Machinistas meet randomistas: useful ML tools for empirical researchers

Esther Duflo

Based on work with: Victor Chernozhukhov, Bruno Crepon, Mert Demirer, Pascaline Dupas, Ivan Fernanzez-Val, Chris Hansen, Elise Huilery, Michael Kremer, William Pariente, Juliette Seban, Paul-Armand Veillon

NBER summer institute, 2018

Find the codes at
`https://github.com/demirermert/MLInference`

Introduction

- ML and RCT are the two most important developments for empirical researchers in the past few years
- ML: A set of constantly evolving prediction tools such as:
  - random forests,
  - boosted trees,
  - lasso,
  - ridge,
  - deep and standard neural nets,
  - gradient boosting,
  - their aggregations,
  - and cross-hybrids.

What do they have to do with each other?

- ML is primarily used for *prediction* across a large number of variables
- RCT for causal effects for a low dimensional parameter (the "treatments")
- It seems they would have different applications. In development, ML tools have been used for:
  - predicting which region or person is poor (Blummenstock et al)
  - classifying urban landscapes (Naiak)
  - ... and growing

But they have actually some things in common

- ML tools are developed for causal estimation of a low dimensional parameter. That can be compared with RCT.
- RCT face some problems which are high dimensional
  - What control variables to chose?
  - What are the relevant dimensions of heterogeneity?
  - Multiple treatments and multiple outcomes.

Outline

1. *Lalonde redux: Comparing RCT to ML estimate of causal effects*
2. Choosing control variables
3. Assessing heterogeneity: method and applications

RCT as benchmark for ML tools for causal effects

Chernozhukhov at al (2018) "Double Machine learning ..."

- Main goal: Estimate and construct confidence intervals for a low-dimensional parameter (θ_0) in the presence of high-dimensional nuisance parameter (η_0)
- Now we have a causal question, akin to the questions that are asked by RCT. And the possible suggestion to use a rich array of control variables.
- Lots of work using double Lasso (e.g. Belloni, Chernozukhov, Hansen)
- This paper: **"double/di-biased" ML or "orthogonalized" ML** and sample splitting that can be used with any tools
- Method has started to be used a lot (by Amazon, Microsoft, etc) but does it work and when? We need more "Lalonde" style studies to tell us...

Inference using Modern Nonlinear Regression Methods

- Inference question: how does the predicted value of Y change if we increase a regressor D by a unit, holding other regressors Z fixed?
- We answer this question within the context of the partially linear model, which reads:

$$Y = \beta D + g(Z) + \epsilon, \quad \mathrm{E}[\epsilon \mid Z, D] = 0,$$

  where Y is the outcome variable, D is the regressor of interest, and Z is a high-dimensional vector of other regressors or features, called "controls".
- The coefficient β provides the answer to the inference question.
- (Note) This approach can be extended to ATE in heterogenous models

- We can rewrite the model in the partialled-out form as:

$$\tilde{Y} = \beta \tilde{D} + \epsilon, \quad \mathrm{E}(\epsilon \tilde{D}) = 0, \tag{1}$$

where $\tilde{Y}$ and $\tilde{D}$ are the residuals left after predicting Y and D using Z, namely,

$$\tilde{Y} := Y - \ell(Z), \quad \tilde{D} := D - m(Z),$$

where $\ell(Z)$ and $m(Z)$ are defined as conditional expectations of Y and D given Z:

$$\ell(Z) := \mathrm{E}[Y \mid Z], \quad m(Z) := \mathrm{E}[D \mid Z].$$

The equation $\mathrm{E}(\epsilon \tilde{D}) = 0$ above is the Normal Equation for the population regression of $\tilde{Y}$ on $\tilde{D}$. This implies the following result:

[Frisch-Waugh-Lovell for Partially Linear Model] The population regression coefficient β can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\beta = \arg\min_{b} \mathrm{E}(\tilde{Y} - b\tilde{D})^2 = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}\tilde{D}\tilde{Y},$$

where β is uniquely defined if D cannot be perfectly predicted by Z, i.e. $\mathrm{E}\tilde{D}^2 > 0$.

So β can be interpreted as a regression coefficient of *residualized* Y on *residualized* D, where the residuals are defined by taking-out the conditional expectation of Y and D given Z, from Y and D.

Estimation of β: The DML Procedure

- Our estimation procedure for β in the sample will mimic the partialling out procedure in the population.
- In order to avoid the possibility of overfitting we rely on sample splitting. We have data $(Y_i, D_i, Z_i)_{i=1}^n$. We randomly split the data into two halves: one half will serve as an **auxilliary sample**, which will be used to estimate the best predictors of Y and D, given Z, and then estimate the residualized Y and residualized D. Another half will serve as the **main sample** and will be used to estimate the regression coefficients.
- Let A denote the set of observation names in the auxiliary sample, and M the set of observations names in the main sample.

Estimation of β via DML Procedure

Step 1: using auxiliary sample, we employ modern nonlinear regression methods to build estimators $\widehat{\ell}(Z)$ and $\widehat{m}(Z)$ of the best predictors $\ell(Z)$ and $m(Z)$. Then, using the main sample, we obtain the estimates of the residualized quantities:

$$\check{Y}_i = Y_i - \widehat{\ell}(Z_i), \quad \check{D}_i = D_i - \widehat{m}(Z_i), \quad \text{for each } i \in M,$$

and then using ordinary least squares of $\check{Y}_i$ on $\check{D}_i$ obtain the estimate of β, denoted by $\widehat{\beta}^1$ and defined by the formula:

$$\widehat{\beta}^1 = \arg\min_b \sum_{i \in M} (\check{Y}_i - b\check{D}_i)^2.$$

Step 2: we reverse the roles of the auxiliary and main samples, repeat Step 1, and obtain another estimate of β, denoted by $\widehat{\beta}^2$.

Step 3: we take the average of the two estimates from Steps 1 and 2 obtaining the final estimate:

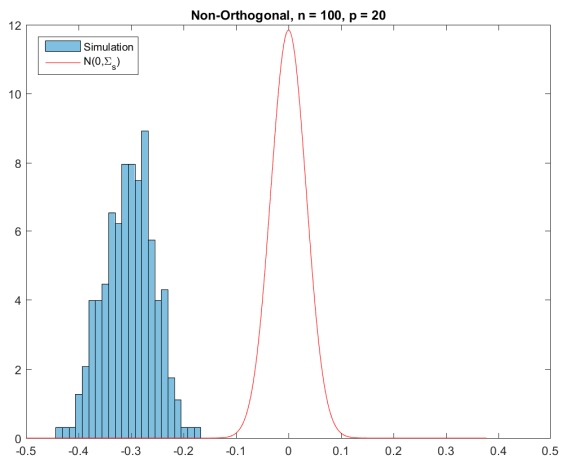
$$\widehat{\beta} = \frac{1}{2}\widehat{\beta}^1 + \frac{1}{2}\widehat{\beta}^2.$$

Point I. Without Othogonalization, "Naive" or Prediction-Based ML Approach is Bad

- Predict Y using D and Z, and obtain

$$D\widehat{\beta}_0 + \widehat{g}_0(Z)$$

- For example, estimate by alternating minimization. Given initial guesses, run Random Forest of $Y - D\widehat{\beta}_0$ on Z to fit $\widehat{g}_0(Z)$, and then Ordinary Least Squares of $Y - \widehat{g}_0(Z)$ on D to fit $\widehat{\beta}_0$. Repeat until convergence.
- Excellent prediction performance! BUT the distribution of $\widehat{\beta}_0 - \beta_0$ looks like this:



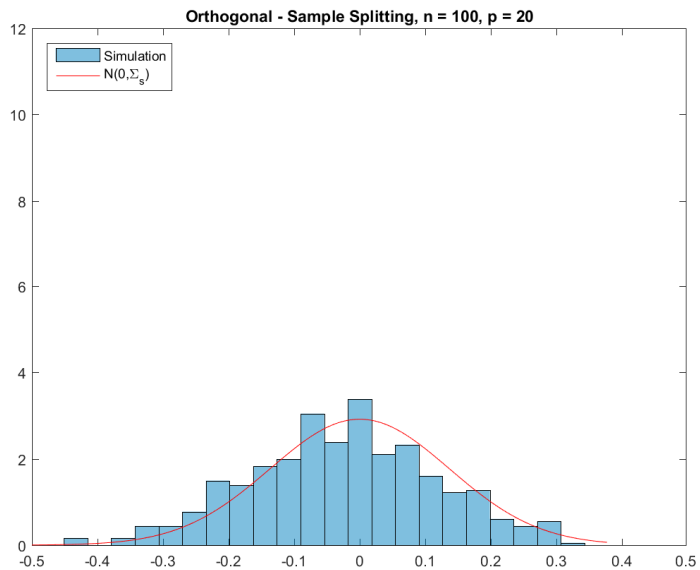
Point II. The "Double" ML Approach is Good

1. Predict Y and D using Z with

$$\widehat{\mathrm{E}[Y|Z]} \text{ and } \widehat{\mathrm{E}[D|Z]},$$

   obtained using the Random Forest or other "best performing ML" tools.
2. Residualize $\widehat{W} = Y - \widehat{\mathrm{E}[Y|Z]}$ and $\widehat{V} = D - \widehat{\mathrm{E}[D|Z]}$
3. Regress $\widehat{W}$ on $\widehat{V}$ to get $\widehat{\beta}_0$.

- Frisch-Waugh-Lovell (1930s) style. The distribution of $\check{\beta}_0 - \beta_0$ looks like this:



Putting the approach to the test

- No magic here: This is just a disciplined way to use a large number of covariates. Frish-Waugh for the 21st century.
- But the causal estimate is only as good as the covariates.
- To assess the potential of these methods with a potentially large amount of control variables, need to compare to the "truth".

