Machine Learning for Causal Inference

Mitchell J. Lovett[1]

August 10, 2019

[1] The slides and related chapter benefited greatly from assistance by Nan Li, Bowen Luo, and Xavi Vidal-Berastain.

Agenda

1. Introduction
2. Lasso IV
   - Introduction to Lasso IV
   - Model and Proposed Solution
   - Generalizations and Extensions
   - Illustration and Applications
   - Example Code and Monte Carlo Data
3. Synthetic Controls
   - Model and Methods
   - Generalizations to Multiple Treatment Units
   - Illustrations and Applications
   - Example Code and Monte Carlo Data
4. Conclusion

Agenda

1. Introduction
2. Lasso IV
    - Introduction to Lasso IV
    - Model and Proposed Solution
    - Generalizations and Extensions
    - Illustration and Applications
    - Example Code and Monte Carlo Data
3. Synthetic Controls
    - Model and Methods
    - Generalizations to Multiple Treatment Units
    - Illustrations and Applications
    - Example Code and Monte Carlo Data
4. Conclusion

Introduction: Digital Field Experiments Are The Future!

- Recent large scale experiments provide tight causal inference
  - Yahoo!: Lewis and Rao 2015
  - Google: Johnson, Lewis, and Nubbemeyer 2017b
  - Facebook: Gordon et al. 2018
- . . .But maybe not the whole story
- Many marketing activities not digital or not easily manipulated in online experiments

Introduction: Machine Learning Is the Future!

- Many ML Tools
  - Trees & Random Forests
  - Lasso and Ridge
  - Support Vector Machines and Neural Networks
  - Hybrid models - boosting, bagging, ensemble methods
- Applied to many marketing domains
  - Churn (e.g., Vafeiadis et al 2015; Neslin et al 2006)
  - Ad clicking (e.g., Perlich et al 2014)
  - Product recommendations (e.g., Huang et al 2007)
  - Peer effects (e.g., Bailey et al 2019)
  - and more broadly used to extract features from pictures, video, audio, and text

Introduction: Why Machine Learning is the Future

- ML uses more flexible functions of input to output variables
- Implies many, many parameters to calibrate, often too many!
- Two key concepts from Machine Learning:

Regularization

- Tunes the set of parameters to be smaller than the potential set considered.
- Adjusts the objective function to add regularization term (penalty)

Sample Splitting and Cross-validation

- Splitting sample into subsamples (e.g., train and test)
- Subsamples used to tune regularization
- Typically tuning to out-of-sample prediction performance

- Most techniques developed for prediction of outcome, not for the effect of a decision on an outcome (Mullainathan and Speiss 2017)

Introduction: Why ML Is Not Enough?

- Consider a simple time-series analysis
  - y_t as target DV, say sales
  - x_{t-1} as decision variable, say advertising.
- Loosely, "Granger causality" tests whether x_{t-1} has an effect after controlling for y_{t-1}.
- Doesn't establish *causality* as we normally mean it

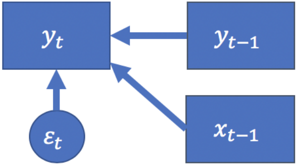


Figure: Time Series Relationships

Introduction: Why ML Is Not Enough? (2)

- ML has a similar problem to time series analysis
- Can help us understand correlations very well, since tuned to predict out-of-sample
- But not tuned to obtain causal inferences
- Known to be biased for measuring causal effects (Leeb et al. 2006; Leeb and Potscher 2008)

Introduction: Why ML Is Not Enough? (2)

- ML $=$ OLS $+$ steroids
- ML $\neq$ silver bullet

Introduction: What do we mean by causality?

- This is a deeply philosophical question
- Loosely, offer two perspectives related to the methods we will discuss
  1. Pearl Causality: A causal effect is measured as the difference between the outcome when you force a variable to a level via a "Do" operator versus to a different level, as in an experiment.
  2. Rubin Counterfactual Outcomes: A causal effect is the difference between the treated unit and an unobservable, counterfactual situation had you not treated the unit
- Ultimately, causal effects are about a special kind of prediction, predicting the effect of an action, not simply the level of an outcome

Introduction: ML For Causal Inference Is the Future!

- At least for the next hour or so!
- We will review two methods and present applications
- Causal Analytics: techniques leverage power of machine learning methods to measure causal effects

Introduction: Causal Analytics

1. Obtain Causal Inference

Predict the effect of an action or outcome under a different action

2. From Observational Data

Observe data, but can't directly manipulate the action in an experiment

3. Leverage Machine Learning

Take advantage of advances in machine learning techniques to improve causal inferences

Introduction: The Two Techniques Today

1. Lasso IV (and double-machine learning)

Tools to obtain causal inference through instrumental variables (Pearl)

2. Synthetic controls (and matrix completion)

Tools to obtain causal inference similar to diff-in-diff approach (Rubin)

Agenda

1 Introduction

2 Lasso IV
- Introduction to Lasso IV
- Model and Proposed Solution
- Generalizations and Extensions
- Illustration and Applications
- Example Code and Monte Carlo Data

3 Synthetic Controls
- Model and Methods
- Generalizations to Multiple Treatment Units
- Illustrations and Applications
- Example Code and Monte Carlo Data

4 Conclusion

Introduction to Lasso IV

What is Lasso IV

A machine learning approach to getting causal effects for endogenous variables, like advertising or price

When can you apply Lasso IV

1. For any focal variable (discrete or continuous)
2. When some exogenous instruments are available

Introduction to Lasso IV

- For this discussion, we will first work with a simple model:

$$y_i = \alpha d_i + \varepsilon_i, \tag{1}$$

  where

  - y_i is the dependent variable
  - d_i is the focal variable (e.g., advertising)
  - α is the causal effect we desire to measure

- Notes:
  - Can easily add controls to this model
  - This section draws primarily from Belloni et al (2014) and Belloni et al (2012)

Visual of Simple Model

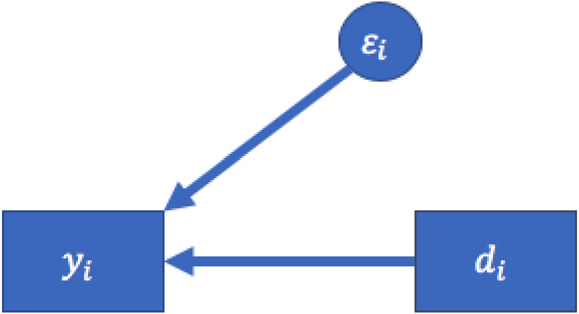


Figure: Causal Diagram for Simple Model. Lines Represent Relationships.

