ML for Causal Inference

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Motivation

Motivation

- In this course our objective is prediction

Motivation

Motivation

- In this course our objective is prediction
- But since we are economists, inference is always there

Motivation

Motivation

- In this course our objective is prediction
- But since we are economists, inference is always there
- Can we use some of these models to do causal inference?

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Treatment Effects Review

- We observe a sequence of triples $\{(W_i, Y_i, X_i)\}_i^N$, where

Treatment Effects Review

Unfortunately, in our data we can only observe one of these two potential outcomes.

Education (X_i)	Treated W_i	No Subsidy $Y_i(0)$	Subsidy $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
High	1	?	$Y_1(1)$	?
High	0	$Y_2(0)$	?	?
Low	0	$Y_3(0)$	?	?
Low	1	?	$Y_4(1)$	?

Treatment Effects Review

- Before proceeding we need to make a couple of assumptions
- Assumption 1: Unconfoundedness

$$Y_i(1), Y_i(0) \perp W_i \mid X_i \tag{1}$$

- Assumption 2: Overlap

$$\forall\, x \in \text{supp}\,(X), \qquad 0 < P\,(W = 1 \mid X = x) < 1 \tag{2}$$

Average Treatment Effects Review

- Computing the difference for each individual is impossible.
- But we can get the Average Treatment Effect (ATE):

$$\tau := E[Y_i(1) - Y_i(0)] \tag{3}$$

- When our above assumptions are true we have:
- Conditional Average Treatment Effect (CATE)

$$\tau(x) := E[Y_i(1) - Y_i(0)|X_i = x] \tag{4}$$

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Model Selection When the Goal is Causal Inference

Let's start with the following model

$$y_i = \alpha + \beta W_i + g(X_i) + \zeta_i \tag{5}$$

were

- W_i is the treatment/policy variable of interest,
- X_i is a set controls
- $E[\zeta_i | W_i, X_i] = 0$

Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects X_i

Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects X_i
- Problem: mistakes can occur.
- Same if they use an "automatic" model selection approach.
- Why?

Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects X_i
- Problem: mistakes can occur.
- Same if they use an "automatic" model selection approach.
- Why?
- It can leave out potentially important variables with small coefficients but non zero coefficients out

Model Selection When the Goal is Causal Inference

- The omission of such variables then generally contaminates estimation and inference results based on the selected set of variables. (e.g. OVB)
- The validity of this approach is delicate because it relies on perfect model selection.
- Because model selection mistakes seem inevitable in realistic settings, it is important to develop inference procedures that are robust to such mistakes.
- Solution here: Lasso

Model Selection When the Goal is Causal Inference

- Using Lasso is useful for prediction
- However, naively using Lasso to draw inferences about model parameters can be problematic.
- Part of the difficulty is that these procedures are designed for prediction, not for inference
- Leeb and Pötscher 2008 show that methods that tend to do a good job at prediction can lead to incorrect conclusions when inference is the main objective
- This observation suggests that more desirable inference properties may be obtained if one focuses on model selection over the predictive parts of the economic problem

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Approximate sparse models

- To fix ideas suppose we have the following model and we want to predict y based on X

$$y_i = g(X_i) + \zeta_i \tag{6}$$

with

  - $E(\zeta_i | g(x_i)) = 0$
  - $i = 1, \ldots, n$ are iid
  - To avoid over-fitting and produce good out of sample prediction we will need to regularize $g(.)$
  - Belloni's et. all approach focuses on an approach that treats $g(X_i)$ as a high-dimensional but that we can approximate linearly

Approximate sparse models

$$g(X_i) = \sum_{j=1}^{p} \beta_j x_{ij} + r_{pi} \tag{7}$$

- where $p >> n$ and r_{pi} is small enough
- Approximate sparsity of this high-dimensional linear model imposes the restriction that linear combinations of only s, where $s << n$; x_{ij} variables provide a good approximation to $g(X_i)$

Approximate sparse models

- We can use Lasso that is slightly modified

$$L(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=2}^{p} |\beta_j|\gamma_j \tag{8}$$

  - where $\lambda > 0$ is the penalty level chosen using Belloni, Chen, Chernozhukov, and Hansen (2012)
  - γ_j are *penalty loadings*
  - *penalty loadings* are chosen to insure equivariance of coefficient estimates to rescaling of x_{ij} and can also be chosen to address heteroskedasticity, clustering, and non-gaussian errors

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Inference with Selection among Many Controls

- Under the approximate sparse models assumption, we can write our model

$$y_i = \alpha + \beta W_i + X_i'\theta_y + r_{yi} + \zeta_i \tag{9}$$

- where $E[\zeta_i | W_i, x_i, r_{yi}] = 0$
- X_i is a p-dimensional vector with $p >> n$, but approximately sparse
- r_{yi} is an approximation error

Inference with Selection among Many Controls

- How do we proceed?

