Big Data for Public Policy

8. Machine Learning for Econometrics

Elliott Ash & Malka Guillot

Where we are

- Past weeks:
  - w1: Overview and motivation
  - w2: Finding datastests using webcrawling and API
  - w3: Intro to supervised Machine Learning (ML) - regressions
  - w4: Text analysis fundamentals
  - w5: Supervised learning - classifications
  - w6: Unsupervised ML
  - w7: ensemble explanations [Ash]
- This week (w8):
  - econometrics and ML [Guillot]
- Next (last):
  - w9: AI policies [Ash]

Outline

Machine Learning and Public Policy Evaluation

- Anwser question about the world using observational data *"our goal as a field is to use data to solve problems"* Breiman (2001)
- The goal of social-science research with big data is the same as other social-science research:
  - provide credible tests of social-science hypotheses
  - estimate policy parameters to inform policymakers

How machine learning can be applied in research designs targeting public policy analysis?

Context: Machine Learning vs. Causal Inference

ML strength and weakeness:

- Powerful & flexible for doing predictions
- Allow for nonlinearities
- Can deal well with large nb of variables
- Unstable estimation of parameters

Context: Machine Learning vs. Causal Inference

ML strength and weakeness:

- Powerful & flexible for doing predictions
- Allow for nonlinearities
- Can deal well with large nb of variables
- Unstable estimation of parameters

ML approach:

- ML methods do not directly apply to estimating **causal** parameters
- Solve the problem first, worry about all the theory later

Context: Machine Learning vs. Causal Inference

ML strength and weakeness:

- Powerful & flexible for doing predictions
- Allow for nonlinearities
- Can deal well with large nb of variables
- Unstable estimation of parameters

ML approach:

- ML methods do not directly apply to estimating **causal** parameters
- Solve the problem first, worry about all the theory later

Causal inference approach:

- Causality: what identifies the causal parameters of interest?
- Statistics: how to estimate the identified target parameters the “best way”?

Context: Machine Learning for Policy Analysis

For policy analysis **both prediction and inference** are important.

3 ways ML can contribute:

1. pre-processing
2. prediction in policy
3. prediction in the service of estimation

Context: Machine Learning for Policy Analysis

For policy analysis **both prediction and inference** are important.

3 ways ML can contribute:

1. pre-processing $\rightarrow$ **unsupervised learning**
2. prediction in policy
3. prediction in the service of estimation

Context: Machine Learning for Policy Analysis

For policy analysis **both prediction and inference** are important.

3 ways ML can contribute:

1. pre-processing
2. prediction in policy$\rightarrow$ **supervised learning**
3. prediction in the service of estimation

Context: Machine Learning for Policy Analysis

For policy analysis **both prediction and inference** are important.

3 ways ML can contribute:

1. pre-processing
2. prediction in policy
3. **prediction in the service of estimation** $\rightarrow$ today

Context: Machine Learning for Policy Analysis

For policy analysis **both prediction and inference** are important.

3 ways ML can contribute:

1. pre-processing
2. prediction in policy
3. **prediction in the service of estimation**
    - Double ML
    - Heterogenous treatment effect
    - Synthetic control method

Model for Causal Inference

- **Causal question**: we want to know what would happen if a policy-maker changes a policy
  - E.g. increase minimum wage, raise a price, administer a drug
  - Fundamental pb of causal inference:
    - we do not see the same unit at the same time with alternative counterfactual policies

$\rightarrow$ Expressing some inference tasks as prediction problems

Model for Causal Inference

- **Causal question**: we want to know what would happen if a policy-maker changes a policy
  - E.g. increase minimum wage, raise a price, administer a drug
  - Fundamental pb of causal inference:
    - we do not see the same unit at the same time with alternative counterfactual policies

$\rightarrow$ Expressing some inference tasks as prediction problems

- **Potential Oucome Framework**
  - Rubin causal model
  - $Y(w) =$ the outcome the treatment w
  - For binary treatment, treatment effect is $\tau = Y(1) - Y(0)$
  - We also observe X_i, individual characteristics not affected by the treatment w

General References

- Textbook treatment of causal inference methods: Imbens and Rubin (2015)
- Lecture notes for *Machine Learning for Econometrics*, by L'Hour and Gaillac
- Athey and Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11.
- Mullainathan and Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87-106.