Problem: d_i is endogenous, i.e., $E(\varepsilon_i d_i) \neq 0$

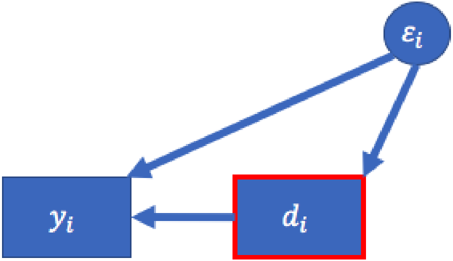


Figure: Causal Diagram for Simple Model. Lines Represent Relationships. Red Box Represents Endogenous Variable

- Can arise in almost any regression, but is likely when
  - can't/didn't manipulate the focal variable directly
  - don't believe that we have "unspecified" exogenous variation

What is an Instrument?

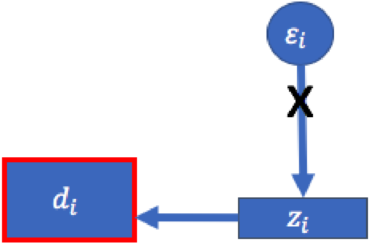


Figure: Causal Diagram for Simple Model. Lines Represent Relationships. Red Box Represents Endogenous Variable

- An instrument, z_i, must be
  - Exogenous: Not related to ε_i, i.e., $E(\varepsilon_i z_i) = 0$ or put differently, affects y_i only through d_i (not directly)
  - Relevant: Explains d_i, i.e., $E(z_i d_i) >> 0$

How does an Instrument Work? (1)

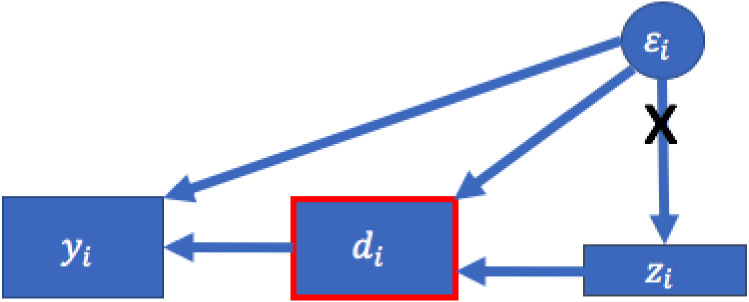


Figure: Causal Diagram for Simple Model. Lines Represent Relationships. Red Box Represents Endogenous Variable

How does an Instrument Work? (2)

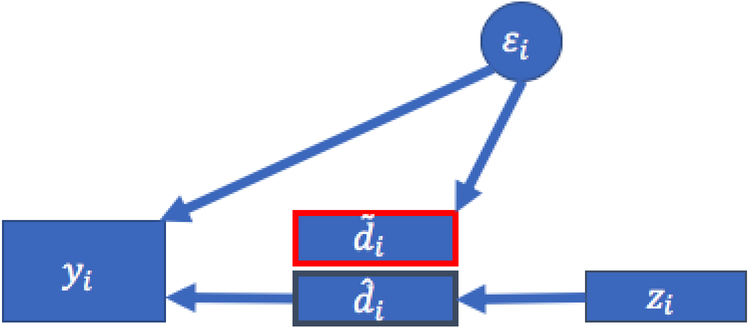


Figure: Causal Diagram for Simple Model. Lines Represent Relationships. Red Box Represents Endogenous Variable

Using instruments to estimate causal effects

- With such an instrument, we can conduct various methods of estimating the causal effect with the instrument
- For example, two-stage least squares proceeds as follows:
    1. Run first stage regression: $d_i = \nu z_i + \omega_i$
    2. Construct fitted values $\hat{d}_i = \nu z_i$
    3. Run second stage regression: $y_i = \tilde{\beta}\hat{d}_i + \tilde{\varepsilon}_i$[2]
- This approach introduces the equation for the first stage,

$$d_i = \nu z_i + \omega_i \tag{2}$$

[2] If obtaining the estimates this way, you must correct the standard errors!

Weak Instruments Problem

- When instruments do not explain d_i well (fails relevance)
  - Can produce a weak instruments problem (bias!)
  - Or simply low efficiency (large second stage standard errors)
- One solution is to find more/stronger instruments

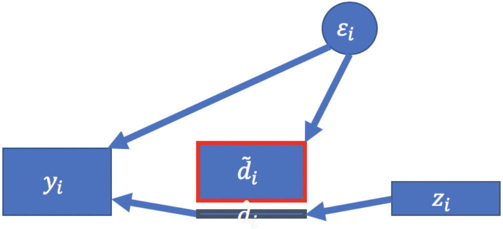


Figure: Causal Diagram for Simple Model. Lines Represent Relationships. Red Box Represents Endogenous Variable

How do we get more instruments?

- Although sometimes we struggle to identify any instruments, other times we have access to many options
  - Component cost shifters
  - Instruments in aggregate logit estimation (Berry et al 1995; Nevo 2000) include summaries of characteristics of other markets/periods
  - Waldfogel instruments allow including (many) demand shifters for excluded market if the decision can't vary across the focal and excluded markets (Li et al 2019).

