# Synthetic Control

### Frankensteining a Clone

Fernando Rios-Avila

## Introduction

- One more last time. What is the Goal of Causal Analysis?

> **!** Important
>
> **The goal of Causal Analysis is to identify how a treatment affects the outcome by itself, once all other factors are kept constant or controlled for.**

From a theoretical point of view, that is very easy. You simply compare two Potential outcomes:

$$TE_i = y_i(1) - y_i(0)$$

and aggregate those outcomes as needed:

$$ATT = E(TE_i|D=1); ATU = E(TE_i|D=0); ATE = E(TE_i); ATX = E(TE_i|X)$$

Unfortunately, we only observe one outcome. You are either treated or untreated...So how do we fix this?

You need to find counterfactuals so both observed ($X$) and unobserved ($e$) are the same (or close) between treated and contro group.

1. **RCT**: Gold Standard, You randomize treatment and compare means. If correctly done, $X's$ and $e's$ will be comparable across groups, and ATE's can be identified.

2. **Reg + FE**: For other cases, we just work with observational data. First method, Regression (OLS?). Adding covariates controls for their presence, working as a pseudo balancing approach.

   You could also add *fixed effects*, to control for factors that are fixed (across time), but you do not observe. (requires Panel data).

   It works if Treatment occurs at the same time for everyone treated. and if Unobserved are "fixed"

3. **Instrumental variables**: 2nd Best to RCT. It uses IV to generate a small randomization process that can be used for estimating ATT. Technically it compares the effect among those potentially affected by the random instrument. Requires Randome instrument, and no-defiers. Its a Local ATE
4. **Matching and Reweigthing**. Similar to Regression, but better to balance characteristics. The goal is to find units with similar characteristics for all treated units. You can estimate ATE, ATT or ATU. Depends on how well Matching is done
5. **RDD**. If you have data where treatment depends on a single variable and a threshold, you can use this to identify TE for those "Near" the threshold. They Key assumption, treatment assigment is as good at random at the threshold.

6. **DD**. Differences in differences uses variation across time and across individual to identify treatment effects. Under PTA, and SUTVA

   Dif. within individuals eliminates common time trends, Dif across time, eliminates individual fixed effects. DD provide you with ATT's for the treated, after treatment.

$$ATT = (Y_{g=1,t=1} - Y_{g=1,t=0}) - [(Y_{g=0,t=1} - Y_{g=0,t=0})]$$

   Can be generalized to Many periods and many groups, but requires stronger assumptions (no anticipation and no change in treatment status), and further aggregation.

   Or combined with Matching for even better results.

## Synthetic Control: Special case

As previous Cases, Synthetic control aims to identify treatment constructing appropriate "counterfactuals".

It is said that Synthethic controls may be even MORE credible methodology, because the treated group is by construction Exogenous…but how?

> The treated group is a Case Study.
>
> An isolated event or unit that is affected by a treatment, and should not affect other units !

In this sense, the treatment is exogenous, because it affected a single unit.

But what about the counterfactual?

- In other methods (in particular Matching), our "conterfactual" mean to look for observations that had the same characteristics as the treated observation.
- Some times, we needed to settle to use a single "bad" control, because we couldnt find one better. (people are very different).
  - Using Stricter criteria would make it unfeasible.
  - More relax and we have lots of biases.
- **SC** is different. You have **MANY** controls, so why settle with only one?
- **SC** is like Dr Frankenstein, where you "build" a single comparison group by averaging information of all controls.
- You build the synthetic control getting "weighted averages".
- But…we assume you can see all units across time (panel data)

## This is a very popular method

Where has this method been used:

- effects of right-to-carry laws (Donohue et al., 2019),
- legalized prostitution (Cunningham and Shah, 2018),
- immigration policy (Bohn et al., 2014),
- corporate political connections (Acemoglu et al., 2016),
- taxation (Kleven et al., 2013),

- organized crime (Pinotti, 2015)

Just to name a few.

