Machine Learning Tools for RCTs

2) Post-Double-Lasso

Bruno Fava

November 27, 2023

Post-Double-Lasso

Thanks to Matteo Ruzzante for collaboration on making the slides!

Post-Double-Lasso

- In previous sections, we learned about Lasso and Post-Lasso (OLS after Lasso)
  - ↪ Novelty: double!
- Goal is to learn the Average Treatment Effect of intervention under SOO
  - ↪ Selection on Observables: treatment independent of outcomes conditional on covariates
- Original paper: “Inference on Treatment Effects after Selection among High-Dimensional Controls”
- by Alexandre Belloni, Victor Chernozhukov and Christian Hansen, Review of Economic Studies (2014)
- Note: not about RCTs! In fact, learning ATE is straightforward in RCTs...
- So... why to use PDL in RCT? More on this later

Motivation

Motivation

- Empirical researchers often want to estimate the *causal effect* of a policy
  - E.g., impact of some government program on an economic outcome of interest
- Often rely on conditional-on-observables identification strategy
  - Assumption: even without randomization, treatment as good as random conditional on X
  - Common techniques: propensity-score matching, adding controls in OLS, etc.
  - Not common in state-of-the-art Development Economics
- Common problem for econometric analysis:

  Which controls to include? Which functional form?

  1. Use economic intuition
  2. Show *ad hoc* sensitivity analysis

How Did We Get Here?

- Leamer critique: *Let's Take the Con Out of Econometrics* (AER, 1983)
  - "Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone else's takes anyone data analysis seriously"
  - Remedy proposed: do sensitivity analysis

    $\Longrightarrow$ Show how results vary with changes in *specification* or *functional form*

- In RCTs, it's still common practice to include a set of controls in the regression model
  - "Even in experiments, it's customary to control for covariates to correct for chance associations between treatment status and applicant characteristics and to increase precision" (Angrist and Pischke, 2008, p. 288)
  - BUT lack of guidance on how to select covariates (more on RCTs later)

This Paper

"Post-Double-Selection" (PDS)

Proposes a novel method for estimating and performing **inference** on the effect of a treatment variable (conditionally exogenous on observables) on an outcome of interest in the presence of high-dimensional regressors, possibly larger than the sample size

- Differs from usual post-model-selection by relying on three steps

  1. Select set of control variables that are useful at predicting the treatment
  2. Select *additional* control variables that predict the outcome
  3. Estimate the treatment effect by linear regression on the treatment and union of the variables selected in two previous steps

Success of PDS

- 1,648 citations on Google Scholar since April 2014
- Rapidly increasing popularity of this inference method in empirical economic research
- Especially in experimental designs with large set of covariates
  - E.g.: need to select variables from baseline survey data
  - Extra benefit: transparent way to pick controls (robust to *p*-hacking)
- Widely available on statistical softwares: `pdslasso` in Stata, `hdm` in R

Framework

Framework

Exogenous Treatment Given Controls

- The model:

$$y_i = d_i\alpha_0 + z_i'\beta + \zeta_i, \qquad \mathbb{E}\left[\zeta_i | d_i, z_i\right] = 0 \tag{1}$$

- $d_i \sim$ policy/treatment variable → does NOT need to be binary
- $z_i \sim$ set of p (potentially $> n$) covariates, including controls and powers/interactions
  ↪ "Nonparametric" approximation to smooth functions!

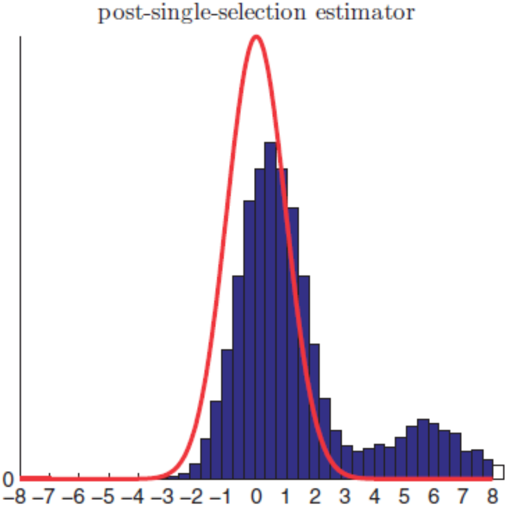
- LASSO can deal with $p > n$, but estimate of α will be biased
- Motivates use of Post-LASSO: OLS after LASSO