How do we get more instruments? (2)

- Also can construct many instruments from even a small set of base instruments
  - Interactions between exogenous variables, e.g., 2-way or 3-way
  - Non-parameteric functions of a single instrument, e.g., polynomial approximations

- Such instruments need not be atheoretical
  - Approximating a non-linear threshold relationship between political advertising as it crowds out commercial advertising (Lovett et al 2019)
  - Allowing different demographic groups to respond differently to breaking their mobile phone, such as approximating income or budget constraints (Bailey et al 2019)

Many instruments creates a new problem

- With many instruments most are not expected to be relevant, but fitting noise may allow better approximation to d_i
- This leads to a "many weak instruments problem" (Stock et al 2002)
  - Problem exists even if some instruments are strong!
  - Many weak instruments invalidate the estimator (bias)
  - Biases second stage estimates toward the OLS

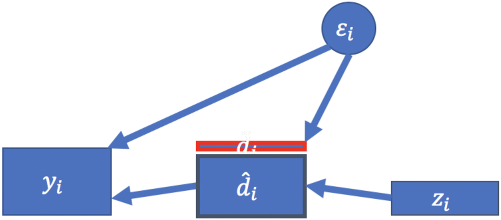


Figure: Causal Diagram for Simple Model. Lines Represent Relationships. Red Box Represents Endogenous Variable

Lasso IV Solution

- **Goal**: Balance between risk of fitting noise and increasing explained variation in d_i, $\hat{d}_i$
- **How?**: Achieve using ML tuned for causal inference!

Lasso IV Model

- Let the base k_x instruments be in the vector x_i, which are *assumed* to be exogenous
- The p dimensional instrument vector, z_i is characterized by

$$z_i = \left(f_1\left(x_i\right), f_2\left(x_i\right), \ldots, f_p\left(x_i\right)\right), \tag{3}$$

  where f_m are functions of the original or base instrument vector, x_i.
- z_i can be
  - exactly x_i, so that $f_m\left(x_i\right) = x_i$
  - or z_i can be formed by taking various functions of x_i so that $p > k_x$
- In either case, p is allowed to be very large, even $p > n$

Lasso IV Model (2)

- We rewrite the model set-up as follows

$$y_i = \alpha d_i + \varepsilon_i, \tag{4}$$
$$d_i = z_i' \Pi + r_i + v_i \tag{5}$$

where

r_i is approximation error and we assume that $E\left[\varepsilon_i | x_i\right] = E\left[\varepsilon_i | z_i\right] = E\left[v_i | z_i, r_i\right] = 0$, but that $E\left[\varepsilon_i v_i\right] \neq 0$

- Goal of z_i is approximating optimal instrument, $D\left(x_i\right)$,
- $D\left(x_i\right)$ minimizes the asymptotic variance (Amemiya 1974)
- Approximation error is relative to this $D\left(x_i\right)$

Lasso IV Model: Further Assumptions and Estimator

- Under two additional key assumptions
  - Strong optimal instrument, $D\left(x_i\right)$
  - Sparsity, i.e., a subset s of the p instruments are non-zero, and $s << n$
- And some additional technical conditions related to regularity, growth rates, and selection of the Lasso penalty parameter
- Let $\hat{D} = z_i'\Pi$, $\hat{Q} = \left[\hat{D}\hat{D}'\right]$, and $\hat{\varepsilon} = y - d\hat{\alpha}$, where $\hat{\alpha}$ is the causal parameter estimate

$$\hat{\alpha} = \left[\hat{D}d\right]^{-1}\left[\hat{D}y\right] \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\varepsilon}\hat{\varepsilon}' \tag{7}$$

$$\text{asymVar}\left(\hat{\alpha}\right) = \hat{\sigma}^2\hat{Q}^{-1} \tag{8}$$

Lasso IV Algorithm (one version)

- The algorithm as presented in Belloni et al (2012):
  1. Run post-Lasso with conservative penalty level and loadings
  2. Compute residuals to approximate the optimal penalty loadings
  3. Repeat (1)-(2) K times, but using the refined penalty levels in step (1)
  4. Are no instruments are selected or the approximation to the regularity condition is near singular?
     - Yes: switch to alternative approach
     - No: compute IV estimator according to (6)-(8).
- Note multiple packages implement Lasso IV, so you don't need to.
- These include R and Stata packages, we will introduce R one shortly.

Lasso IV Methods

- The algorithm tunes the penalty parameters of the lasso selection for causal inference, not prediction alone!
- The idea is to avoid two types of mistakes arising from the nuisance parameters in the first stage
  1. Including too many variables, leading to many weak instruments bias
  2. Including too few variables, potentially leading to inefficiency or weak instruments problem
- Will return to this later

Lasso IV: Extensions

- Can include multiple endogenous variables, i.e., d_i can be a vector
- Can include controls, i.e., model specification becomes

$$y_i = \alpha d_i + \beta w_i + \varepsilon_i, \tag{9}$$

$$d_i = z_i' \Pi + \lambda w_i + r_i + v_i \tag{10}$$

Double Machine Learning and Related Generalizations

- Can extend related ideas to other general problems:

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U, \quad & E[U|X,Z] = 0, \qquad (11) \\ D &= m_0(X) + V, \quad & E[V|X] = 0, \qquad (12) \end{aligned}$$

  where m_0 and g_0 are true functions, D exogenous conditional on X

- Double Machine Learning Approach (Chernozhukov et al 2017):
  1. Apply machine learning methods to estimate m and g
  2. Form moments by multiplying the two predicted residuals
- Two keys to the approach:
  1. Neyman orthogonality: derivative of moment function in terms of nuisance parameters vanishes at true parameters

     (avoids problems with two predicted residuals leading to bias)
  2. Sample splitting: Use with ML methods estimating m and g

Illustrations and Applications

1. Illustrate with simple one endogenous variable case developed from Lovett et al (2019)
2. Discuss full application with multiple endogenous variables in Lovett et al (2019)
3. Illustration of solving many weak instruments problem in more complex application of Gordon et al (2019)

Application 1: Simple Illustration of Adv on WOM

- Illustration is simplification of the setting in Lovett et al (2019)
- For illustration only, don't recommend this specification!
- Evaluating causal effect of a single endogenous variable
- Focal relationship is between TV advertising expenditures and total WOM
  - TV Adv: Measured as $\log(TVAdv + 1)$ from Kantar Media's Ad$pender product
  - WOM: Measured as $\log(WOM + 1)$ from Keller-Fay Groups TalkTrack via daily 24 hour self-reports
- Model includes brand fixed effects plus 115 controls
  - category-year dummies (category time effects)
  - cubic of month-in-year (seasonality controls)
- Have 46,722 month-brand observations

