Machine Learning for policy evaluation: supervised learning

Jérémy Do Nascimento Miguel*

*BSE, Univ. Bordeaux, jdnmiguel@u-bordeaux.fr

Master APP - EADD; Univ. Bordeaux - Fall 2023

Overview

Economic applications

Machine Learning for prediction

Dealing with high-dimensionality in Causal inference

- High dimensional methods and inference
- Model selection when the goal is inference
- PDS-LASSO
- Double/Debiased Machine Learning

Unpack the treatment effect heterogeneity

- Honest estimation
- Causal trees and forests
- Generic ML inference

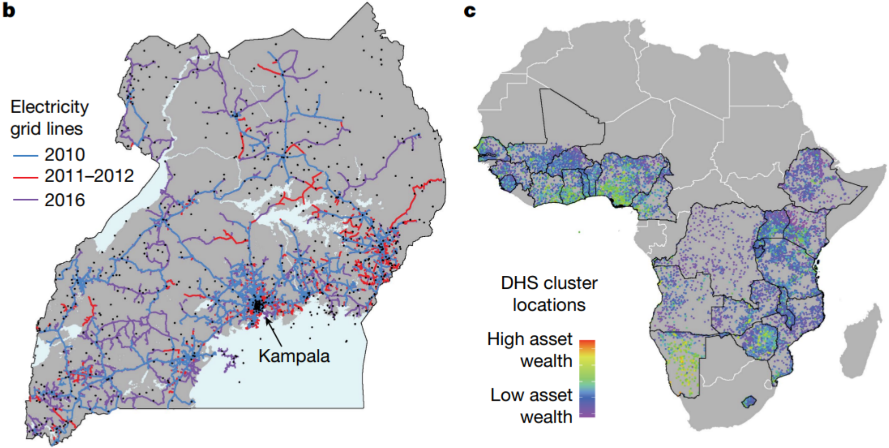
Empirical application

Assess the impact of electricity access

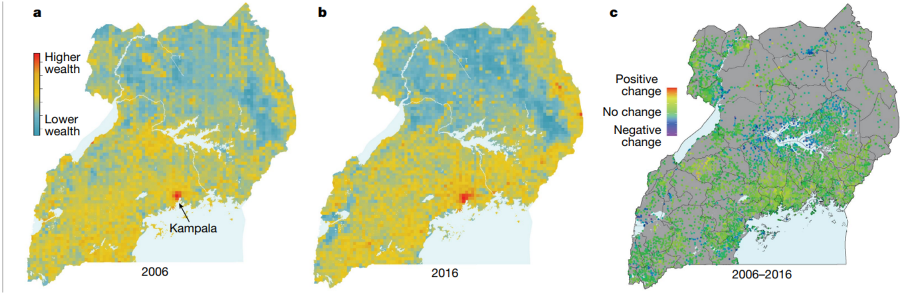
Ratledge et al. (2022):use satellite imagery and machine learning to estimate causal effect of electrical grid expansion on livelihood in rural Uganda

- Digitize electric distribution map from 2005 to 2018 and combine them with satellite-based prediction of local level asset wealth
- Trained a machine learning model to predict household asset from DHS across 25 SSA countries during 13 years (641,621 HH)

Electricity lines and wealth distribution



Wealth prediction



Causal impact

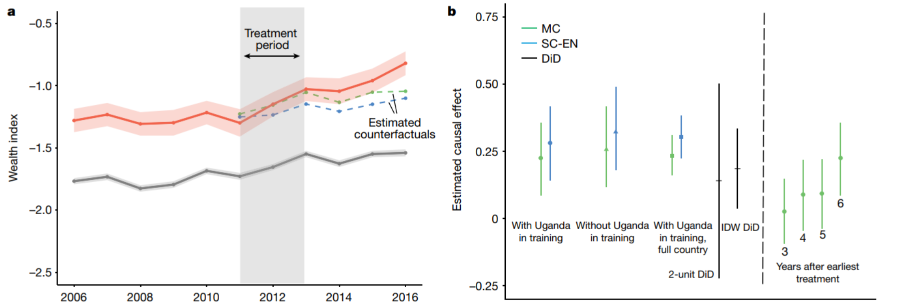


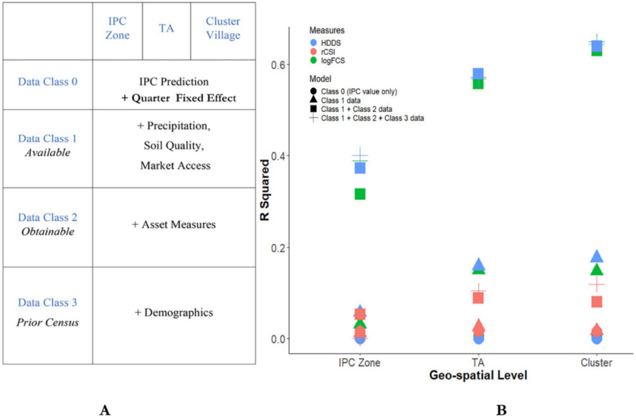
Fig. 3 | Electricity access increases household wealth in electrified communities compared with unelectrified communities. a, Solid lines show average asset wealth in untreated (grey) and treated (red) locations and shaded bands represent the standard error of the mean. Untreated counterfactuals for the treated group, as predicted by MC (green) and SC-EN (blue), are shown by dotted lines. **b**, Estimated causal effect of electrification on wealth by the end of the sample (2016). Error bars represent 95% CIs. CIs for each of the ML estimates are based on 500 bootstrapped model runs. We find similar statistically significant positive impacts in each of our three ML-based estimates (with Uganda, without Uganda and full country). '2-unit' represents a repeated cross-section DiD run using only DHS estimates. 'IDW' reflects DiD results from an inverse distance weighting approach. The four numbered lines on the right represent the causal estimates in the third to sixth years (2013–2016) after the initial treatment year, with CIs based on 100 bootstraps.