Framework

Exogenous Treatment Given Controls

- Post-LASSO is not perfect: OLS is unbiased only if LASSO selects all relevant variables

↪ Finite-sample distribution is bi-modal



Framework

Exogenous Treatment Given Controls

- Belloni et al., 2014 propose the solution: Post-**Double**-Selection LASSO
- Remember: what causes omitted-variable-bias in OLS?
  $\hookrightarrow$ Omitted variable must be relevant **and** correlated with treatment
- Idea of Double Lasso: include all *X* correlated with outcome or treatment
- Intuition: two chances to include variable → less likely to omit!
- Only omit if coefficient small in **both** equations → bias is small
- Run Lasso twice:

$$y_i = z_i'\beta + \zeta_i, \qquad \mathbb{E}\left[\zeta_i | d_i, z_i\right] = 0$$

$$d_i = z_i'\gamma + \nu_i, \qquad \mathbb{E}\left[\nu_i | z_i\right] = 0$$

- Pick vars relevant in any of the two. Then, run OLS.

PDS Estimator

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \mathbb{E}_n \left[\left(y_i - d_i \alpha - x_i' \beta \right)^2 \right] : \beta_j = 0, \forall j \notin \hat{I} \right\}$$

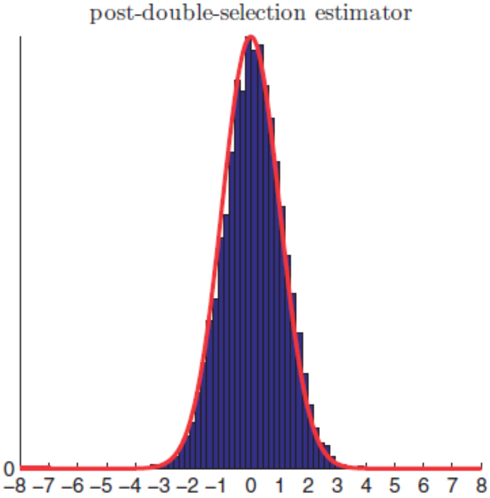
$$\hat{I} = \hat{I}_1 \cup \hat{I}_2 \cup \hat{I}_3$$

- Allows for researcher to choose to add arbitrary regressors for robustness: $\hat{I}_3$

Framework

High-Dimensional-Sparse Model

- Post-Double-Selection LASSO has much better finite-sample properties



Connection with Double Machine Learning

- Double Lasso is “precursor” of what is now called “Double Machine Learning”

- In linear model, recover ATE if learn **either** (i) $\mathbb{E}[Y|X]$ or (ii) propensity score $p(X)$
  $\hookrightarrow$ Note if d is binary, regression of d on z consists of learning the propensity score!

- Learning both leads to robustness! (less bias) And better rates of convergence

- Beyond linear model and ATE, many parameters can be estimated with similar benefits

- Learning both $\mathbb{E}[Y|X]$ and $p(X)$ leads to better rates of convergence and “double-robustness”

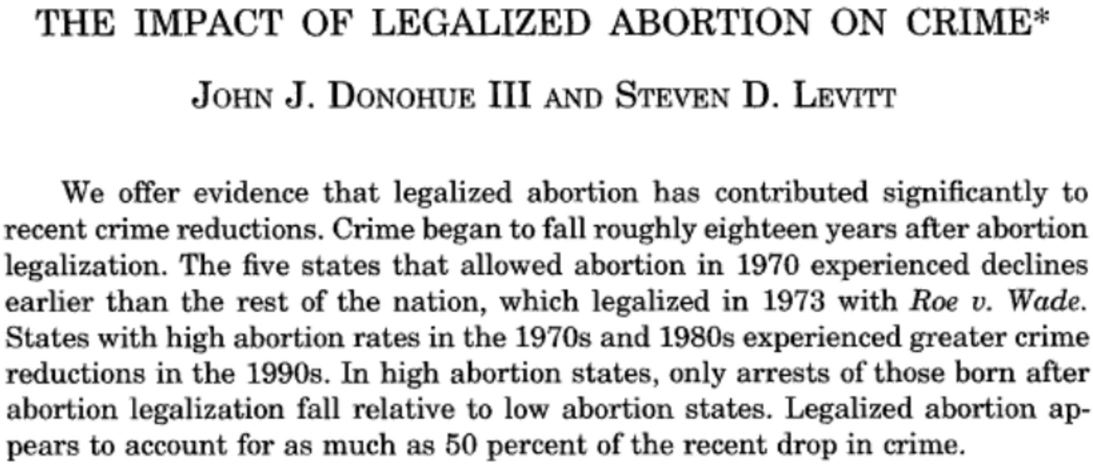
- Double Machine Learning denotes this generalization

