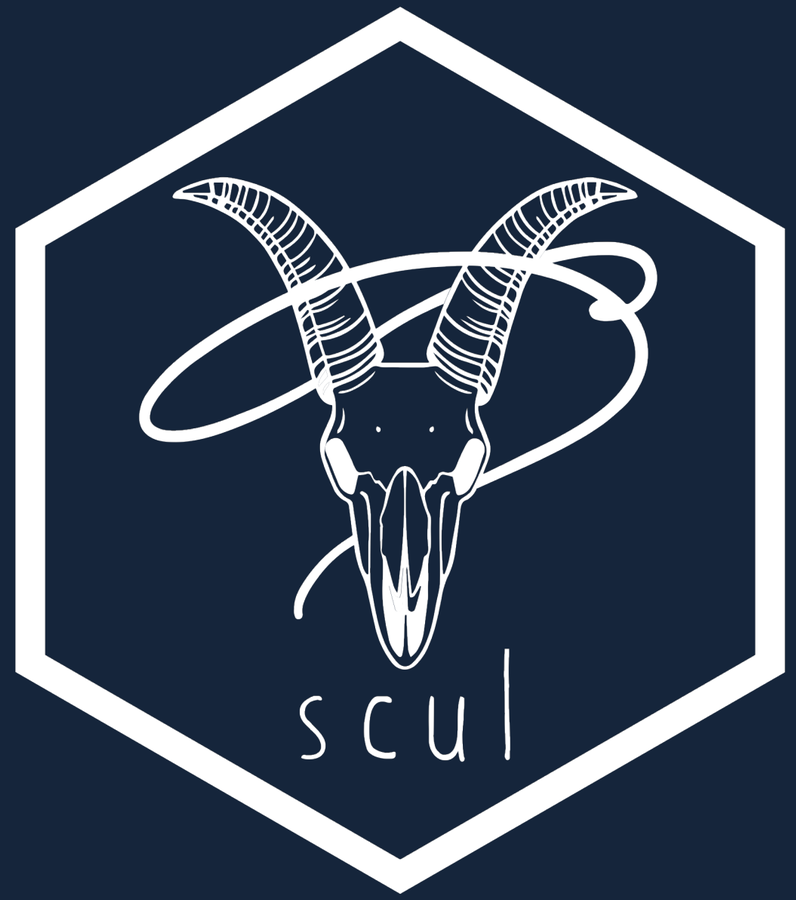
Tactics for design and inference in synthetic control studies

An applied example using high-dimensional data

Alex Hollingsworth and Coady Wing

O'Neill School of Public and Environmental Affairs

Indiana University

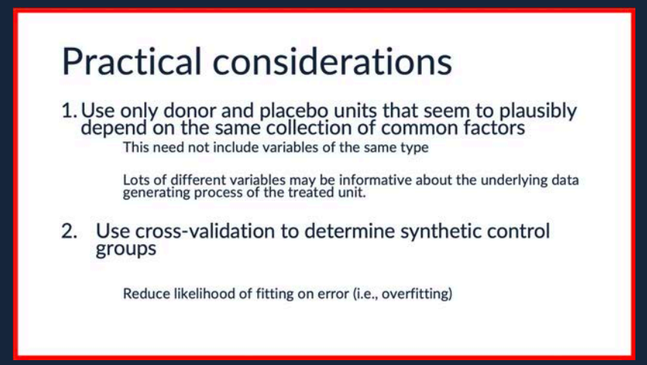


Roadmap

- What is a synthetic control?
- Identification assumptions
- Applied example
  - High-dimensional data (enter lasso)
  - Overview of implementation
  - Practical tips/advice for common issues

Roadmap

- What is a synthetic control?
- Identification assumptions
- Applied example
  - High-dimensional data (enter lasso)
  - Overview of implementation
  - Practical tips/advice for common issues



Practical considerations

1. Use only donor and placebo units that seem to plausibly depend on the same collection of common factors

   This need not include variables of the same type

   Lots of different variables may be informative about the underlying data generating process of the treated unit.

2. Use cross-validation to determine synthetic control groups

   Reduce likelihood of fitting on error (i.e., overfitting)

What is a synthetic control?

What is a synthetic control?

A strategy for estimating causal treatment effects

What is a synthetic control?

A strategy for estimating causal treatment effects

Time series outcomes for a treated unit

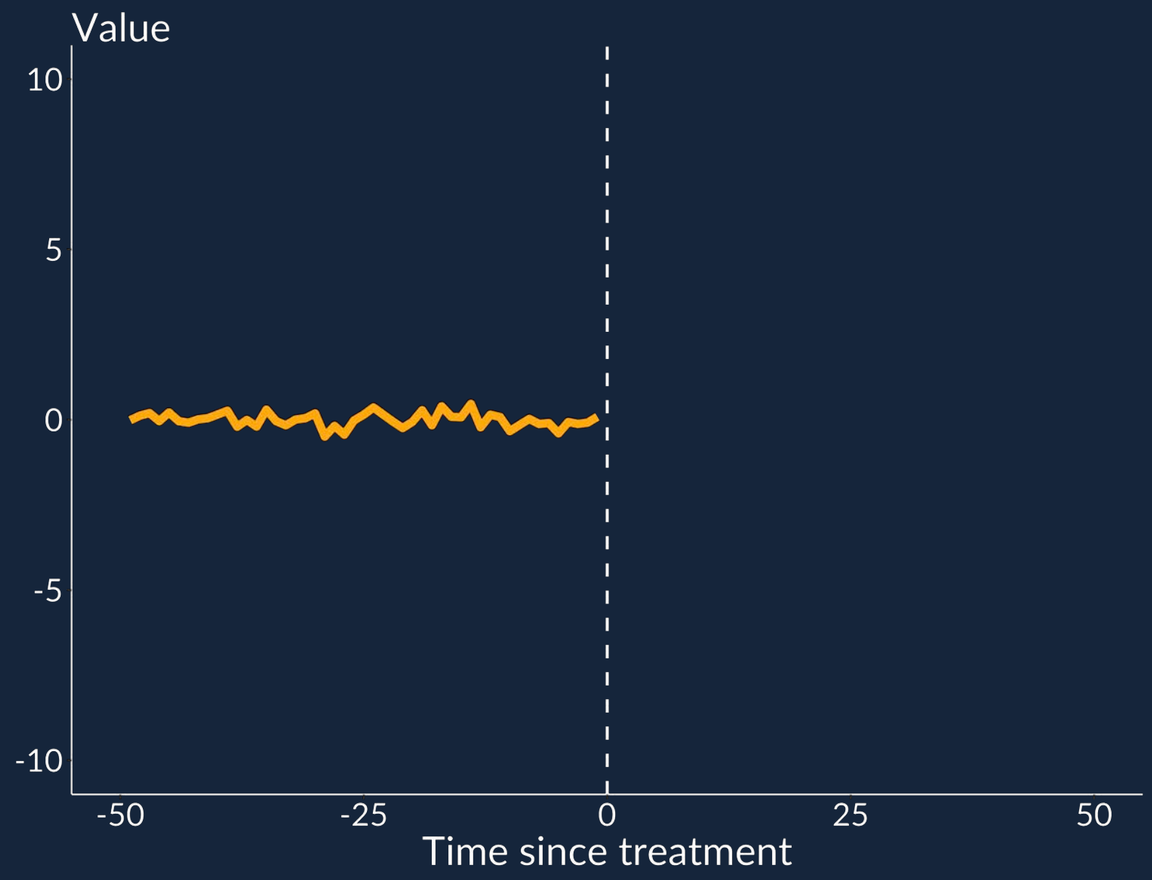
Time series outcomes for a number of untreated units (i.e., the donor pool)

What is a synthetic control?

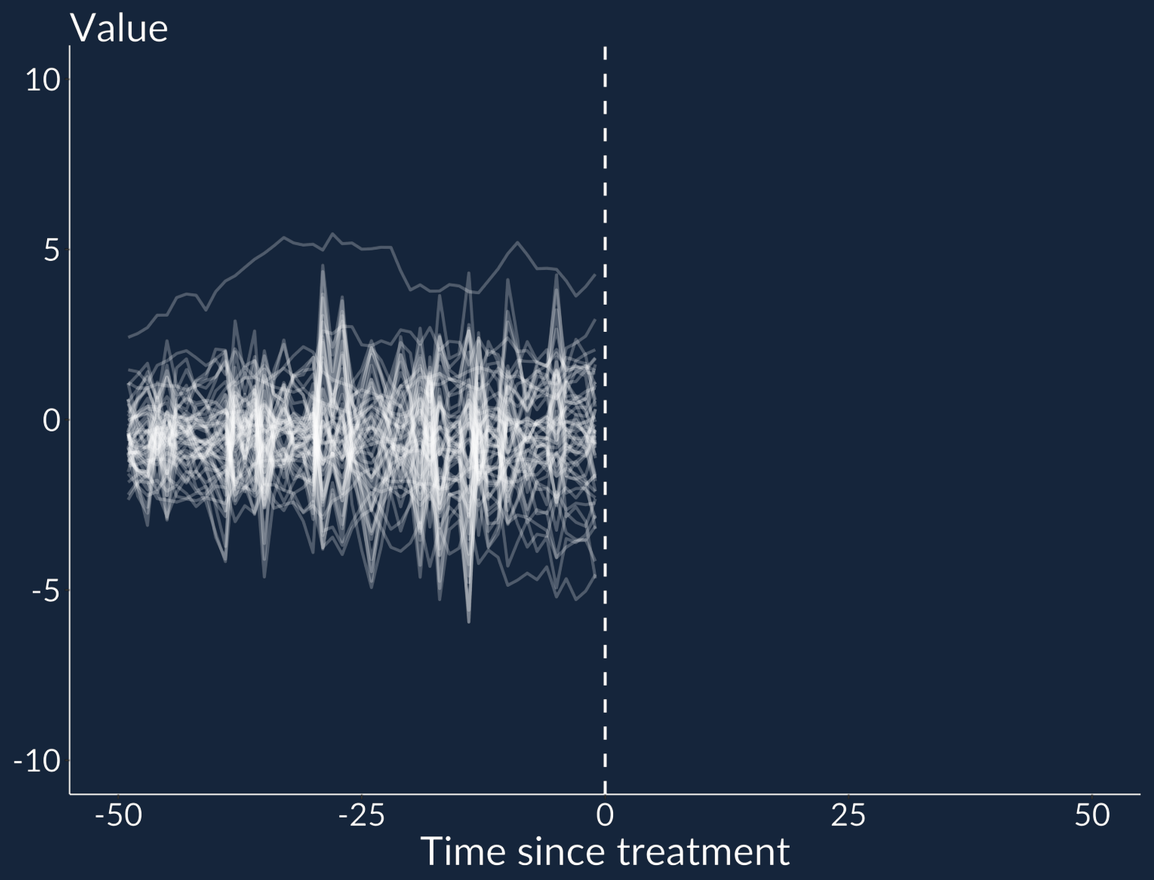
A weighted average of the untreated series is used as a counterfactual estimate of the treated series