One example: Secondary education in Ghana

Duflo, Dupas, Kremer

- Ghana. Ongoing longitudinal study started in 2008
- Sampled 2,064 students admitted to local secondary school, but had not enrolled (mostly due to lack of funds) by end of Term 1 of school year 2008/2009
- Age 17 on average when start, we follow them until age 26
- 682 (randomly selected) received 4-year scholarships to attend local secondary school

The Lalonde exercise

We use DML to estimate (in the control group)

$$Y = \beta D + g(Z) + \epsilon, \quad \mathrm{E}[\epsilon \mid Z, D] = 0,$$

where D is secondary education. We compare with IV estimate, where D is instrumented by T, received scholarship (and "naive" OLS controlling for some obvious control variables, in particular JHS score).

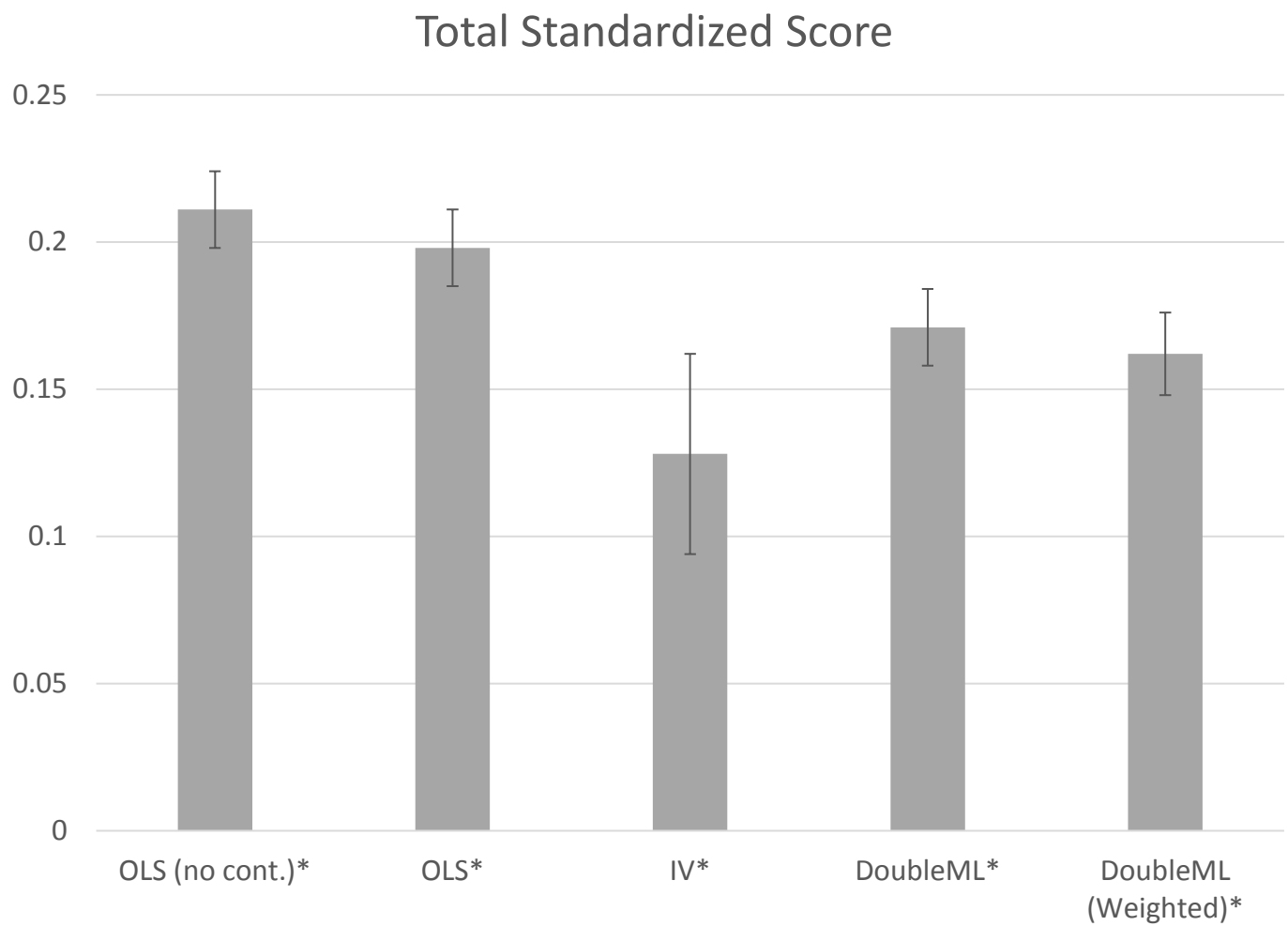
Caveat (1): is IV good measure of treatment effect?

- Direct impact of getting scholarship on outcome:
  - Financial (for inframarginals who would have paid anyway)
  - Self confidence
  - Psychological incentive effects.
- A few children in the control group go to technical institute:
  - To the extent that quality is lower, returns to education are under-estimated.

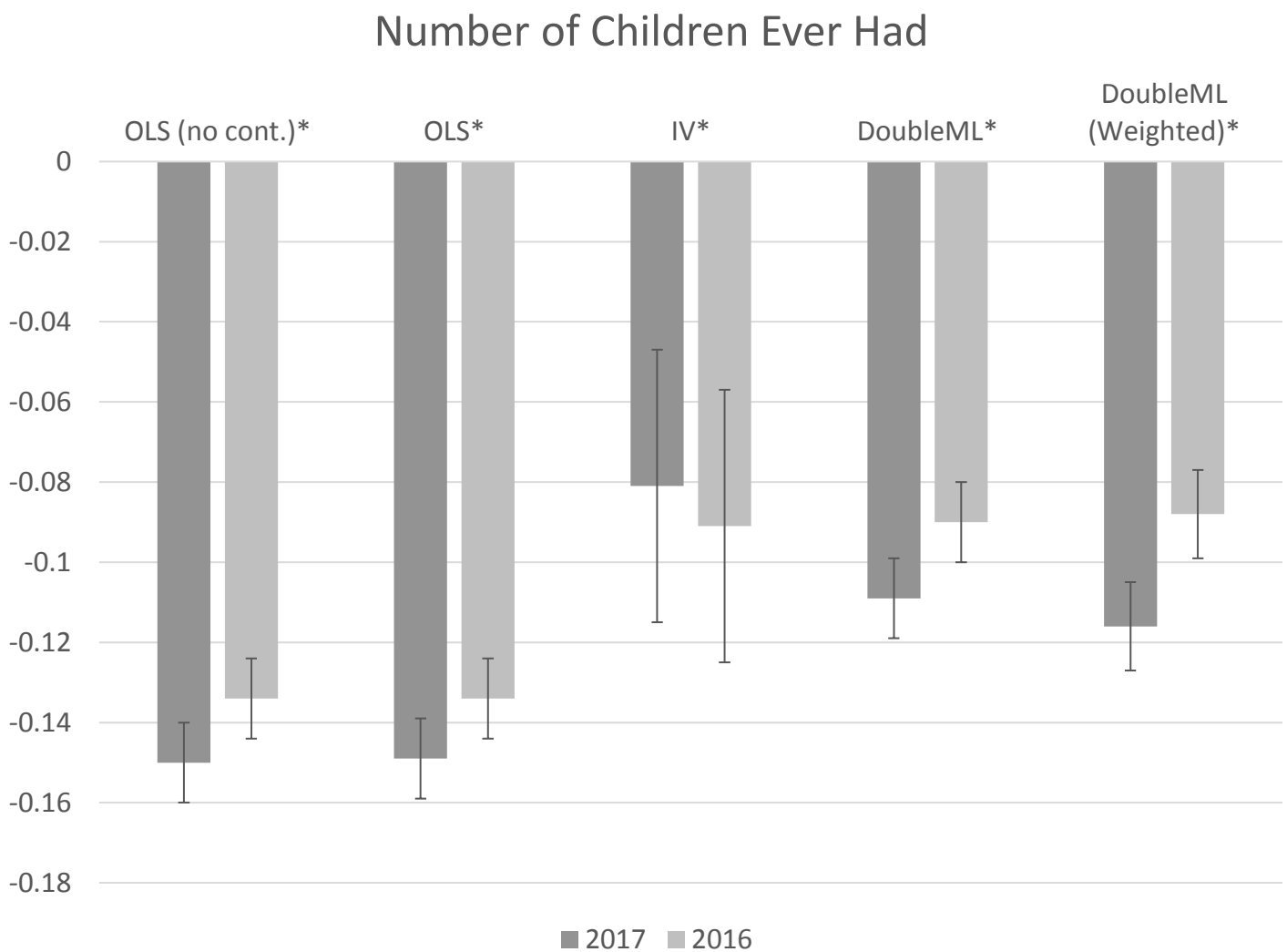
Caveat (2): IV estimate returns for compliers

- Solution:
  - Use ML to estimate treatment effect heterogeneity in the first stage (see below)
  - Estimate a weighted regression, so that the resulting estimates represent the effect for people who are (in terms of their Z) like the compliers

Test scores


Total Standardized Score
0.25
0.2
0.15
0.1
0.05
0
OLS (no cont.)*
OLS*
IV*
DoubleML*
DoubleML (Weighted)*

Fertility


Number of Children Ever Had
OLS (no cont.)*
OLS*
IV*
DoubleML*
DoubleML (Weighted)*
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
2017
2016