### Assumptions:

1. The Donor Pool should be a good match for the treated unit. Thus, the synthethic control should be Zero before treatment.
    - This is similar to PTA, but stronger. Before treatment, there should be no difference between Treated and synthetic control
2. SUTVA. Only the treated group is affected by treatment. The control group should be unaffected (no spill over effects).
3. There should be NO other "event" in the period of analysis. (Thus we only capture treatment impact)

### How does it work.

Recall, we want to estimate TE for the single untreated unit:

$$ATT_{1t} = Y_{1t} - Y(0)_{1t}$$

but we do not observe $Y(0)_{1t}$. We only know that before treatment

$$ATT_{1t} = Y_{1t} - Y(0)_{1t} = 0$$

We could construct a synthetic control:

$$\hat{Y}_{1t}(0) = \sum_{i=2}^{N} w_i Y_{it}$$

At the very least, the weights $w$ should be such that before treatment ($G$):

$$Y_{1t} = \sum_{i \neq 1} w_i Y_{it} \ \forall \ t < G$$

At the very least, the weights $w$ should be such that before treatment ($G$):

$$Y_{1t} = \sum_{i \neq 1} w_i Y_{it} \ \forall \ t < G$$

Havent we seen seen something like this Before? OLS:

$$y = x\beta + e$$
$$y_1^t = a_0 + y_i^t w + e$$

$$\begin{bmatrix} y_1^1 \\ y_2^1 \\ \dots \\ y_{G-1}^1 \end{bmatrix} = a_0 + \begin{bmatrix} y_1^2 & y_1^3 & \dots & y_1^k \\ y_2^2 & y_2^3 & \dots & y_2^k \\ \dots & \dots & \dots & \dots \\ y_{G-1}^2 & y_{G-1}^3 & \dots & y_{G-1}^k \end{bmatrix} \begin{bmatrix} w_2 \\ w_3 \\ \dots \\ w_k \end{bmatrix} + e$$

$$y_1^t = a_0 + y_i^t w + e$$

- In this Specification, each row (observation) is a "pre-treatment" period of observed data.
- and each control unit (from the many controls) will be a variable.

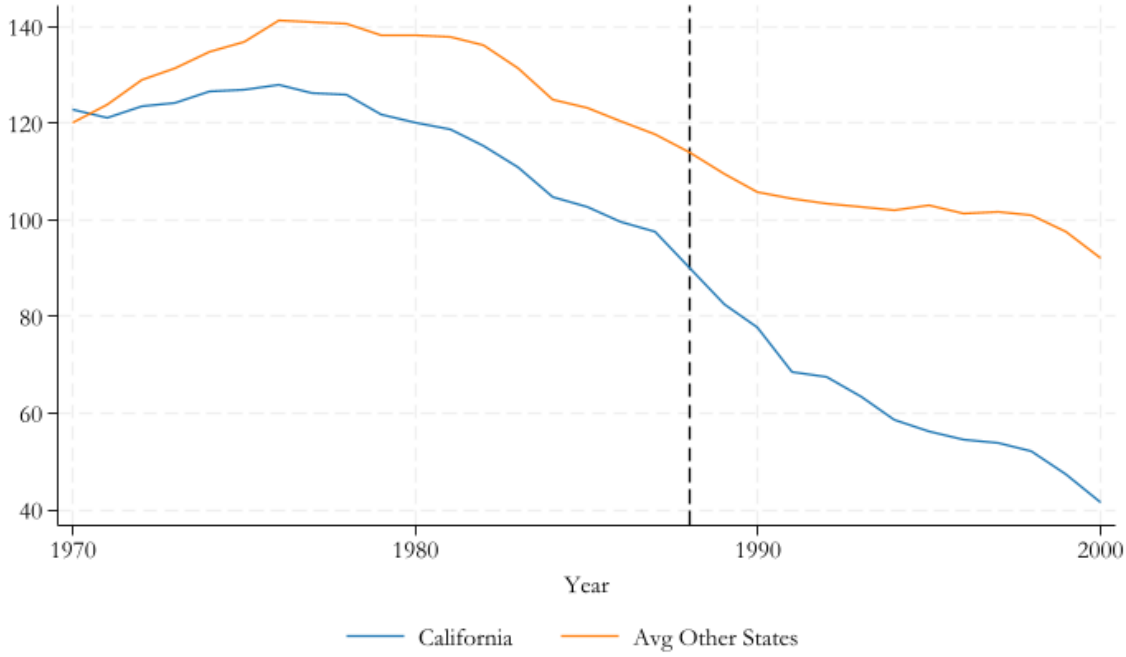
OLS can help you find the weights, which can then be used for obtaining the "Synthetic" control