As if treatment had not occurred

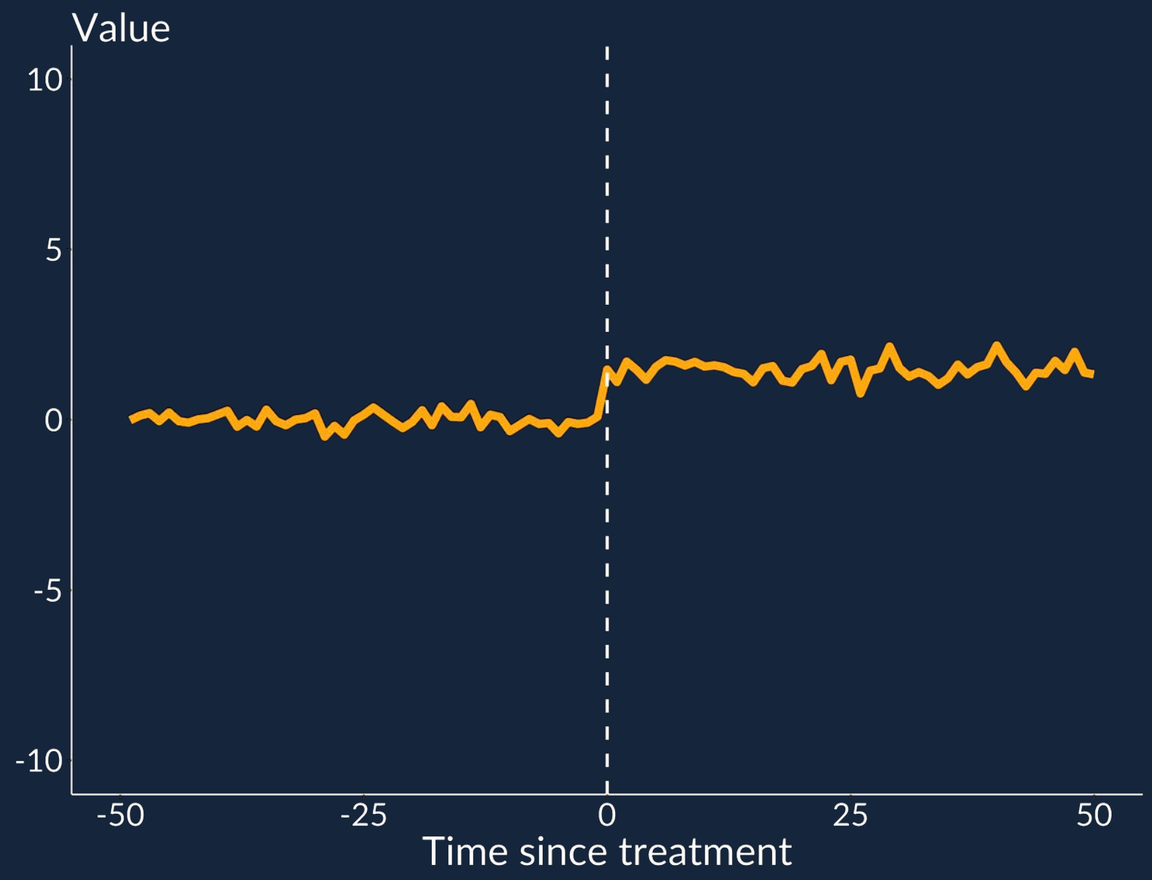
Consider this target series



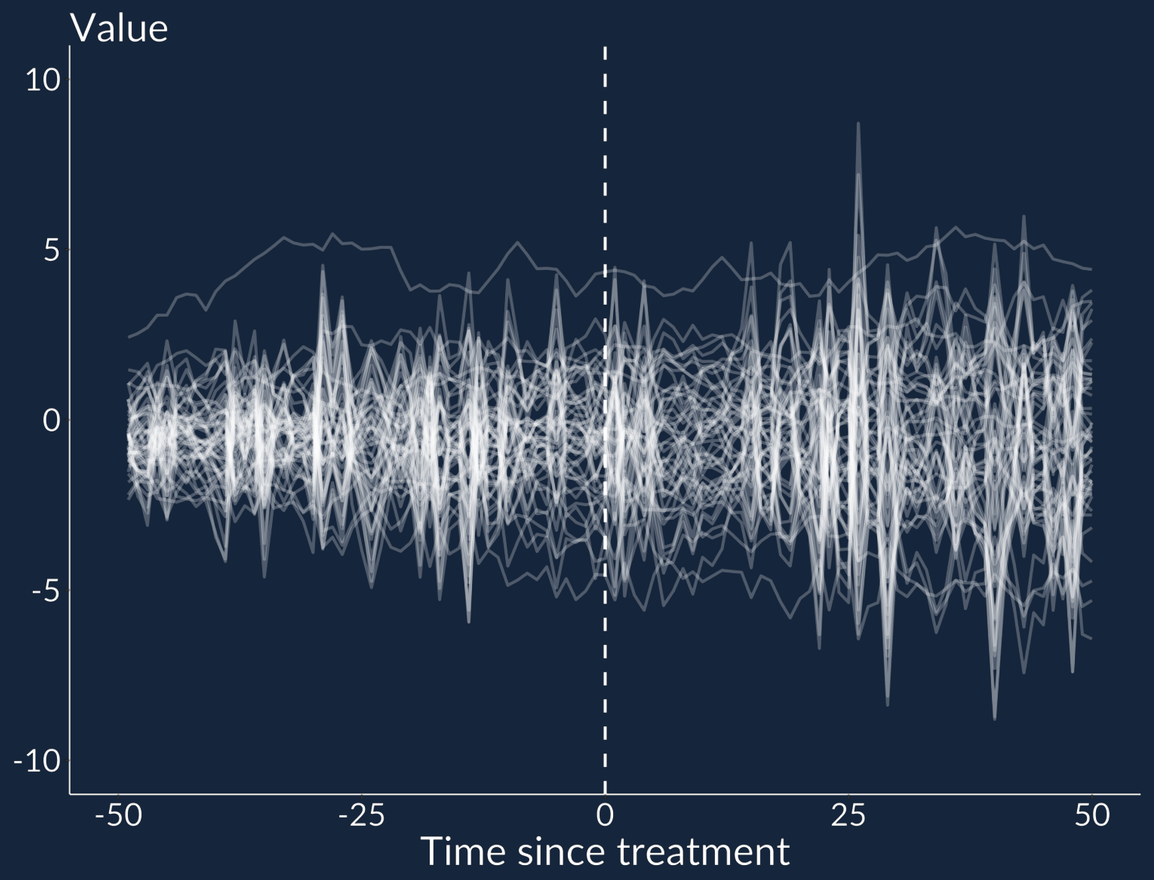
and these untreated series



Some treatment occurs



Never treated



Think of it as matching

Synthetic controls tries to match a target series to untreated donor series based on the unobserved factors that determine the data generating process before a treatment occurs.

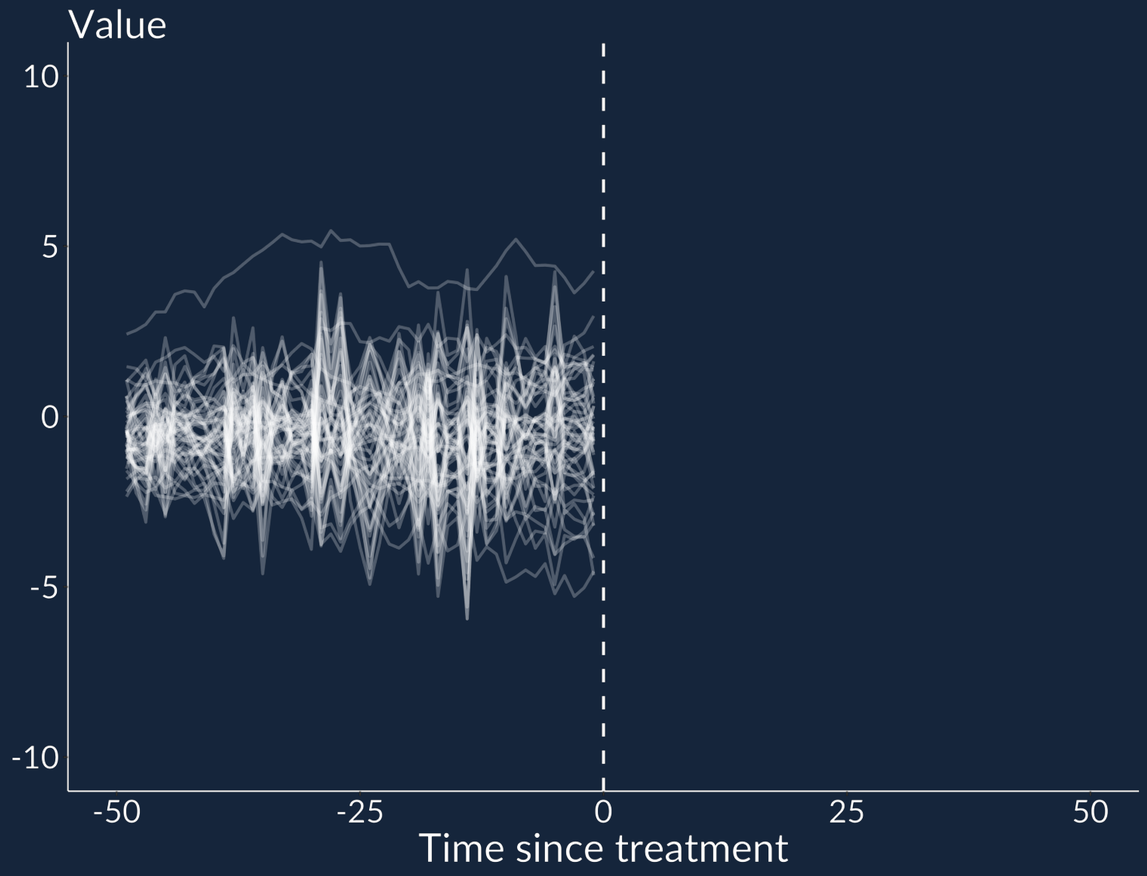
Think of it as matching

Every time series has it's own data generating process

$$y_{it} = \delta_t \alpha_i + \epsilon_{it}$$

y_{it}: outcome
δ_t: common factor
α_i: unit-specific coefficient on δ_t
ϵ_{it}: error

Observed Unobserved



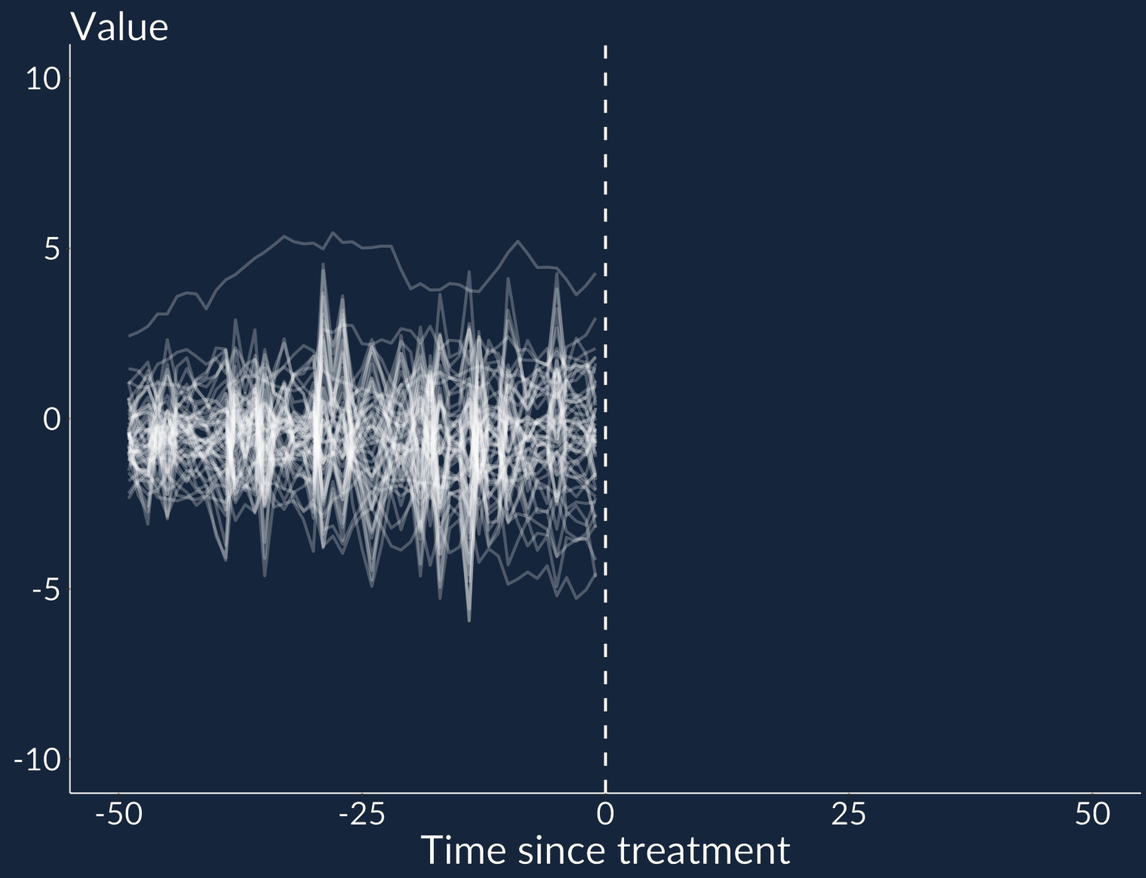
Think of it as matching

Every time series has it's own data generating process

$$y_{it} = \delta_t \alpha_i + \epsilon_{it}$$

y_{it}: outcome
δ_t: common factor
α_i: unit-specific coefficient on δ_t
ϵ_{it}: error

Observed Unobserved



Think of it as matching

Every time series has it's own data generating process

$$y_{it} = \delta_t \alpha_i + \epsilon_{it}$$

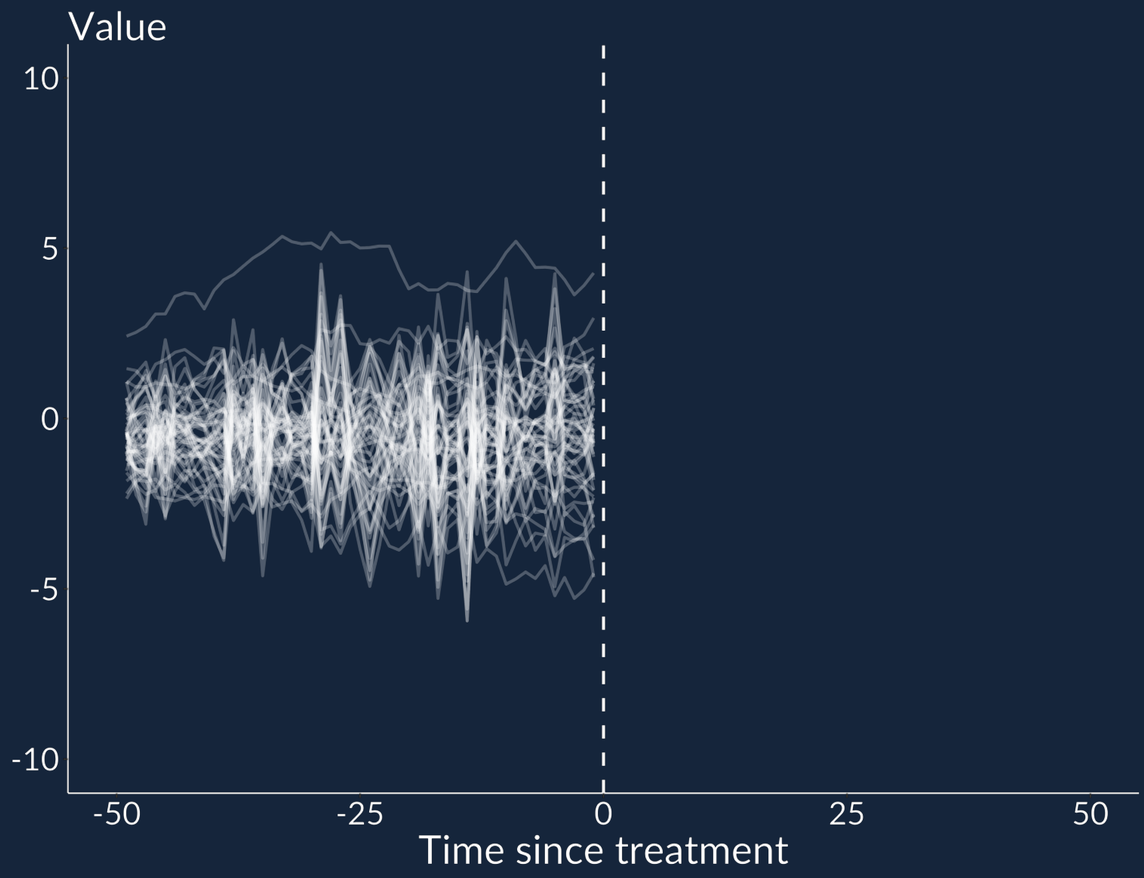
y_{it}: outcome

δ_t: common factor

α_i: unit-specific coefficient on δ_t

ϵ_{it}: error

Observed Unobserved



Think of it as matching

Data generating process

$$y_t^* = \delta_t \alpha_i + \epsilon_{it}$$
$$y_{it} = \delta_t \alpha_i + \epsilon_{it}$$

Ideally you could match on the unit-specific coefficients: α_i

but these are unobserved.