Outline

Introduction

Double ML

Heterogeneous Treatment Effect

Synthetic Control Method

Wrap-up

Double Machine Learning: Motivation

ML models perform well for prediction tasks, but it does not necessarily translate into good performance for estimation of **causal parameters**

$\rightarrow$ Double ML aims to construct high-quality estimates for causal parameteres

Idea: **solving 2 prediction problems**, and using the results to construct the estimate of the low dimensional parameter

Chernozhukov, Chetverikov, Duflo, Hansen, Demirer, and Newey (2017)

$$Y = \theta T + g(X) + \epsilon$$

- Y outcome variable
- T: low-dimensional treatment
- θ target parameter of interest
- $g(.)$ unkown, complicated function
- X: high-dimensional set of (observed) confounders

Chernozhukov, Chetverikov, Duflo, Hansen, Demirer, and Newey (2017)

$$Y = \theta T + g(X) + \epsilon$$

- Y outcome variable
- T: low-dimensional treatment
- θ target parameter of interest
- $g(.)$ unkown, complicated function
- X: high-dimensional set of (observed) confounders
- Confounders s.t. $T = m(X) + \eta$, $E(\eta|X) = 0$.

Chernozhukov, Chetverikov, Duflo, Hansen, Demirer, and Newey (2017)

$$Y = \theta T + g(X) + \epsilon$$

- Y outcome variable
- T: low-dimensional treatment
- θ target parameter of interest
- $g(.)$ unkown, complicated function
- X: high-dimensional set of (observed) confounders
- Confounders s.t. $T = m(X) + \eta$, $E(\eta|X) = 0$.
- Because of confounders $\hat{Y} = \hat{\theta} T + \hat{g}(X)$ will be biased.

How to use modern ML techniques to estimate $g(.)$ and carry out inference about low-dimensional θ?

Double ML method

1. using any ML method, predict:
    1.1 Y given X: $\hat{Y}(X)$
    1.2 T given X: $\hat{T}(X)$

Double ML method

1. using any ML method, predict:
   1.1 Y given X: $\hat{Y}(X)$
   1.2 T given X: $\hat{T}(X)$
2. Form residuals $\tilde{Y} = Y - \hat{Y}(X)$ and $\tilde{T} = T - \hat{T}(X)$

Double ML method

1. using any ML method, predict:
   1.1 Y given X: $\hat{Y}(X)$
   1.2 T given X: $\hat{T}(X)$
2. Form residuals $\tilde{Y} = Y - \hat{Y}(X)$ and $\tilde{T} = T - \hat{T}(X)$
3. Regress $\tilde{Y}$ on $\tilde{T}$ to learn $\hat{\theta}$.

Double ML method

1. using any ML method, predict:
    1.1 Y given X: $\hat{Y}(X)$
    1.2 T given X: $\hat{T}(X)$
2. Form residuals $\tilde{Y} = Y - \hat{Y}(X)$ and $\tilde{T} = T - \hat{T}(X)$
3. Regress $\tilde{Y}$ on $\tilde{T}$ to learn $\hat{\theta}$.

- **Sample split**:
    - Run (1) on sample a, then run (2) and (3) on sample b, to estimate $\hat{\theta}_a$
    - and vice versa (run (1) on sample b, and (2/3) on sample a), to learn a second estimate for $\hat{\theta}_b$.

Double ML method

1. using any ML method, predict:
   1.1 Y given X: $\hat{Y}(X)$
   1.2 T given X: $\hat{T}(X)$
2. Form residuals $\tilde{Y} = Y - \hat{Y}(X)$ and $\tilde{T} = T - \hat{T}(X)$
3. Regress $\tilde{Y}$ on $\tilde{T}$ to learn $\hat{\theta}$.

- **Sample split**:
  - Run (1) on sample a, then run (2) and (3) on sample b, to estimate $\hat{\theta}_a$
  - and vice versa (run (1) on sample b, and (2/3) on sample a), to learn a second estimate for $\hat{\theta}_b$.
- **Cross-fitting**
  - average them to get a more efficient estimator:$\hat{\theta} = \frac{1}{2}(\hat{\theta}_a + \hat{\theta}_b)$.

$\rightarrow$ cf. cross-validation in ML

Empirical Application: Job Training Program

Bléhaut, D'Haultfœuille, L'Hour, and Tsybakov (2017).

- Revisit Lalonde (1986) evaluation of the impact of the National Supported Work (NSW) program:
  - transitional, subsidized work experience program targeted towards people with longstanding employment problems
- **Treated group**: people who were randomly assigned to this program from the population at risk ($n_1 = 185$).
- Take advantage of the **2 possible control groups**:
  1. experimental data from a RCT ($n_0 = 269$)
  2. survey data from the Panel Study of Income Dynamics (PSID) (sample size $n_0 = 2490$).