### Small Example

```
qui:frause smoking, clear
color_style tableau
bysort year:egen mean_cig=mean(cigsale) if state!=3
two (line cigsale year if state ==3) (line mean_cig year if state==1), ///
    legend(order(1 "California" 2 "Avg Other States") pos(6) col(2)) xline(1988)
```

```
<IPython.core.display.HTML object>
```

```
(31 missing values generated)
```




```
drop mean_cig
qui:reshape wide cigsale lnincome beer age15to24 retprice , i(year) j(state)
ren cigsale1 mcigsale
```

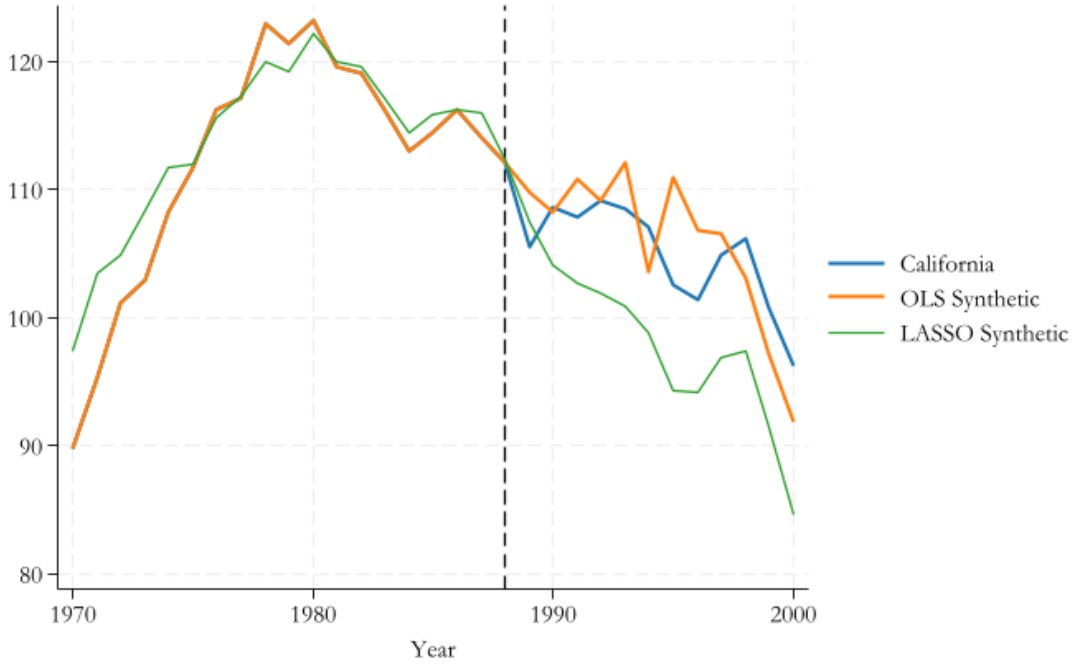
- Now we have...38 variables, (other States but California)
- And 31 periods (only 19 Before treatment)
- Can we estimate the weights using OLS?

...

- Nop. N<K !

```
qui:reg mcigsale cigsale* if year<=1988, nocons
predict mcigh1
qui: lasso linear  mcigsale cig* if year<=1988, nocons
predict mcigh2
 two (line mcigsale year, lw(0.5)  ) ///
  (line mcigh1 mcigh2   year, lw(0.5)  ) , ///
  legend(order(1 "California" 2 "OLS Synthetic" 3 "LASSO Synthetic")) xline(1988)
```

```
(option xb assumed; fitted values)
(options xb penalized assumed; linear prediction with penalized coefficients)
```




- OLS Not appropriate (specially if N<K)
- Lasso Better, because of regularization, but not great.
- We are not controlling for other factors either (controls)
- Other Details we cover next