Simple Illustration: Instruments

- Base Instruments include
  - Cost per advertising unit (from Kantar Media)
  - Political advertising (as crowd out argument, also collected from Kantar Media)
- For political advertising, the crowd out argument suggests a nonlinear relationship (use a cubic function).
- Interact these with the category-year dummies
- Leads to 448 potential instruments
- Estimate OLS, 2SLS, and Lasso IV
- Lasso IV routine is rlassoIV from the hdm package in R

Simple Illustration: Results

Table: Example of post-Lasso IV for TV Advertising on WOM

			Model		
	OLS	IV-2	IV-4	IV-484	LassoIV-484
Instruments Included	None	Linear Cost & Political Ads	Lin. Cost & Cubic Pol. Ads	Full Set	Lasso Selected
TV Adv. Exp.	0.037 (0.001)	-0.087 (0.022)	-0.086 (0.022)	0.044 (0.007)	0.075 (0.036)
First Stage Matches Expected Sign		No	Only Pol.	Varies	Varies
First Stage F-Stat (for Instr. Matching Sign)		83.96 (NA)	44.18 (2.92)	2.92	31.51 (26.47)

Simple Illustration: Discussion

- OLS has small positive significant effect
- With two linear instruments (what people typically include)
  - First stage indicates costs and political ads increase TVAdv
  - Partial F-test appears strong, but exogeneity story fails
  - Large and significant negative effect of TVAdv on WOM
- Cubic for political ads does not help much
- With all of the instruments
  - First stage indicates some category-years match expectations
  - Partial F-test indicates ones matching sign are not relevant
  - Sign and magnitude are similar to OLS (positive)
  - Reflective of the many weak instruments bias
- The Lasso IV
  - Selects a small subset and most of these have the expected sign
  - Partial F-test (no longer strictly valid) indicates relevant instruments
  - Larger point estimate, but not significantly different from OLS

Application 2: Full Paid Media on Earned Media Analysis

- The full analysis in Lovett et al (2019) uses a more complex model
  - They study the effect of TV, Internet, and Other advertising on Total and Online WOM
  - that has many additional controls for brand heterogeneity, lagged effects, etc.
  - Main model makes conditional independence assumption,
- The main finding: advertising effects on WOM are small on average
- Check robustness of conditional independence assumption
  - Run IV and LassoIV methods
  - Base instruments are the same as illustration
  - Use brand interactions, not category-year for full set

Paid Media on Earned Media: Results

Table: Example of post-Lasso IV from Lovett et al (2019)

	OLS	Model IV-All*	LassoIV
TV Adv. Exp.	0.018	0.020	0.021
	(0.001)	(0.003)	(0.033)
Internet Adv. Exp.	0.016	0.026	0.017
	(0.002)	(0.004)	(0.043)
Other Adv. Exp.	0.014	0.009	0.000
	(0.002)	(0.004)	(0.031)

* Instruments are weak in the IV-All model.

Application 3: Disentangling Positive and Negative Political Advertising

- Gordon et al (2019) study the effect of positive and negative political advertising
- Different ad tones could affect turnout and relative candidate choice differently
- Incorporate own and opponent effects of each advertising variable
- Utility formulation becomes:

$$\begin{aligned} u_{ijm} &= \alpha_{P,Own} PAd_{jm} + \alpha_{N,Own} NAd_{jm} && (13) \\ &+ \alpha_{P,Opp} PAd_{j',m} + \alpha_{N,Opp} NAd_{j',m} && (14) \\ &+ \beta X_{jm} + \xi_{jm} + \varepsilon_{ijm} && (15) \end{aligned}$$

Positive and Negative Political Ads: Instruments

- Need instruments that disentangle the positive and negative advertising effects on both turnout and choice
- Cost instruments (Gordon and Hartmann) cannot as they are constant across tones
- Add Waldfogel instruments. Logic is
    - If decisions are exogenously fixed across a set of markets M,
    - then for market $m \in M$ use excluded demand shifters from markets m', where $m' \neq m$ and $m' \in M$ as instruments
- Include many interactions and functions of these instruments
- Interactions with incumbency status, which are known to shape negativity (Lovett and Shachar 2011)
- Leads to more than 1000 potential instruments
- A natural candidate for Lasso IV

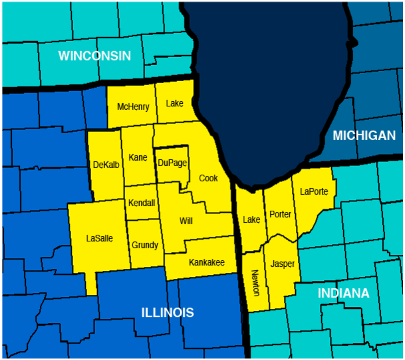


Figure: Illustration of Waldfogel Instruments

Positive and Negative Political Ads: Results Illustration

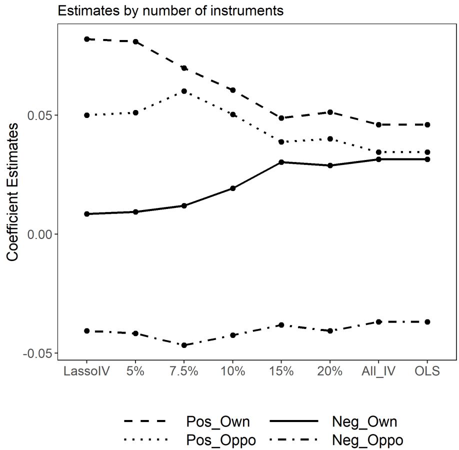


Figure: Point estimates of the parameters from the four endogenous variables under various models.

Applications Wrap-up

- Illustrated use in simple and more complex settings
  - How to obtain many instruments
  - How Lasso IV can improve efficiency and reduce bias over poorly specified linear instruments assumptions
  - How Lasso IV helps reduce the many weak instruments problem
- Now turn to example code with Monte Carlo data

Agenda

1. Introduction
2. Lasso IV
   - Introduction to Lasso IV
   - Model and Proposed Solution
   - Generalizations and Extensions
   - Illustration and Applications
   - Example Code and Monte Carlo Data
3. Synthetic Controls
   - Model and Methods
   - Generalizations to Multiple Treatment Units
   - Illustrations and Applications
   - Example Code and Monte Carlo Data
4. Conclusion

Introduction

- Synthetic controls is

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

- How do synthetic controls relate to other methods?
  - An advancement over the more commonly used differences-in-differences estimator (e.g., Card 1990)
  - Has similarities to matching estimators (Abadie and Cattaneo 2018)
  - But also something new that in most applications in marketing will rely on machine learning

- Basic idea:
  - Compare test case against a a synthesized control case, constructed as a weighted average of untreated “donor” cases

When Can you Apply Synthetic Controls?