Treatment Effect on LaLonde (1986)

- 172 variables:
  - age, education, black, hispanic, married, no degree, income in 1974 and in 1975, no earnings in 1974 and in 1975
  - 2x2 interactions between dummies &
    - continous variables ; up to a 5 order polynomial in the continuous variables
- Continuous variables are linearly rescaled to $[0, 1]$.

	Estimator:		
	Experimental	*Cross-fitting*	*Cross-fitting w. 20 partitions*
	(1)	(2)	(3)
OLS	1,794		
	(633)		
Lasso		2,305	2,403
		(676)	(685)
Random Forest		7,509	1,732
		(6,711)	(1,953)

Conclusion on Double ML and cross-fitting

- In Bléhault et al.:
  - Estimates are compared to the experimental OLS
  - Lasso perform well in both cross-fitting
  - Random forests performs poorly and standard-erros are very high

Conclusion on Double ML and cross-fitting

- In Bléhault et al.:
  - Estimates are compared to the experimental OLS
  - Lasso perform well in both cross-fitting
  - Random forests performs poorly and standard-erros are very high

→ Test several ML algorithms when possible

→ Consider many data splits so the results do not depend so much on the partitions.

Outline

Motivation (1)

Why is it important to uncover heterogeneity in the impact of a policy on individuals' outcomes?

Motivation (1)

Why is it important to uncover heterogeneity in the impact of a policy on individuals' outcomes?

- How can the government design a policy under budget constraint?
  - E.g: health & education expenditures

Motivation (1)

Why is it important to uncover heterogeneity in the impact of a policy on individuals' outcomes?

- How can the government design a policy under budget constraint?
  - E.g: health & education expenditures
- Government wants to
  - Determine a fixed allocation of treatment resources to a target population,
  - While maximizing the population mean outcome

Motivation (1)

Why is it important to uncover heterogeneity in the impact of a policy on individuals' outcomes?

- How can the government design a policy under budget constraint?
  - E.g: health & education expenditures
- Government wants to
  - Determine a fixed allocation of treatment resources to a target population,
  - While maximizing the population mean outcome

$\rightarrow$ Identify the populations that benefit the most/the least from the treatment, in order to maximize the social welfare

- Objective: estimating optimal policy assignments

Heterogeneous Treatment Effects

- ML can be useful for **uncovering treatment effect heterogeneity**
    - Where we focus on heterogeneity with respect to observable covariates.
- Which individuals benefit most from a treatment? For which individuals is the treatment effect positive?

Adaptation of the Rubin Causal Model

- $Y_i(w) =$ the outcome the treatment w would have had if assigned to unit i
- For binary treatment, treatment effect is $\tau_i = Y_i(1) - Y_i(0)$
- In this framework, τ_i can be different for each unit i

$\rightarrow$ Heterogeneous treatment effects

Adaptation of the Rubin Causal Model

- $Y_i(w) =$ the outcome the treatment w would have had if assigned to unit i
- For binary treatment, treatment effect is $\tau_i = Y_i(1) - Y_i(0)$
- In this framework, τ_i can be different for each unit i

$\rightarrow$ Heterogeneous treatment effects

- **Fundamental problem of causal inference**: we never observe both $Y_i(0)$ and $Y_i(1)$ for the same individual.

$\rightarrow$ cannot directly use ML methods to estimate τ_i.

Conditional Average Treatment Effect (CATE)

Under the assumption of selection on the observables
[i.e. all variation in treatment is random once the covariates X are fixed]

$$\begin{aligned}\tau(\mathbf{x}) &= E[Y(1)|\mathbf{X}=\mathbf{x}] - E[Y(0)|\mathbf{X}=\mathbf{x}] \\ &= \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\end{aligned} \tag{1}$$

If $\tau(.)$ is estimated by using separate ML estimations of $\mu_1(.)$ and $\mu_0(.)$, the estimation bias can accumulate

$\rightarrow$ ML solution to jointly estimate $\mu_1(.)$ and $\mu_0(.)$

Motivation (2)

- Why not
  - Stratifying the data and estimate the ATE within each strata
  - Most straightforward and popular unbiased estimation of the CATE

Motivation (2)

- Why not
  - Stratifying the data and estimate the ATE within each strata
  - Most straightforward and popular unbiased estimation of the CATE
- Problem: false discovery (data snooping, "p-hacking", "fishing")