Predict food insecurity crisis

Lentz et al. (2019) develop a near-real time model to forecast food security using market, remote sensing, and household data.

- Class 0: Integrated Food Security Phase Classification System (IPC)
- Class 1= readily available data
  - Remotely collected and widely available, LSMS in certain case
  - Precipitation, market prices, soil quality, geographic variables
- Class 2: likely to be available but require additional work to be accessed and processed
  - Household roof type, cellphone ownership
- Class 3: infrequently gathered but publicly available household-level data including demographics and assets
  - large census, DHS, LSMS
  - demographic data,

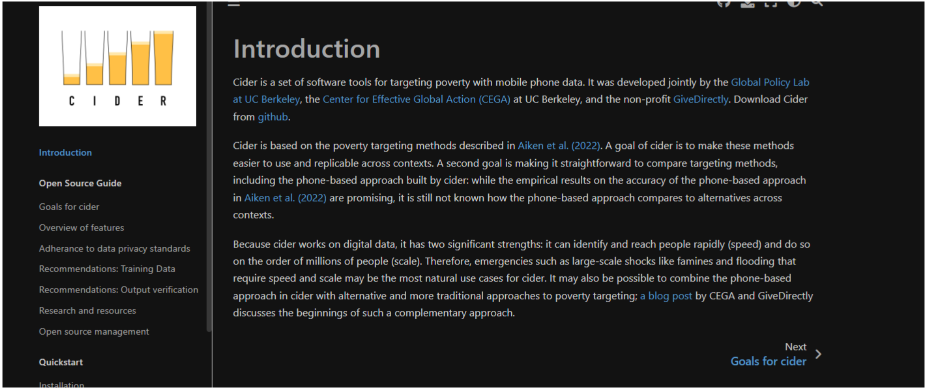
Results



	IPC Zone	TA	Cluster Village
Data Class 0	IPC Prediction + Quarter Fixed Effect		
Data Class 1 *Available*	+ Precipitation, Soil Quality, Market Access		
Data Class 2 *Obtainable*	+ Asset Measures		
Data Class 3 *Prior Census*	+ Demographics		

Fig. 2. The share of variation in out-of-sample cluster-level food security predicted by our models improves with greater spatial granularity and richer data. (A) We predict food security outcomes using three levels of spatial granularity and 4 classes of models. Class 0 data include the IPC early warning value only. Class 1 data contains: past IPC values, precipitation, market prices, market access measures, and soil quality. Class 2 data contains: share of households owing cellular phone and share of dwellings with metal versus thatch roof. Class 3 data contains: household demographics and assets. (B) The explanatory power of the models, measured as R-squared, increases with the Classes of data and the spatial granularity of data used. Our best model explains 64% of the variation in cluster-averages of household dietary diversity (HDDS) and 65% of logged Food Consumption Scores (logFCS). However, at the most spatially disaggregated level, the cluster level, additional household and demographic variables add little additional information.

Cider: creating poverty predictions based on non-traditional data sources



Introduction

Open Source Guide

Goals for cider

Overview of features

Adherance to data privacy standards

Recommendations: Training Data

Recommendations: Output verification

Research and resources

Open source management

Quickstart

Installation

Introduction

Cider is a set of software tools for targeting poverty with mobile phone data. It was developed jointly by the Global Policy Lab at UC Berkeley, the Center for Effective Global Action (CEGA) at UC Berkeley, and the non-profit GiveDirectly. Download Cider from github.

Cider is based on the poverty targeting methods described in Aiken et al. (2022). A goal of cider is to make these methods easier to use and replicable across contexts. A second goal is making it straightforward to compare targeting methods, including the phone-based approach built by cider: while the empirical results on the accuracy of the phone-based approach in Aiken et al. (2022) are promising, it is still not known how the phone-based approach compares to alternatives across contexts.

Because cider works on digital data, it has two significant strengths: it can identify and reach people rapidly (speed) and do so on the order of millions of people (scale). Therefore, emergencies such as large-scale shocks like famines and flooding that require speed and scale may be the most natural use cases for cider. It may also be possible to combine the phone-based approach in cider with alternative and more traditional approaches to poverty targeting; a blog post by CEGA and GiveDirectly discusses the beginnings of such a complementary approach.

Next
Goals for cider ›

Overview

Economic applications

Machine Learning for prediction

Dealing with high-dimensionality in Causal inference

- High dimensional methods and inference
- Model selection when the goal is inference
- PDS-LASSO
- Double/Debiased Machine Learning

Unpack the treatment effect heterogeneity

- Honest estimation
- Causal trees and forests
- Generic ML inference

Empirical application

Why do we want to do prediction?