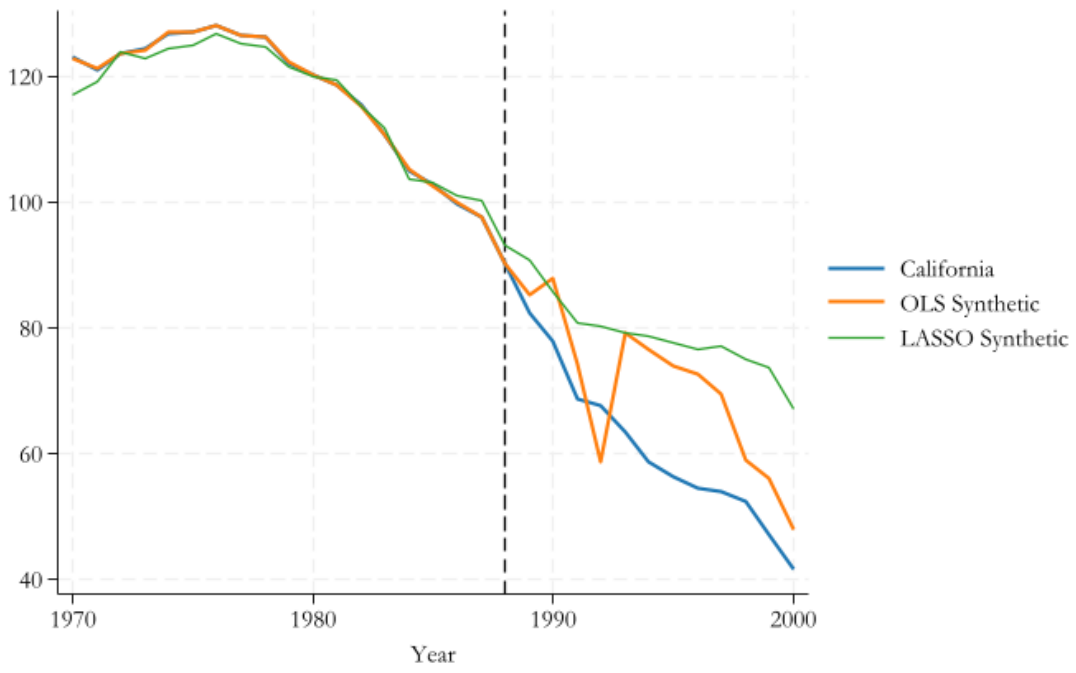
### Allowing for Covariates:

- As with other methodologies, one should also considered controlling for covariates.
- Specifically, more covariates can be allowed by Stacking them:

$$\begin{bmatrix} y_1^t \\ x_1^t \\ z_1^t \end{bmatrix} = (a_0 = 0) + \begin{bmatrix} y_2^t & y_3^t \dots & y_k^t \\ x_2^t & x_3^t \dots & x_k^t \\ z_2^t & z_3^t \dots & z_k^t \end{bmatrix} \begin{bmatrix} w_2 \\ w_3 \\ \dots \\ w_k \end{bmatrix} + e$$

```
qui:frause smoking, clear
ren (cigsale lnincome beer age15to24 retprice) ///
      (var1    var2     var3 var4      var5)
qui: reshape long var, i(state year)    j(new)
qui: reshape wide var, i(year new)  j(state)
label define new 1 "cigsale" ///
                 2 "lnincome" ///
                 3 "beer" ///
                 4 "age15to24" ///
                 5 "retprice", modify
label values new new
ren var3 cal_out
qui:reg cal_out var* if year<=1988, nocons
predict mcigh1
qui: lasso linear   cal_out var* if year<=1988, nocons
predict mcigh2
 two (line cal_out year if new==1, lw(0.5)  ) ///
  (line mcigh1 mcigh2   year if new==1, lw(0.5)  ) , ///
  legend(order(1 "California" 2 "OLS Synthetic" 3 "LASSO Synthetic")) xline(1988)
```

```
(option xb assumed; fitted values)
(32 missing values generated)
(options xb penalized assumed; linear prediction with penalized coefficients)
```




### What else to keep in mind

- With More Variables, the goal is still to be able to choose $w's$ that best explain the observed outcomes (and characteristics) of the "treated unit".

$$w = \min_{w} \sum_{m=1}^{K} \left[ v_m \left( X_{1t} - \sum_{j=2}^{J} w_j X_{jt} \right)^2 \right]$$

However, we also need to impose restrictions on Weights:

1. $w_j \geq 0$ Weights cannot be negative.
2. $\sum w_j = 1$ They should sum up to 1.
3. $v_m$ can be used to increase, or reduce the relative importance of factors in the model. (lower bound at 0) The constant is zero.