Instead, we match y_t^* on observed outcomes to y_{it}

In the limit, a good match on y_{it} will be matching on α_i

The “match” is a weighted combination of the donor pool

$$y_t^* = \sum_{i=1}^{N} y_{it}\, \pi_i$$

Weights are determined using pre-treatment data and are held fixed over the whole time period

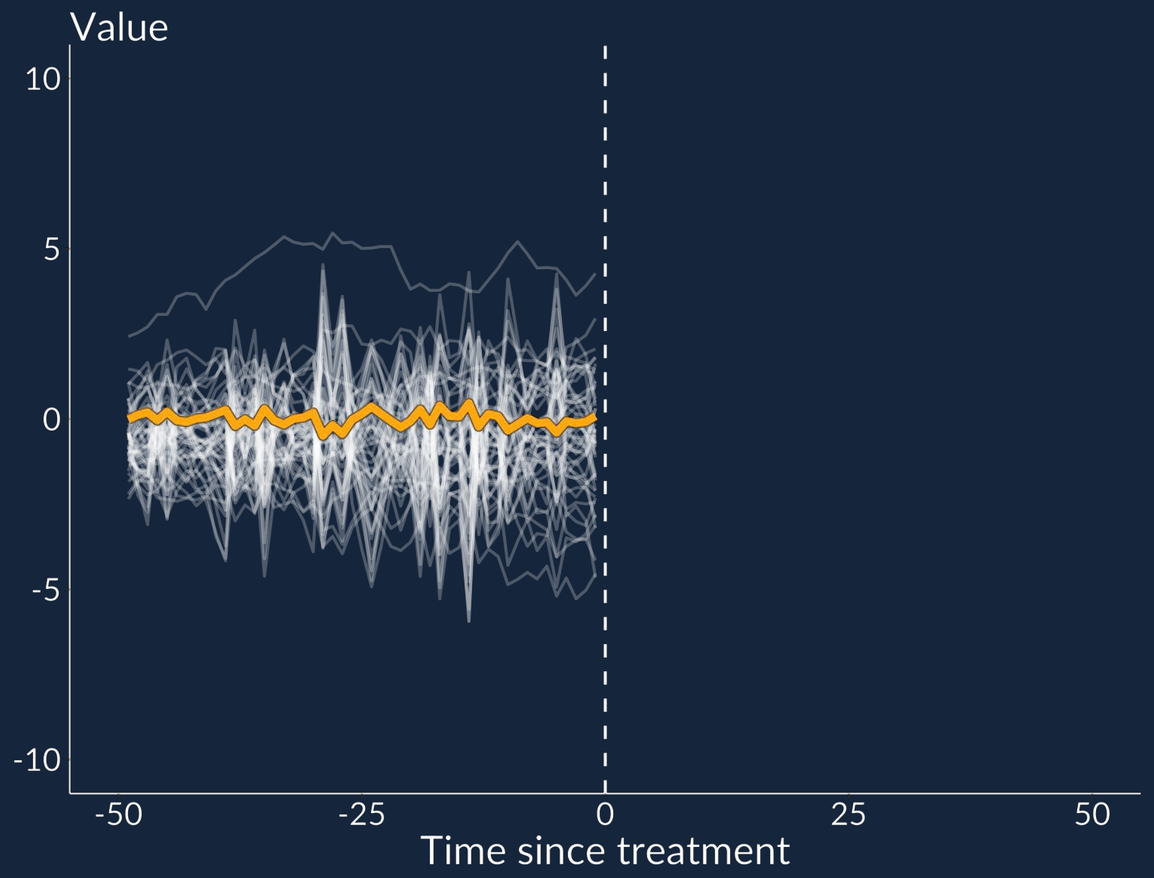
The “match” is a weighted combination of the donor pool

$$y_t^* = \sum_{i=1}^{N} y_{it}\,\pi_i$$

How these weights are determined differs by the synthetic control method.

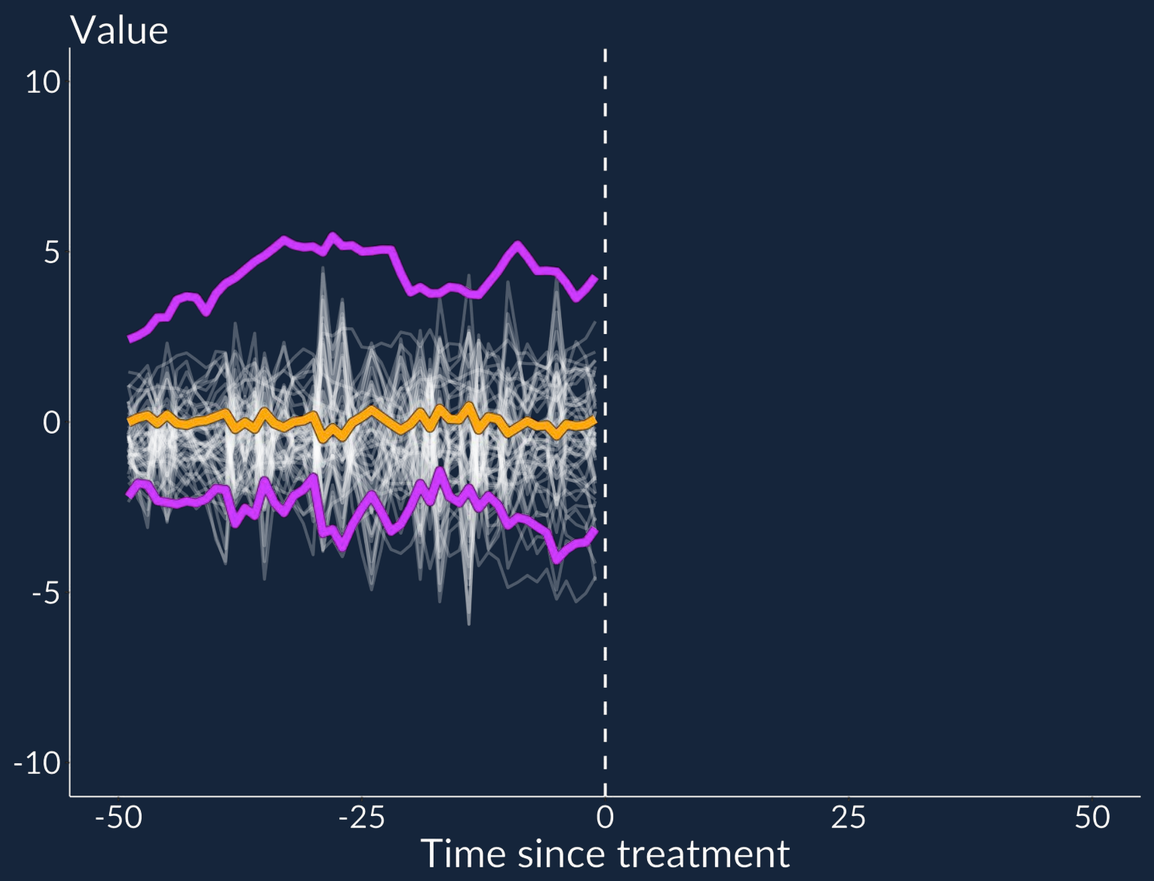
Can be zero.
Can be negative.
Need not sum to one.

Is there a combination of donor units that is a good match for the treated unit during the pre-treatment period?



Is there a combination of] donor units that is a good match for the treated unit during the pre-treatment period?

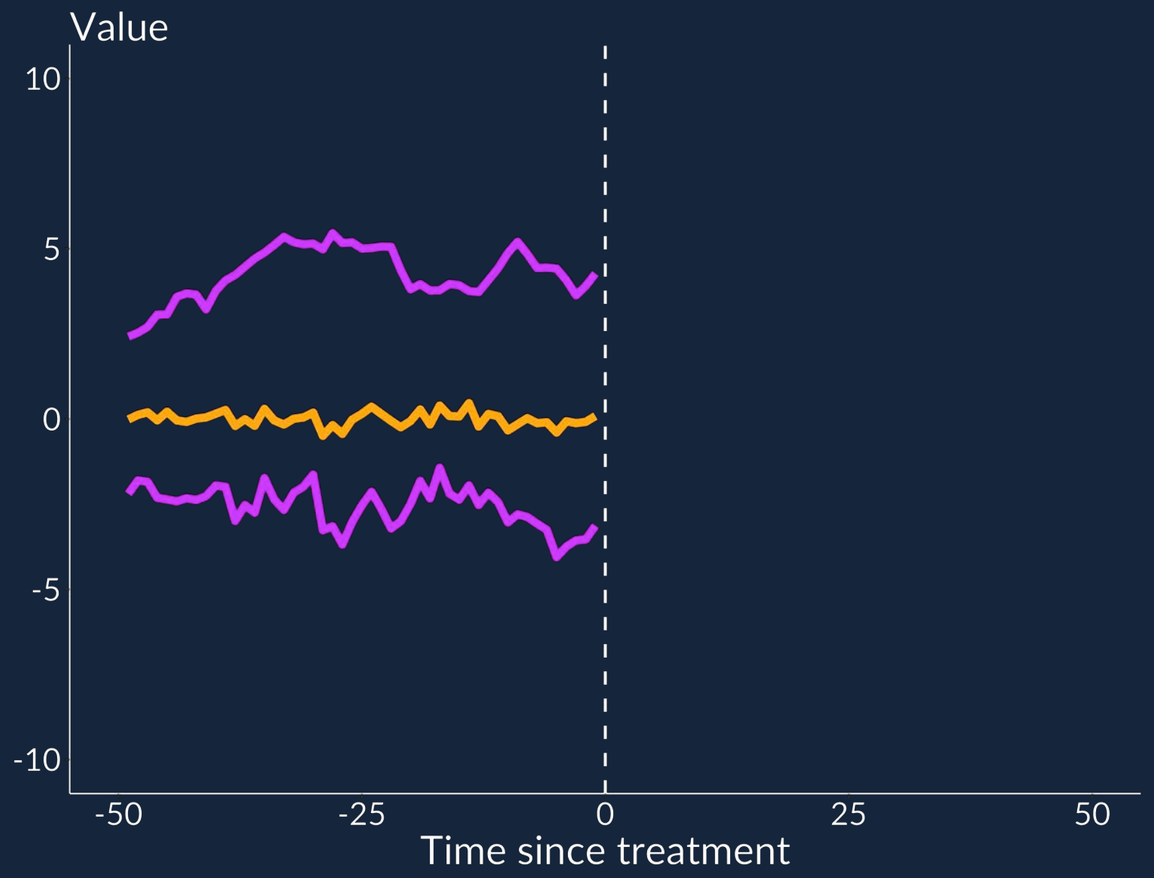
Yes!



The “match” is a weighted combination of the donor pool

- Each of the N donors receives weight π_i
- The synthetic control for the target series in time period t is

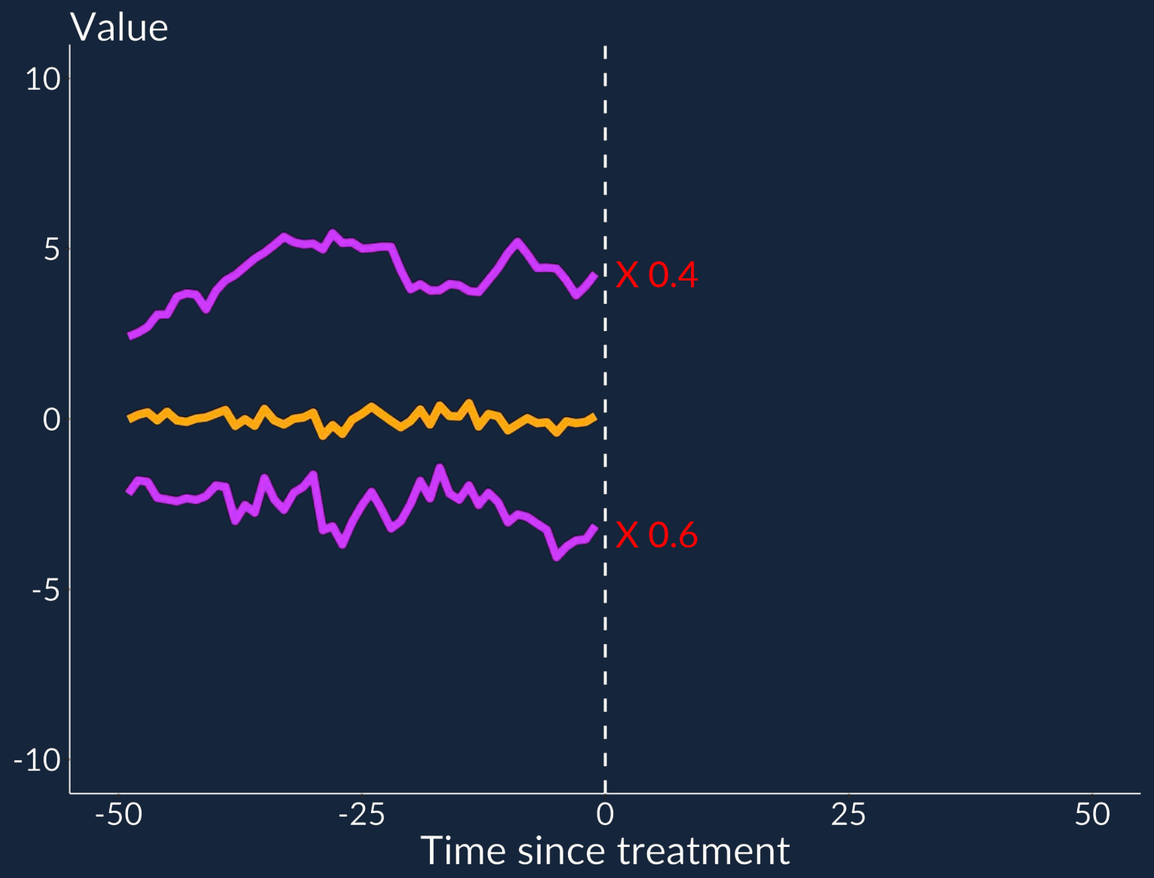
$$y_t^* = \sum_{i=1}^{N} y_{it}\,\pi_i$$



The “match” is a weighted combination of the donor pool

- Each of the N donors receives weight π_i
- The synthetic control for the target series in time period t is

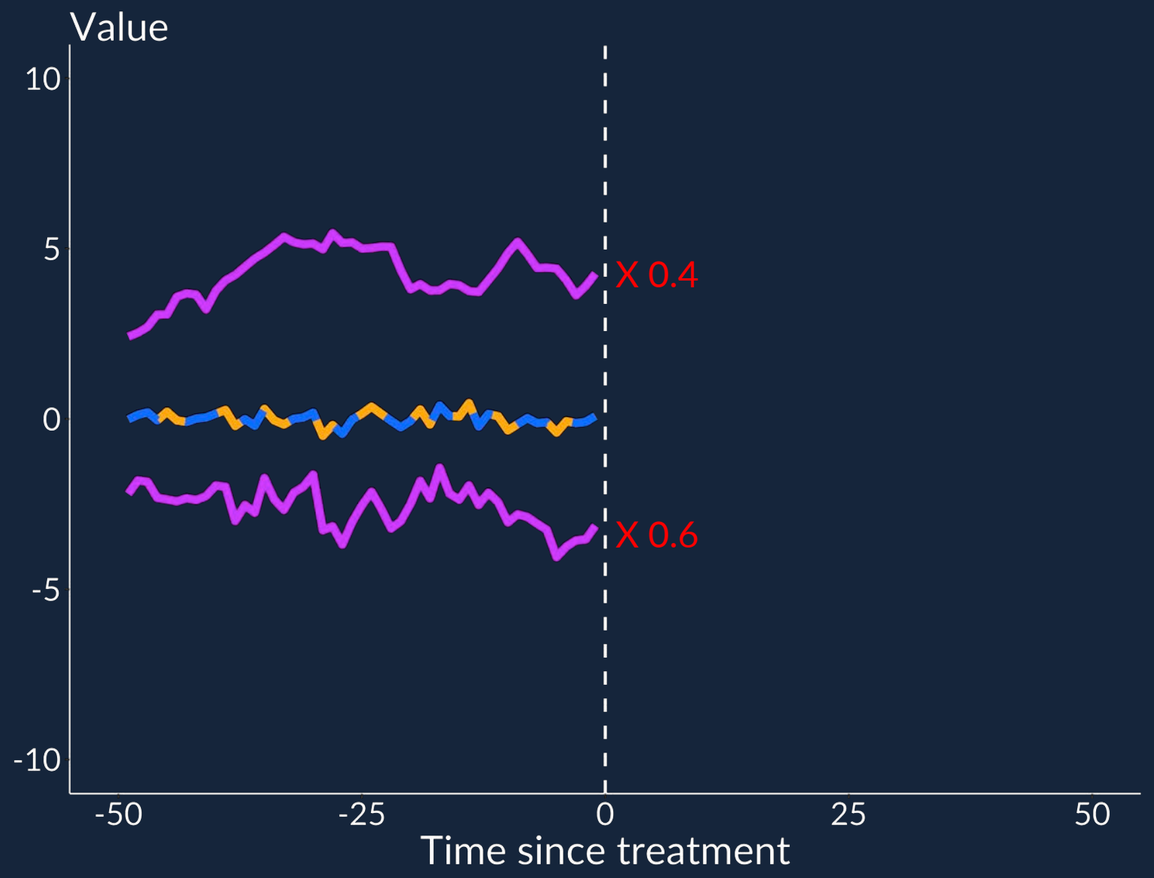
$$y_t^* = \sum_{i=1}^{N} y_{it}\,\pi_i$$