Reliable and quantitative data are essential for economic policy, but scarce in developing countries

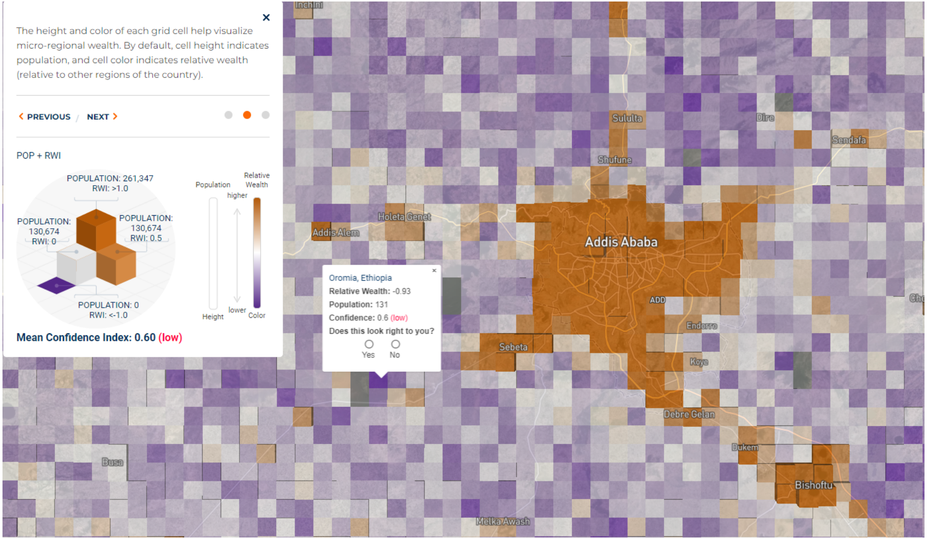
- Censuses and household surveys are rare
- Spatial disaggregated data often do not exist
- Leverage new sources of data to predict poverty and wealth, aid targeting, or infrastructure access
- See Aiken et al. (2022), Blumenstock et al. (2015), Oshri et al. (2018)

Predicting the main determinants of a given outcome of interests

- Predicting successful entrepreneurs? (McKenzie and Sansone, 2019)
- Price wheat farmers get on market day (Do Nascimento Miguel, 2022)

While useful, ML is only a tool and should not be perceived as the ideal solution for all our questions

Microestimates of wealth for all low-and middle-income countries (Aiken et al., 2022)



The height and color of each grid cell help visualize micro-regional wealth. By default, cell height indicates population, and cell color indicates relative wealth (relative to other regions of the country).

POP + RWI

Mean Confidence Index: 0.60 (low)

Oromia, Ethiopia

Relative Wealth: -0.93

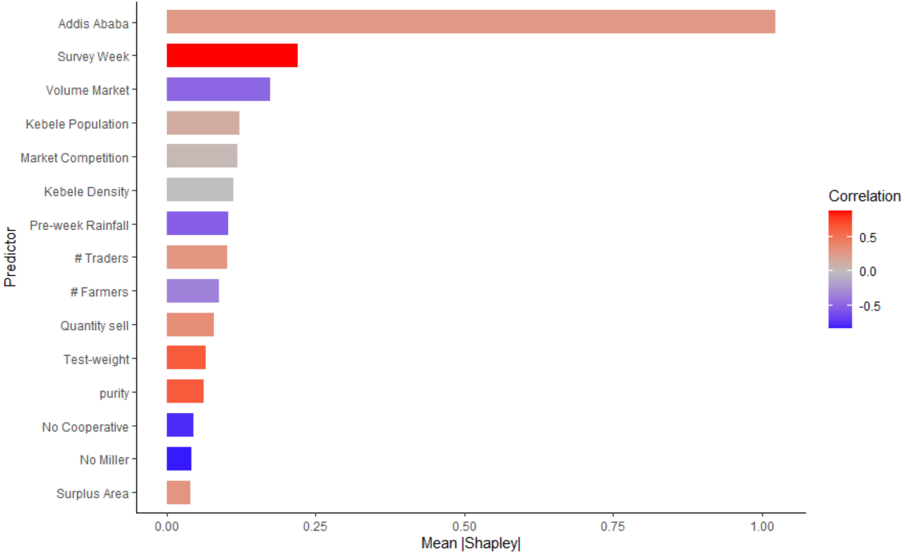
Population: 131

Confidence: 0.6 (low)

Does this look right to you?

○ Yes ○ No

Correlation between covariates and predicted what price (Do Nascimento Miguel, 2022)



Overview

Economic applications

Machine Learning for prediction

Dealing with high-dimensionality in Causal inference

- High dimensional methods and inference
- Model selection when the goal is inference
- PDS-LASSO
- Double/Debiased Machine Learning

Unpack the treatment effect heterogeneity

- Honest estimation
- Causal trees and forests
- Generic ML inference

Empirical application

Traditional way to do causal inference

$$Y_i = \delta D_i + \beta X_i + \epsilon_i$$

1. Regress Y_i on X_i' compute the residuals

$$\tilde{Y}_i = Y_i - \hat{Y}_i^{OLS}$$
$$\hat{Y}_i^{OLS} = X_i'(X'X)^{-1}X'Y$$

2. Regress D_i on X_i compute the residuals

$$\tilde{D}_i = D_i - \hat{D}_i^{OLS}$$
$$\hat{D}_i^{DML} = X_i'(X'X)^{-1}X'D$$

3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$

OLS might not be the right way to go if (i) **high dimensionality** ; (ii) non-linear relationship between X and D or Y

Curse of dimensionality in causal inference

High-dimensional data arise through a combination of 2 phenomena

1. High-dimensional settings: many characteristics per observation
2. Rarely know the exact functional form with which a variable enter the model

We have seen in previous lectures few way to deal with this issue when we are interested in prediction model.