Main outcomes, 2016

	OLS(no cont)	OLS	IV	DoubleML	DoubleML(Weighted
Total standardized score	0.211	0.198	0.128	0.171	0.162
	(0.013)*	(0.013)*	(0.034)*	(0.013)*	(0.014)*
Ever enrolled in tertiary education	0.049	0.046	0.021	0.04	0.039
	(0.004)*	(0.004)*	(0.011)	(0.005)*	(0.005)*
Number of children ever had	-0.134	-0.134	-0.091	-0.09	-0.088
	(0.01)*	(0.01)*	(0.027)*	(0.009)*	(0.01)*
Inv. hyperbolic sine earnings	-0.101	-0.101	0.225	-0.151	-0.147
	(0.039)*	(0.039)*	(0.108)*	(0.039)*	(0.041)*
Log earnings last month if positive	0.022	0.022	-0.016	0.006	0.002
	(0.018)	(0.018)	(0.04)	(0.019)	(0.02)
Positive earnings	-0.019	-0.019	0.041	-0.022	-0.022
	(0.007)*	(0.007)*	(0.019)*	(0.007)*	(0.007)*
Total earnings last month	-4.703	-4.44	5.402	-5.727	-4.624
	(2.775)	(2.874)	(8.073)	(3.148)	(3.381)
Index of risky sexual behavior	-0.062	-0.065	-0.031	-0.037	-0.038
	(0.009)*	(0.009)*	(0.022)	(0.009)*	(0.009)*
Preventative health behavior	0.008	0.007	0.078	0.01	0.017
	(0.012)	(0.012)	(0.029)*	(0.012)	(0.013)

Main outcomes, 2017

	OLS I	OLS II	IV	DoubleML	DoubleML(Weighted)
Combined Wages Both	14.636	14.039	23.482	13.408	19.886
	(6.097)*	(6.161)*	(15.74)	(6.538)*	(7.185)*
Combined Wages Both Hst	0.022	0.018	0.249	0	0.035
	(0.041)	(0.42)	(0.112)*	(0.043)*	(0.046)*
Number of Children Ever	-0.15	-0.149	-0.081	-0.109	-0.116
	(0.01)*	(0.01)*	(0.033)*	(0.01)*	(0.011)*
Unwanted Pregnancy	-0.069	-0.07	-0.032	-0.049	-0.047
	(0.006)*	(0.006)*	(0.017)	(0.006)*	(0.006)*
Log Earnings	0.05	0.046	0.001	0.048	0.046
	(0.018)*	(0.018)*	(0.042)	(0.018)*	(0.02)*
Positive earnings	-0.005	-0.005	0.02	-0.01	-0.006
	(0.007)	(0.007)	(0.018)	(0.007)	(.0.008)
Wage Worker	0.015	0.014	0.047	0.011	0.014
	(0.006)*	(0.006)*	(0.017)*	(0.006)	(0.007)

Summary

- For many outcomes DML seems to come closer to the IV estimate, sometimes quite close.
- For some it is still very different.
- It would be very useful to do this in more straightforward applications when you directly compare the RCT treatment effect with the DML estimate (for example in our setting, effect of scholarships in the control group). I could not immediately find one but there should some...

Outline

1. Lalonde redux: Comparing RCT to ML estimate of causal effects
2. *Choosing control variables*
3. Assessing heterogeneity: method and applications

Choosing control variables in RCT

- In principle, it would not be necessary to control for anything while running an RCT
- It may even be problematic (see Athey-Imbens discussion) (stratify, don't control)
- And potentially open the way to specification searching.
- But in practice, applied researchers often do, especially when it happens by chance that some variables are imbalanced (worry that this may bias the results one way or the other, concern for precision).
- Unscientific survey of the applied researcher practice: try to think about what variables may affect the outcomes of interest. Then include it if it turns out that it is imbalanced.
- It turns out that this 'method' formalizes one for one by the Belloni, Chernozhukhov, and Hansen double lasso approach.

Post double selection lasso: a primer

Belloni et al

- Lasso: choose from z's to predict x in a linear regression
  1. Solve the following minimization to obtain $\widehat{\beta}$:

$$\min_{\beta} \mathbb{E}_n[(x_i - z_i'\beta)^2] + \frac{\lambda}{n}||\widehat{\Psi}\beta||_1 \tag{2}$$

  2. Use any z_{ij} with $\widehat{\beta}_j \neq 0$
- Post double selection lasso: choose from z's to control for when estimating treatment effect of d on y
  1. Solve (2) with $x_i = y_i$ to obtain $\widehat{\beta}_1$
  2. Solve (2) with $x_i = d_i$ to obtain $\widehat{\beta}_2$
  3. Use any z_{ij} with $\widehat{\beta}_{1,j} \neq 0$ *or* $\widehat{\beta}_{2,j} \neq 0$
- Stata command "pdslasso" (Ahrens et al (2018)) implements

What we do

1. Gather all potential control variables from baseline
2. Apply cleaning procedure to controls, which includes:
   - Converting categorical variables into sets of indicators
   - Adding the square of each variable
     - Optional: adding two-way interactions of all variables
   - Creating indicators for each variable $=1$ if the variable is missing, then replacing missing values with 0
   - Dropping one from any pair of perfectly collinear variables
   - Standardizing variables
3. (Re)estimate treatment effects using post double selection
   - Partial out strata fixed effects prior to lasso estimation

Empirical Examples

- Olken et al (2014)
- Duflo et al (2015)

Olken et al (2014): Health Outcomes

	(1) Prenatal Visits	(2) Midwife Delivery	(3) Postnatal Visits	(4) Iron Tablets	(5) Percent Immunize	(6) Weight Checks	(7) Vitamin A	(8) Percent Malnour
Panel A: Key Treatment Effects from Paper								
Block Grants	-0.274	0.040	-0.056	0.051	0.012	0.069	0.005	0.011
	(0.201)	(0.027)	(0.120)	(0.081)	(0.018)	(0.049)	(0.055)	(0.015)
Incentives	0.608***	-0.004	-0.104	0.078	0.015	0.096*	-0.013	-0.027*
	(0.220)	(0.025)	(0.140)	(0.081)	(0.018)	(0.053)	(0.058)	(0.015)
Panel B: Key Treatment Effects with Post Double Selection Lasso								
Block Grants	-0.245	0.035	-0.096	0.042	0.022	0.081	0.011	0.019
	(0.224)	(0.027)	(0.115)	(0.084)	(0.018)	(0.051)	(0.055)	(0.015)
Incentives	0.420*	0.005	0.054	0.106	0.013	0.097*	-0.014	-0.026
	(0.240)	(0.025)	(0.135)	(0.081)	(0.018)	(0.055)	(0.059)	(0.016)
N	3840	2763	2763	3791	3521	4804	2758	4749

Olken et al (2014): Education Outcomes

	(1)	(2)	(3)	(4)
	Participation Rate		Gross Attendance	
	Age 7-12	Age 13-15	Age 7-12	Age 13-15
Panel A: Key Treatment Effects from Paper				
Block Grants	0.004	-0.050**	0.002	-0.065***
	(0.006)	(0.023)	(0.005)	(0.024)
Incentives	-0.004	0.016	-0.001	0.025
	(0.006)	(0.024)	(0.006)	(0.025)
Panel B: Key Treatment Effects with Post Double Selection Lasso				
Block Grants	0.003	-0.049**	0.001	-0.064***
	(0.006)	(0.022)	(0.006)	(0.022)
Incentives	-0.005	0.008	-0.003	0.019
	(0.006)	(0.022)	(0.006)	(0.023)
N	4962	1856	4952	1853

Duflo et al (2015): Boys

	(1) Dropped Out	(2) Attendance Rate	(3) Ever Married	(4) Ever Pregnant	(5) Ever Pregnant, Never Married	(6) Ever Married, Never Pregnant
Panel A: Key Treatment Effects from Paper						
Educ Sub Only	-0.024**	-0.001	-0.008*	-0.002	0.001	-0.004
	(0.011)	(0.008)	(0.004)	(0.003)	(0.001)	(0.003)
HIV Educ Only	0.010	-0.021***	0.000	-0.002	0.000	0.001
	(0.010)	(0.008)	(0.005)	(0.002)	(0.001)	(0.004)
Both	-0.015	0.000	-0.010**	-0.006**	-0.000	-0.004
	(0.010)	(0.008)	(0.004)	(0.002)	(0.001)	(0.003)
Panel B: Key Treatment Effects with Post Double Selection Lasso						
Educ Sub Only	-0.021**	-0.007	-0.007*	-0.002	0.001	-0.003
	(0.010)	(0.008)	(0.004)	(0.003)	(0.001)	(0.003)
HIV Educ Only	0.010	-0.022***	0.002	-0.002	-0.000	0.003
	(0.009)	(0.008)	(0.004)	(0.003)	(0.001)	(0.003)
Both	-0.021**	0.001	-0.009**	-0.005**	-0.001	-0.004
	(0.010)	(0.007)	(0.004)	(0.002)	(0.001)	(0.003)
N	9461	8985	9393	9433	9382	9382

Duflo et al (2015): Girls

	(1) Dropped Out	(2) Attendance Rate	(3) Ever Married	(4) Ever Pregnant	(5) Ever Pregnant, Never Married	(6) Ever Married, Never Pregnant
Panel A: Key Treatment Effects from Paper						
Educ Sub Only	-0.031**	-0.002	-0.026**	-0.027**	-0.004	-0.002
	(0.012)	(0.006)	(0.010)	(0.011)	(0.006)	(0.003)
HIV Educ Only	0.003	-0.008	0.011	-0.007	-0.014**	0.005*
	(0.011)	(0.006)	(0.009)	(0.011)	(0.006)	(0.003)
Both	-0.016	0.000	-0.000	-0.011	-0.013**	-0.001
	(0.012)	(0.006)	(0.009)	(0.010)	(0.006)	(0.003)
Panel B: Key Treatment Effects with Post Double Selection Lasso						
Educ Sub Only	-0.016	-0.001	-0.020**	-0.020*	-0.003	-0.001
	(0.012)	(0.006)	(0.010)	(0.011)	(0.006)	(0.003)
HIV Educ Only	0.006	-0.007	0.011	0.001	-0.013**	0.003
	(0.011)	(0.006)	(0.009)	(0.011)	(0.006)	(0.003)
Both	-0.010	0.002	0.001	-0.007	-0.012**	-0.001
	(0.011)	(0.006)	(0.009)	(0.011)	(0.006)	(0.003)
N	9116	8232	9107	9072	9072	9072

Outline

1. Lalonde redux: Comparing RCT to ML estimate of causal effects
2. Choosing control variables
3. *Assessing heterogeneity: method and applications*

Heterogeneous Effects in Randomized Experiments

- Very often we want to know how effects vary with covariates
  - to explore mechanisms
  - to predict what the result may be in a specific population
- We often have many potential covariates: again, risk of specification searching
- Solution 1: pre-register. But that is very unsatisfactory. This amounts to throwing away lots of data.
- Solution 2: use machine learning to guide prediction. This has gained traction with empirical researchers:
  - Hussam, Rigol and Roth paper you will see later: compare ML prediction of who should be more affected by a grant with human prediction.
  - Davis and Heller: predicting summer job (both use Wager and Athey (2017))

The problem...