Motivation (2)

- Why not
    - Stratifying the data and estimate the ATE within each strata
    - Most straightforward and popular unbiased estimation of the CATE
- Problem: false discovery (data snooping, "p-hacking", "fishing")
- Solution: Pre-register hypotheses and analyses

Motivation (2)

- Why not
  - Stratifying the data and estimate the ATE within each strata
  - Most straightforward and popular unbiased estimation of the CATE
- Problem: false discovery (data snooping, “p-hacking”, “fishing”)
- Solution: Pre-register hypotheses and analyses
- Pre-registration solves commitment and transparency problems
- It does not solve the statistical problem

Motivation (2)

- Why not
  - Stratifying the data and estimate the ATE within each strata
  - Most straightforward and popular unbiased estimation of the CATE
- Problem: false discovery (data snooping, "p-hacking", "fishing")
- Solution: Pre-register hypotheses and analyses
- Pre-registration solves commitment and transparency problems
- It does not solve the statistical problem

$\rightarrow$ **data-driven approach avoiding strong modelling assumption**

Wager and Athey (2017)

- Use **random forest** to estimate heterogeneous treatment effects

Wager and Athey (2017)

- Use random forest to estimate heterogeneous treatment effects
- Principle:
  1. Allocate "similar as possible" units, treated and non-treated, to the trees' leafs
  2. Read-off the treatment effect within the leaf

Wager and Athey (2017)

- Use random forest to estimate heterogeneous treatment effects
- Principle:
  1. Allocate "similar as possible" units, treated and non-treated, to the trees' leafs
  2. Read-off the treatment effect within the leaf
- The **similarity** is defined with respect to outcome or probability of treatment

Wager and Athey (2017)

- Use random forest to estimate heterogeneous treatment effects
- Principle:
    1. Allocate "similar as possible" units, treated and non-treated, to the trees' leafs
    2. Read-off the treatment effect within the leaf
- The **similarity** is defined with respect to outcome or probability of treatment
- **Causal trees** make sure each leaf has at least k observations from both treated and non-treated groups

Wager and Athey (2017)

- Use random forest to estimate heterogeneous treatment effects
- Principle:
    1. Allocate "similar as possible" units, treated and non-treated, to the trees' leafs
    2. Read-off the treatment effect within the leaf
- The **similarity** is defined with respect to outcome or probability of treatment
- **Causal trees** make sure each leaf has at least k observations from both treated and non-treated groups

$\rightarrow$ **Causal forests** are random forests for treatment effect estimation

Application: Davis and Heller (2017) [Context]

- Estimate the benefits from two youth employment program in Chicago
- 2 Randomized Control Trials (RCTs): same summer job program in 2012 and 2013
  - Sample sizes: 1,634 and 5,216 observations
  - Large set of covariates

Application: Davis and Heller (2017) [Context]

- Estimate the benefits from two youth employment program in Chicago
- 2 Randomized Control Trials (RCTs): same summer job program in 2012 and 2013
    - Sample sizes: 1,634 and 5,216 observations
    - Large set of covariates
- The program
    - provides disadvantaged youth aged 14-22 with 25 hrs/week of employment and an adult mentor.
    - Participants are paid at Chicago's minimum wage.

Application: Davis and Heller (2017) [Context]

- Estimate the benefits from two youth employment program in Chicago
- 2 Randomized Control Trials (RCTs): same summer job program in 2012 and 2013
  - Sample sizes: 1,634 and 5,216 observations
  - Large set of covariates
- The program
  - provides disadvantaged youth aged 14-22 with 25 hrs/week of employment and an adult mentor.
  - Participants are paid at Chicago's minimum wage.
- 2 outcomes:
  - violent-crime arrests within two years of random assignment
  - an indicator for ever being employed during the six quarters after the program.