- Perform pretty well in doing accurate prediction
- **BUT** perform poorly when we want to make inference about model parameters

Solution?

The magic weapon: regularization

Why do we need regularization? Let consider when we have as many observations as variables. Problem?

The magic weapon: regularization

Why do we need regularization? Let consider when we have as many observations as variables.

- OLS will fit the data perfectly: R^2=1
- Poor out-of-sample prediction, why?

Approximately Sparse regression models

Many potential predictor/control variables of which only a few are important for predicting Y

- Challenge: good out-of-sample prediction of Y (and/or treatment)
- We do not ex-ante which variables are important predictors

$$y_i = g(w_i) + \epsilon_i E[\epsilon_i | z_i,] = 0 \tag{1}$$

Regularize $g(w_i)$ to avoid over-fitting. Let consider $g(w_i)$ as a high-dimensional approximately linear model:

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{i,j} + \rho_{p,i} \tag{2}$$

Main assumption: only a subsample of variables have β_j different from 0.
Does it ring a bell?

Return of the Lasso

Remember the drawbacks?

How do we achieve inference?

Difficulties in drawing inferences after model selection

- Designed for prediction
- Model selection mistakes may occur: **OVB** contaminates estimaton

Intuition to overcome these biases:

- Focus on few variables for which no model selection will be done
- Model selection is done only over "nuisance" part (i.e., other covariates)
- Estimation for main variables is done using equation equations orthogonalized from nuisance

Inference with many instruments

$$y_i = \alpha d_i + \epsilon_i \tag{3}$$

$$d_i = z_i^{'}\Pi + r_i + v_i \tag{4}$$

where $E[\epsilon_i v_i]$ is not 0 (i.e., endogeneity)
Solution to estimation and inference about α

- Variable selection limited at to the first stage: select a small number of IV from z_i. Pure predictive problem

- Estimate the second-stage

Inference with selection among many controls

Consider a linear model where a treatment variable, d_i, is exogenous

$$y_i = \alpha d_i + x_i^{'}\theta + r_{yi} + \eta_i \tag{5}$$

α: Treatment effect on the outcome .

Naive approaches

1. Run Lasso while forcing d_i to remain in the OLS:
   - Problem: Lasso targets prediction and will drop any variables highly correlated to the treatment. **OVB risk**
   - Omit relationship between treatment variable and controls: can model this reduced form

2. Use only eq. from previous slide or the reduced form
   - Assume: no errors in variable selection
   - Issue: tend to pick variables with larger value correlated with y_i and miss variables witch have large correlation with d_i

Post-double selection LASSO

Belloni et al. (2014)'s paper. They consider a regression model with treatment indicator and control variables on the RHS:

- Remember previous lecture, we hare in a case with large number of *potential* controls $(x_1, \ldots, x_p)$. Potentially in a case were $p >> n$
- Objective is to select a small sub-sample of variables that matters, s

They propose a "double selection procedure" involving 3 steps:

1. Select covariates for outcome y_i using LASSO (first LASSO)
2. Select covariates for treatment status D_i using LASSO (second LASSO)
3. Use covariates relevant in one of the cases as control variables

PDS-LASSO: formal framework

$$y_i = D_i\alpha_0 + g(z_i) + \epsilon_i \quad E[\epsilon_i|z_i, D_i] = 0 \tag{6a}$$

$$D_i = m(z_i) + \eta_i \quad E[\eta_i|z_i] = 0 \tag{6b}$$

y_i is the outcome, D_i the treatment variable, z_i represents confounding factors, which affect D_i via $m(z_i)$ and y_i via $g(z_i)$.

However, form of $g()$ and the identity of z are unknown. Therefore we approximate (1a) and (1b) using a linear combinations of all potential control variables (x_i):

$$y_i = D_i\alpha_0 + x_i'\beta_{g0} + r_{gi} + \epsilon_i \tag{7a}$$

$$D_i = x_i'\beta_{m0} + r_{mi} + \eta_i \tag{7b}$$

$x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ are approximations to $g(z_i)$ and $m(z_i)$; r are the corresponding approximation errors.

PDS-LASSO

Having many controls creates a challenge for estimation and inference. A key condition allowing us to perform estimation and inference is sparsity:

- Sparsity: exist approximations $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ to $g(z_i)$ and $m(z_i)$ that require only a small number of non-zero coeff. to make r small relative to estimation error

Relies on two conditions:

1. At most $s << n$ elements of $\beta_m 0$ and $\beta_g 0$ are non zero

2. The resulting approximation errors are small compared to the estimation error

$\Rightarrow$ Relying on sparsity, they implement a selection method to select approximately the right set of controls and then estimate the treatment effect α_0

PDS Lasso: The recipe

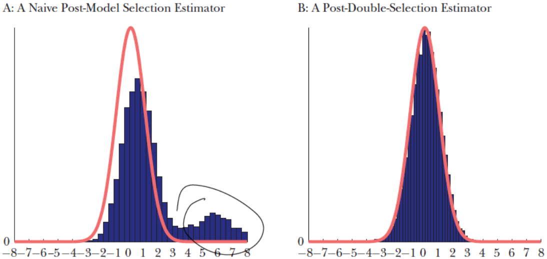
Practically, PDS is implemented in three steps:

1. Lasso Y_i on X_i, collect selected controls in X_i^Y
2. Lasso D_i on X_i, collect selected controls in X_i^D
3. Regress Y_i on D_i and $X_i^Y \cup X_i^D$

Caveats and considerations:

- Standardizing controls pre-lasso is important
- Cross-validation works fine for selection λ
- Inference: Just use robust standard errors from step 3

Does it work?