- Once again, generically, ML tools are great at prediction but it is much more difficult to obtain valid inference
- Several papers (e.g. Athey and Wager, Athey and Imbens) make progress by focusing on some methods (e.g. trees or forest) or assumptions that guarantee consistency may be satisfied
- We build on the DML approach above to build tools that can work with any ML method you like and provide valid confidence intervals
- The key will be to give up on estimating all the possible heterogeneity but focus on a limited number of core features (is there heterogeneity? what are the characteristics of those with the largest treatment effect?)

Methodology: the set up

- Let $Y(1)$ and $Y(0)$ be the potential outcomes in the treatment state 1 and the non-treament state 0. Let Z be a vector of covariates. The main causal functions are the baseline conditional average:

$$b_0(Z) := \mathrm{E}[Y(0) \mid Z],$$

and the conditional average treatment effect:

$$s_0(Z) := \mathrm{E}[Y(1) \mid Z] - \mathrm{E}[Y(0) \mid Z].$$

- Suppose the treatment variable D is randomly assigned conditional on Z, with probability of assignment depending only on a subvector of stratifying variables Z_1 in Z, namely $D \perp\!\!\!\perp (Y(1), Y(0)) \mid Z$, and the propensity score is known and is given by

$$p(Z) := \mathrm{P}[D = 1 \mid Z] = \mathrm{P}[D = 1 \mid Z_1].$$

- The observed outcome is given by $Y = DY(1) + (1 - D)Y(0)$. Under the stated assumptions, the causal functions coincide with the components of the regression function of Y given D, Z:

$$Y = b_0(Z) + Ds_0(Z) + U, \quad \mathrm{E}[U \mid Z, D] = 0,$$

that is,

$$b_0(Z) = \mathrm{E}[Y \mid D = 0, Z]$$

and

$$s_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z].$$

- We assume that the propensity score is bounded away from zero or unity:

$$p(Z) \in [p_0, p_1] \subset (0, 1).$$

- We observe $\mathrm{Data} = (Y_i, Z_i, D_i)_{i=1}^N$, consisting of i.i.d. copies of random vector (Y, Z, D) having probability law P.

Properties of Machine Learning Estimators of $s_0(Z)$

- Work well in practice for prediction purposes, much better than classical methods in the high-dimensional settings, albeit many tuning parameters. Real implementations produced by a huge engineering effort.
- Justification is very often heuristic and practice based. Theoretical justification is available in some cases, existence type results. There exist tuning parameters that make some of these methods work under assumptions that are hard to verify in practice.
- Often there are not known theoretical guarantees for real implementations with the real tuning parameters (exception: Lasso)
- Consequence: We don't know how to do uniformly valid confidence bands based on $z \mapsto \widehat{s}_0(z)$.

Our (Agnostic) Approach

- We propose two strategies for inference about

$$\textit{key features of} \quad s_0(Z) \text{ rather than } s_0(Z).$$

- Both rely on the random data splitting into the main sample, indexed by M, and an auxiliary sample, indexed by A.
- From the auxiliary sample A, we obtain Machine Learning estimates of the baseline and treatment effects, which we call proxy scores

$$z \mapsto B(z) = B(z; \mathrm{Data}_A)$$

and

$$z \mapsto S(z) = S(z; \mathrm{Data}_A),$$

which are possibly biased and noisy predictors of $b_0(z)$ and $s_0(z)$.
- We condition on the auxiliary sample, so we consider these maps as frozen.

Target Parameters

- We target and develop valid inference about *key features of* $s_0(Z)$ rather than $s_0(Z)$, which include
  - (1) Best linear predictor (BLP) of $s_0(Z)$ using $S(Z)$;
  - (2) Average of $s_0(Z)$ (ATE) by heterogeneity groups induced by $S(Z)$;
  - (3) Average characteristics of the most and least affected units.

- Our approach is *generic* with respect to the Machine Learning method being used, and is *agnostic* about its properties.

BLP of $s_0(Z)$ on $S(Z)$: First Strategy

Consider the weighted linear projection:

$$Y = \alpha' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S - \mathrm{E}S) + \epsilon, \ \ \mathrm{E}[w(Z)\epsilon X] = 0,$$

where $S := S(Z)$,

$$w(Z) = \{p(Z)(1 - p(Z))\}^{-1}, \quad X := (X_1, X_2)$$

$$X_1 := X_1(Z), \quad \text{e.g. } X_1 = (1, B(Z)),$$

$$X_2 := (D, (D - p(Z))S(Z)).$$

The first main result is

$$\beta_1 + \beta_2(S(Z) - \mathrm{E}S) = \mathrm{BLP}[s_0(Z) \mid S(Z)],$$

in particular $\beta_1 = \mathrm{E}S_0(Z)$ and
$\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

Special Cases

- If $S(Z)$ is a perfect proxy for $s_0(Z)$, then

$$\beta_2 = 1.$$

- In general, $\beta_2 \neq 1$, correcting for noise in $S(Z)$.
- If $S(Z)$ is complete noise, uncorrelated to $s_0(Z)$, then $\beta_2 = 0$.
- If there is no heterogeneity, that is $s_0(Z) = s$, then

$$\beta_2 = 0.$$

- Rejecting the hypothesis

$$\beta_2 = 0$$

means that there is heterogeneity and $S(Z)$ is its relevant predictor.

Average $s_0(Z)$ by Groups

- The target parameters are

$$\mathrm{E}[s_0(Z) \mid G],$$

  where G is an indicator of a group membership.
- We build the groups to explain as much variation in $s_0(Z)$ as possible

$$G_k = \{S \in I_k\}, \quad k = 1, ..., K,$$

  where $I_k = [\ell_{k-1}, \ell_k)$ are non-overlaping intervals that divide the support of S into regions $[\ell_{k-1}, \ell_k)$ with equal or unequal masses:

$$-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty.$$

- The parameters of interest are

$$\mathrm{E}[s_0(Z) \mid G_k]$$

Average $s_0(Z)$ by Groups: First Strategy

- Consider the weighted linear projection:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k \cdot (D - p(Z)) \cdot 1(S \in I_k) + \nu, \quad \mathrm{E}[w(Z)\nu W] = 0, \tag{3}$$

for $B := B(Z)$, $S := S(Z)$,

$$W = (W_1', W_2')' = (X_1', \{(D - p(Z))1(S \in I_k)\}_{k=1}^{K})'.$$

- $D - p(Z)$ in the interaction $(D - p(Z))1(S \in I_k)$ orthogonalizes this regressor relative to all other regressors that are functions of Z.
- X_1, e.g. B, is included to improve precision, but can be omitted.

- The second main result is

$$\gamma_k = \mathrm{E}[s_0(Z) \mid G_k].$$

Classification Analysis

- Focus on the "least affected group" G_1 and "most affect group" G_K.
- Let $g(Y, Z)$ be a vector of characteristics of a unit.
- The parameters of interest are the average characteristics of the most and least affected groups:

$$\delta_1 = \mathrm{E}[g(Y, Z) \mid G_1] \quad \text{and} \quad \delta_K = \mathrm{E}[g(Y, Z) \mid G_K].$$

- Compare δ_K and δ_1 to quantify differences between the most and least affected groups.
- δ_K and δ_1 are identified because they are directly observed.

Inference: Target

Let θ denote a generic target parameter or functional, e.g.,

- $\theta = \beta_2$ is the heterogeneity loading parameter;
- $\theta = \beta_1 + \beta_2(S(z) - \mathrm{E}S)$ is the personalized BLP of $s_0(Z)$;
- $\theta = \gamma_k$ is the expectation of $s_0(Z)$ for the group $\{S \in I_k\}$;
- $\theta = \gamma_K - \gamma_1$ is the difference in the expectation of $s_0(Z)$ between the most and least affected groups;
- $\theta = \delta_K - \delta_1$ is the difference in the expectation of the characteristics between the most and least affected.

Quantification of Uncertainty: Two Sources

- Two sources:
  - (I) Estimation uncertainty regarding the parameter θ, conditional on the data split;
  - (II) Uncertainty induced by the data splitting.
- We develop confidence intervals that take both sources into account

Application to Morocco Data (Crépon et al (2015))

- The effect of access to microfinance services, experiment from Morocco.
- 162 villages in rural areas of Morocco are divided into 81 pairs.
- One treatment and one control village were randomly assigned within each pair.
- In treated villages, a microfinance institution opened branches.
- Introduced in 2006, outcomes from follow-up surveys in 2009.

- Y is financial and non-financial outcomes, D is indicator of offering access to microfinance services, and Z are 22 household characteristics including the number of household members, number of adults, head age, and 81 pair dummies.
- We use stratified sample splitting where the strata are village pairs.