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data
2. split the sample into **training and estimation sample** (50-50)

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data
2. split the sample into training and estimation sample (50-50)
3. train the **causal forest** on the **training sample**

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data
2. split the sample into training and estimation sample (50-50)
3. train the causal forest on the training sample
4. use the tree from (3) to compute **treatment effets in the estimation sample at each terminal leaf**

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data
2. split the sample into training and estimation sample (50-50)
3. train the causal forest on the training sample
4. use the tree from (3) to compute treatment effets in the estimation sample at each terminal leaf
5. propagate these estimate to the full sample using relevant leaf estimates

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data
2. split the sample into training and estimation sample (50-50)
3. train the causal forest on the training sample
4. use the tree from (3) to compute treatment effets in the estimation sample at each terminal leaf
5. propagate these estimate to the full sample using relevant leaf estimates
6. repeat previous steps $25,000$ times

Application: Davis and Heller (2017) [Method - step 1]

1. draw a random sample for the full data
2. split the sample into training and estimation sample (50-50)
3. train the causal forest on the training sample
4. use the tree from (3) to compute treatment effets in the estimation sample at each terminal leaf
5. propagate these estimate to the full sample using relevant leaf estimates
6. repeat previous steps $25,000$ times
7. Compute the **average prediction for each obseration across trees** $=$CATE

Application: Davis and Heller (2017) [hyperparameters]

- the number of trees,

Application: Davis and Heller (2017) [hyperparameters]

- the number of trees,
- the minimum number of treatment and control observations in each leaf,
    - bias-variance trade-off
    - bigger leaves make results more consistent across different samples but predict less heterogeneity

Application: Davis and Heller (2017) [hyperparameters]

- the number of trees,
- the minimum number of treatment and control observations in each leaf,
    - bias-variance trade-off
    - bigger leaves make results more consistent across different samples but predict less heterogeneity
- the subsample size
    - smaller subsamples reduce dependence across trees but increase the variance of each estimate

Application: Davis and Heller (2017) [Method - step 2]

Question: if we divide the sample into a group predicted to respond positively to the program and one that is not, would we successfully **identify youth with larger treatment effects**?

i.e. standard subgroup analysis, but with **subgroups determined by the high-dimensional combination** of covariates captured

Test this by:

- Split the sample in 2 halves: S_{in} and S_{out}
- Run step 1 on S_{in}, use the model on S_{out}
- Group youth by whether they are predicted to have a positive or negative treatment effect
- regressing each outcome on the group indicator and covariates

Application: Davis and Heller (2017) [Results]

Subgroup	No. of violent crime arrests	Any formal employment
Panel A. In sample		
$\hat{\tau}_i^{CF}(x) > 0$	0.22	0.19
	(0.05)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	–0.05	–0.14
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.00	0.00
Panel B. Out of sample		
$\hat{\tau}_i^{CF}(x) > 0$	–0.01	0.08
	(0.05)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	–0.02	–0.01
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.77	0.02

Outline

Synthetic Control Method

Athey and Imbens (2017) : “arguably the most important innovation in policy evaluation literature in the last 15 years.”

Athey and Imbens (2017) : “arguably the most important innovation in policy evaluation literature in the last 15 years.”

General framework: Comparing the evolution of an outcome variable for one treated region relative to the evolution of this variale in this region as if it had not been treated

Athey and Imbens (2017) : “arguably the most important innovation in policy evaluation literature in the last 15 years.”
General framework: Comparing the evolution of an outcome variable for one treated region relative to the evolution of this variale in this region as if it had not been treated

$\rightarrow$ an alternative to difference-in-differences when only **aggregate data** are available

Synthetic Control Method

Athey and Imbens (2017) : "arguably the most important innovation in policy evaluation literature in the last 15 years."
General framework: Comparing the evolution of an outcome variable for one treated region relative to the evolution of this variale in this region as if it had not been treated

$\rightarrow$ an alternative to difference-in-differences when only **aggregate data** are available

Seminal papers:

- Abadie and Gardeazabal (2003), *American Economic Review*
- Abadie, Diamond, Hainmueller (2010), *Journal of American Statistical Association*

Matching / Synthetic Control

- **Matching**: use covariates to find matching individuals
- **Synthetic control**: construct a synthetic "match" from a weighted average of other individuals (based on covariates).
- Note:
  - Equivalent to controlling for many observed confounders.
- Can imagine the text documents associated with individual or groups as a set of covariates for matching
  - e.g., text features from the criminal history of each defendant.

