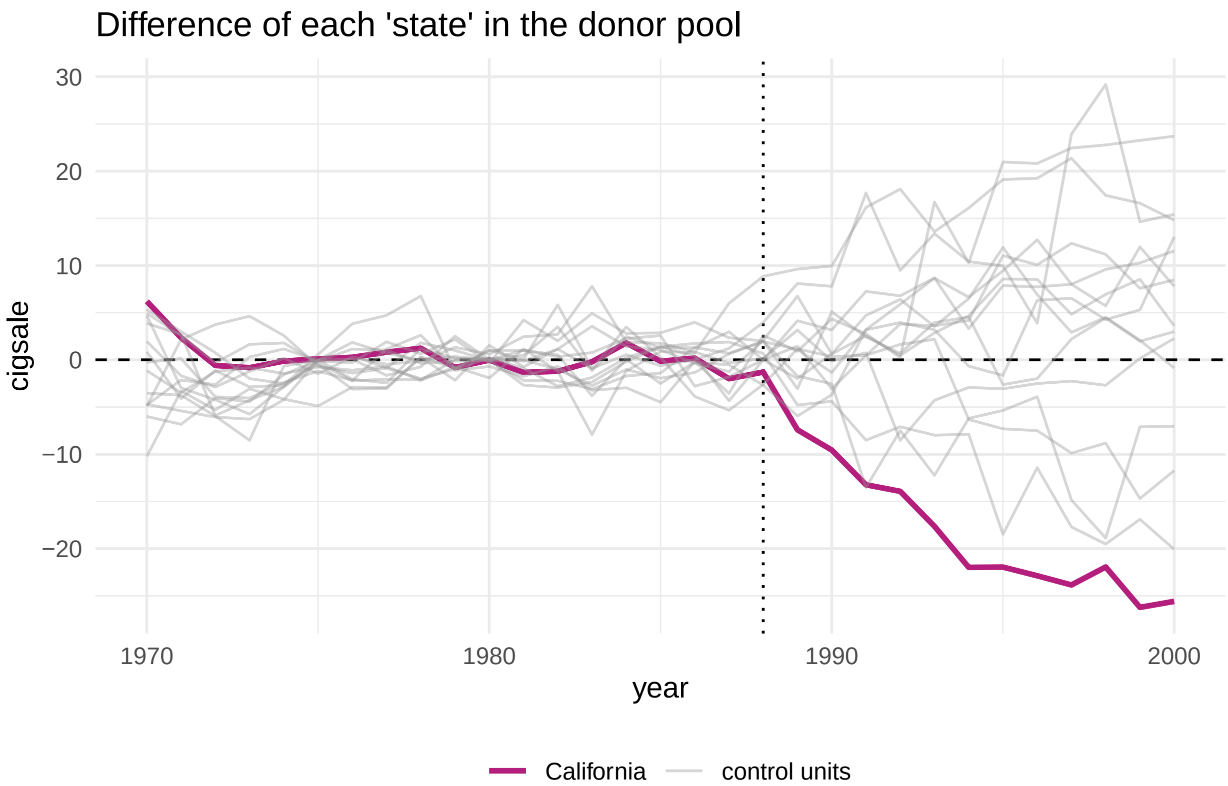


Inference

For inference, the method relies on repeating the method for every donor in the donor pool exactly as was done for the treated unit — i.e. generating placebo synthetic controls). By setting generate_placebos = TRUE when initializing the synth pipeline with synthetic_control(), placebo cases are automatically generated when constructing the synthetic control of interest. This makes it easy to explore how unique difference between the observed and synthetic unit is when compared to the placebos.

Note that the plot_placebos() function automatically prunes any placebos that poorly fit the data in the pre-intervention period. The reason for doing so is purely visual: those units tend to throw off the scale when plotting the placebos. To prune, the function looks at the pre-intervention period mean squared prediction error (MSPE) (i.e. a metric that reflects how well the synthetic control maps to the observed outcome time series in pre-intervention period). If a placebo control has a MSPE that is two times beyond the target case (e.g. "California"), then it's dropped. To turn off this behavior, set prune = FALSE.

```
plot_placebos(smoking_out)
```