A Glance on Theory: Regularity Conditions

- To ensure inference is valid asymptotically, relies on three classes of assumptions

- Approximate Sparse Treatment Effects
  - We can have $p > n$, but only few variables matter
- Sparse Eigenvalues
  - Behavior of $\mathbb{E}_n[x_i x_i']$
- Structural Moments
  - Moment conditions on the structural errors and regressors

Empirical Application



THE IMPACT OF LEGALIZED ABORTION ON CRIME*

JOHN J. DONOHUE III AND STEVEN D. LEVITT

We offer evidence that legalized abortion has contributed significantly to recent crime reductions. Crime began to fall roughly eighteen years after abortion legalization. The five states that allowed abortion in 1970 experienced declines earlier than the rest of the nation, which legalized in 1973 with *Roe v. Wade*. States with high abortion rates in the 1970s and 1980s experienced greater crime reductions in the 1990s. In high abortion states, only arrests of those born after abortion legalization fall relative to low abortion states. Legalized abortion appears to account for as much as 50 percent of the recent drop in crime.

The Study

- Examines the effect of abortion on crime rates in the US

- Two key arguments for a causal channel

  1. More abortion among a cohort → smaller cohort size 20 years later → lower crime
  2. Abortion → women have more control over timing of fertility
     → childbirth in a more favorable environment
     → lower crime holding fertility rates constant

- **Empirical challenge:** state-level abortion rates are NOT randomly assigned

  - There are factors associated to both abortion rates and the current crime rate
  - Any association is likely to be spurious

- Need adequate controls to establish causal link

Empirical Evidence

- **Identification** relies on controlling for state & year FEs, a vector of time-varying state-specific controls
  - Prisoners & police per capita (lagged), a range of variables capturing state economic conditions, such as unemployment, income and poverty rate, state welfare generosity at $t - 15$, concealed handgun laws, beer consumption per capita

- **Main findings:** Abortion legalization reduces crime
  - Crime began to fall roughly 18 years after abortion legalization
  - Legalized abortion account for as much as 50 percent of the recent drop in crime rates

Comments

- Paper was received with some criticisms
  - Identifying assumptions are not well discussed in the paper
  - Claim of causality is rather debatable

- **Foote & Goetz (2008)**: results are NOT robust to allowing for differential state trends based on state-wide crime rates that predate the period when abortion could have had a causal effect on crime

- Donohue & Levitt (2008): measurement error in the abortion proxy explains the smaller estimates

  ↪ Using a more carefully constructed measure of abortion that better links birth cohorts to abortion exposure fixes it

What if We Use PDS?

- Belloni, Chernozhukov & Hansen (2014) allow for a much richer set of controls
  - Higher order terms & interactions
  - Initial conditions & differences
  - Within-state averages
- Features of a state may be associated both with its growth rate in abortion and crime

  ↪ 284 control variables to select among

Caveats:

- PDS is about inference, but does not grant causality (conditional exogeneity assumed)
- Specification relies on a large number of district cross time fixed effects, which does not immediately fit into the regularity conditions of PDS

New Results

- PDS gives a small set of 8 to 12 variables
- Generally related to non-linear trends that depend on initial state-level characteristics
- Differ substantially from 8 variables used in original paper
  - ↪ Once this set is linearly controlled for, estimated effect of abortion is imprecise
  - ↪ Same conclusion as Foote & Goetz's comment,
    which was based on intuitive grounds