This is a maximization problem with constrains. Restrictions ensure the prediction is based on a "convex" set, avoiding extrapolation.

## Is it noise? or Causal?

- When using **SC**, you essentially have a sample $n = 1$ to estimate an effect. How do you know that effect is significant? and not just noise?
  - You can do a randomization experiment! and answer:

"how unusual is this estimate under the null hypothesis of no policy effect?".

- How does this work?

## Randomization

1. Excluding the treated unit, estimate the pseudo effect of every other unit in the dataset. These are placebos, and you should expect the effect to be zero for them…but you may see some positive and negative effects.
   - This may be consider the sampling distribution of the estimated effect.
2. Calculate the pre- and post- treament *Root mean squared prediction error* for all units (treated and placebos).
   - Pre-RMSPE provides a statistic of how well the model fits before treatment.
   - Post-RMSPE provides a statistic of how unusual is the outcome after the "treatment date". The largest it is, the more unpredictable (or stronger treatment effect) it would be.

$$RMSPE_i^{pre} = \sqrt{\frac{1}{g-1}\sum_{t=1}^{g-1}(y_{i,t} - \sum_{j \neq i} w_j^i y_{j,t})^2}$$

$$RMSPE_i^{post} = \sqrt{\frac{1}{T-g+1}\sum_{t=g}^{T}(y_{i,t} - \sum_{j \neq i} w_j^i y_{j,t})^2}$$