Now use these donors to form a prediction of the treated unit

$$y_t^* = \sum_{i=1}^{N} y_{it}\,\pi_i$$

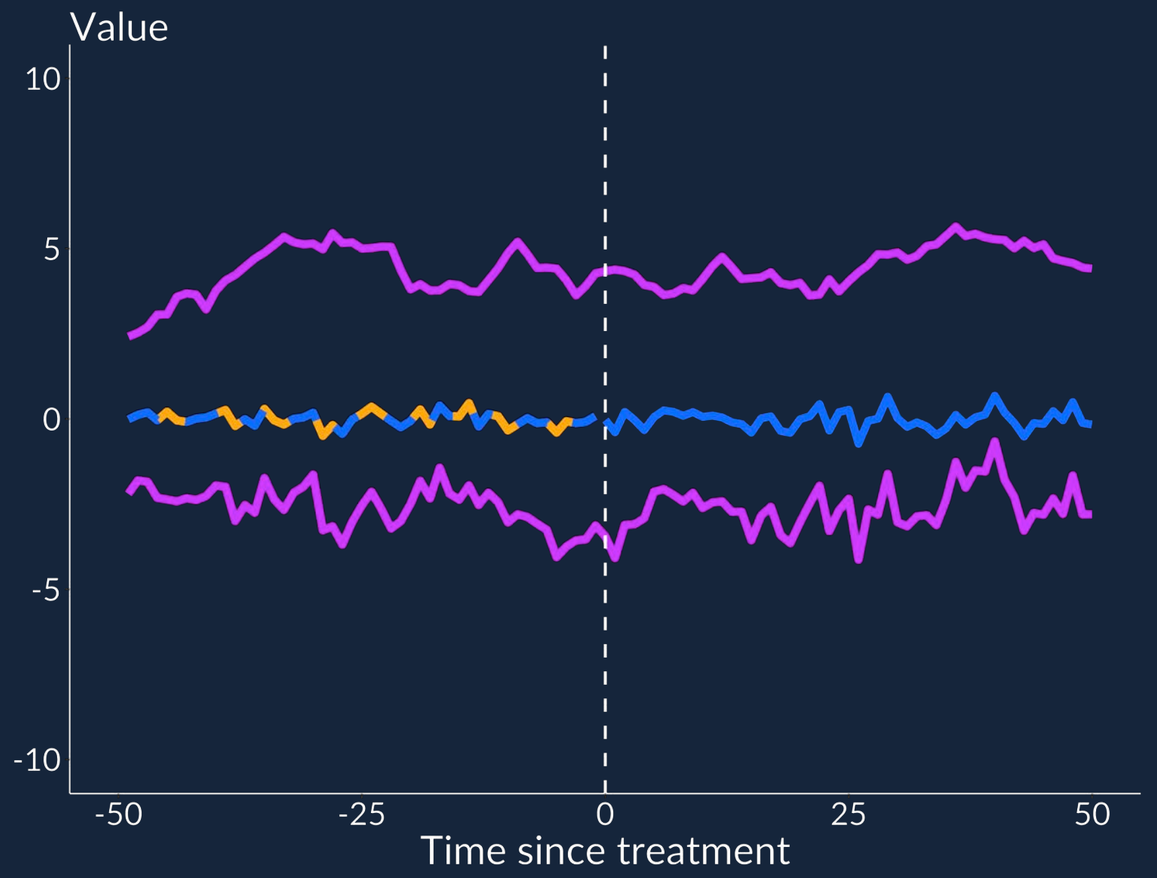
$$y_t^* = \widehat{y}_t$$



And extend the prediction into the post-treatment period

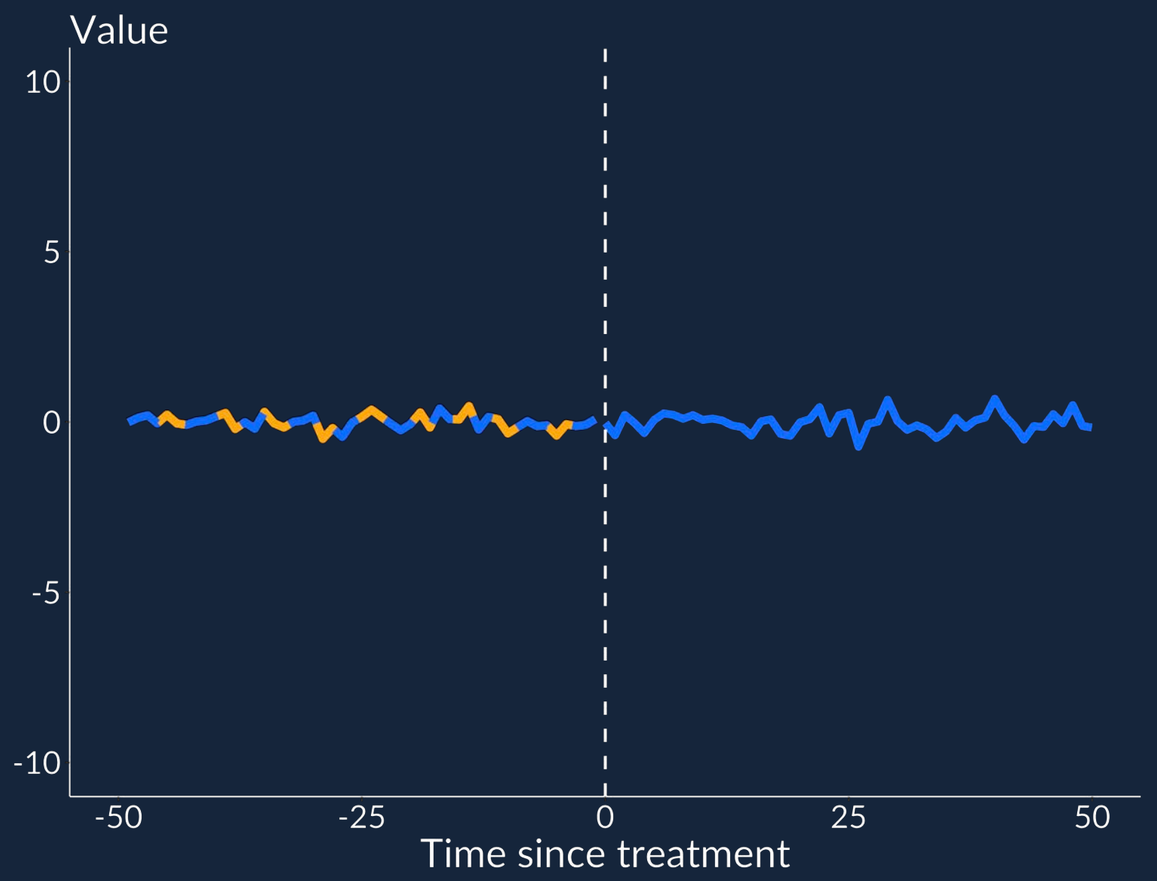
$$y_t^* = \sum_{i=1}^{N} y_{it}\,\pi_i$$

$$y_t^* = \hat{y}_t$$

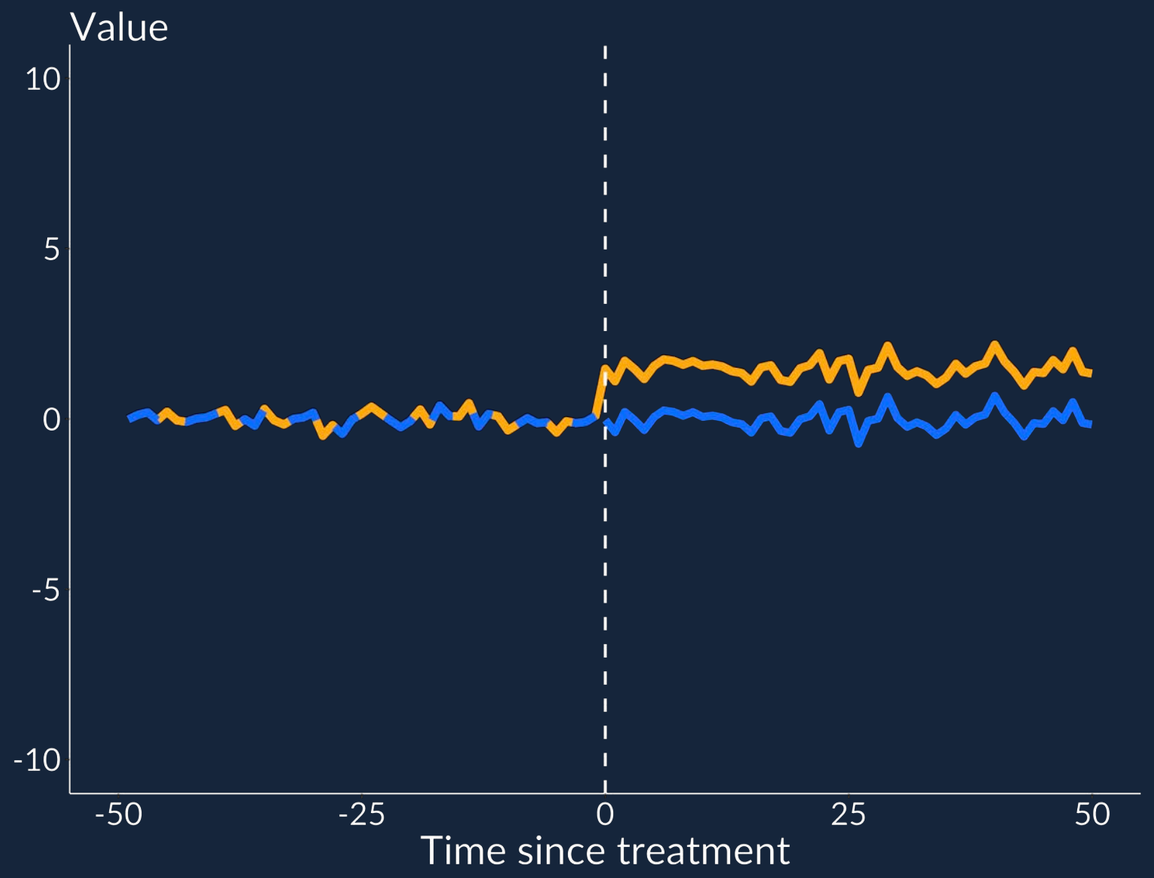


And extend the prediction into the post-treatment period

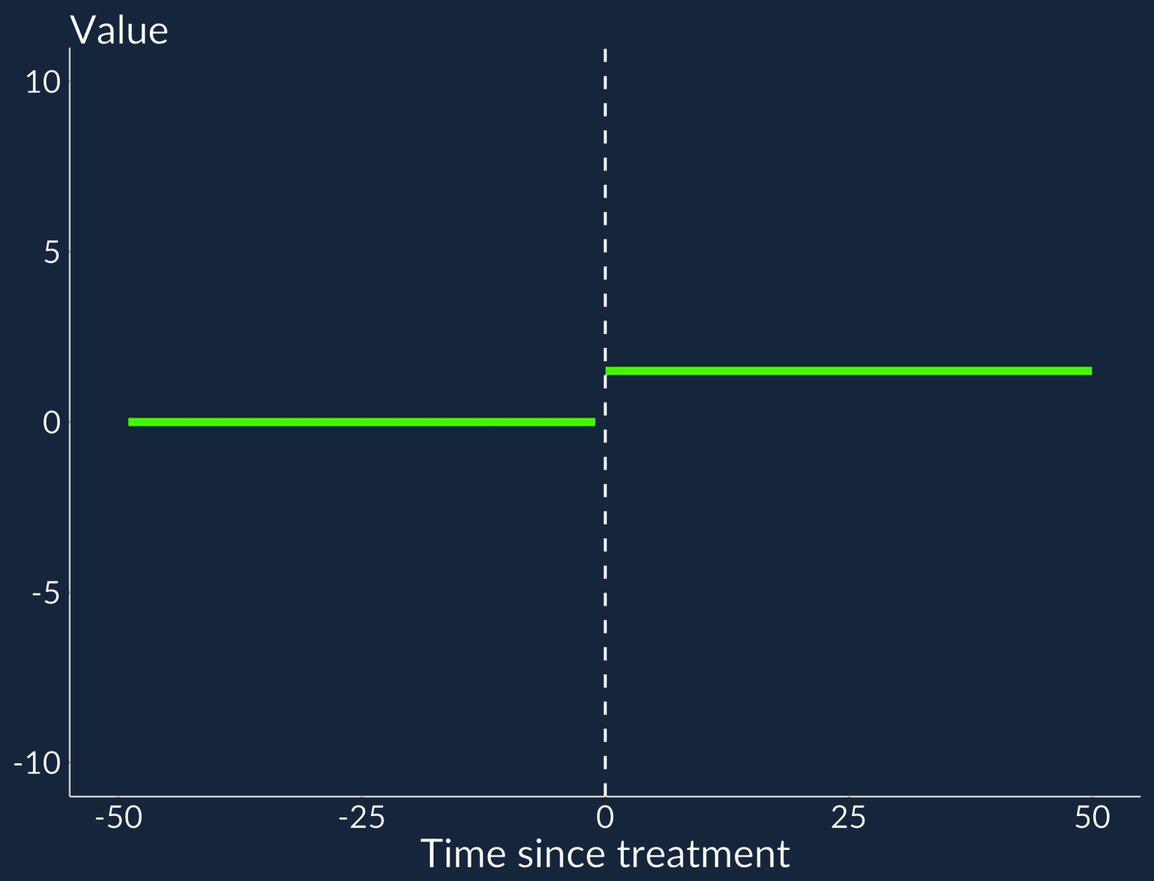
- A counterfactual estimate as if treatment had not occurred.



Now compare the counterfactual estimate to the treated series.