A: A Naive Post-Model Selection Estimator

B: A Post-Double-Selection Estimator

Source: Belloni, Chernozhukov, and Hansen (forthcoming).
Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

DML: the basics

$$Y_i = \tau D_i + g(X_i) + \epsilon_i$$

1. Predict Y_i using X_i with ML and compute the residuals

$$\tilde{Y}_i = Y_i - \hat{Y}_i^{DML}$$
$$\hat{Y}_i^{DML} = \text{prediction generated by ML}$$

2. Predict D_i using X_i with ML and compute the residuals

$$\tilde{D}_i = D_i - \hat{D}_i^{DML}$$
$$\hat{D}_i^{DML} = \text{prediction generated by ML}$$

3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$

$\tilde{Y}_i$ and $\tilde{D}_i$ should be generated by ML algorithms trained on a set of observation **that does not contain** i. Use cross-fitting to estimate τ.

DML: Recipe

1. Divide the sample into K folds

2. For $k = 1, \ldots, K$
   - Train a model to predict y given x, leaving out obs. i in fold k: $\hat{Y}^{-k}(x)$
   - Train a model to predict d given x, leaving out obs. i in fold k: $\hat{D}^{-k}(x)$

3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$

Do not forget to CV to choose tuning parameters. For inference consideration we rely on standard errors from last step.

Overview

Economic applications

Machine Learning for prediction

Dealing with high-dimensionality in Causal inference

- High dimensional methods and inference
- Model selection when the goal is inference
- PDS-LASSO
- Double/Debiased Machine Learning

Unpack the treatment effect heterogeneity

- Honest estimation
- Causal trees and forests
- Generic ML inference

Empirical application

Basic causal inference summary

Target is to estimate the average treatment

$$ATE = E[Y_i(1) - Y_i(0)] = E[\tau_i]$$

The key identifying assumption:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | X_i$$

Heterogeneous Treatment Effect

Uncovering treatment effect heterogeneity with respect to some covariates

- Individuals benefit most from a treatment?
- How do treatment effect change with covariates?

Estimate the overall average effect:

$$Y_i = \tau D_i + \epsilon_i$$

Explore heterogeneity by gender:

$$Y_i = \tau^{male} D_i + \beta \text{Female}_i + \gamma D_i \times \text{Female}_i + \epsilon_i$$

More generally,

$$\begin{aligned} Y_i =& \tau D_i + \beta X_i' + \gamma D_i \times X_i' + \epsilon_i \\ \tau(x) =& \tau + x'\gamma \end{aligned}$$

Conditional average treatment effect (CATE): $\tau(x) = E[\tau_i | X_i = x]$

Challenge with traditional heterogeneity analaysis

$$Y_i = \tau D_i + \beta X_i' + \gamma D_i \times X_i' + \epsilon_i$$

- Functional form: treatment effects may not vary linearly with X_i
- Curse of dimensionality
- Luckily we can rely on ML to solve these issues

Heterogeneous treatment effects in randomized experiments

In most of our analysis, we want to know how treatment effects vary with covariates to explore mechanisms or predict the effect in a specific subpopulation

Problem? We often have many covariates and can find significant effects by "chance" = cherry picking

- Solution 1: register a pre-analysis plan (PAP). But that is partially satisfactory, we are throwing away lots of data we are collecting
- Solution 2: use machine learning to guide prediction as complement to theory

My advice would be to combine both: identify sub-categories of interest that make sense from a theoretical standpoint in a PAP, and dig deeper with ML

Honest estimation

Problem: do not know how to make inference given the output of a tree (i.e., no CI)

A simple solution: "Honest" estimation via sample splitting:

- Split sample into training and test sample
- Build tree on the training samples using CV
- Use tree grown to regress y on region or leaf dummies in testing sample
- Perform inference as usual

Causal trees (Athey and Imbens (2016))

Key element of this paper: splitting involves picking something "unusual" relative to the data generating process. This causes bais.

- Estimation of means: can split on using a cutoff. For instance if $\bar{Y_1} - \bar{Y_0} > c$
- On average each one is consistent. BUT if split only when the difference is large, you will ended up with bias subset compared to the population

Their inution is to follow a honnest approach:

- Decide where to split using the training sample
- Calculate means and assess accuracy using testing sample

Causal trees

Objective: $\tau(\hat{X}) = E[Y(1) - Y(0)|X]$

- Reveals heterogeneity: different treatment effects in different leaves
- Caution: any heterogeneity discovered here is not causal, but can test its significance using an honest estimation

Steps to grow a causal tree:

1. Split sample into training and testing sample of roughly same size
2. Build tree on training sample: grow the tree until stopping criterion is reached, prune the tree, do CV
3. Using OLS in testing sample, regress Y on leaf dummies based on the tree built interacted with treatment indicator

Drawbacks:

- Only half of the data is used: loss of efficiency $\Rightarrow$ increase in MSE
- Can't do cross-fitting
- Heterogeneity often very much depends on the random split

Causal Forests

Same intuition as in the previous lecture: tree has high variance, so we can take the avg. of many trees to plan a causal forest

1. Draw a sub-sample without replacement n_b
2. Split n_b into training and testing sample of roughly same size
3. Grow tree to minimize parameters of interest on training sample until reaching stopping criterion
4. Compute $\hat{\tau}(x)$ on testing sample: $\hat{\tau}_b(x)$
5. Repeat B times and compute $\hat{\tau}_{CF}(x) = \dfrac{1}{B}\sum_{b=1}^{B} \hat{\tau}_b(x)$

Causal Forests: pros and cons

- Pros: useful for targeting because we treat individuals with marge $\hat{\tau}$ given value of a specific covariate

- Cons: loosing in interepretability because we are estimating a different parameter, $\tau(x) = E[(Y(1) - Y(0)|X = x]$

I recommend to read Wager and Athey (2018) and Athey and Imbens (2019) for further details about causal forests. Personally it is not my favourite one to go.

GATES and CLAN (Chernozhukov et al., 2020)

Problem in causal tree approaches: inference does not account for uncertainty in covariates binning

- Why? Rely on split to choose the bin, then CATE within those bins work roughly as well as regression approach
- But fails to consider bins variation across samples

Generic ML inference: the set up

$Y(1)$ and $Y(0)$ are the potential outcomes in the treatment and control group. Z is a vector of covariates. We assume that:

$$Y(1), Y(0) \perp D|Z$$

The observed outcome is $Y = DY(1) + (1-D)Y(0)$ which given D, Z can be rewritten such as:

$$Y = b_0(Z) + Ds_0(Z) + U, \quad E[U|Z, D] = 0$$

The main causal functions are the baseline conditional average:

$$b_0(Z) = E[Y(0)|Z]$$

and the conditional average treatment effect:

$$s_0(Z) = E[Y_1 - Y_0|Z]$$

We are interested in $s_0(Z)$, the CATE, but **we cannot observe it**

The problem and the solution

We do not know how to do uniformly valid confidence based on z. Therefore, Chernozhukov et al. (2017) propose to estimate and make inference on *key features* of $s_0(Z)$ rather than $s_0(Z)$.

- Identify how much heterogeneity there is in the estimates and then find out those groups characteristics with HE
- Randomly split sample into main and auxiliairy sample
- Apply ML in $Data_A$ to estimate control mean, B(Z), and treatment effect, S(Z)
- Use test sample to predict actual values

Key features of $s_0(Z)$

1. Best Linear Predictor of $s_0(Z)$ given $S(Z)$
2. Sorted Group Average Treatment Effects (GATES)
3. Classification Analysis (CLAN)

Best Linear Predictor (BLP)

The BLP of $s_0(Z)$ given S(Z) is given by:

$$BLP(s_0(Z)|S(Z)] = \beta_1 + \beta_2(S(Z) - ES(Z))$$

but we can’t run the regression of $s_0(Z)$ on S(Z), why?

Best Linear Predictor

Instead, we can run the following model in $Data_A$

$$Y = \alpha' X_1 + \beta_1 (D - p(Z)) + \beta_2 (D - p(Z)(S(Z) - ES) + \epsilon$$

- X_1 is a vector of ones, $B(Z)$, and and variables used to improve precision (e.g., district FE)
- $p(Z)$ is the probability of treatment conditional on covariates

We obtain $\beta*_1$ (ATE) and $\beta*_2$, the treatment effect heterogeneity on the proxy predictor (HET)

Then, we can estimate the same model using $Data_M$ and test for:

$$H0 : \beta_2 = 0$$

$H0$ corresponds to the combined hypothesis that:

1. No treatment effect heterogeneity based on Z ($s_0(Z) = s$)
2. or $S(Z)$ is completely noise, uncorrelated to $s_0(Z)$

Moving on from BLP

If we reject $H0 : \beta_2 = 0$ from the BLP, we have identified heterogeneity. Now we would like to know which covariates drive it. Main issue: high-dimensional settings

Sorted Group Average Treatment Effects (GATES)

Can use the covariates in Z to **GROUP** individuals together into k groups according to the ATE size $S(Z)$:

$$\gamma_k = E[s_0(Z)|G_k], \qquad k = 1, \ldots, K$$

Practically, we estimate the following specification:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k (D - p(Z)) G_k + \epsilon$$

with G_k is a dummy variable for being in group k and test for

$$H_0 : \gamma_k - \gamma_1 = 0$$

Test whether the GATES differ between the most and least affected groups

From GATES to CLAN

Issue we still have not solve: we observe heterogeneity in treatment effects across groups but we still do not know which covariates drive these results

Classification Analysis (CLAN)

Focus on the "least affected group" G_1 and the "most affected group" G_K

Objective: test for differences along covariates comprising Z between G_1 and G_K

- Let $g(Y, Z)$ be a vector of characteristic of a unit
- Parameters of interest are the avg. characteristics (e.g., age, income) of the G_1 and G_K:

$$\delta_1 = E[g(Y,Z)|G_1] \quad and \delta_K = E[g(Y,Z)|G_K]$$

Test the following hypothesis:

$$H0 : \delta_K - \delta_1 = 0.$$

Rejecting $H0$ for a given characteristic in Z means there is a difference between the most and least affected groups for this characteristic