Application to Morocco Data (Crépon et al (2015))

- For each iteration, split sample into main (M) and auxiliary (A)
- Tune and train ML method to learn B and S using A
- Estimate the BLP parameters by weighted OLS

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1 (D_i - p(Z_i)) + \widehat{\beta}_2 (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad i \in M$$

- Estimate the GATES parameters by weighted OLS

$$Y_i = \widehat{\alpha}' X_{1i} + \sum_{k=1}^{K} \widehat{\gamma}_k \cdot (D_i - p(Z_i)) \cdot 1(S_i \in I_k) + \widehat{\nu}_i, \quad i \in M,$$

- Estimate the CLAN parameters by

$$\widehat{\delta}_1 = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_1] \quad \text{and}$$

$$\widehat{\delta}_K = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_K],$$

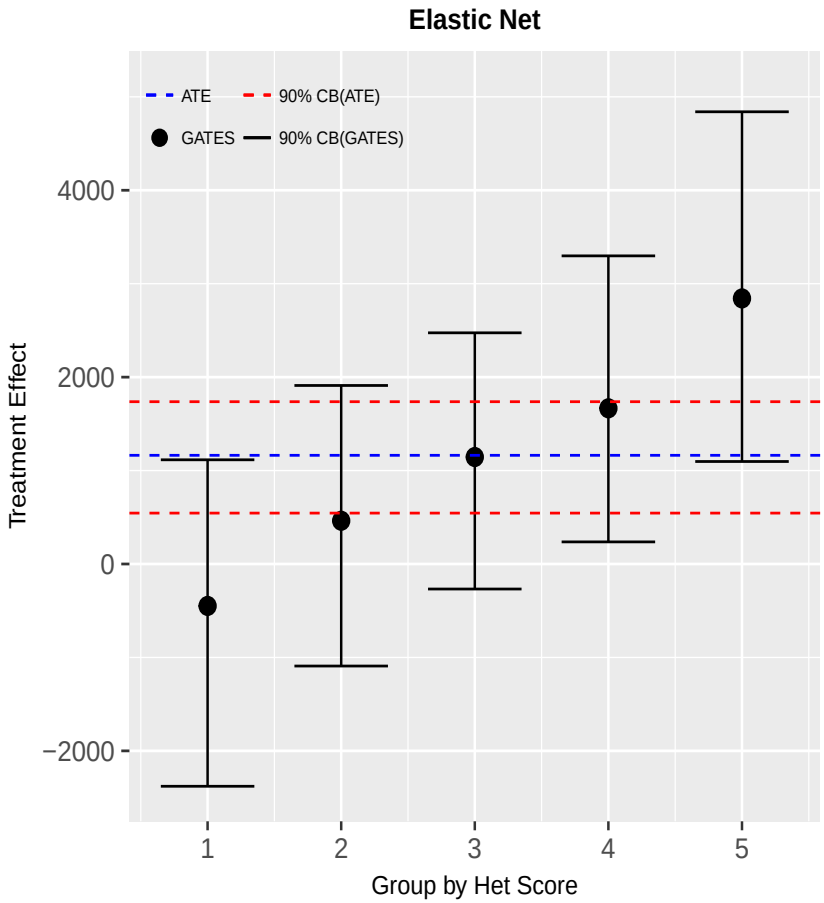
X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$, and village pair fixed effects. Standard errors are clustered at the village level.

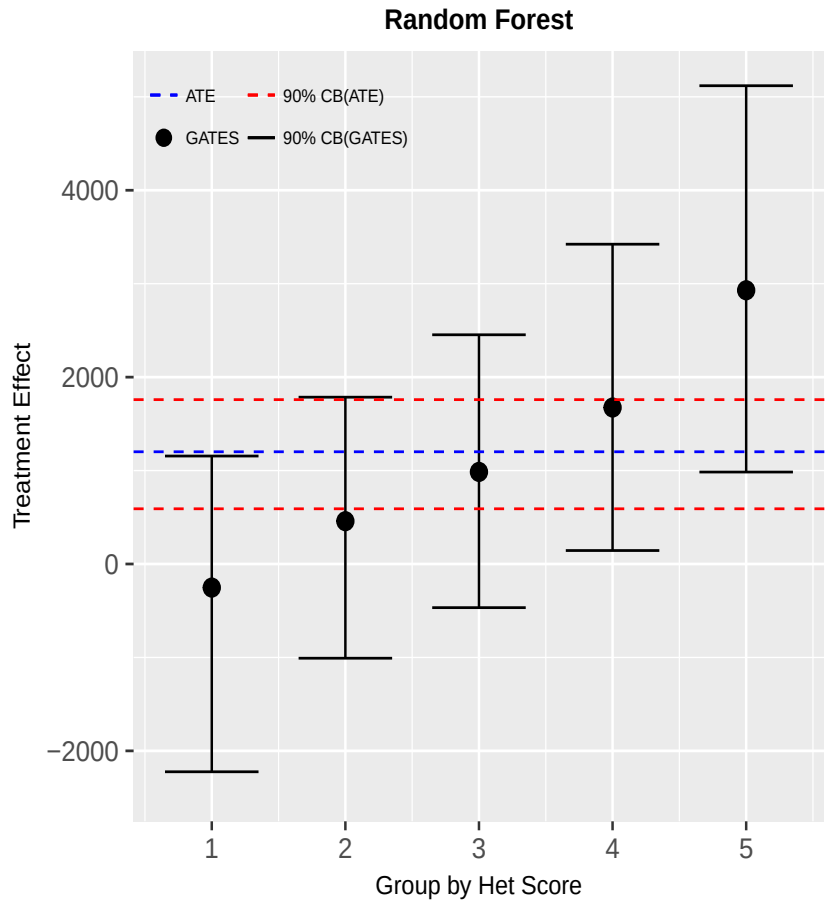
BLP of Conditional Average Treatment Effect

	Elastic Net		Random Forest	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Amount of Loans	1,163 (544,1736) [0.000]	0.238 (0.021,0.448) [0.060]	1,185 (561,1771) [0.000]	0.375 (0.028,0.774) [0.069]
Output	5,095 (232,10033) [0.079]	0.262 (0.085,0.433) [0.008]	5,027 (-89,10194) [0.109]	0.192 (-0.100,0.508) [0.391]
Profit	1,553 (-1344,4389) [0.584]	0.244 (0.079,0.416) [0.008]	1,603 (-1276,4536) [0.521]	0.279 (0.046,0.518) [0.039]
Consumption	-59.1 (-161.5,44.2) [0.514]	0.157 (-0.058,0.385) [0.278]	-58.6 (-166.6, 43.3) [0.508]	0.196 (-0.160,0.574) [0.553]

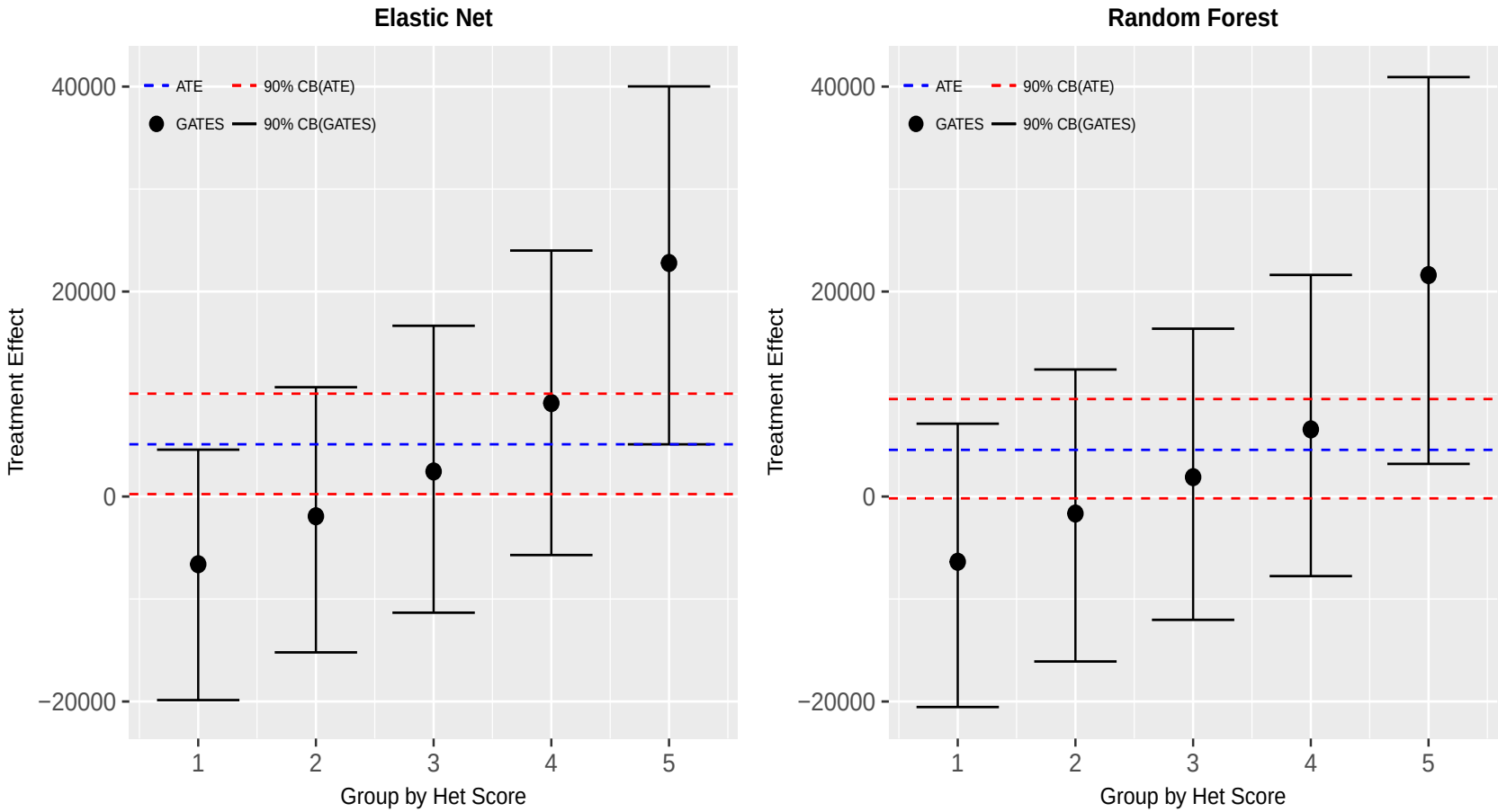
Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Sorted effects: Amount of Loans