The difference between the counterfactual estimate and the treated series is our treatment effect estimate



Identification Assumptions

1. Conditional Independence

Once you conditional on α_i, treatment is as good as randomly assigned.

Same motivating assumption underlying propensity score matching and regression

1. Conditional Independence

Omitted variable bias is a violation of this assumption

Occurs when treated series is generated from different time-varying factors than the donor-series

$$y_t^* = \delta_t \alpha_i + \theta_t \sigma_i + \epsilon_{it}$$
$$y_{it} = \delta_t \alpha_i + \epsilon_{it}$$

1. Conditional Independence

Most plausible in settings where the donor pool consists of outcomes that likely respond to a similar collection of time-varying common factors

Raises concerns of overfitting

- Do not want to match on idiosyncratic error
- More of an issue in short-panels

2. Structural Stability

Assumption that the data generating process for the untreated outcomes is the same in the pre−period and the post−period

2. Structural Stability

You assume that $y_{it} = \delta_t \alpha_i + \epsilon_{it}$ for all t

Your match and therefore post-period prediction will no longer be valid if

$$y_{it} = \delta_t \alpha_i \times 1(t \leq T_0) + \delta_t \mathrm{A}_s \times 1(t > T_0) + \epsilon_{it}$$

3. No Dormant Factors

δ must independently vary during pre-period

If some elements of δ are dormant (i.e., perfectly collinear/no variance) during the pre-period, then match on outcomes does not imply perfect match on α_i

3. No Dormant factors

Low frequency events

- Presidential election in the pre-period
- Seasonality

The pre-treatment period may not be long enough to capture the α for a low-frequency event

Especially problematic if the post-treatment period includes such events

You do not want factors that "wake up"

- One concern is that some units may adopt new policies or experience novel/unique economic or social events.
  - Either through a new δ
  - or a change in α

Practical considerations

1. Use only donor and placebo units that seem to plausibly depend on the same collection of common factors

    This need not include variables of the same type

    Lots of different variables may be informative about the underlying data generating process of the treated unit.

2. Use cross-validation to determine synthetic control groups

    Reduce likelihood of fitting on error (i.e., overfitting)

Practical considerations

3. Use only donors and placebos where no known violations of the dormant factors have occurred

    No policy or other changes to the data generating process

    Be on the lookout for events that occur in the post-treatment period that do not have much precedent in the pre-treatment data.

Practical considerations

4. Use a longer pre-treatment time period when possible

   Trade-off between dormant factor and structural stability

   Both length and frequency of dataset which help mitigate dormant factors

   But going back too far may result in an entirely different data generating process.

Practical considerations

5. Show that the pre-treatment difference between the synthetic control estimator and the target variable is small and centered around zero

   Use a unit-free measure of fit to determine “what is small”

Applied Example

How does marijuana legalization affect the sale of alcohol and over-the-counter pain medication?

Recreational Marijuana in Colorado

- Possession Legal
  - One month following the vote (December 2012)
  - Grow marijuana for personal use.
  - Decriminalized for possession of Marijuana from a homegrown source.
- Transactions Legal
  - January 2014
  - Licensed stores can legally sell Marijuana for personal recreational use.

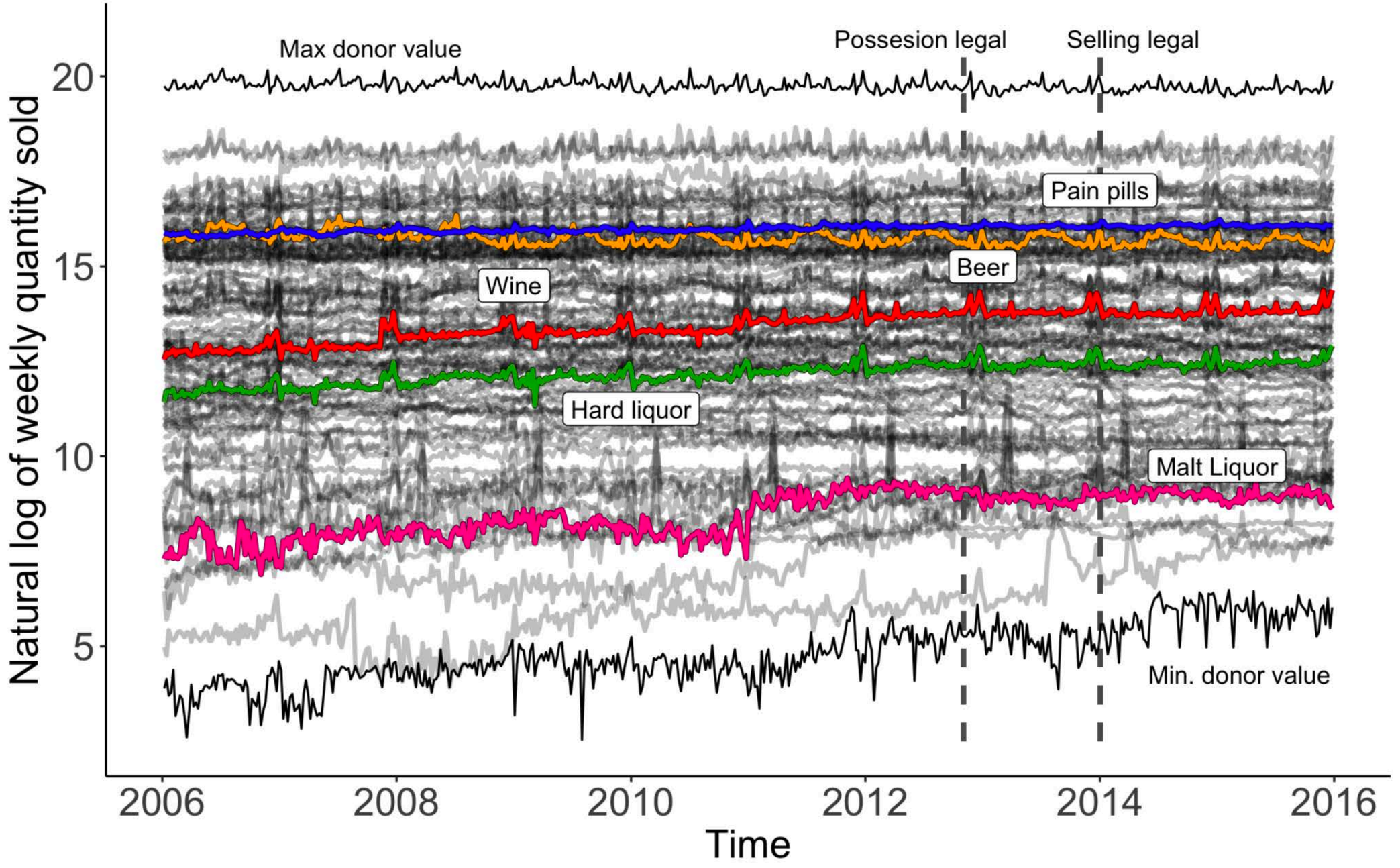
Data

Retail scanner database that provides store by week level information on the sales of a large set of products

Build comparison groups using a synthetic control estimator

- combines information from weekly data on the sale of a basket of goods from states where marijuana is not legal

Traditional SCM will not work.


Max donor value
Possesion legal
Selling legal
Pain pills
Beer
Wine
Hard liquor
Malt Liquor
Min. donor value
Natural log of weekly quantity sold
20
15
10
5
2006
2008
2010
2012
2014
2016
Time

Synthetic Control Using Lasso

SCUL

What is SCUL?

Using only pre-period data.

Choose weights to satisfy:

$$argmin_{\beta}\left\{\frac{1}{2N}\sum_{t=1}^{T_0}\left(Y_{1t}-\sum_{s=2}^{S}\beta_s Y_{st}\right)^2+\lambda\left(\sum_{s=2}^{S}|\beta_s|\right)\right\}$$

The first term is just regular OLS.

What is SCUL?

$$argmin_{\beta}\left\{\frac{1}{2N}\sum_{t=1}^{T_0}\left(Y_{1t}-\sum_{s=2}^{S}\beta_s Y_{st}\right)^2+\lambda\left(\sum_{s=2}^{S}|\beta_s|\right)\right\}$$

The second term is the Lasso penalty function.

λ is a parameter that controls the penalty.

When $\lambda = 0$ you have OLS.

When $\lambda > 0$ you shrink the coefficients towards zero and sometimes you set some coefficients to zero. (Sparsity)

Why penalized OLS?

- Sparsity
  - OLS may overfit data → poor out-of-sample forecasts
  - Fewer coefficients → Interpretable
- Allows for more donors than observations
- Allows for the same model selection procedure and thought to be put into placebo analyses as was done in target analyses (removes researcher degrees of freedom)
- Allows for negative and non-convex weights

Choose weights using cross-validation

λ is a parameter that controls the penalty → controls weights

λ can be so large that no donors survive.

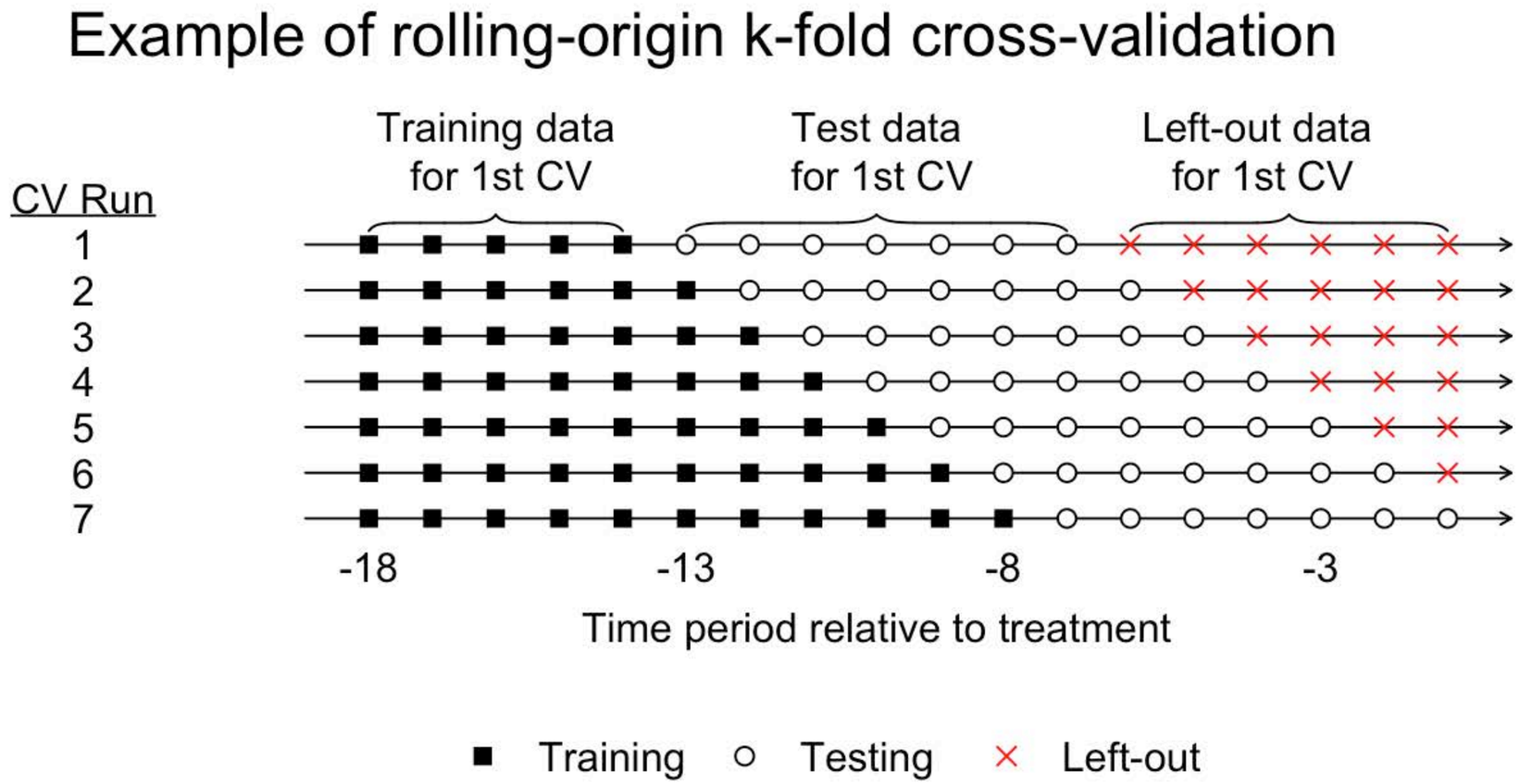
λ can be so small that the model is the same as OLS

For each unique λ weights are different

Choose λ using cross-validation to avoid over-fitting. Use rolling-origin cross-validation to avoid autocorrelation from creeping in.