Inference

Let θ denote the generic parameter of interest, for instance:

- $\theta = \beta_2$ the heterogeneity loading parameter
- $\theta = \beta_1 + \beta_2(S(z) - ES)$ BLP of $s_0(Z)$
- $\theta = \gamma_K$ is the expectation of $s_0(Z)$ for a given group
- $\theta = \gamma_K - \gamma_1$ difference in expectation of $s_0(Z)$ between the most and least affect gps
- $\theta = \delta_K - \delta_1$ difference in expectation of the characteristics

Two sources of uncertainty:

1. Estimation uncertainty regarding θ conditional on the data split (as we seen before)
2. Uncertainty induced by the data splitting (new)

$\Rightarrow$ develop confidence intervals taking into account both sources: repeated random splits of the initial data into $Data_M$ and $Data_A$. Compute individual split CI, θ, and p-value. Then compute sample median over a certain number of random splits (rather than all possible splits).

Overview

Economic applications

Machine Learning for prediction

Dealing with high-dimensionality in Causal inference

- High dimensional methods and inference
- Model selection when the goal is inference
- PDS-LASSO
- Double/Debiased Machine Learning

Unpack the treatment effect heterogeneity

- Honest estimation
- Causal trees and forests
- Generic ML inference

Empirical application

Application using (Crépon et al., 2015)

Experiment in Morocco estimating the effect of access to microfinance services on financial and non-financial outcomes

- 162 *villages* in rural Morocco, divided into 81 *pairs*, 5,551 households
- Within each pair: 1 treatment and 1 control village
- Treatment: opening of a microfinance institution
- Started in 2006 with follow-up surveys in 2009
- Use stratified sample splitting with village pairs as strata
- Variables:
  - Y = financial and non-financial outcommes. We are focusing on
  - D = indicator of being in a village where a microfinance institution opened
  - Z = vector of household characteristics, and pair fixed effects

Best Linear Predictors of Conditional Average Treatment Effect

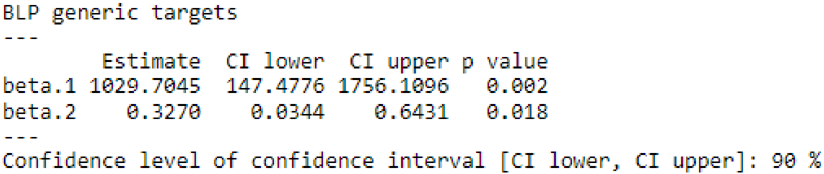
Experiment in Morocco estimating the effect of access to microfinance services on financial and non-financial outcomes

- 162 *villages* in rural Morocco, divided into 81 *pairs*, 5,551 households
- Within each pair: 1 treatment and 1 control village
- Treatment: opening of a microfinance institution
- Started in 2006 with follow-up surveys in 2009
- Use stratified sample splitting with village pairs as strata
- Variables:
    - Y = financial and non-financial outcommes. We are focusing on total amount borrowed
    - D = indicator of being in a village where a microfinance institution opened
    - Z = vector of household characteristics, and pair fixed effects

- Rely on R package "Generic$_M L''$ Main results: low take-up (17% in T group) and significant effect on total borrowing: ATE of MAD 1,193 (Table 2, col 7)

Best Linear Predictor (BLP)

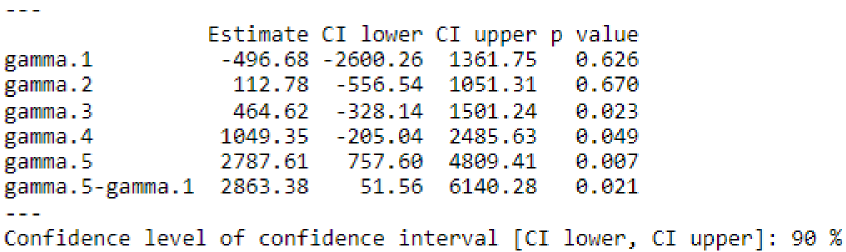
Recall: BLP estimates some β_1 and β_2 via OLS, with $\beta_1 = E_{s0}(Z)$ is the ATE and $\beta_2 \neq 0$ if there is heterogeneity in $s0(Z)$ and $S(Z)$ predicts it well



```
BLP generic targets
---
         Estimate  CI lower  CI upper p value
beta.1 1029.7045  147.4776 1756.1096   0.002
beta.2    0.3270    0.0344    0.6431   0.018
---
Confidence level of confidence interval [CI lower, CI upper]: 90 %
```

Sorted Group Average Treatment Effects (GATES)