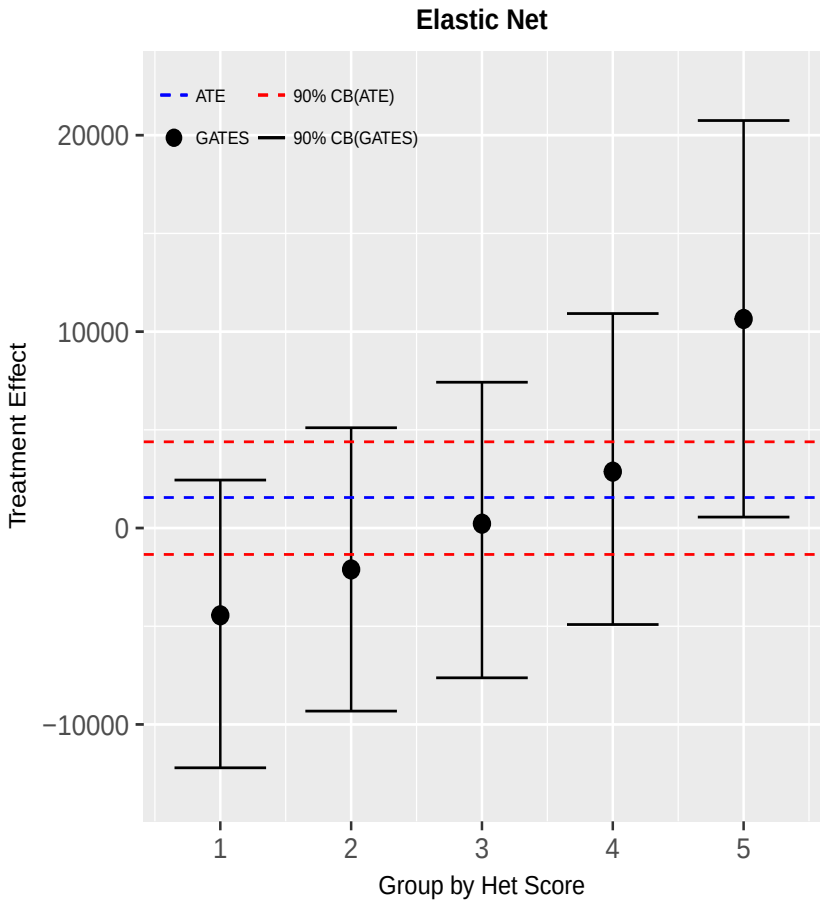


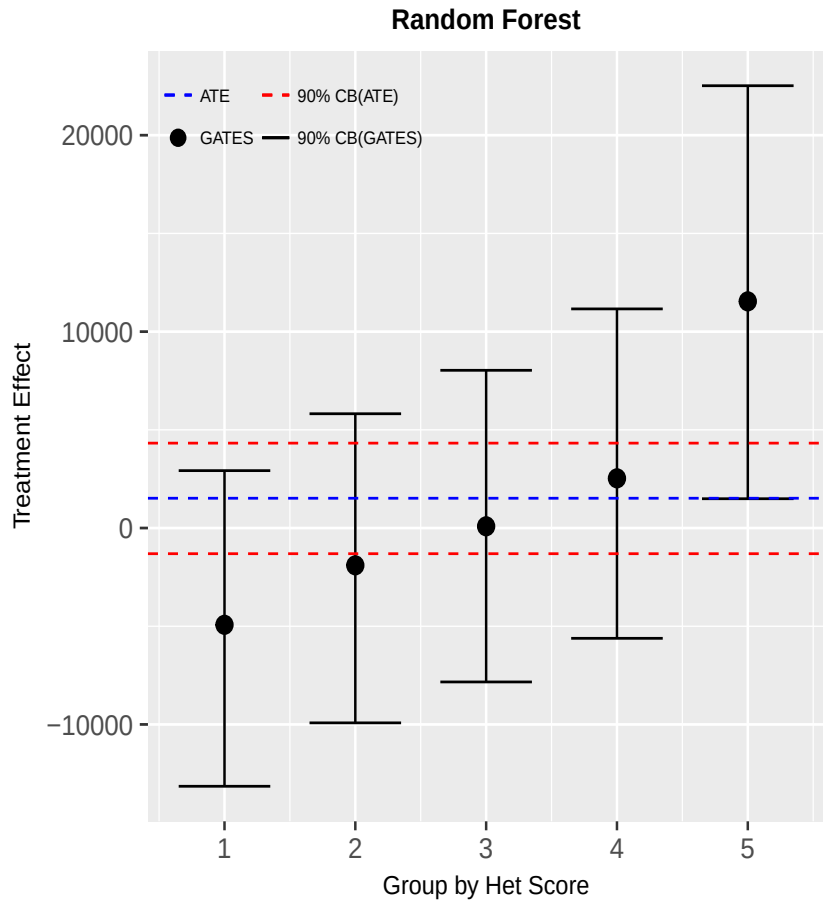


Sorted effects: Output

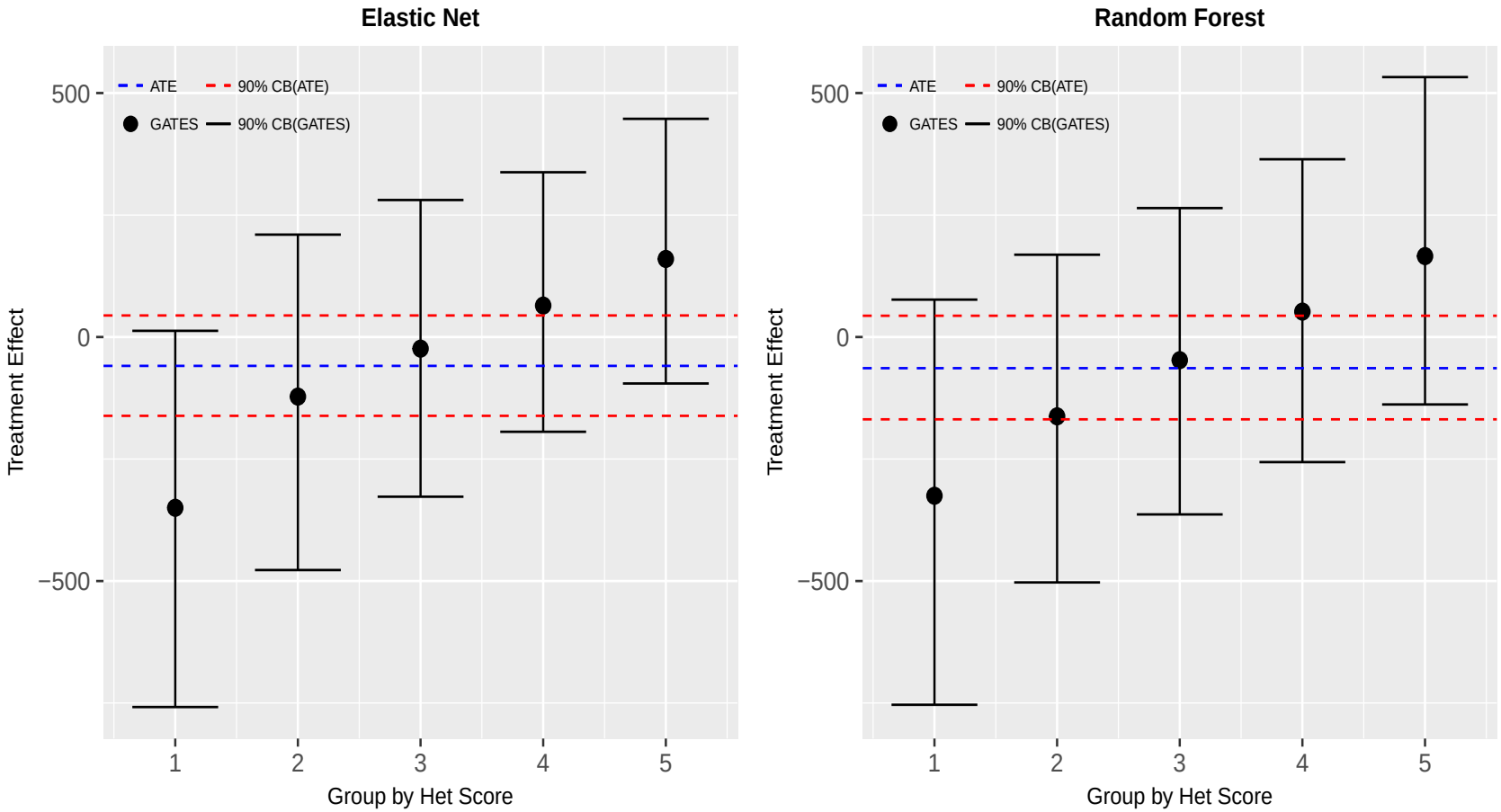


Sorted effects: Profit





Sorted effects: Consumption



Sorted Group Average Treatment Effects: Microfinance

Table: GATES of 20% Most and Least Affected Groups

	Elastic Net			Random Forest		
	20% Most (γ_5)	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)	20% Most (γ_5)	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)
Amount of Loans	2,677	-197	2,995	2,870	94.707	2,814
	(1298,4076)	(-1835,1307)	(945,5103)	(1149,4587)	(-1663,1723)	(503,5193)
	[0.000]	[1.000]	[0.008]	[0.002]	[1.000]	[0.032]
Output	22,367	-3,039	25,088	21,606	626	21,035
	(7678,36920)	(-12546,6535)	(7028,42698)	(5862,38022)	(-11871,13529)	(125,43170)
	[0.007]	[1.000]	[0.015]	[0.015]	[1.000]	[0.097]
Profit	10,644	-1,152.242	11,768	11,540	-2,031	14,037
	(2146,19096)	(-7250,4952)	(1077,22422)	(2965,20955.576)	(-8721,4796)	(2459,25833)
	[0.028]	[1.000]	[0.061]	[0.014]	[1.000]	[0.037]
Consumption	66.4	-333	383	62	-300	332
	(-166.2,289.8)	(-695.6,23.2)	(-38.0,805.6)	(-271,346)	(-683,66)	(-196,835)
	[1.000]	[0.140]	[0.152]	[1.000]	[0.228]	[0.429]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Classification Analysis: Microfinance