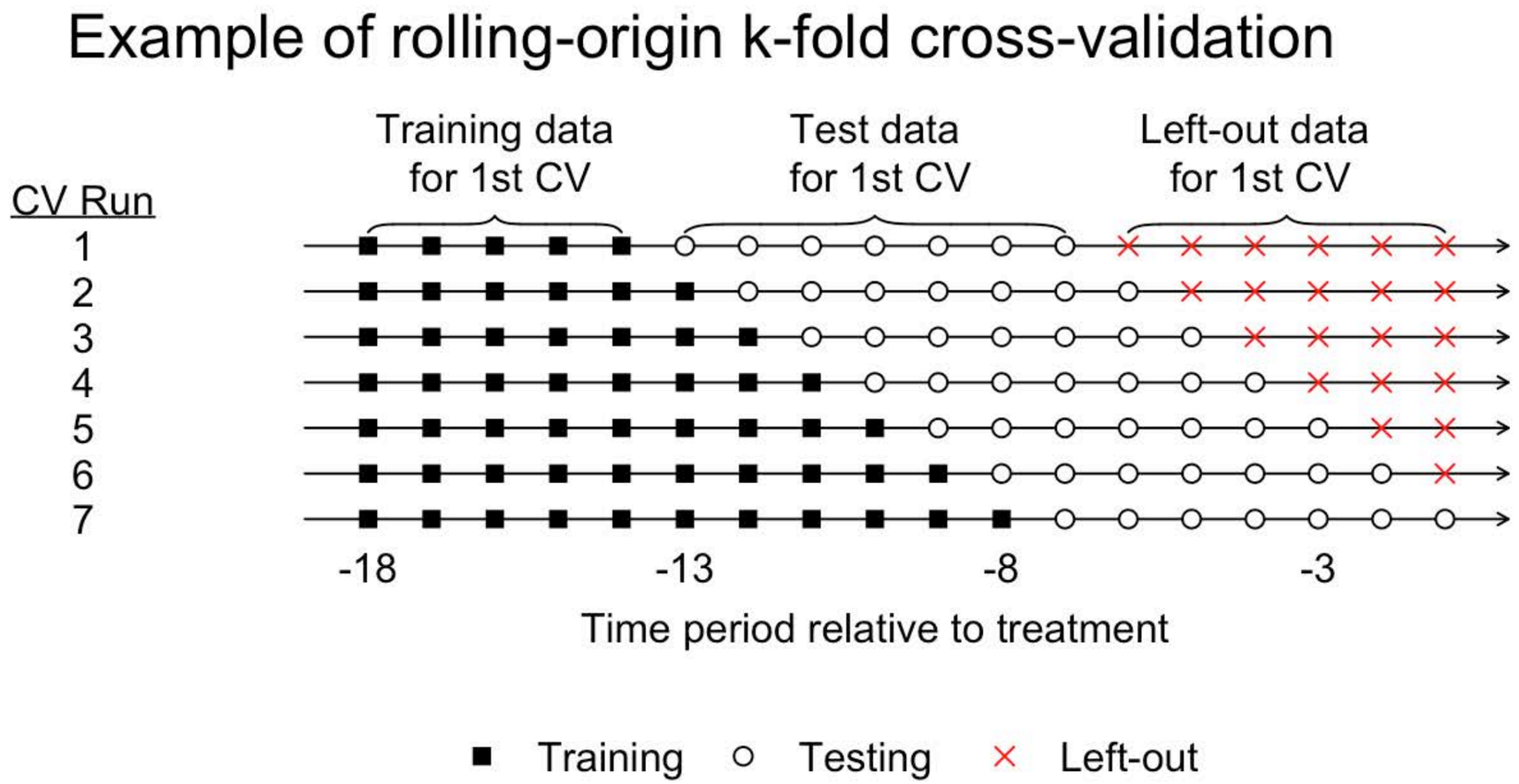
SCUL chooses λ using rolling-origin cross-validation

Avoids over-fitting and autocorrelation issues

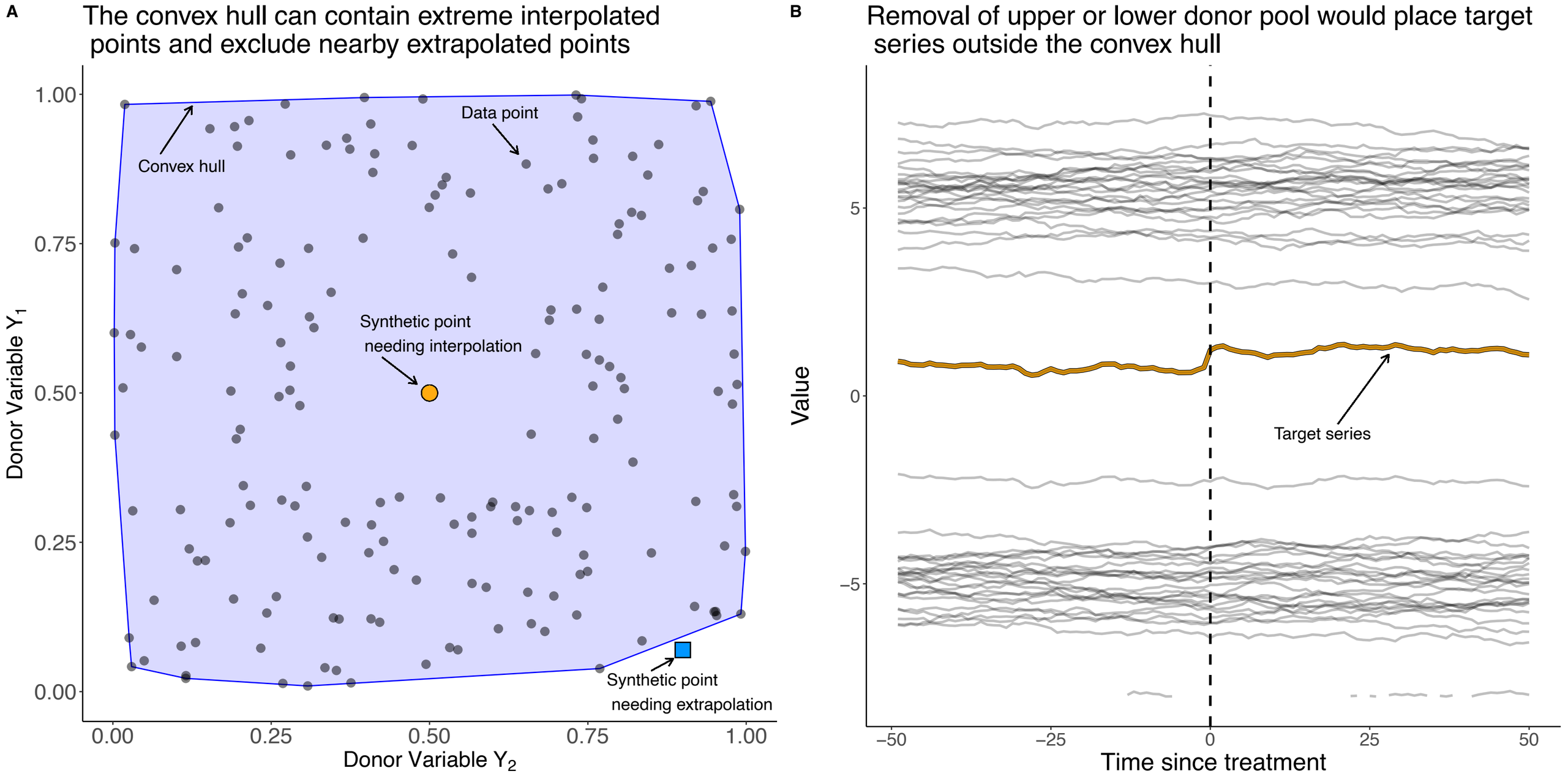


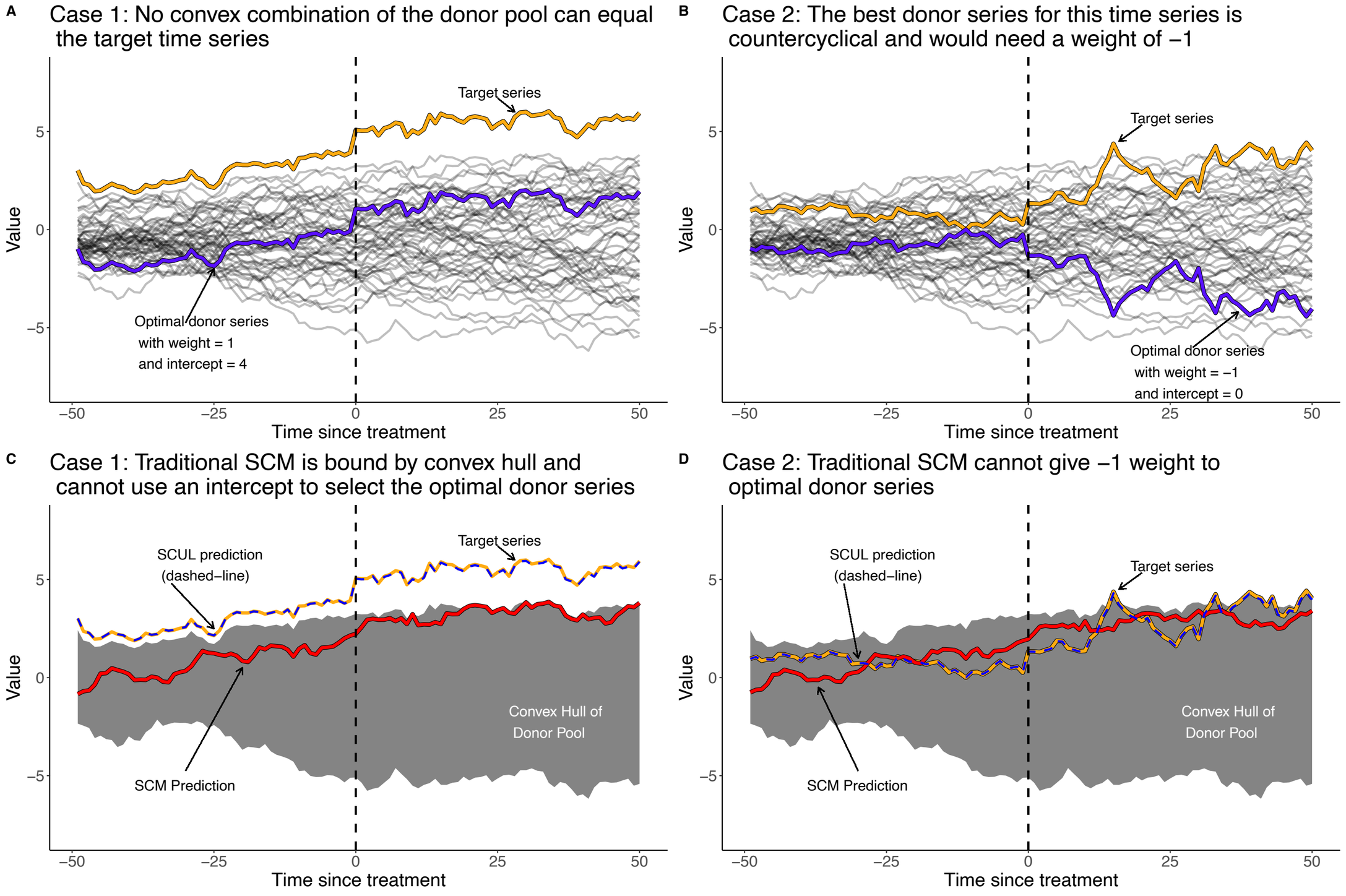
SCUL chooses λ using rolling-origin cross-validation

We choose the median λ across all C.V.



What is the convex hull?


A
The convex hull can contain extreme interpolated points and exclude nearby extrapolated points
Convex hull
Data point
Synthetic point needing interpolation
Synthetic point needing extrapolation
Donor Variable Y_1
Donor Variable Y_2
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
B
Removal of upper or lower donor pool would place target series outside the convex hull
Target series
Value
Time since treatment
5
0
−5
−50
−25
0
25
50


A
Case 1: No convex combination of the donor pool can equal the target time series
Target series
Optimal donor series with weight = 1 and intercept = 4
Value
Time since treatment
B
Case 2: The best donor series for this time series is countercyclical and would need a weight of −1
Target series
Optimal donor series with weight = −1 and intercept = 0
Value
Time since treatment
C
Case 1: Traditional SCM is bound by convex hull and cannot use an intercept to select the optimal donor series
SCUL prediction (dashed−line)
Target series
Convex Hull of Donor Pool
SCM Prediction
Value
Time since treatment
D
Case 2: Traditional SCM cannot give −1 weight to optimal donor series
SCUL prediction (dashed−line)
Target series
Convex Hull of Donor Pool
SCM Prediction
Value
Time since treatment

What do synthetic control weights mean?

Interpreting weights

- Typical synthetic control weights only report the fraction of the total weight that is given to a particular donor series;
  - they do not reflect the size and variability of the outcome for each unit across time periods.
  - $y_t^* = \sum_{i=1}^{N} y_{it}\, \pi_i$

Interpreting weights

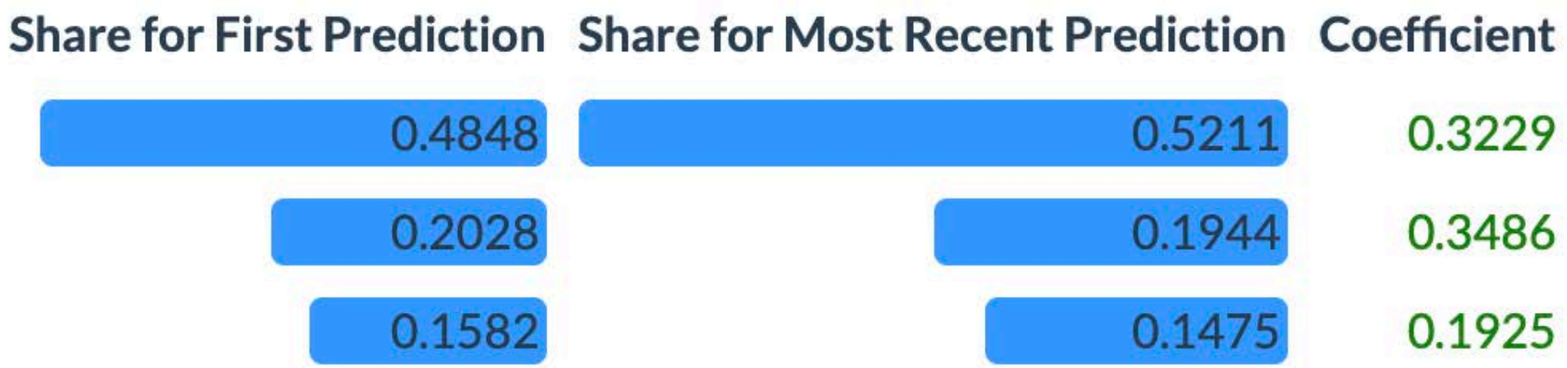
- Suppose, for example that
  - there are two donor series, A and B,
    - A = 10 and B = 1
  - each unit receives a weight equal to 0.5

- The synthetic prediction is 5.5 it is
  - Y* = 0.5*A+ 0.5*B
  - Y* = 5.5

- Despite being the same weight, 91% of the prediction came from A because of the large nominal value

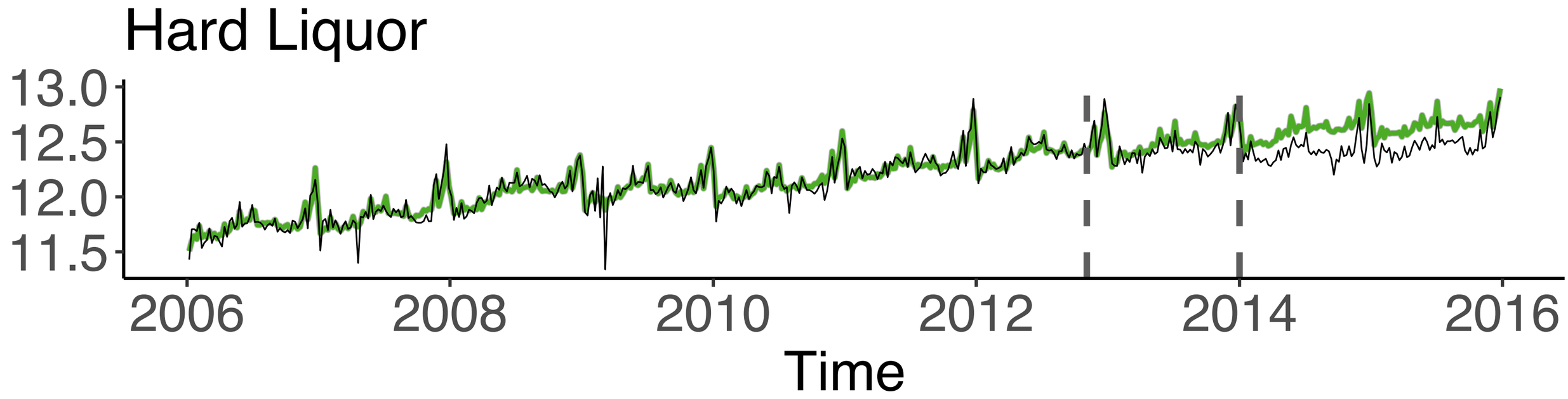
Practical advice

- Report the weight from the model AND the share of the contribution to the prediction