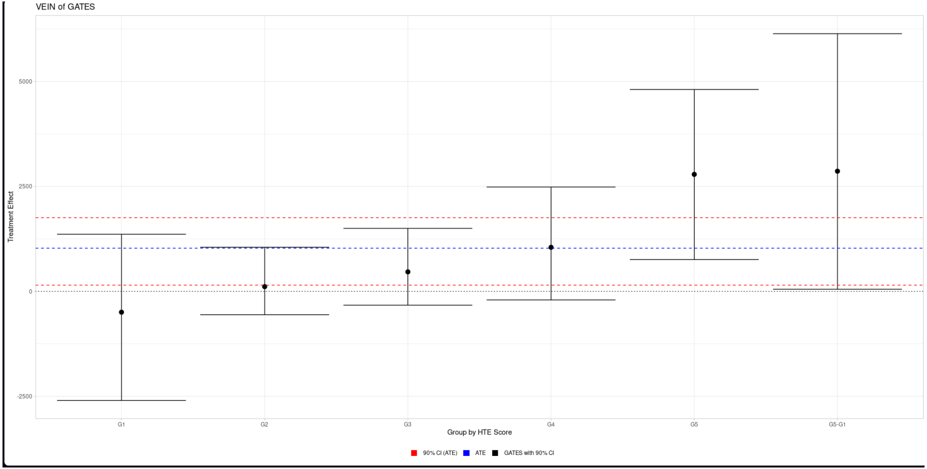
Build groups, here 5 and estimate the group-ATE $\gamma_k = E[s_0(Z)|G_k]$ via OLS



```
---
                 Estimate CI lower CI upper p value
gamma.1           -496.68 -2600.26  1361.75   0.626
gamma.2            112.78  -556.54  1051.31   0.670
gamma.3            464.62  -328.14  1501.24   0.023
gamma.4           1049.35  -205.04  2485.63   0.049
gamma.5           2787.61   757.60  4809.41   0.007
gamma.5-gamma.1   2863.38    51.56  6140.28   0.021
---
Confidence level of confidence interval [CI lower, CI upper]: 90 %
```

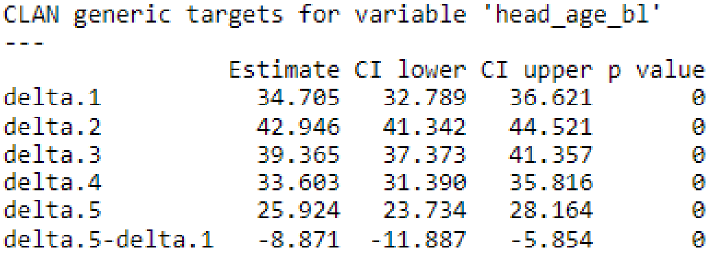
Sorted Group Average Treatment Effects (GATES)

Build groups, here 5 and estimate the group-ATE $\gamma_k = E[s_0(Z)|G_k]$ via OLS



Classification Analysis (CLAN)

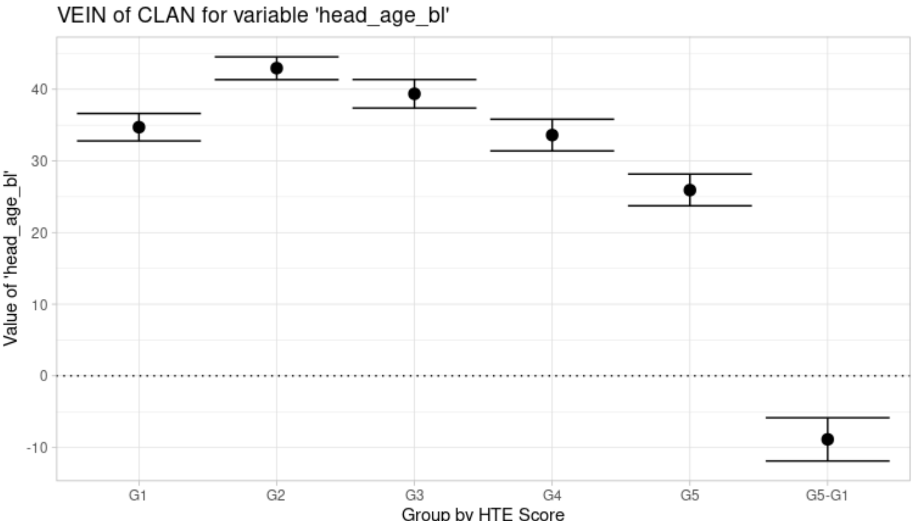
Observed within-group averages, δ_k, of a given variable for groups G_k. Here household head age.



```
CLAN generic targets for variable 'head_age_bl'
---
              Estimate CI lower CI upper p value
delta.1         34.705   32.789   36.621       0
delta.2         42.946   41.342   44.521       0
delta.3         39.365   37.373   41.357       0
delta.4         33.603   31.390   35.816       0
delta.5         25.924   23.734   28.164       0
delta.5-delta.1 -8.871  -11.887   -5.854       0
```

Classification Analysis (CLAN)

Observed within-group averages, δ_k, of a given variable for groups G_k. Here household head age.



References I

Emily Aiken, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E Blumenstock. Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903): 864–870, 2022.

Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

References II

Bruno Crépon, Florencia Devoto, Esther Duflo, and William Parienté. Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. *American Economic Journal: Applied Economics*, 7(1):123–150, 2015.

Jérémy Do Nascimento Miguel. *Return to quality in rural agricultural markets: evidence from wheat markets in Ethiopia*. Intl Food Policy Res Inst, 2022.

Erin C Lentz, Hope Michelson, Katherine Baylis, and Yang Zhou. A data-driven approach improves food insecurity crisis prediction. *World Development*, 122:399–409, 2019.

David McKenzie and Dario Sansone. Predicting entrepreneurial success is hard: Evidence from a business plan competition in nigeria. *Journal of Development Economics*, 141:102369, 2019.

Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 616–625, 2018.

References III

Nathan Ratledge, Gabe Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke. Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611 (7936):491–495, 2022.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.