	Elastic Net			Random Forest		
	10% Most (δ_{10})	10% Least (δ_1)	Difference ($\delta_{10} - \delta_1$)	10% Most (δ_{10})	10% Least (δ_1)	Difference ($\delta_{10} - \delta_1$)
Amount of Loans						
Head Age	29.3	35.2	-6.6	23.0	33.8	-10.4
	(26.3,32.4)	(32.2,38.2)	(-10.9,-2.4)	(19.8,26.2)	(30.5,36.9)	(-14.9,-6.0)
	-	-	[0.004]	-	-	[0.000]
Non-agricultural self-emp.	0.199	0.068	0.123	0.134	0.118	0.022
	(0.159,0.238)	(0.030,0.108)	(0.069,0.178)	(0.096,0.173)	(0.076,0.156)	(-0.033,0.075)
	-	-	[0.000]	-	-	[0.875]
Borrowed from Any Source	0.144	0.169	-0.038	0.109	0.217	-0.107
	(0.099,0.189)	(0.124,0.212)	(-0.101,0.025)	(0.064,0.153)	(0.175,0.262)	(-0.164,-0.050)
	-	-	[0.448]	-	-	[0.001]
Output						
Head Age	36.280	36.708	-0.896	29.090	30.831	-1.925
	(33.4,39.1)	(33.6,39.6)	(-5.242,3.432)	(25.8,32.3)	(27.5,34.1)	(-6.648,2.799)
	-	-	[1.000]	-	-	[0.849]
Non-agricultural self-emp.	0.275	0.050	0.226	0.215	0.088	0.130
	(0.233,0.315)	(0.007,0.093)	(0.169,0.285)	(0.172,0.257)	(0.045,0.129)	(0.070,0.190)
	-	-	[0.000]	-	-	[0.000]
Borrowed from Any Source	0.193	0.215	-0.033	0.165	0.189	-0.024
	(0.142,0.241)	(0.167,0.262)	(-0.102,0.034)	(0.121,0.208)	(0.146,0.234)	(-0.086,0.039)
	-	-	[0.687]	-	-	[0.895]
Profit						
Head Age	34.1	40.4	-6.5	29.2	33.7	-5.8
	(31.2,37.0)	(37.5,43.4)	(-10.7,-2.5)	(25.7,32.6)	(30.390,37.108)	(-10.566,-1.217)
	-	-	[0.003]	-	-	[0.029]
Non-agricultural self-emp.	0.181	0.108	0.082	0.153	0.099	0.051
	(0.140,0.222)	(0.068,0.149)	(0.022,0.138)	(0.113,0.192)	(0.058,0.139)	(-0.003,0.105)
	-	-	[0.014]	-	-	[0.129]
Borrowed from Any Source	0.180	0.257	-0.091	0.144	0.162	-0.032
	(0.130,0.230)	(0.207,0.307)	(-0.160,-0.022)	(0.098,0.190)	(0.122,0.206)	(-0.095,0.029)
	-	-	[0.020]	-	-	[0.578]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

A second application: comparing two interventions

Cream skimming and the comparison between different interventions.

Bruno Crepon, Esther Dulo, Elise Huillery, William Pariente, Juliette Seban, Paul-Armand Veillon

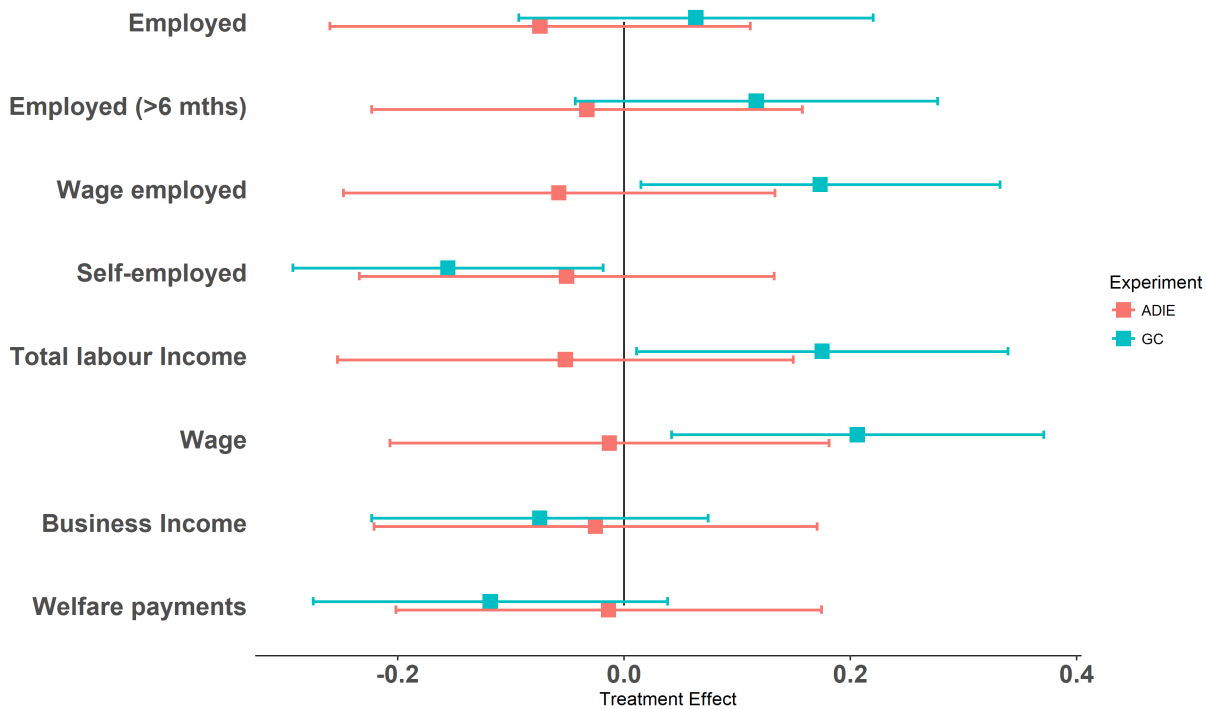
- Increasing number of RCT on a same topic: active literature to explore how results change
  - Same intervention in different contexts or populations (Allcott (2015)) or close to similar interventions (Imbens and Hotz (2005))
  - Many meta-analyses (Meager (2016), Card et al. (2018), Grimm et al. (2015))
  - Sometimes programs are different but selection is also different

This project compares two interventions with the same goal

- ADIE: launched by a microcredit agency; GC: launched by the social services "one stop shop" for the youth.
- Both target unemployed youth with a self employment project.
- Both aim to put them back in employment.
- Some differences
  - ADIE selects; GC takes everyone
  - ADIE really emphasizes the self employment project. GC uses it as leverage but would be happy to place people with salaried jobs instead.

GC is effective, ADIE is not

Figure: Comparison of ITT(ADIE) and ITT(GC)



The error bars display the 95% confidence interval. ▸ Details

	ADIE=0	GC=0	ADIE=GC
Wald Test	0.94	0.034	0.226

Is GC really more effective than ADIE?

- Does GC target a different population?
- Or is the content of the program more effective?
- We will compare ADIE and GC populations and evaluate GC impact on ADIE population
  - Matching on the baseline characteristics

The populations are different on unobservables

- After controlling for observables, ADIE control group still does much better than the control group in GC (68% of ADIE control group employed vs 44% in GC)
- ADIE: interviews candidates to explicitly look for motivated people.
- ADIE: good at selecting people based on their observable and unobservable variables (non-cognitive and cognitive skills)
- As stressed in Heckman et al. (2002), cream-skimming is problematic when the impact is heterogeneous and decreases with the propensity to be enrolled

Can we do better?

- Limited set of variables available in both data sets (X_0).
- However, a broad set of variables in GC initial survey (X_1): cognitive skills, their employment history, their project progress, their motivation...

$\Rightarrow$ What would have been the effect of the GC program if they had picked people who were as likely to find a job anyway as the ADIE population (based on this rich set of variables)?

- Strategy to match participants based on their potential probability to be employed, E^0:
  - We estimate the probability to be employed $\widetilde{E}(0) = E(E(0) \mid X_0, X_1, D = 0, T^{GC} = 0)$ without treatment
  - We look at the heterogenity according to $\widetilde{E}(0)$
  - We estimate GC effects in the population with high $\widetilde{E}(0)$ (so that they are similar to ADIE).

Steps of our strategy

1. Estimate the heterogeneous component

$$\widetilde{E}(0) = E(E(0) \mid X_0, X_1, D = 0, T^{GC} = 0)$$

2. Explore heterogeneity of GC impact with respect to $\widetilde{E}(0)$

3. Estimate and interpret weights computed so that

$$E(w^* \widetilde{E}(0) | D = 0) = E(E(0) | D = 1, T^{ADIE} = 0)$$

4. Estimate impacts of the GC program using weights w^*, and compare to ADIE impact and GC impact on full sample

Prediction of E^0

Three constraints to apply ML methods (Hastie et al. (2008)):

- Only 294 observations in GC control group: cannot train too complex ML algorithms (NN, boosting..)
- More covariates than observations (curse of dimensionality)
- Cross-folding procedure to avoid overfitting: our prediction relies on the random partitioning of the sample.