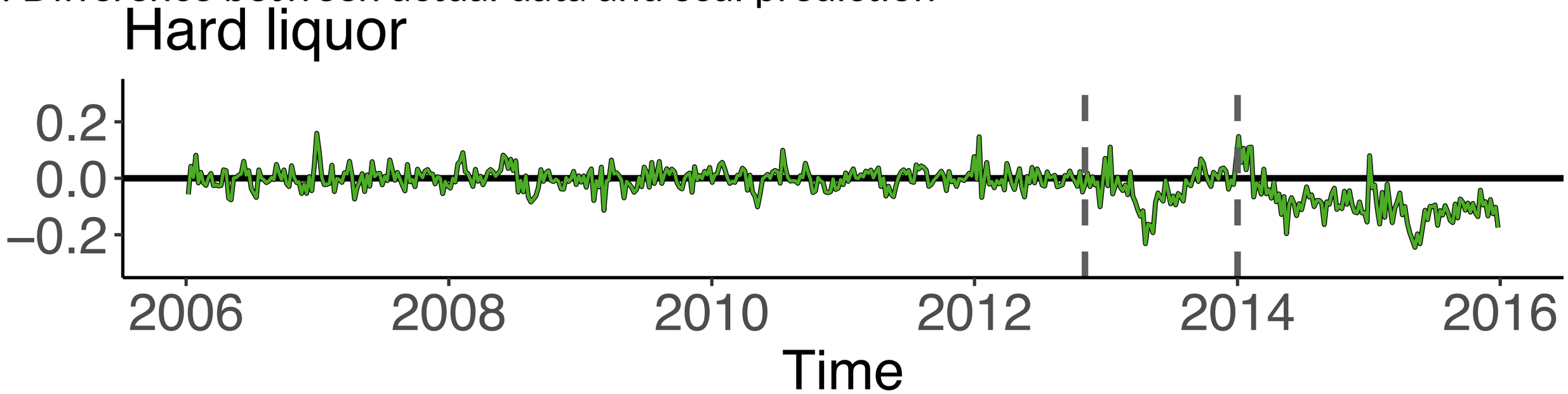


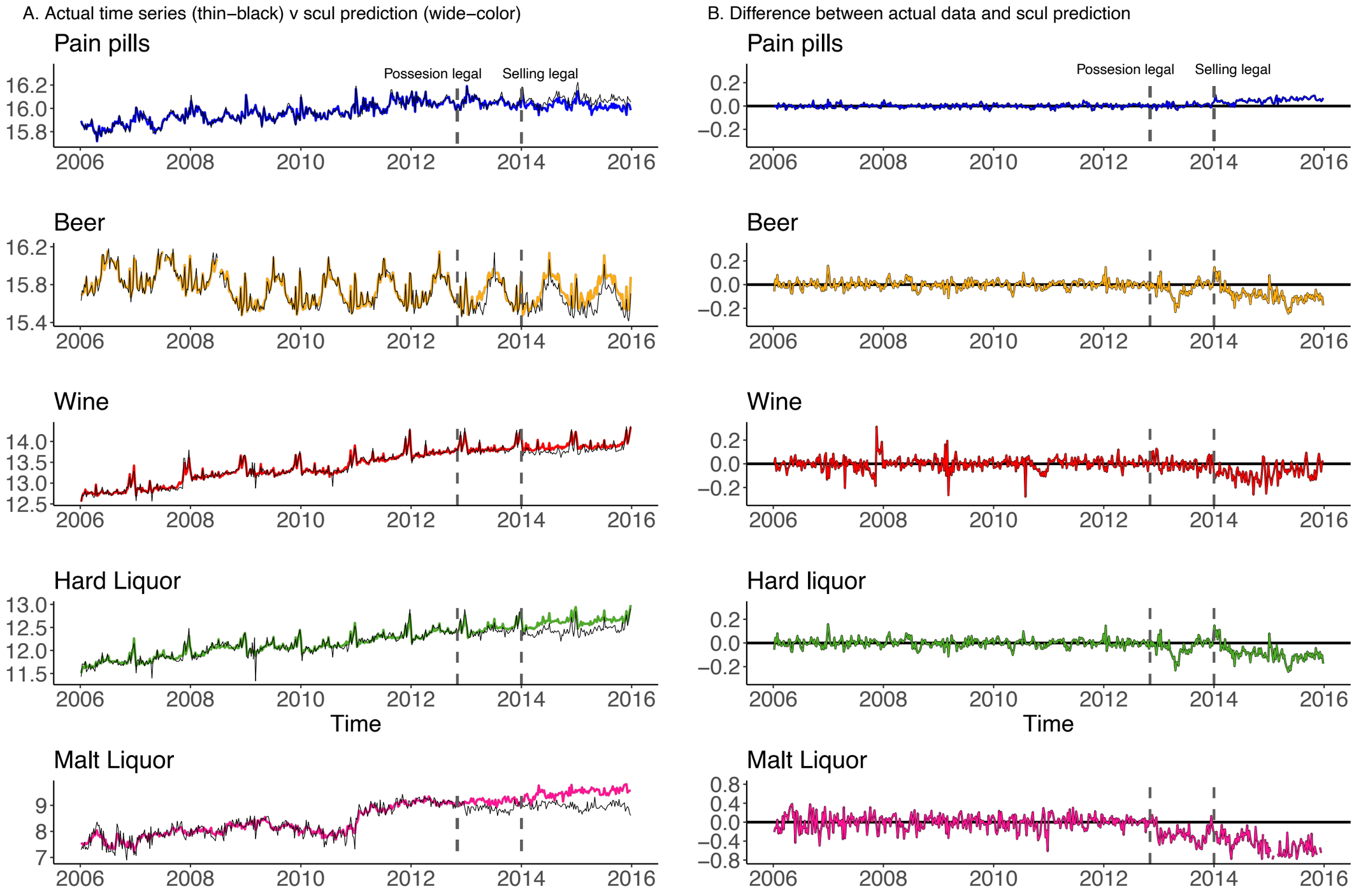
How do I know if I have a good synthetic control?

A. Actual time series (thin-black) v scul prediction (wide-color)



B. Difference between actual data and scul prediction




A. Actual time series (thin−black) v scul prediction (wide−color)
B. Difference between actual data and scul prediction
Pain pills
Possesion legal
Selling legal
16.2
16.0
15.8
0.2
0.0
−0.2
Beer
16.2
15.8
15.4
Wine
14.0
13.5
13.0
12.5
Hard Liquor
Hard liquor
13.0
12.5
12.0
11.5
Time
Malt Liquor
9
8
7
0.8
0.4
0.0
−0.4
−0.8
2006
2008
2010
2012
2014
2016

Cohen's D

Average pre-treatment fit in standard deviation units

$$\frac{1}{T_0}\sum_{t=1}^{T_0}\left|\frac{y_{st}-y_{st}^{*}}{\sigma_s}\right|$$

where

$$\sigma_s = \sqrt{\frac{1}{T_0}\sum_{t=1}^{T_0}(y_{st}-\overline{y_s})^2}$$

Adequate fit is < 0.25

	Pre-treatment fit 2006-2012
Pain pills	0.15
Beer	0.15
Wine	0.12
Hard liquor	0.18
Malt liquor, 0-40oz.	0.22

Practical advice

Eliminate any donor or target series that has poor fit based upon pre-determined, unit-free threshold

- Using fit for the target series as the ”maximum threshold” biases the target series to be an outlier
  - When rank based p-values are used this attenuates p-values
- Using RMSE penalizes donors with large nominal variance

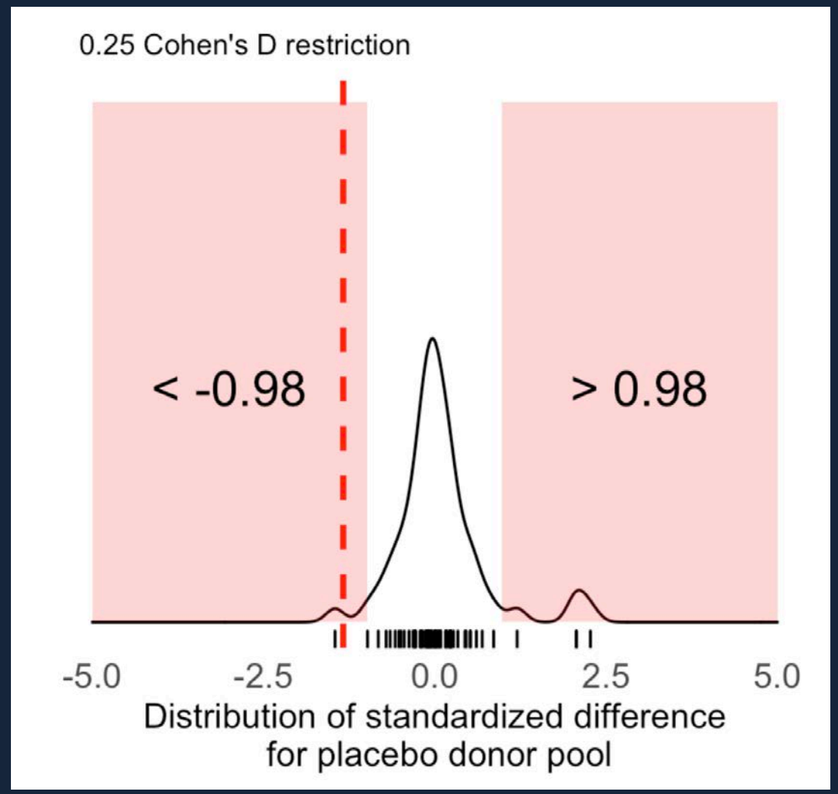
How should I think about statistical inference and power in synthetic controls?

Placebo-analyses

- Make a rank-based, two-sided p-value using randomization inference
- Compare the absolute value of the standardized treatment estimate to the absolute value of the standardized estimate from a number of placebo series

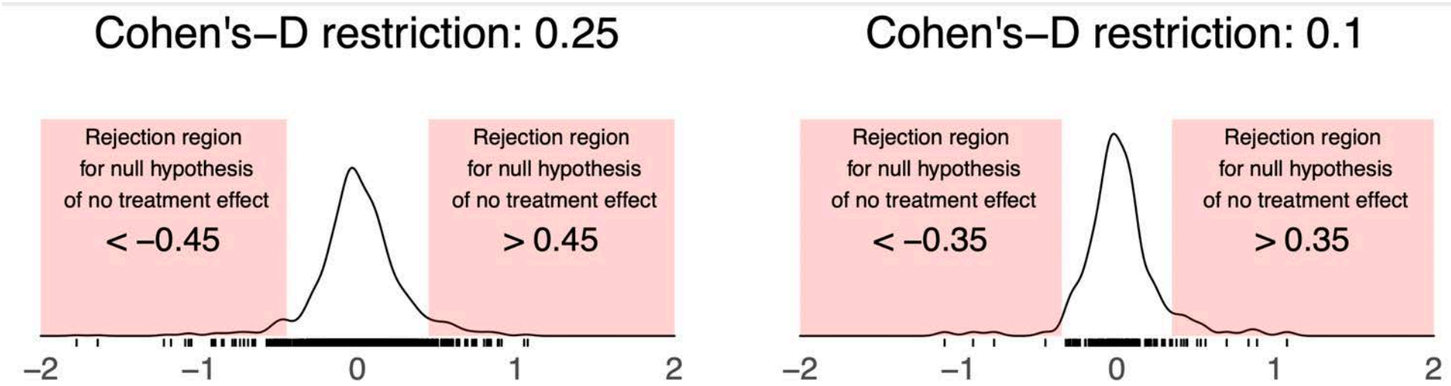
Placebo-analyses

- The estimates from the placebo distribution serve as the null distribution that assumes no treatment effect.

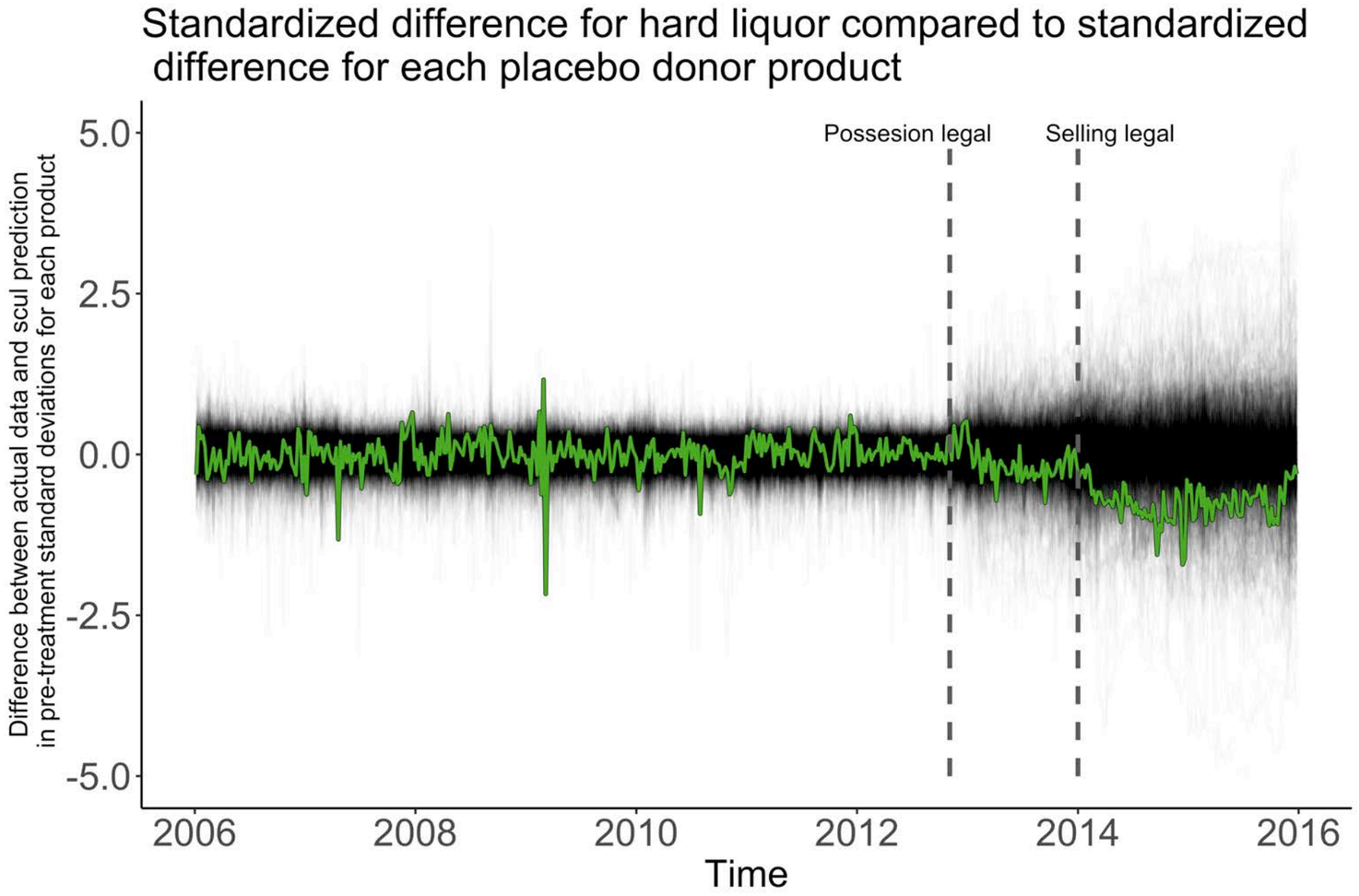


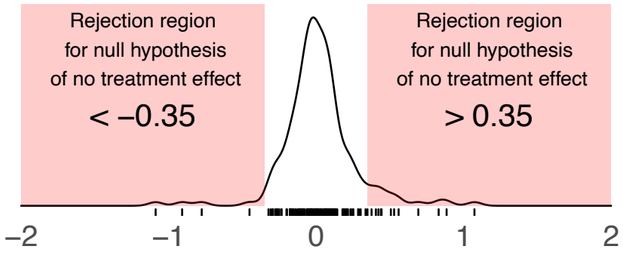
Practical Advice

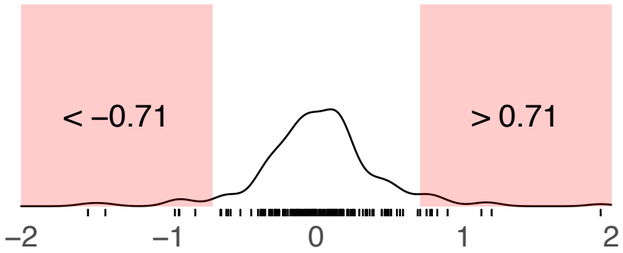
- Be sure to apply the same selection rules to your placebo pool as you did the treatment series
- Compare based on unit-free measure of fit.
  - We use post-treatment Cohens' D