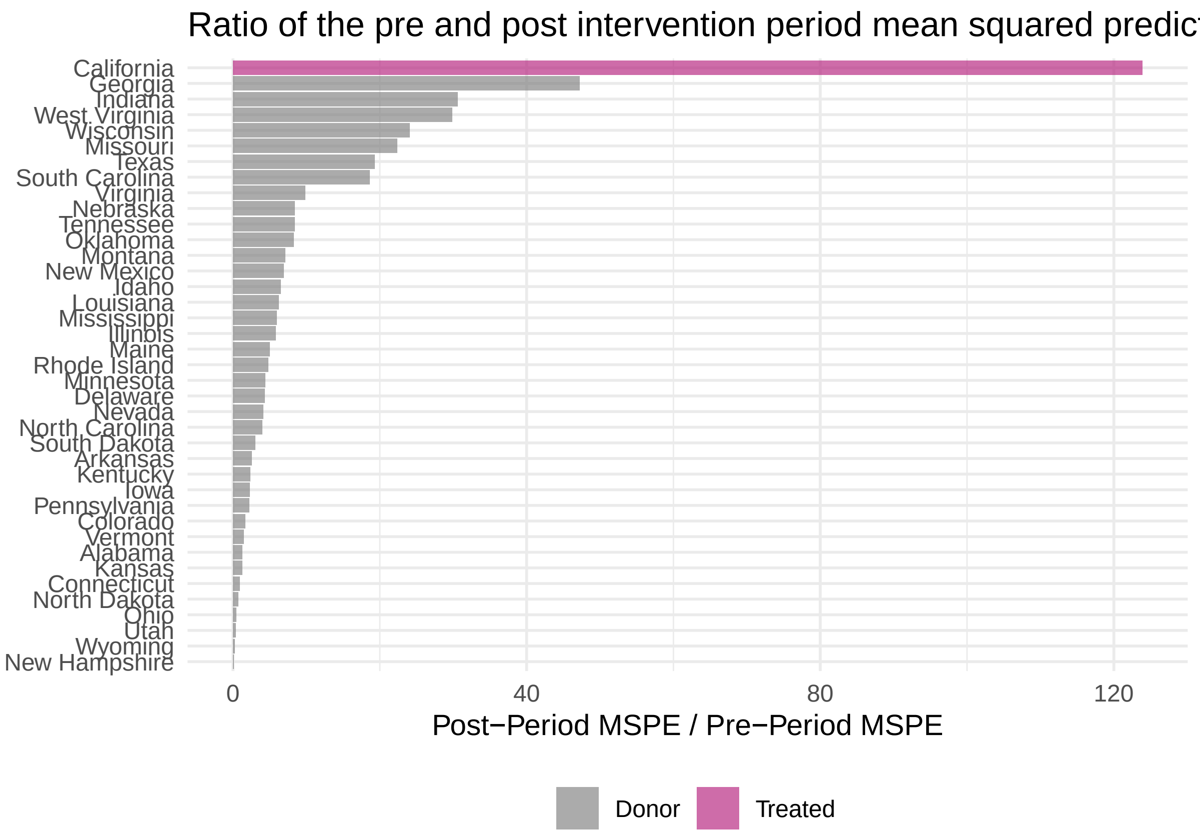


Pruned all placebo cases with a pre−period RMSPE exceeding two times the treated unit's pre−period RMSPE.

Finally, Adabie et al. 2010 outline a way of constructing Fisher's Exact P-values by dividing the post-intervention MSPE by the pre-intervention MSPE and then ranking all the cases by this ratio in descending order. A p-value is then constructed by taking the rank/total.1 The idea is that if the synthetic control fits the observed time series well (low MSPE in the pre-period) and diverges in the post-period (high MSPE in the post-period) then there is a meaningful effect due to the intervention. If the intervention had no effect, then the post-period and pre-period should continue to map onto one another fairly well, yielding a ratio close to 1. If the placebo units fit the data similarly, then we can't reject the hull hypothesis that there is no effect brought about by the intervention.

This ratio can be easily plotted using plot_mspe_ratio(), offering insight into the rarity of the case where the intervention actually occurred.

```
plot_mspe_ratio(smoking_out)
```