3. Estimate the ratio between Pre and Post RMSPE, and rank them.

$$Ratio_i = \frac{RMSPE_i^{post}}{RMSPE_i^{pre}}$$

4. The p-value for the treatment is proportional to the Rank:

$$pvalue_i = \frac{rank(i)}{Tot}$$

**Lets continue the example:**

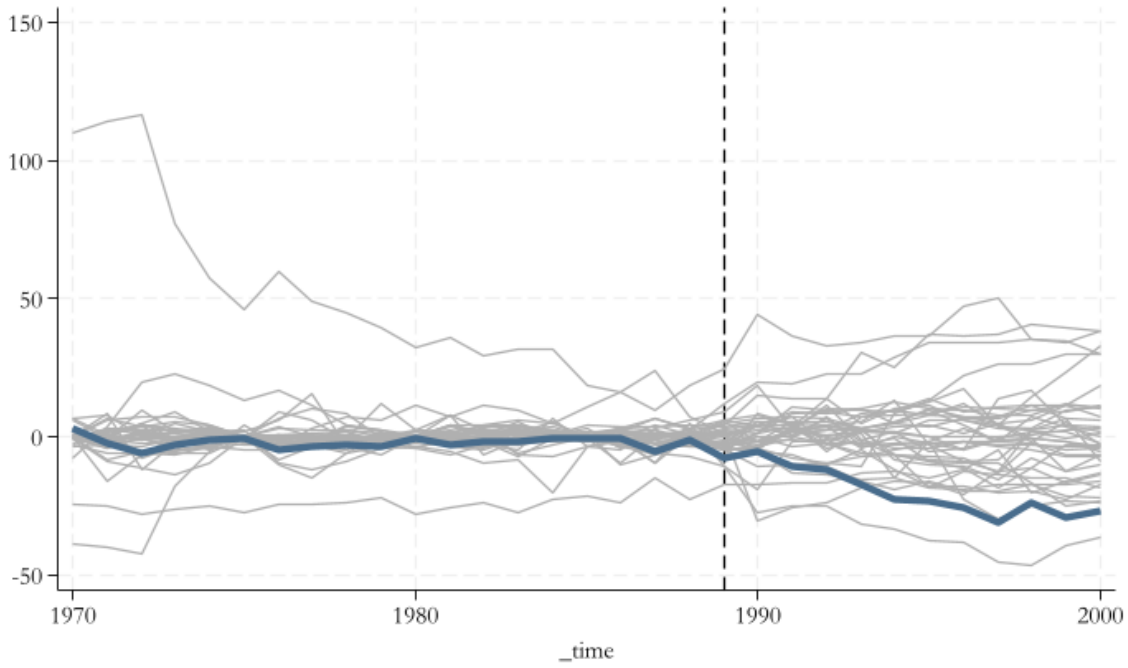
```
qui:frause smoking, clear
xtset state year
tempfile sc3
** For California
synth cigsale cigsale(1970) cigsale(1975) cigsale(1980) cigsale(1985) cigsale(1988), trunit(3
** Same Specification for All other States excluding California
forvalues i =1/39{
    if `i'!=3 {
        local pool
        foreach j of local stl {
            if `j'!=3 & `j'!=`i' local pool `pool' `j'
        }
        tempfile sc`i'
        synth cigsale cigsale(1970) cigsale(1975) cigsale(1980) cigsale(1985) cigsale(1988),
        trunit(`i') trperiod(1989) keep(`sc`i'') replace counit(`pool')
    }
}
** Some data cleaning and prepration
forvalues i =1/39{
    use `sc`i'' , clear
    gen tef`i' = _Y_treated - _Y_synthetic
    egen sef`i'a =mean( (_Y_treated - _Y_synthetic)^2) if _time<=1988
    egen sef`i'b =mean( (_Y_treated - _Y_synthetic)^2) if _time>1988
  replace sef`i'a=sqrt(sef`i'a[1])
    replace sef`i'b=sqrt(sef`i'b[_N])
    drop if _time==.
    keep tef`i' sef`i'* _time
    save `sc`i'', replace
}
**
** Merging all together, and getting ready to plot
**

use `sc1', clear
forvalues i = 2/39 {
    merge 1:1 _time using `sc`i'', nogen
}
global toplot
global toplot2
```

```
forvalues i = 1/39 {
    global toplot $toplot (line tef`i' _time, color(gs11) )
  if (sef`i'a[1])<(2*sef3a[1]) {
        global toplot2 $toplot2 (line tef`i' _time, color(gs11) )
    }
}
```

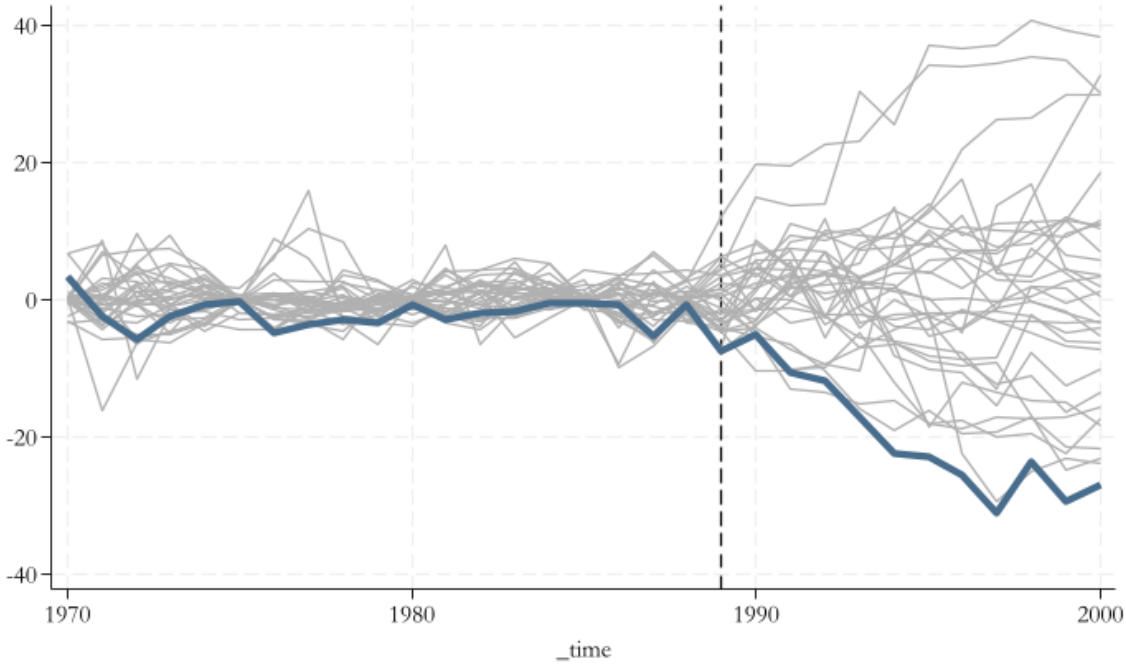
**All Cases**

```
two $toplot (line tef3 _time, lw(1) color(navy*.8)), xline(1989) legend(off)
```