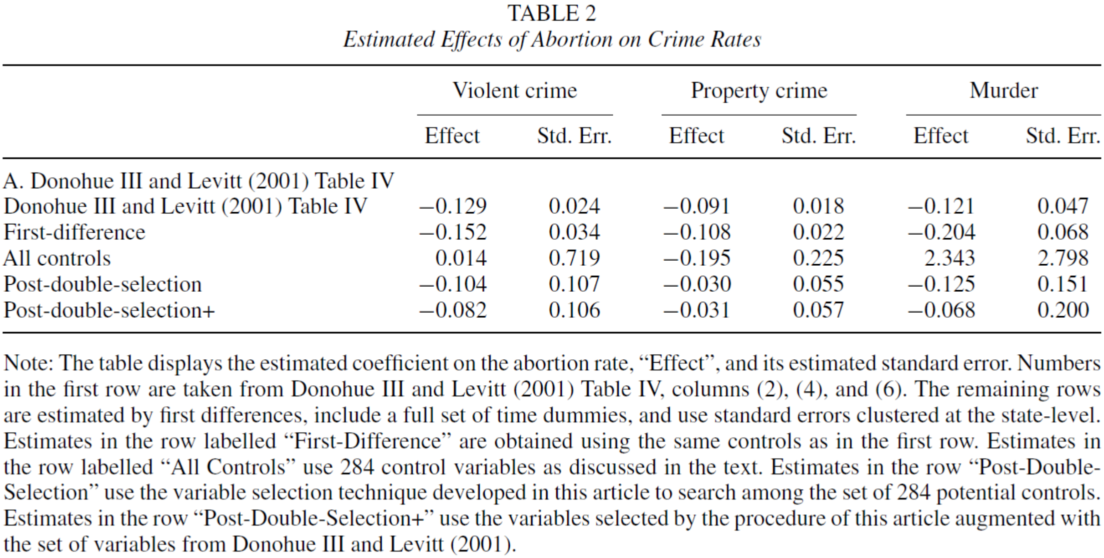


TABLE 2

Estimated Effects of Abortion on Crime Rates

	Violent crime		Property crime		Murder	
	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
A. Donohue III and Levitt (2001) Table IV						
Donohue III and Levitt (2001) Table IV	−0.129	0.024	−0.091	0.018	−0.121	0.047
First-difference	−0.152	0.034	−0.108	0.022	−0.204	0.068
All controls	0.014	0.719	−0.195	0.225	2.343	2.798
Post-double-selection	−0.104	0.107	−0.030	0.055	−0.125	0.151
Post-double-selection+	−0.082	0.106	−0.031	0.057	−0.068	0.200

Note: The table displays the estimated coefficient on the abortion rate, “Effect”, and its estimated standard error. Numbers in the first row are taken from Donohue III and Levitt (2001) Table IV, columns (2), (4), and (6). The remaining rows are estimated by first differences, include a full set of time dummies, and use standard errors clustered at the state-level. Estimates in the row labelled “First-Difference” are obtained using the same controls as in the first row. Estimates in the row labelled “All Controls” use 284 control variables as discussed in the text. Estimates in the row “Post-Double-Selection” use the variable selection technique developed in this article to search among the set of 284 potential controls. Estimates in the row “Post-Double-Selection+” use the variables selected by the procedure of this article augmented with the set of variables from Donohue III and Levitt (2001).

What about RCTs?

What about RCTs?

- Post Double Lasso became standard in Development for choosing controls to include

- But... why do we include controls in RCTs to begin with?

- Reason 1: Power.
  - More covariates $\rightarrow$ smaller variance of residuals $\rightarrow$ smaller SEs
  - But irrelevant covariates can **increase** the variance! PDL picks only "relevant" covariates

- Reason 2: Robustness.
  - If randomization carried well, covariates should not change results!

- Reason 3: Correct for chance associations
  - If RCT not perfectly balanced, covariates can control for imbalance and potentially correct for "in-sample bias"

Post Double Lasso in RCTs

- Why use Post Double Lasso to choose covariates in RCT?

- Automatic: using default choice of λ, reduces possibility of p-hacking

- But: note that in general first stage (treatment on covariates) never selects with RCT

- Double Lasso is often equivalent to just Lasso in well-randomized RCTs

Thank you!