When You Have. . .

1. Discrete causal variables or treatments (e.g., advertise or not)
2. Treatments are relatively rare/isolated in time
3. You observe (many) untreated cases in addition to the treated cases

Diff-in-Diff

- To explain Synthetic Controls, we begin with the simpler Differences-in-Differences (Diff-in-Diff)
- In Diff-in-Diff, you construct two differences
  1. before treatment versus after
  2. treatment group versus control group
- The effect is then measured by
  - The treatment compared to control during the post-treatment period
  - Alternatively and equivalently, the interaction effect from the post-period and treatment dummies

Diff-in-Diff Model

- The model for the treated, $y_{1,it}$ and untreated $y_{0,it}$ cases is

$$\begin{aligned} y_{1,it} &= d_{it}\tau_{it} + \mu_i + \delta_t + \varepsilon_{it}, \\ y_{0,it} &= \mu_i + \delta_t + \varepsilon_{it}, \end{aligned} \tag{16}$$

- $\tau_{it} = y_{1,it} - y_{0,it}$ is the treatment effect on the treated case i and averaging these gives the ATET or ATT
- unobserved confounders μ_i and δ_t are "differenced" away
- in practice, often assume $\tau_{it} = \tau$ and get ATE

Diff-in-Diff Unconfoundedness Assumption: Parallel Trends

- For Diff-in-Diff to work, you require an assumption of "parallel trends"
- This assumption requires that the control and treatment (or test) groups face the same time effects, δ_t
- Evaluate by examining pre-treatment time series to determine whether they are parallel

Diff-in-Diff Parallel Trends

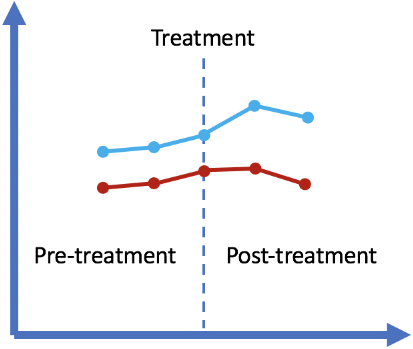


Figure: Diff-in-Diff Parallel Trends.

Diff-in-Diff Non-Parallel Trends

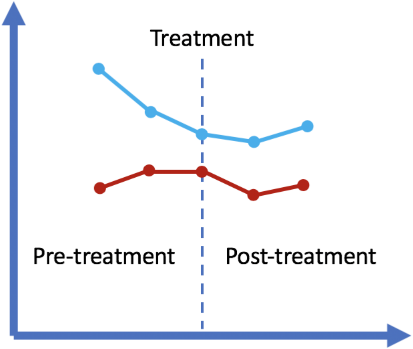


Figure: Diff-in-Diff Non-Parallel Trends.

Diff-in-Diff Real Example

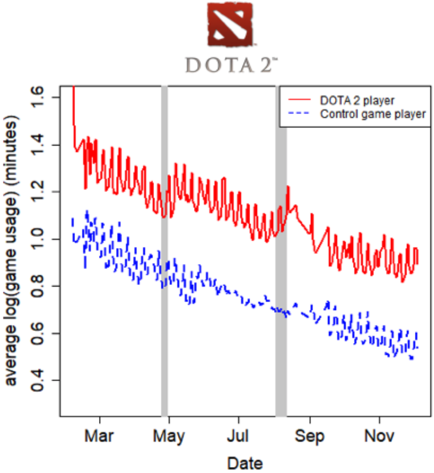


Figure: Diff-in-Diff Parallel Trends Evaluation.

Can We Do Better Than Diff-in-Diff?

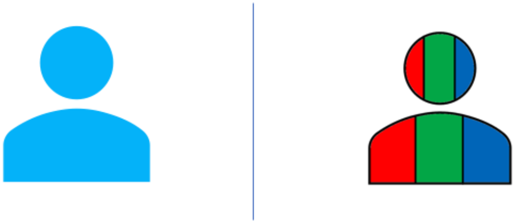


Figure: Diff-in-Diff Compares Combined Treated Cases against Combined Control

Can We Do Better? Yes We Can

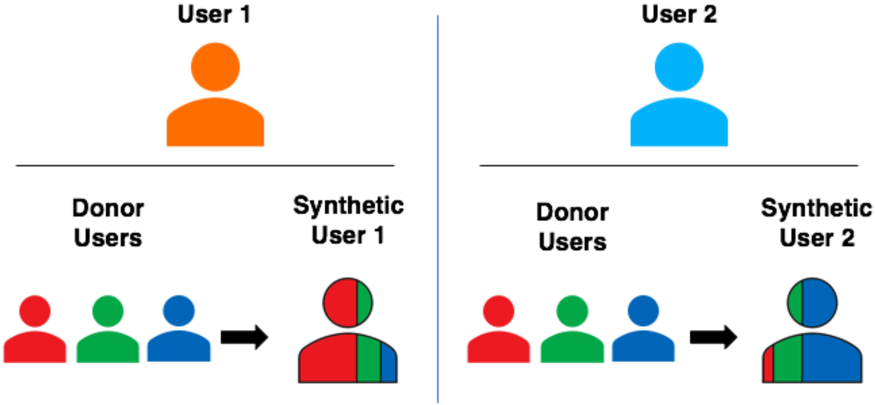


Figure: Synthetic Controls Compares Each Unit to a Synthesized Control to Match That Unit

Synthetic Controls Concept

- Synthetic controls views the control case as an unobserved counterfactual that is missing data
- The goal is to impute this missing data
- The method imputes a synthetic version from untreated cases
- But not just any cases, the synthetic control should be constructed
    - from a subset of the untreated cases
    - that look very similar to the treated case