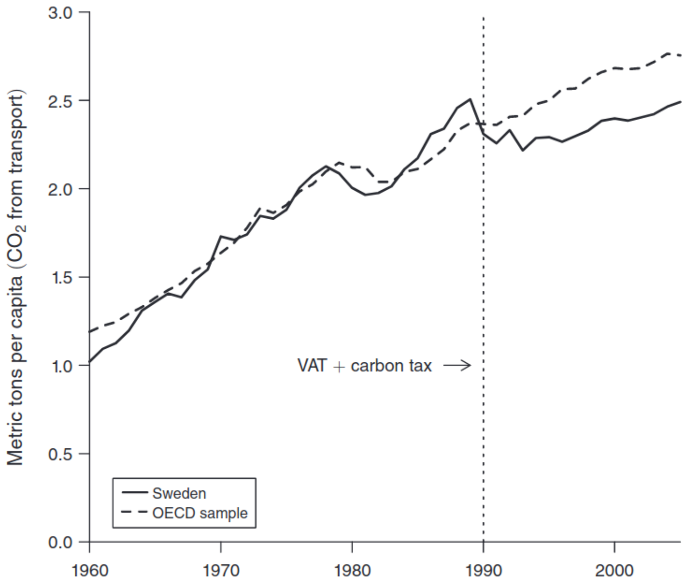
J. J. Andersson, 2019, "Carbon Taxes and CO2 Emissions: Sweden as a Case Study", *American Economic Journal: Economic Policy* 2019, 11(4): 1-3

Research question: What is the impact of carbon taxation on CO2 emission?

Method: construct a "synthetic Sweden" using SCM

Natural experiment: introduction of a carbon tax and a value-added tax on transport fuel in Sweden (1990)

Carbon Tax in Sweden



Carbon Tax in Sweden - Results

In 2005, carbon emissions are 12.5% smaller thanks to carbon taxation

Total reduction of emission for 1990-2005 : 2.5Mt CO_2/year in avg.(ATE), 40Mt CO_2 in total.

Why not a Difference-in-Differences strategy?

DiD is good for eliminating the influence of inobservables using the difference between pre and post treatment

Assumption:

- Effects of inobservable do not vary with time
- Every macroeconomic shock is common to treatment and counterfactual

$\rightarrow$ Parallel trend assumption

$\rightarrow$ Not easy to check

SCM Intuition

SCM solves 2 pbs: the difficulties to

- check the parallel trend assumption
- find a suitable control unit whose characteristics are close to the treated unit

$\rightarrow$ the method combines control unit to build a counterfactual

Setting

$J+1$ units (or regions) and T periods:

- Only region 1 is treated from period T_0 onward
- Unit $i=2,...,J+1=$ donor pool

Setting

$J+1$ units (or regions) and T periods:

- Only region 1 is treated from period T_0 onward
- Unit $i = 2, ..., J+1$ = donor pool

Potential outcome framework:

- $Y_{i,t}(0)$: the potential outcome for unit i at time t if it is not treated
- $Y_{i,t}(1)$: the potential outcome for i if it is exposed to the intervention

Setting

$J+1$ units (or regions) and T periods:

- Only region 1 is treated from period T_0 onward
- Unit $i = 2, ..., J+1 =$ donor pool

Potential outcome framework:

- $Y_{i,t}(0)$: the potential outcome for unit i at time t if it is not treated
- $Y_{i,t}(1)$: the potential outcome for i if it is exposed to the intervention
- We observe exposure to the treatment $D_{i,t}$ and the realized outcome $Y_{i,t}^{obs}$ defined by:

$$Y_{i,t}^{obs} = Y_{i,t}(D_{i,t}) = \begin{cases} Y_{i,t}(0) \text{ if } D_{i,t} = 0 \\ Y_{i,t}(1) \text{ if } D_{i,t} = 1 \end{cases}$$

Setting

Quantity of interest:

$$\tau_t = Y_{1,t}(1) - Y_{1,t}(0) \text{ for } t \in \{T_0, ..., T\}$$

Objective: estimate the effect of the intervention on unit 1 for $t \in \{T_0, ..., T\}$

Asumption: the outcome variable in non treated regions is not impacted byt the treatment in the treated region $\rightarrow$ no interference

Setting

Quantity of interest:

$$\tau_t = Y_{1,t}(1) - Y_{1,t}(0) \text{ for } t \in \{T_0, ..., T\}$$

Objective: estimate the effect of the intervention on unit 1 for $t \in \{T_0, ..., T\}$

Asumption: the outcome variable in non treated regions is not impacted byt the treatment in the treated region $\rightarrow$ no interference

Solution:

$$\hat{\tau}_t = Y_{1,t}(1) - \sum_{j=2}^{J+1} w_j^* Y_{j,t} \text{ for } t \in \{T_0, ..., T\}$$

A Practical Note

In most papers:

- Time dimension:
  - T is relatively large
  - Different than in panel data where many treated units but small T (usually < 12)
- Unit of interest: a city, a region or even a country

Why does it work?