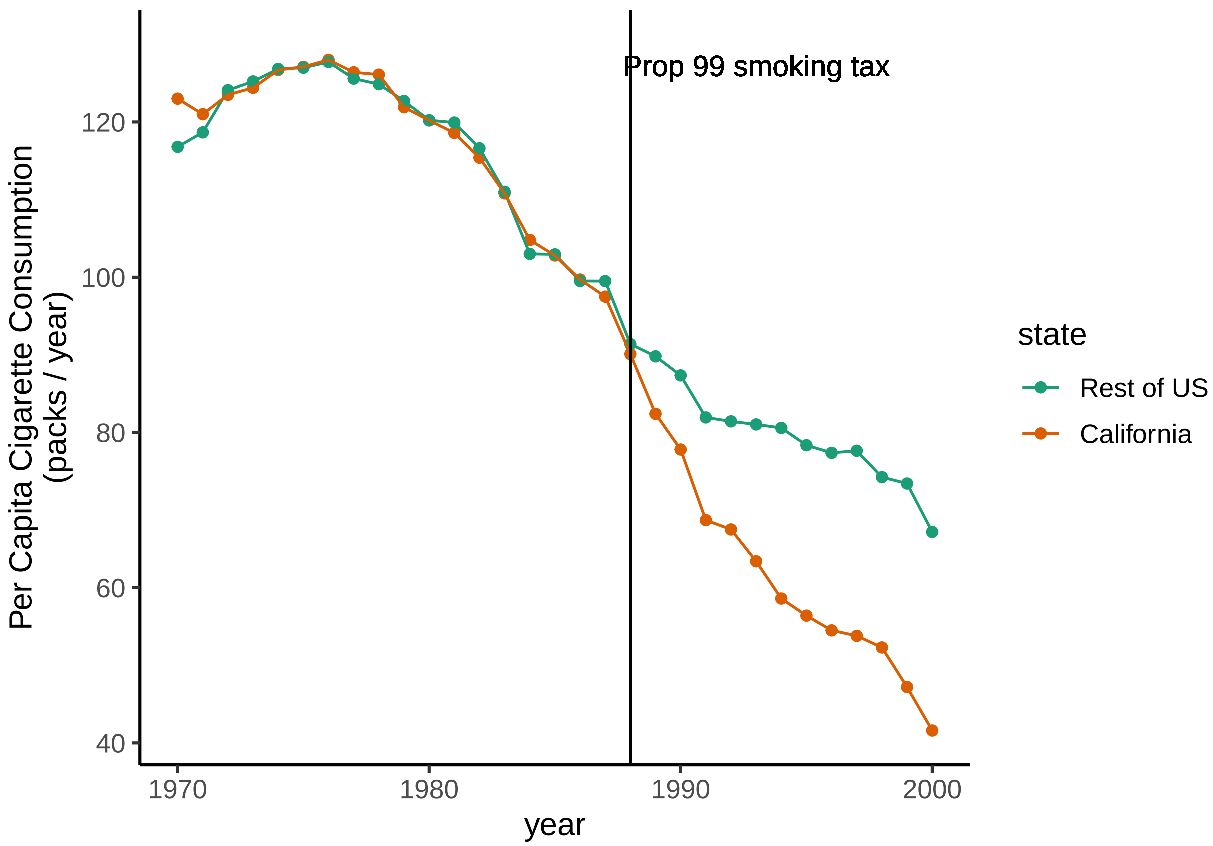
Prep data and rerun ITS model

```
sc <- grab_synthetic_control(smoking_out)

dreal <- sc[c("time_unit","real_y")]
dreal$state = rep("California", nrow(dreal))
names(dreal) <- c("year","cigsale","state")
dsynth <- sc[c("time_unit","synth_y")]
dsynth$state = rep("Rest of US", nrow(dsynth))
names(dsynth) <- c("year","cigsale","state")
ds <- as.data.table(rbind(dreal, dsynth))

ds[, ":="(
  treated = year >= 1988,
  years_after_intervention = ifelse(year < 1988, 0, year - 1988),
  year0 = year - 1988
)]

ds$state <- factor(ds$state, levels = c("Rest of US", "California"))

smoking.ds.means <- ds[,
  lapply(.SD,
         mean,
         na.rm=TRUE),
  by=c("state","treated"),
  .SDcols=c("cigsale")]
```

```r
smoking.ds.means
```

```
##          state treated   cigsale
##         <fctr>  <lgcl>     <num>
## 1: California   FALSE 117.66111
## 2: California    TRUE  62.63846
## 3: Rest of US   FALSE 117.34447
## 4: Rest of US    TRUE  80.13355
```

```r
ggplot(data = ds,
       aes(x = year,
           y = cigsale,
           color = state,
           group = state)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1988) +
  labs(y = "Per Capita Cigarette Consumption\n(packs / year)") +
  geom_text(aes(x=1993,
                label="\nProp 99 smoking tax",
                y=130),
            # angle=90,
            color="black") +
  scale_color_brewer(palette = 'Dark2') +
  theme_classic(base_size = 12)
```




```r
m3 <- lm(cigsale ~
           year0
```

```r
         + treated
         + years_after_intervention
         + state
         + year0:state
         + treated:state
         + years_after_intervention:state
         , data = ds)
summary(m3)
```

```
##
## Call:
## lm(formula = cigsale ~ year0 + treated + years_after_intervention +
##     state + year0:state + treated:state + years_after_intervention:state,
##     data = ds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5947  -3.1379   0.5003   3.8963   7.6481
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            103.8843     2.4807  41.878  < 2e-16
## year0                                   -1.4169     0.2292  -6.182 8.68e-08
## treatedTRUE                            -13.7535     3.6255  -3.794 0.000377
## years_after_intervention                -0.2494     0.4386  -0.569 0.572007
## stateCalifornia                         -1.2477     3.5082  -0.356 0.723487
## year0:stateCalifornia                   -0.1647     0.3241  -0.508 0.613472
## treatedTRUE:stateCalifornia             -4.9249     5.1273  -0.961 0.341069
## years_after_intervention:stateCalifornia -1.7224     0.6202  -2.777 0.007524
##
## (Intercept)                            ***
## year0                                  ***
## treatedTRUE                            ***
## years_after_intervention
## stateCalifornia
## year0:stateCalifornia
## treatedTRUE:stateCalifornia
## years_after_intervention:stateCalifornia **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.044 on 54 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9616
## F-statistic: 219.4 on 7 and 54 DF,  p-value: < 2.2e-16
```