Synthetic Controls Model

- The model is very similar to the Diff-in-Diff model but with one term different
- Instead of μ_i we have $\lambda_t \mu_i$
- The model becomes

$$\begin{aligned} y_{1,it} &= d_{it}\tau_{it} + x_i\theta_t + \delta_t + \lambda_t\mu_i + \varepsilon_{it}, \\ y_{0,it} &= x_i\theta_t + \delta_t + \lambda_t\mu_i + \varepsilon_{it} \end{aligned} \tag{17}$$

Synthetic Controls Flexibility

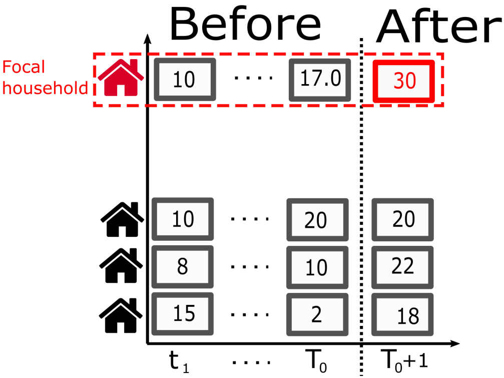
- What can $\lambda_t \mu_i$ accommodate that Diff-in-Diff cannot?
  - Seasonality that affects units/cases differently
  - Time trends that differ by units/cases
  - Time varying unobserved variables that affect cases differently

- Some interesting examples for marketing that $\lambda_t \mu_i$ cover
  - Brand advertising or promotions
  - Households baseline shopping habits
  - Household seasonal product use patterns
  - Consumer learning about product quality and match-value
  - Households with different aged children whose needs change as the children age

Synthetic Controls Is Unbiased

- Abadie et al (2010) proves that the synthetic controls estimate is unbiased under this structural model when
  1. Untreated cases are not affected by the treatment
  2. Number of time periods is large relative to the scale of transitory shocks
  3. Treated outcomes fall in the convex hull of the untreated cases
  4. The factor model component is well-behaved
- Most applications treat the parametric model as a local approximation, not necessarily the actual data generating process

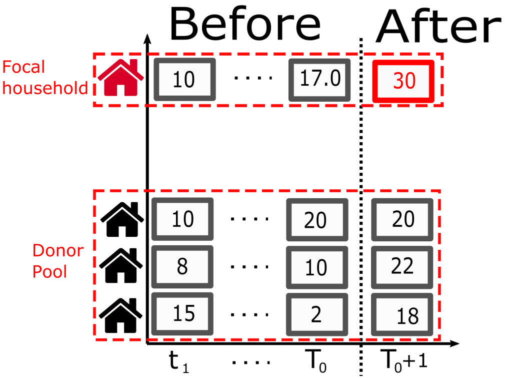
Construction of a synthetic control



	t_1	....	T_0	T_0+1
Focal household	10	....	17.0	30
	10	....	20	20
	8	....	10	22
	15	....	2	18

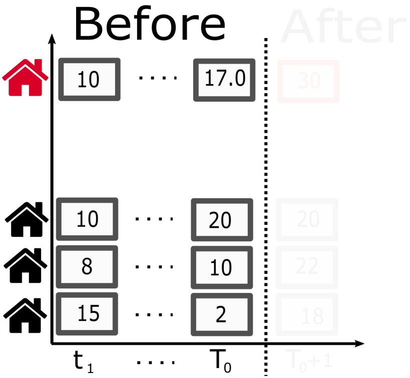
- **Goal**: Create synthetic control for focal household

Construction of a synthetic control



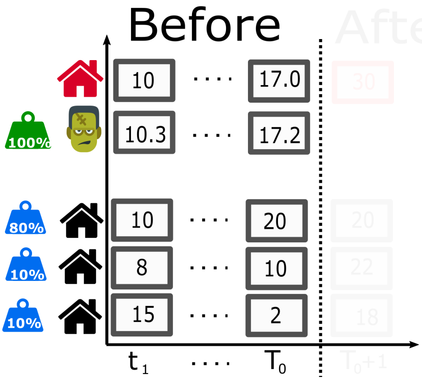
- **Goal**: Create synthetic control for focal household
- **How?** From untreated observations in donor pool

Construction of a synthetic control



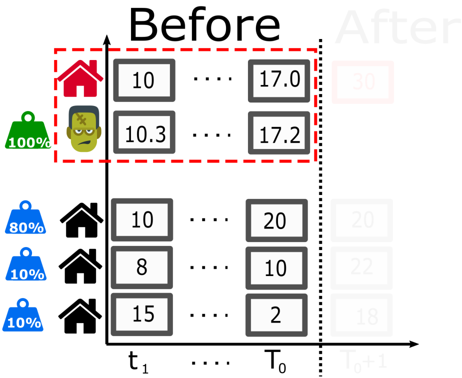
- **Goal**: Create synthetic control for focal household
- **How?** Use pre-treatment period variables

Construction of a synthetic control



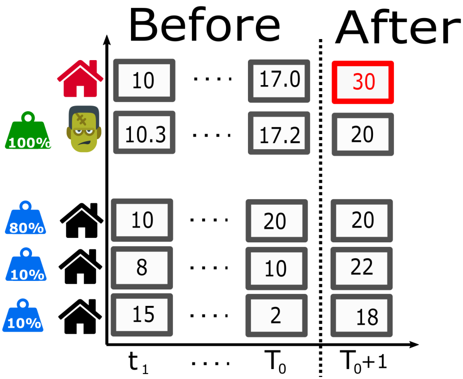
- **Goal**: Create synthetic control for focal household
- **How?** Assign weights to donors to construct weighted average of donors

Construction of a synthetic control



- **Goal**: Create synthetic control for focal household
- **How?** Assign weights to donors to construct weighted average of donors to match pre-treatment values of treated case