**Good Cases** Restricts to States with Good RMSEP (less than 2 California)

```
two $toplot2 (line tef3 _time, lw(1) color(navy*.8)), xline(1989) legend(off)
```

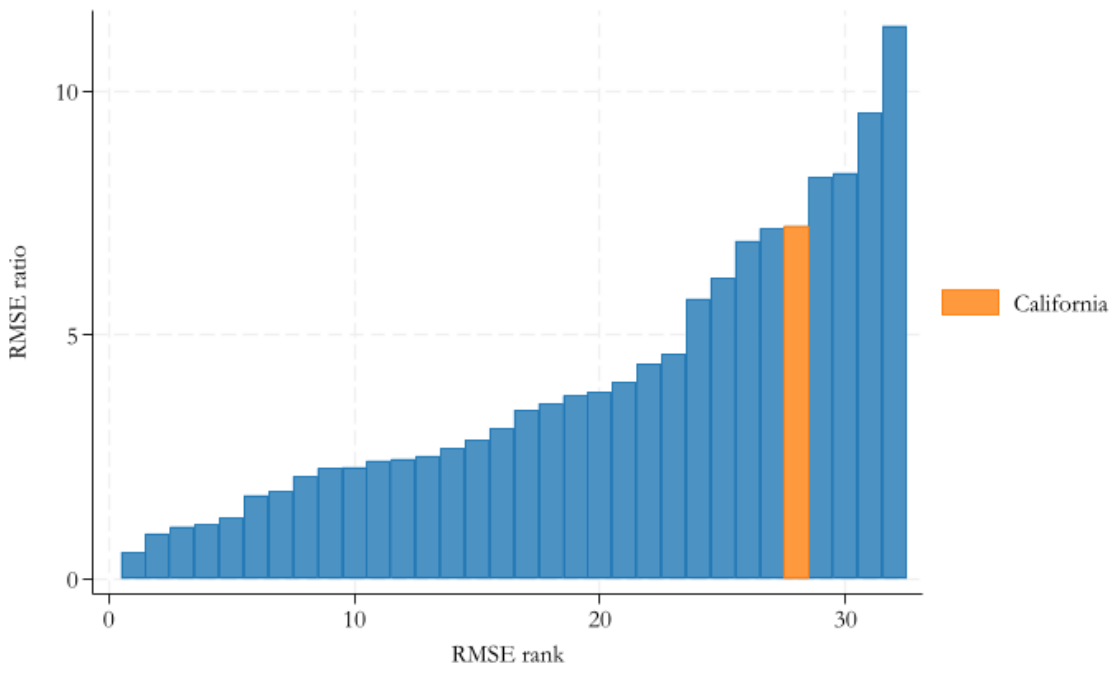



### RMSE Ratio

```
forvalues i = 1/39 {
  if (sef`i'a[1])<(2*sef3a[1]) {
        matrix rt=nullmat(rt)\[`i',sef`i'b[1]/sef`i'a[1]]
    }
}
svmat rt
egen rnk=rank(rt2)

two bar rt2 rnk || bar rt2 rnk if rt1==3 , ///
 legend(order( 2 "California")) ///
  ytitle(RMSE ratio)  xtitle(RMSE rank)
```

```
number of observations will be reset to 32
Press any key to continue, or Break to abort
Number of observations (_N) was 31, now 32.
```