Suppose the outcome under no-treatment is given by the model:

$$Y_{1,t}(0) = \delta_t + Z_i'\theta_t + \lambda_t'\mu_i + \epsilon_{i,t}$$

- δ_t is a time fixed-effect,
- θ_t is a vector of time-varying parameters,
- Z_i are observed covariates,
- λ_t are unobserved common factors of dimension F ,
- μ_i are unobserved factors loadings (dimension F) and
- $\epsilon_{i,t}$ are unobserved transitory shocks.

This is a **factor model**:

- think of λ_t as the underlying macroeconomic dynamics that affect differently each unit through μ_i .

Why does it work?

Consider $W = (w_2, ..., w_{J+1})$ such that $w_j > 0\ \forall j$ and $w_2 + ... + w_{J+1} = 1$

Then

$$\sum_{j=2}^{J+1} w_j Y_{j,t} = \delta_t + \theta_j \sum_{j=2}^{J+1} w_j Z_j' + \lambda_t \sum_{j=2}^{J+1} w_j \mu_j + \sum_{j=2}^{J+1} w_j \epsilon_{j,t}$$

If we can find $(w_2^*, ..., w_{J+1}^*)$ such that

(a) $\forall t \{1, ..., T_0\},\ \sum_{j=2}^{J+1} w_j^* Y_{j,t} = Y_{1,t}$

(b) $\sum_{j=2}^{J+1} w_j^* Z_j = Z_1$

(c) and some other conditions are satisfied

Then $\forall t \{1, ..., T_0\},\ \sum_{j=2}^{J+1} w_j^* Y_{j,t} \approx Y_{1,t}(0)$

Return to Andersson (2019) - Synthetic Sweden

A unit comparable to Sweden made of a combination of countries, with the following **pre-treatment** features:

- with no carbon tax [not treated]
- that look like Sweeden on a number of variables explaining the outcome (Z_j)
- With trends and levels in CO2 emissions close to that of Sweden

Andersson (2019) - Donor pool

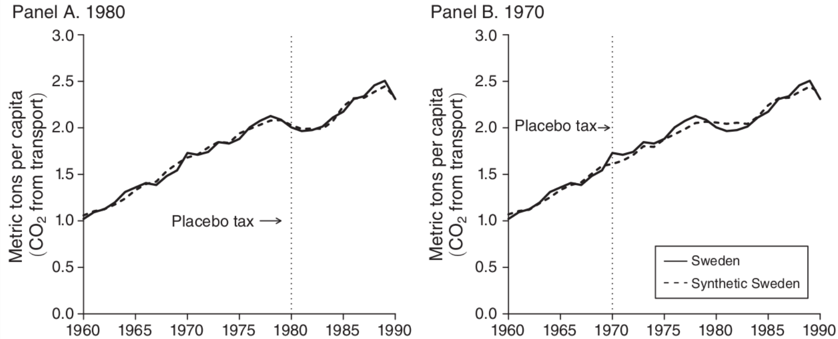
25 OECD countries, less:

- Countries with a carbon tax: Finalande, Norway, the Netherlands
- Countries with reforms of fuel tax: Germany, Italy, UK
- Countries subject to "fuel tourism": Luxembourg, , Austria, Turkey
- Countries too different to Sweden: Irland and Turkey

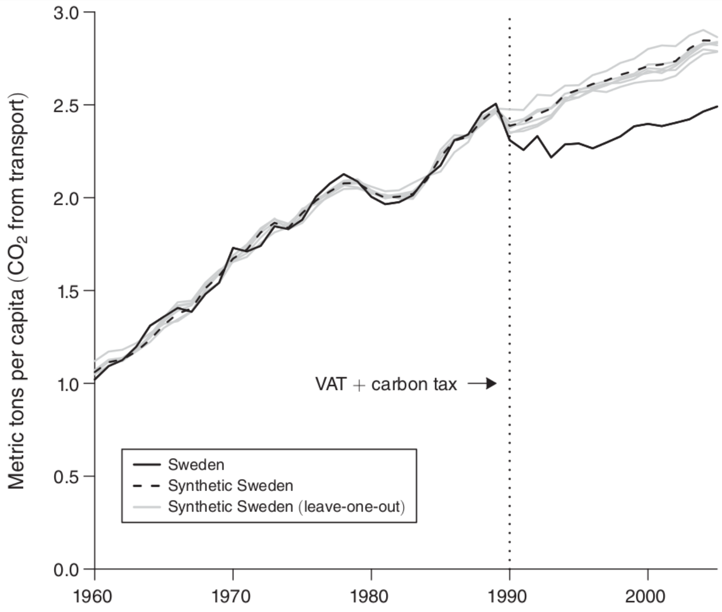
Andersson (2019) - Variables Mean before Treatment