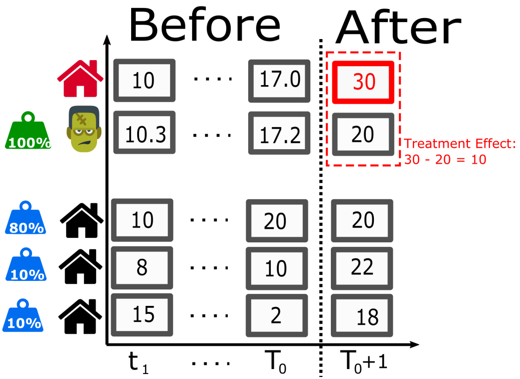
Construction of a synthetic control



	t_1	....	T_0	T_0+1
	10	....	17.0	30
100%	10.3	....	17.2	20
80%	10	....	20	20
10%	8	....	10	22
10%	15	....	2	18

- **Goal**: Create synthetic control for focal household
- **How?** Assign weights to donors to construct weighted average of donors to match pre-treatment values of treated case
- **Why?** To create the counterfactual synthetic control for the post-treatment value

Construction of a synthetic control



- **Goal**: Create synthetic control for focal household
- **How?** Assign weights to donors to construct weighted average synthetic case to match treated case
- **Why?** To measure the treatment effect

Synthetic Controls Methods

- Focusing estimating the treatment effect for a single treated case
- Let pre and post-treatment periods be $t = 1, \ldots, T_0$ and $t = T_0 + 1, \ldots, T$
- Let units be $i = 1, \ldots, N + 1$
- The single treated case is in position 1
- Focus is to predict $y_{0,1t}$ for $t \geq T_0$ had the treated case NOT been treated
- Estimate via

$$\hat{y}_{0,1t} = \sum_{i=2}^{N+1} w_i y_{0,it} \tag{18}$$

- and the ATET is estimated by

$$\hat{\tau}_{it} = y_{1,1t} - \hat{y}_{0,1t} = y_{1,1t} - \sum_{i=2}^{N+1} w_i y_{0,it} \tag{19}$$

How to select the weights, w_i?

- Goal: synthetic control exactly match treated case on a set of variables, z_i

$$z_i = (x_i, y_{i1}, y_{i2}, \ldots, y_{iT_0},)$$

- Let the matrix version of these variables be Z_1 and Z_0 for the treated and untreated (donor) cases, and W for the weights
- Objective is to obtain weights that minimize $Z_1 - Z_0 W$ for some distance measure

How to select the weights, w_i? (2)

- To choose W, Abadie et al (2010) suggest

$$||Z_1 - Z_0W||_V = \sqrt{(Z_1 - Z_0W)'V(Z_1 - Z_0W)} \qquad (20)$$

- where V is an semidefinite matrix of weights of same dimension as Z_1
- V places differential value on the variables in z_i
- V is especially important when using variables with different scales or noise levels
- Abadie et al (2010) suggest a sample splitting approach
    - Divide pre-treatment periods into training and test (prediction) period
    - the optimal $W^*(V)$ is chosen to minimize in (20) using the training period data
    - The optimal V is chosen using the prediction period data

Synthetic Controls Concern: Interpolation Bias

- As the number of donor cases increases relative to the number of time observations, you can be ensured a perfect fit
- In principle, we can match the pre-treatment treated cases with combinations of donor cases far from the treated cases
- But this could lead to interpolation bias, if the true relationship is non-linear
- Because of the local approximation idea, when constructing the Synthetic Control, we'd like to use donor cases that are individually similar to the treated case

Synthetic Controls Solution: Overfit Penalizer

- The solution is to penalize the objective for having donor cases being used to construct the synthetic control that are individually far from the treated case
- This penalization function is sometimes referred to as the overfit penalizer
- This function also regularizes the function
- For example, Vidal-Berastain (2019) obtains weights by solving

$$\min_{w^h} \left\{ \underbrace{\left\| X_h^{T_h^{pre}} - \sum_{s \in D} w_s X_s^{T_h^{pre}} \right\|^2}_{\text{Underfit penalizer}} + \overbrace{\lambda_h^\star}^{\text{hyper-parameter}} \times \underbrace{\sum_{s \in D} w_s \left\| X_h^{T_h^{pre}} - X_s^{T_h^{pre}} \right\|^2}_{\text{Overfit penalizer}} \right\}$$

subject to:

$$w_1 \geq 0, \;\; w_2 \geq 0, \;\; w_3 \geq 0, \;\; \cdots \;\; , w_{Num\ Donors} \geq 0$$
$$w_1 + \cdots + w_{Num\ Donors} = 1$$

Inference in Synthetic Controls

- Uses placebo test as foundation of inference
- Justified via randomization logic similar to Fisher exact test rather than sampling approach
- Evaluate effect size relative to randomly chosen untreated units
- Repeatedly sample random donor from untreated units and perform synthetic control estimation as if the control unit is the treated unit
- The treatment effect for the treated unit is then compared against these effects for the untreated units
- This allows calibration of uncertainty for each treated unit

Generalizations to Multiple Treatment Units

- The model and methods so far are for a *single* treated unit
- Often we want to speak about many such units
- One option: apply synthetic controls approach to each unit and then aggregate
- Some other (better) recent options

Synthetic Controls Methods for Multiple Treatment Units

1. Generalized synthetic controls (Xu 2017)
2. Unbalanced synthetic controls (Vidal-Berastain 2019)
3. Generalized raking procedures (Robbins et al 2017)
4. Matrix completion methods (Athey et al 2017)

Illustrations and Applications

1. Generalized synthetic controls applied to the effect of being a superbowl advertiser on word-of-mouth (Lovett et al 2019)
2. Unbalanced synthetic controls applied to the effect of eSport competitions on video game usage (Li et al 2018)
3. Unbalanced synthetic controls applied to the effect of retail platform entry on consumer shopping habits (Vidal-Berastain et al 2018)

Application 1: Superbowl Advertising and Word-of-Mouth

- Lovett et al (2019) use generalized synthetic controls
  - to estimate impact of being a superbowl advertiser
  - on word-of-mouth (WOM), both online and offline
  - considering different periodicities of data (daily, weekly, monthly)
- As illustration, we consider the daily online analysis
- Data
  - Online WOM come from Nielsen-McKinsey Insight tool
  - Collects public social media posts that include the brand name (and related variants)
  - Over 500 brands across many industries/categories
- Analysis
  - Use 60 days of pre-treatment posts
  - Evaluate effect for 31 days on/after Superbowl
  - Apply *gsynth* R routine for estimation