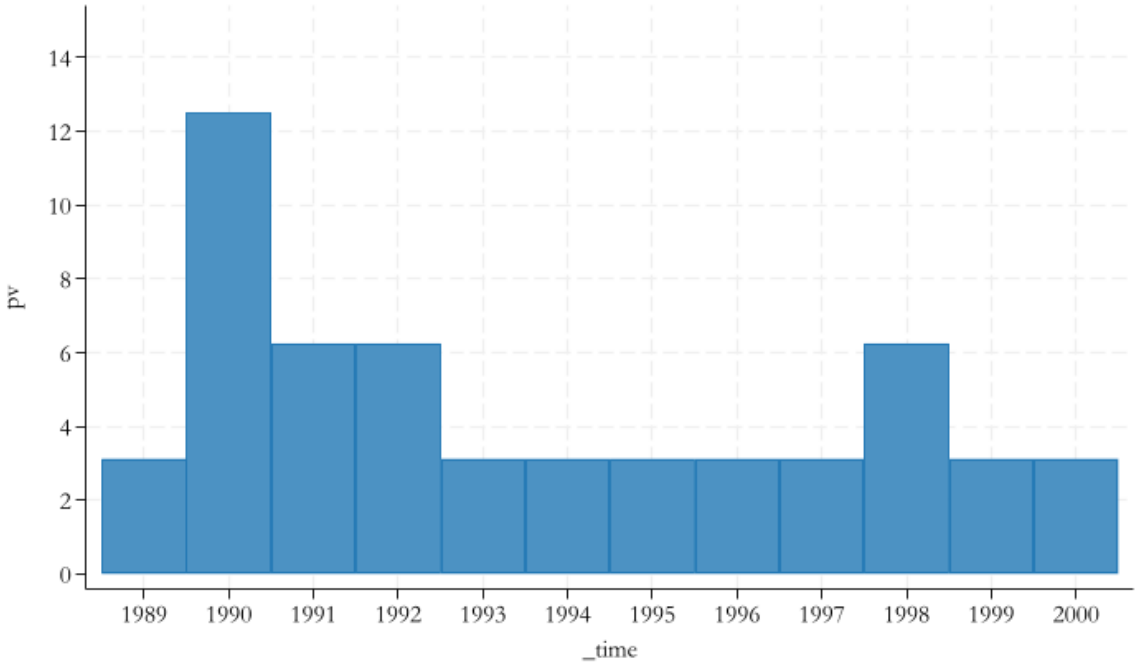

**p-values**

```
 gen rnk2=0
forvalues i = 1/39 {
    if   (sef`i'a[1])<(2*sef3a[1]) {
        local t = `t'+1
        replace rnk2=rnk2+(tef`i'<=tef3)
    }
}
gen pv=rnk2*100/`t'

two bar pv _time if _time>1988 & rnk2<32, ylabel(0(2)15) xlabel(1989/2000)
```

```
(10 real changes made)
(6 real changes made)
(32 real changes made)
(5 real changes made)
(9 real changes made)
(3 real changes made)
(5 real changes made)
```

```
(4 real changes made)
(6 real changes made)
(4 real changes made)
(4 real changes made)
(8 real changes made)
(5 real changes made)
(4 real changes made)
(8 real changes made)
(3 real changes made)
(5 real changes made)
(8 real changes made)
(5 real changes made)
(5 real changes made)
(7 real changes made)
(2 real changes made)
(4 real changes made)
(4 real changes made)
(2 real changes made)
(6 real changes made)
(3 real changes made)
(6 real changes made)
(8 real changes made)
(5 real changes made)
(4 real changes made)
(6 real changes made)
```




## Other Falsification Tests

- Change of treatment Year.
  - In the manual implementation you may want to change the treatment year (to an earler point). One should see no effect between *false* treatment date and the true to be zero.
  - Using `synth` (Stata) you may want to drop some of the controls, so only "pre-false" treatment data is used.
    - Look also into `synth_runner`
- Change of Outcome.
  - One can estimate the effect on alternative outcomes. No effect should be estimated.

## Conclusions

- The basic methodology presented here differs from other strategies because one uses a single treated unit, with pletora of treated groups.
- Instead of comparing single units with the treated group, it aims to compare a weighted average "synthetic control" to do so.

- It will work better than matching because you are focusing on getting the best "weighted" group for a single unit.
- But this methodology is still under development, with extensions toward using dissagregated data, or a combination with DD approaches.
- This may change how much more one can do with the method


## Next week:

Your papers! And Goodluck Friday