Variables	Sweden	Synth. Sweden	OECD sample
GDP per capita	20,121.5	20,121.2	21,277.8
Motor vehicles (per 1,000 people)	405.6	406.2	517.5
Gasoline consumption per capita	456.2	406.8	678.9
Urban population	83.1	83.1	74.1
CO_2 from transport per capita 1989	2.5	2.5	3.5
CO_2 from transport per capita 1980	2.0	2.0	3.2
CO_2 from transport per capita 1970	1.7	1.7	2.8

Inference : in-time Placebo test



Robustness check : Leave-on-out test



Conclusion on SCM

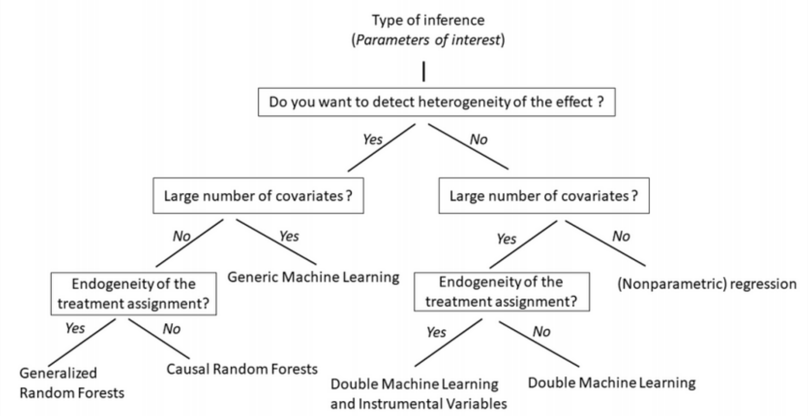
- The construction of the counterfactual is transparent
- Very useful when it is hard to find a counterfactual
- More general than DiD:
  - several weighted controls
  - relax the parallel trend assumption
  - allows effects of inobservables to vary across time
- Easy implementation using Stata and R package [cf. synth package]

More References on SCM

- Original papers that developed the method:
  - Abadie and Gardeazabal (2003); Abadie et al. (2010, 2015).
- Clear and concise presentation:
  - Abadie and Cattaneo (2018)
- Research frontier reviewed by:
  - Athey and Imbens (2017). Angrist and Pischke (2010)
- Some applications:
  - taxation and migration of football players (Kleven et al., 2013),
  - immigration (Bohn et al., 2014),
  - health policy (Hackmann et al., 2015);
  - minimum wage (Allegretto et al., 2013),
  - regional policies (Gobillon and Magnac, 2016);
  - prostitution laws (Cunningham and Shah, 2017),
  - financial value of connections to policy-makers (Acemoglu et al., 2016),

Outline

Road-map



Type of inference
(*Parameters of interest*)
Do you want to detect heterogeneity of the effect ?
Yes
No
Large number of covariates ?
Large number of covariates ?
No
Yes
Yes
No
Endogeneity of the treatment assignment?
Generic Machine Learning
Endogeneity of the treatment assignment?
(Nonparametric) regression
Yes
No
Yes
No
Generalized Random Forests
Causal Random Forests
Double Machine Learning and Instrumental Variables
Double Machine Learning

Wrap-up

1. **Double Machine Learning** and **sample-splitting**:
    - deals with high-dimensional features
    - Any ML algorithm, cross-fitting
2. Heterogenous treatment effects
    - inference about features of the conditional average treatment effect
    - Causal random forest
3. Synthetic control method:
    - an inherently high-dimensional tool particularly useful for policy evaluation with aggregate data

Wrap-up

1. Double Machine Learning and sample-splitting:
   - deals with high-dimensional features
   - Any ML algorithm, cross-fitting
2. **Heterogenous treatment effects**
   - inference about features of the conditional average treatment effect
   - Causal random forest
3. Synthetic control method:
   - an inherently high-dimensional tool particularly useful for policy evaluation with aggregate data

Wrap-up

1. Double Machine Learning and sample-splitting:
    - deals with high-dimensional features
    - Any ML algorithm, cross-fitting
2. Heterogenous treatment effects
    - inference about features of the conditional average treatment effect
    - Causal random forest
3. **Synthetic control method**:
    - an inherently high-dimensional tool particularly useful for policy evaluation with aggregate data