Results: Superbowl Advertising and Word-of-Mouth

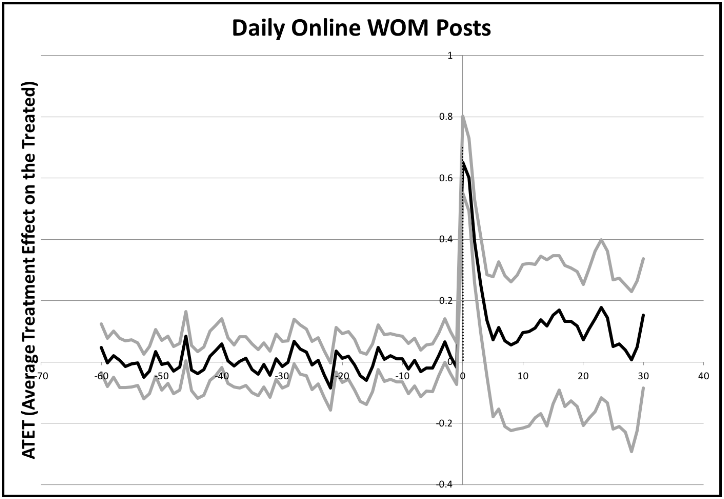


Figure: Treatment Effect of Advertising on Superbowl on Online WOM

Application 2: eSport Competitions on Video Game Usage

- Li et al (2019) examine the effects of sharing and extending intellectual properties related to video games on game usage
- They consider eSport competitions and YouTube video posts about games
- As illustration, we focus on their use of synthetic controls to estimate the effect of eSport competitions on video game usage

eSport Competitions on Video Game Usage

- Data
  - 131,000 users on Steam, the largest digital gaming platform for PC games
  - Collects usage of games on a daily basis
  - Have 27,000 treated users across Counter Strike: Go and DOTA2
  - Have 57,000 potential donor users
- Analysis
  - Compare Diff-in-Diff and Synthetic Controls
  - For Synthetic controls, include only the closest 100 cases
  - Evaluate effects for during, the week after, and the month after the event
  - Due to size of problem, R routines gsynth and synth do not work well
  - Instead apply Vidal-Berastain (2019) approach using a computing cluster

Results: eSport Competitions on Video Game Usage

	Dependent Variable & Model	
	DOTA Diff-in-Diff	DOTA Synthetic Control
During Event	0.057***	0.238***
	(0.010)	(0.019)
Short run (1 wk)	0.131***	0.372***
	(0.014)	(0.020)
Long run (1wk - 1 mo)	0.025**	0.296***
	(0.012)	(0.020)
Synthetic Control	No	Yes
Weekday*Treated	Yes	Yes
Indiv F.E.	Yes	No
Date FE	Yes	Yes

Table: Diff-in-Diff and Synthetic Controls Estimators

Application 3: Retail Platform Entry on Shopping

- Drawn from Vidal Berastain et al (2018)
- Estimates causal effect of geographic entry of retail platforms on household shopping behaviors
- Data
  - Considers 6 retail platforms and thousands of store entries
  - Examines 70,000+ households
  - Evaluates impact on a broad range of shopping behaviors including breadth and frequency of shopping, stockpiling, and private label use
- Methods
  - Use unbalanced synthetic controls (Vidal-Berastain 2019)
  - Innovates on calibration of synthetic control to better balance overfitting and underfitting
  - Applies an ML optimization approach to synthetic controls
- For illustration, we focus on retailers visiting, shopping trips, and total expenditures

Results: Ambiance and unique assortments

- Ambiance & unique assortment: ↑ retailers, trips, spending ⇒ expand the pie.
- Managerial insight: Sharing customers

		Retailers visited	Shopping trips	Number of items	Total expenditure	PL expenditure	PL items
LAR 1	♡	2.1% ↑	4.7% ↑	0.0%	0.0% ↑	3.9%	5.7%
LAR 2	♡	7.7% ↑	3.8% ↑	0.7%	0.9% ↑	2.7%	3.2%
ORG	♡	2.0% ↑	19.7% ↑	1.2%	0.4% ↑	0.9%	0.0%
Sup		-4.1%	2.1%	-0.0%	0.1%	8.7%	6.2%
Club	♡	0.2% ↑	-3.0% ↑	-1.7%	0.3% ↑	2.1%	0.87%
Disc		0.5%	1.7%	0.9%	0.2%	0.72%	0.42%

Example Code

Agenda

1. Introduction
2. Lasso IV
   - Introduction to Lasso IV
   - Model and Proposed Solution
   - Generalizations and Extensions
   - Illustration and Applications
   - Example Code and Monte Carlo Data
3. Synthetic Controls
   - Model and Methods
   - Generalizations to Multiple Treatment Units
   - Illustrations and Applications
   - Example Code and Monte Carlo Data
4. Conclusion

Conclusion: Casual Analytics Methods

1. Lasso IV (and double-machine learning)

Tools to obtain causal inference through instrumental variables (Pearl)

2. Synthetic controls (and matrix completion)

Tools to obtain causal inference similar to diff-in-diff approach (Rubin)

- These methods leverage machine learning for causal inference
- We covered
  - Methods and Models
  - Marketing Applications
  - Simple Code Examples
- These methods illustrate broader concepts in causal analytics
  - Lasso IV: similar to double machine learning, Deep IV, etc.
  - Synthetic Controls: Matrix completion and other imputation methods

Conclusion: When to Use the Methods

- When want a prediction about the effect of an action on an outcome
- When cost is too high (or impossible) to run an experiment

Causal Analytics: Lasso IV

When some exogenous instruments are available and the focal variable is continuous or discrete

Causal Analytics: Synthetic Controls

When the focal variable is discrete, separated in time, and have untreated cases

Causal Analytics

Happy to answer further questions

The Future (of Causal Analytics) is Now In Your Hands!

Machine Learning for Causal Inference

Mitchell J. Lovett[3]

August 10, 2019

[3] The slides and related chapter benefited greatly from assistance by Nan Li, Bowen Luo, and Xavi Vidal-Berastain.