```
#----------------------#
# SMART SUITE
# SPRING 2022
#
# CREATED BY:
#
# Giovanni M. Circo, Ph.D
# University of New Haven
# gcirco@newhaven.edu
#----------------------#

library(Synth)

set.seed(1)

# data comes originally from Scott Cunningham's Book:
# CAUSAL INFERENCE: THE MIXTAPE
# see:
# https://mixtape.scunning.com/synthetic-control.html#prison-
construction-and-black-male-incarceration


# dependent variable is black male incarceration 'bmprison'
# policy goes into effect in 1993

# file url
df <- url('https://www.dropbox.com/s/0cvckcpsggcgtyi/texas_prison.csv?
raw=1')

# read into R
texas <- read.csv(df)


# SYNTH
#----------------------#


# first, need to set up a data object
# using the dataprep() function


# identify pool of control units
# this is a vector of state FIPS codes, excluding texas
controls <- unique(texas$statefip[texas$state != "Texas"])


# set up data using dataprep
# the Synth dataprep function wants a minimum of the following:

# vector of predictor variables (predictors)
```

```
# function to average over predictors (predictors.op)
# name of the dependent variable (dependent)
# vector to index observational units (unit.variable)
# optional vector of names of obs. units (unit.names.variable)
# name of time variable (time.variable)
# numeric variable that IDs the treated unit (treatment.identifier)
# vector that IDs the control units (controls.identifier)
# time periods pre-treatment (time.predictors.prior)
# time periods over which Synth should try and minimize the MSPE
(time.optimize.ssr)
# time periods that should be plotted in the results (time.plot)

synth_prep <-
  dataprep(
    foo = texas,
    predictors = c("poverty", "income", "black", "bmprison"),
    predictors.op = "mean",
    dependent = "bmprison",
    unit.variable = "statefip",
    unit.names.variable = "state",
    time.variable = "year",
    treatment.identifier = 48,
    controls.identifier = controls,
    time.predictors.prior = 1985:1993,
    time.optimize.ssr = 1985:1993,
    time.plot = 1985:2000
  )

# Now we use the dataprep object to fit the model
# depending on size, number of variables can take a while

texas.Synth <- synth(synth_prep)

# To examine the weights, we can call the `synth.tab` function
# where we pass in the model and the dataprep data

Synth.results.tab <- synth.tab(texas.Synth, synth_prep)

# this has a few objects of interest:
# this is the pre-treatment balance
Synth.results.tab$tab.pred

# This is the weights assigned to the predictor variables
# NOTE: measures of pre-treatment outcomes have the highest weight by
far
Synth.results.tab$tab.v

# This is the weights assigned to each control unit
# California, Florida, Louisiana get most of the weight here
Synth.results.tab$tab.w
```

```
# for plotting we will use the built-in `path.plot` and `gaps.plot`
# all we need to specify is the dataprep object and our model

# here, for simplicity, I plot both at the same time, but you can
# also do it individually as well

par(mfrow = c(1,2))
path.plot(dataprep.res =  synth_prep, synth.res = texas.Synth)
abline(v = 1993, col = "red", lty = 2)
gaps.plot(dataprep.res =  synth_prep, synth.res = texas.Synth)
abline(v = 1993, col = "red", lty = 2)


# PLACEBO TESTING
#----------------------#

# Placebo "falsification" tests
# here we assign treatment to every unit in turn
# calculate a treatment effect for each unit, then compare the
observed effect
# against the placebos

# WARNING: This takes a while to run,  because we have to re-run the
model
# for each of the 50 states (about 3-4 minutes on my computer)

# initialize vector of state fips IDS to iterate through
# and a list to hold each of the results
fips <- unique(texas$statefip)
res <- list()

# now use a for loop to iterate through states
for(i in 1:length(fips)) {
  p.prep <-
      dataprep(
        foo = texas,
        predictors = c("poverty", "income", "black", "bmprison"),
        predictors.op = "mean",
        dependent = "bmprison",
        unit.variable = "statefip",
        unit.names.variable = "state",
        time.variable = "year",
        treatment.identifier = fips[i],
        controls.identifier = fips[-i],
        time.predictors.prior = 1985:1993,
        time.optimize.ssr = 1985:1993,
        time.plot = 1985:2000
      )
```

```
  # run synth on prep object
  p.Synth <- synth(p.prep)

  # extract the treatment effect
  res[[i]] <- p.prep$Y1plot - p.prep$Y0plot %*% p.Synth$solution.w
}

# This is just a for loop to plot through placebo fits
# We plot the actual treated unit, Texas, which is observation 44 in
the list
# then we loop through and plot the remaining values in light grey
# NOTE: There are also some tools bundled in the `SCTools` package to
do this
par(mfrow = c(1,1))
plot(x = 1985:2000, y = res[[44]],  type = "l", lwd = 3)
for (i in 1:length(fips)) {
  lines(x = 1985:2000, y = res[[i]], col = rgb(red = .1, green = .1,
blue = .1, alpha = 0.2))
}

# Compare the rank of Texas in terms of post-treatment effect size
# Abadie suggests using the root mean squared error
# we'll use a custom function here:
# this just calculates the RMSE for the pre-period and post-period
# then returns the difference

# root mean squared error of pre t0 and post t1 periods
rmse <- function(x, t0, t1){

  pre <- sqrt(mean(x[1:t0]^2))
  post <- sqrt(mean(x[(t0+1):t1]^2))

  return(post - pre)

}

# get RMSE results in a dataframe
# link to state IDs
res_df <- cbind.data.frame(state = unique(texas$state),
                           tx = unlist(lapply(res, rmse, t0 = 9, t1 =
16)))

# which state has the largest post-treatment RMSE?
# texas is the largest, which is a p-value of
# 1/51 = 0.0196
res_df[which.max(res_df$tx),]

# plot the distribution of post-treatment RMSE
# texas stands out as being the largest, which is 4x larger
# than the next largest placebo
```

```
barplot(res_df$tx, names.arg = res_df$state)
```