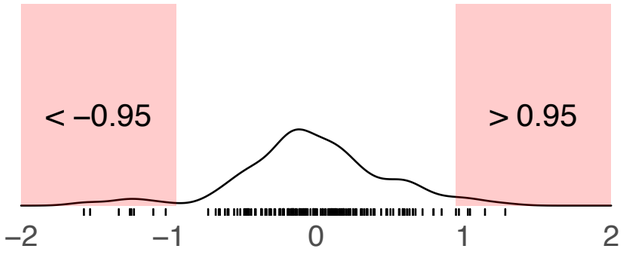


Placebo-analysis inform you of deteriorating model fit





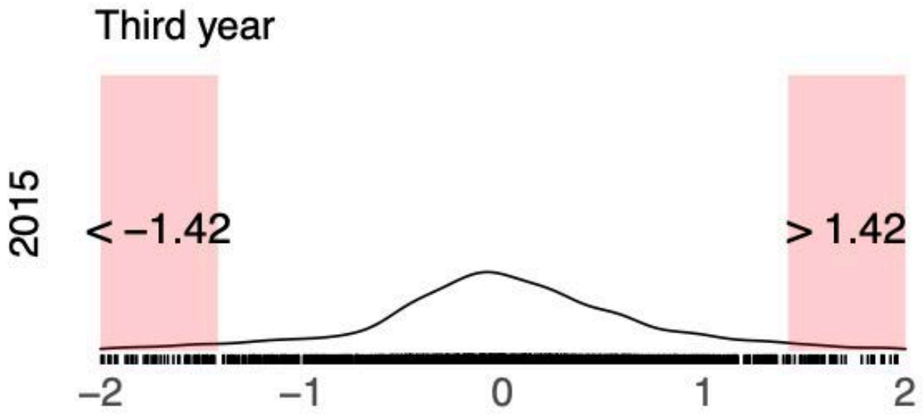




Practical advice

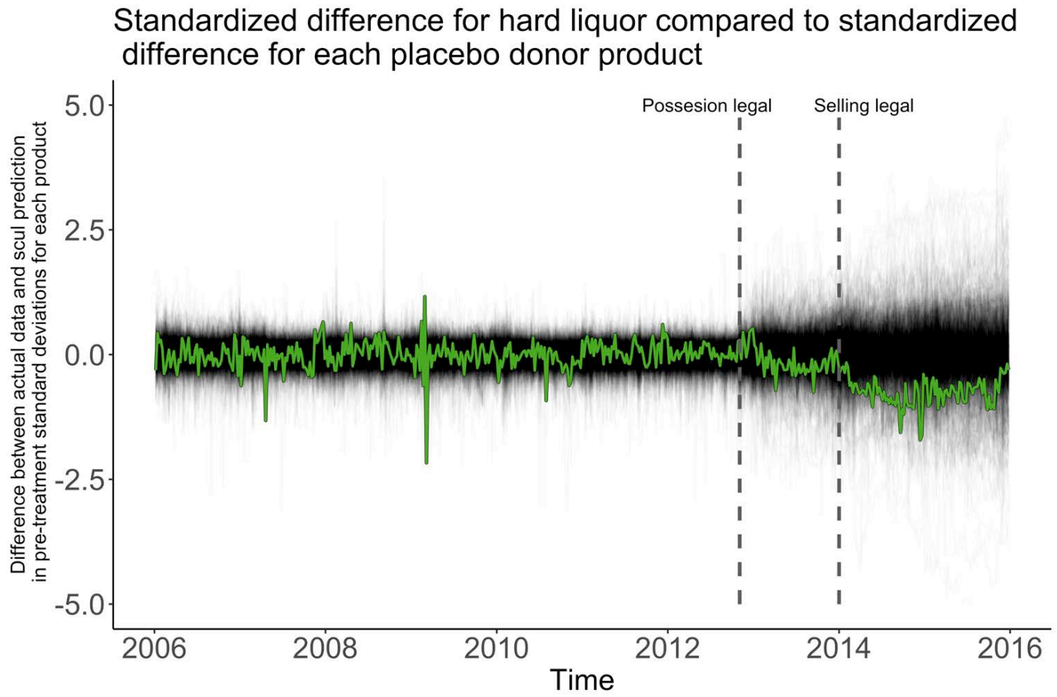
Report the minimum rejection value for your desired significance level.

Perhaps your placebo distribution is so wide, you can only reject outrageous values.



Results

Hard liquor



	Share for First Prediction	Share for Most Recent Prediction	Coefficient
TN_oth_beer_oz_0_40_oz	0.12	0.12	0.28
Intercept	0.12	0.12	3.70
MA_cigarettes_total_cnt	0.08	0.08	-0.15
MI_liquor_oz_handle_oz	0.06	0.06	0.13
IN_liquor_oz_fifth_liter_oz	0.05	0.05	0.12
NH_cigarettes_total_cnt	0.05	0.05	-0.09
AZ_bread_total_oz	0.04	0.05	-0.08
KY_liquor_oz_handle_oz	0.04	0.04	0.10
VA_bread_total_oz	0.04	0.04	-0.08
MS_oth_beer_oz_0_40_oz	0.04	0.04	0.12
NY_wine_total_oz	0.04	0.04	0.08
LA_liquor_oz_handle_oz	0.03	0.03	0.06
NY_liquor_oz_handle_oz	0.03	0.03	0.07

Treatment: Post-2012

Panel A: Treatment begins in 2013 following passage of the recreational marijuana law.

	Pre-treatment fit 2006-2012	First Year 2013	Second Year 2014	Third Year 2015	All Post Treatment 2013-2015
Pain pills	0.15	0.47	3.85	5.96	3.27
		(0.77)	(0.26)	(0.20)	(0.28)
Beer	0.15	–3.44	–6.12	–11.49	–6.82
		(0.33)	(0.35)	(0.21)	(0.28)
Wine	0.12	–0.28	–9.72	–5.48	–4.89
		(0.97)	(0.42)	(0.73)	(0.63)
Hard liquor	0.18	–3.29	–19.44	–17.38	–12.82
		(0.49)	(0.10)	(0.20)	(0.18)
Malt liquor, 0-40oz.	0.22	–24.76	–38.58	–62.75	–41.09
		(0.10)	(0.14)	(0.09)	(0.10)
p-value from joint test of any effect		0.57	0.16	0.19	0.21
p-value from joint test of any alcohol effect		0.15	0.05	0.08	0.06

Treatment: Post-2014

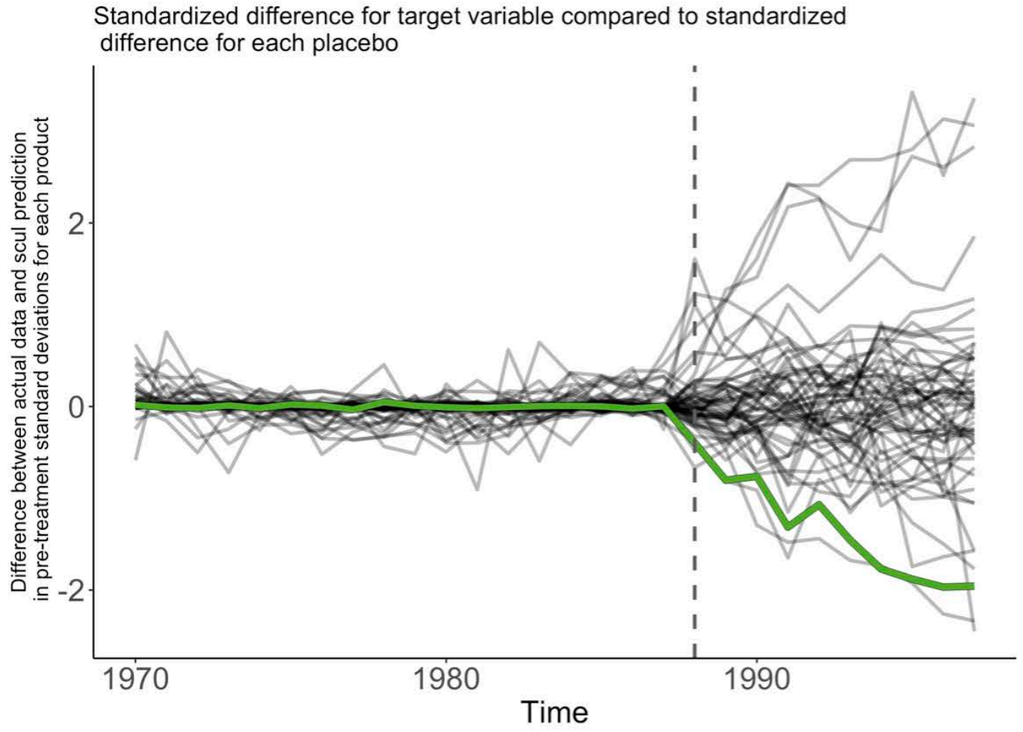
Panel B: Treatment begins in 2014 following opening of recreational marijuana dispensaries.

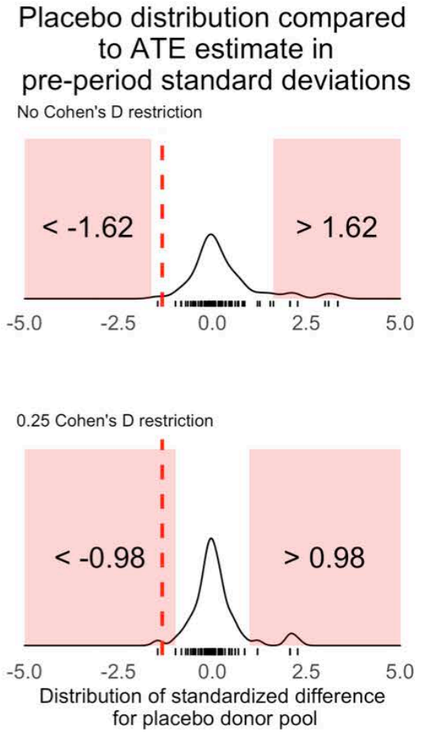
	Pre-treatment fit 2006-2013	First Year 2014	Second Year 2015	All Post Treatment 2014-2015
Pain pills, OTC	0.14	2.32	3.44	2.87
		(0.18)	(0.24)	(0.19)
Beer	0.21	–4.31	–7.93	–6.10
		(0.21)	(0.19)	(0.17)
Wine	0.1	–10.90	–10.19	–10.55
		(0.16)	(0.39)	(0.25)
Hard liquor	0.14	–17.50	–16.75	–17.13
		(0.03)	(0.12)	(0.06)
Malt liquor, 0-40oz.	0.32			
p-value from joint test of any effect		0.04	0.17	0.08
p-value from joint test of any alcohol effect		0.03	0.08	0.06

How do I do this?

Synthetic Control Using Lasso (scul)

This repository contains the R package `scul` that is used in Hollingsworth and Wing (2020) *"Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data."*
https://doi.org/10.31235/osf.io/fc9xt





Links

Browse source code at
https://github.com/hollina/scul/

Report a bug at
https://github.com/hollina/scul/issues/

License

Full license

MIT + file LICENSE

Developers

Alex Hollingsworth
Author, maintainer

Dev status

build passing

lifecycle experimental

License MIT

Installation

```
# Install development version from GitHub (CRAN coming soon) using these two lines of code
if (!require("devtools")) install.packages("devtools")
devtools::install_github("hollina/scul")`
```

SCUL Tutorial

Overview of R package, extended example using publicly available data, and brief comparison to traditional method

Alex Hollingsworth

2020-05-03

Source: `vignettes/scul-tutorial.Rmd`

Contents

Example data

This tutorial uses publicly available data that is similar to the data used in Abadie, Diamond, and Hainmueller (2010). The empirical goal of Abadie, Diamond, and Hainmueller (2010) was to estimate the effects of a California tobacco control policy implemented in 1988.

When in long form, the data are at the state-year level and range from 1970 to 1997 (28 years). For each state and year there are data on cigarette sales per capita (`cigsale`) and the retail price of cigarettes (`retprice`). To be used in the SCUL procedure, the data must be in wide format, where each row is a time-period (e.g., year) and each column is a unit-specific variable. In our data, for each variable the unit is identified by the end of each column name (e.g., variables from the state of California are indicated by `_6` , which is the FIPS code for California.)

The dataset should be sorted by whatever variable you use to index time with the earliest date being first and the most recent date being last.

The `cigarette_sales` dataset is stored in the `data` subdirectory of this package. It should be automatically loaded when the `scul` package is loaded.

Thank you!

Paper: https://doi.org/10.31235/osf.io/fc9xt

R-package: https://hollina.github.io/scul/

Artwork by Amy Jiao: http://www.amyjiaotattoo.com

Twitter:

@ajhollingsworth

@coady_wing

Web: https://alexjhollingsworth.com