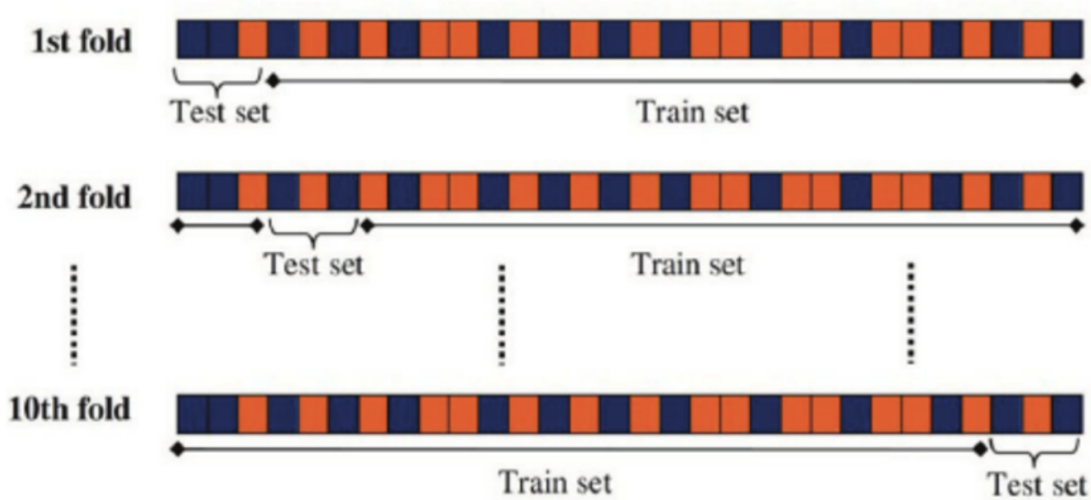


Figure: Cross folding

Prediction of E^0

- Randomly partition our sample in ten splits (stratify by sites)
- For a given split, select the predictors by running a lasso on the nine other splits (tuning parameter by CV).
- Logit model with nine splits used for the lasso regression to predict the outcome for the given split (to remove the bias induced by a direct lasso)
- Repeat this 300 times to get predictions based on different random splitting
- For a given observation, take the median of the 300 predictions produced for this observation.

Is there any Heterogeneity?

- $\widehat{ITT}(GC|ADIE)$ differ from what we obtained only if the treatment is heterogeneous according to $\widetilde{E}(0)$
- We estimate the following specification (Chernozhukov et al. (2017))

$$Y = \alpha + \gamma\widetilde{E}(0) + \beta_1(T_i - P(Z)) + \beta_2(T_i - P(Z))(\widetilde{E}(0) - \overline{\widetilde{E}(0)}) + \epsilon \mid \mathbb{E}(\omega(Z)\epsilon\widetilde{E}(0)) = 0$$

where $P(Z) = \mathbb{P}(T \mid Z)$ and Z a set of covariates, $\omega_i = \frac{1}{p(Z)(1-p(Z)}$.

- Rejecting the hypothesis $\beta_2 = 0$ means that there is heterogeneity according to $\widetilde{E}(0)$.

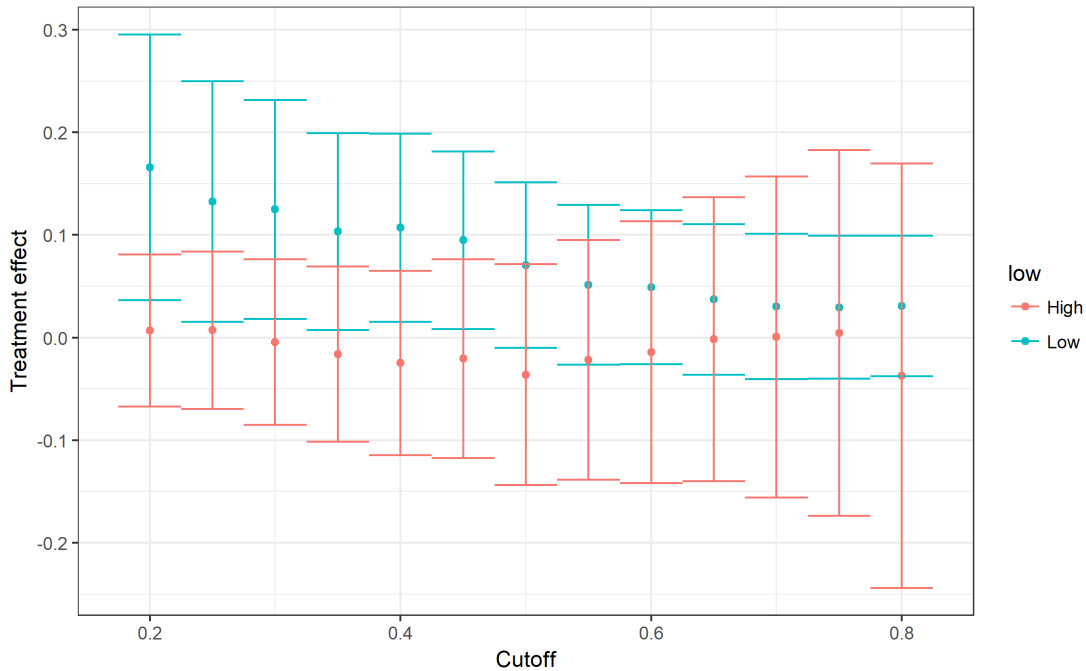
There is heterogeneity which is related to predicted employment

	$(T - P(Z))$	$(T - P(Z)) \times (\widetilde{E}(0) - \overline{\widetilde{E}(0)})$	**Obs.**
	(1)	(2)	(3)
Employed	0.18	-0.34	624
	(0.08)	(0.16)	
Wage employed	0.17	-0.18	624
	(0.08)	(0.16)	
Employed (>6 mths)	0.11	-0.12	624
	(0.07)	(0.16)	
Self-employed	-0.01	-0.07	624
	(0.03)	(0.08)	
Labour Income	176.92	-141.91	619
	(101.15)	(219.06)	
Wage	156.72	-59.3	619
	(100.42)	(217.02)	
Business Income	20.35	-82.76	623
	(22.82)	(68.62)	
Welfare payments	-21.86	-40.79	616
	(58.36)	(121.9)	

The specifications include survey month fixed effects. We display robust stds.

Impact of different level of "cream skimming"

Figure: Weighting schemes



Each orange dot shows the treatment effect when picking the X% applicants predicted to be most likely to get a job anyway.
Each blue dot shows the treatment effect when picking the X% of applicants predicted to be least likely to get a job anyway.

Two weighting schemes that make the GC population look like ADIE's –in terms of employment

Figure: Weighting schemes

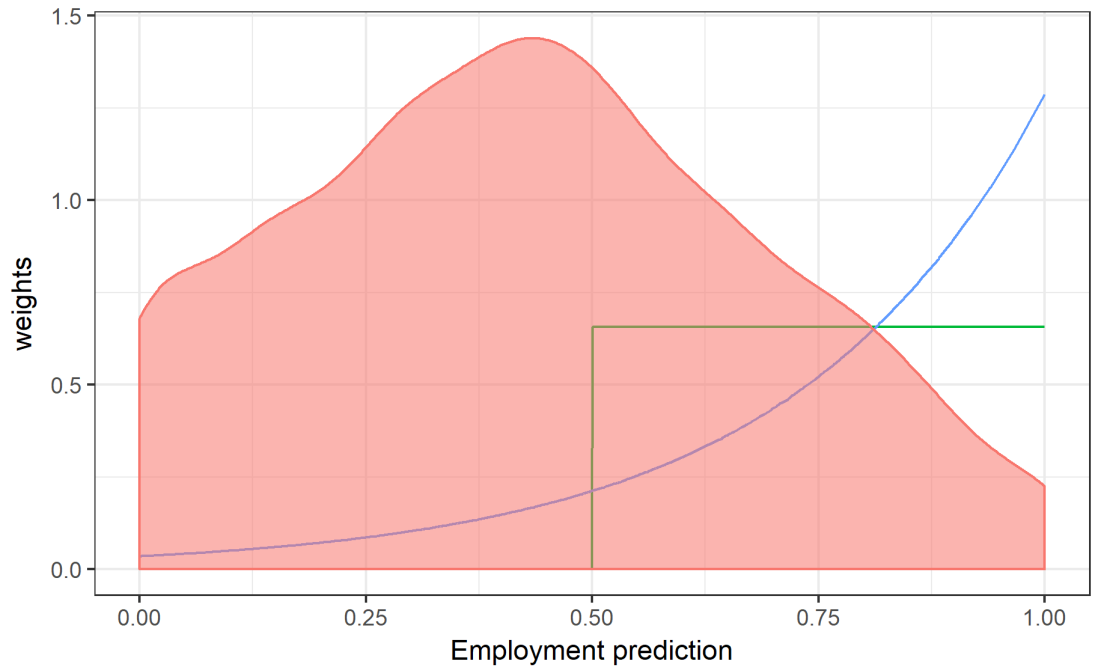


Table: Average treatment effect on ADIE population

	Weighted average			Weighted OLS			
	GC control	**Binary *w***	**Hai. *w***	**ITT(GC)**	**ITT(ADIE)**	**Binary *w***	**Hai. *w***
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)(7)**	
Employed	0.44	0.58	0.58	0.03	-0.04	-0.04	0
				(0.04)	(0.04)	(0.07)	(0.11)
Wage employed	0.36	0.48	0.48	0.08	-0.03	0.04	0.07
				(0.04)	(0.05)	(0.07)	(0.12)
Employed (>6 mths)	0.28	0.39	0.39	-0.02	0	0.02	0.05
				(0.04)	(0.05	(0.06)	(0.12)
Self-employed	0.07	0.09	0.09	-0.04	0.02	-0.06	-0.04
				(0.02)	(0.04)	(0.04)	(0.07)
Labour Income	466.72	637.51	637.51	108.61	-40.09	98.29	73.17
				(52.38)	(79.33)	(94.19)	(176.33)
Wage	432.66	584.82	584.82	124.95	-9.3	119.25	109.11
				(50.85)	(70.09)	(91.4)	(172.75)
Business Income	33.71	52.52	52.52	-16.16	-11.16	-20.96	-35.91
				(16.44)	(44.01)	(41.02)	(80.54)
Welfare payments	225.79	241.34	241.34	-38.89	-4.75	-79.12	-58.46
				(26.24)	(33.39)	(44.92)	(84.29)

The specifications include survey month fixed effects. We diplay robust standard errors.

Thank you!

Find the codes at
`https://github.com/demirermert/MLInference`