Inference with Selection among Many Controls

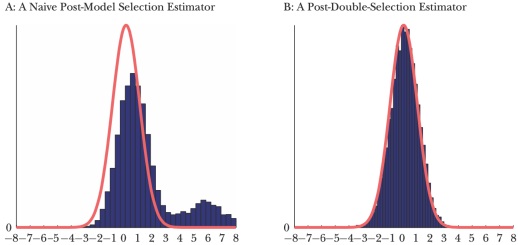
- To prevent model selection mistakes, it is important to consider both equations for selection.
- We apply variable selection methods to each of the two reduced form equations and then use all of the selected controls in estimation of α.
- We select
  1. A set of variables that are useful for predicting y_i, say X_{yi}, and
  2. A set of variables that are useful for predicting W_i, say X_{di}.
- We then estimate β by ordinary least squares regression of y_i on W_i and the union of the variables selected for predicting y_i and W_i, contained in X_{yi} and X_{di}.

Inference with Selection among Many Controls

Figure 1

The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)

(distributions of estimators from each approach)



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.

- We are making sure that we use variables that are important for either of the two predictive relationships to guard against OVB

Example: Green and Kern, 2012

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Heterogeneous Treatment Effects Review

- Detecting heterogeneous treatment effects, i.e., differential effects of an intervention for certain subgroups of the population, can be very valuable in many areas of research.
  - For example in drug trials, researchers might be interested in finding which subgroup of patients benefits most from getting a certain drug.
- Before conducting trials, researchers must often pre-register their analysis in order to prevent ex-post data-mining

Heterogeneous Treatment Effects Review

Concerns

- Issues:
  - Ad hoc searches for particularly responsive subgroups may mistake noise for a true treatment effect.
  - Concerns about ex-post "data-mining" or p-hacking
    - preregistered analysis plan can protect against claims of data mining
    - But may also prevent researchers from discovering unanticipated results and developing new hypotheses
  - But how is researcher to predict all forms of heterogeneity in an environment with many covariates?

Heterogeneous Treatment Effects Review

Concerns

- Pre-analysis plans help prevent ex-post data-mining
- However, researchers may find these plans restrictive since they cannot publish results for subgroups they did not anticipate before running the experiment.
- In such cases, having a data-driven way to find out where is the relevant heterogeneity that still produces usable estimates can be very convenient.
- This was one of the motivations for Athey and Imbens (2016)'s causal trees, a method for estimating heterogeneity in causal effects, and for conducting hypothesis tests about the magnitude of the differences in treatment effects across subsets of the population.

Agenda

1. Treatment Effects Review
   - Causality Review: ATE, CATE
2. Lasso for Causality
   - Approximate sparse models
   - Inference with Selection among Many Controls
3. Causal Trees
   - Causal Trees: Theory Details

Set up: Causal Trees

- Let's define

$$\mu(w, x) = E\left[Y_i(w)|X_i\right]$$

- The goal:

$$\tau(x|\Pi) = E\left[Y_i(1) - Y_i(0)|X_i \in l(x|\Pi)\right]$$

Set up: Causal Trees

- Let's define

$$\mu(w, x) = E\left[Y_i(w)|X_i\right]$$

- The goal:

$$\tau(x|\Pi) = E\left[Y_i(1) - Y_i(0)|X_i \in l(x|\Pi)\right]$$

- Approach 1:Analyze two groups separately, using machine learning
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i = 1$
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i = 0$

Set up: Causal Trees

- Let's define

$$\mu(w, x) = E\left[Y_i(w)|X_i\right]$$

- The goal:

$$\tau(x|\Pi) = E\left[Y_i(1) - Y_i(0)|X_i \in l(x|\Pi)\right]$$

- Approach 1:Analyze two groups separately, using machine learning
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i = 1$
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i = 0$
- Problem: Estimation and cross-validation not optimized for goal

Set up: Causal Trees

- Given a tree Π, define for all x and both treatment levels w the population average outcome

$$\mu(w, x|\Pi) = E\left[Y_i(w)|X_i \in l(x|\Pi)\right]$$

- The CATE

$$\tau(x|\Pi) = E\left[Y_i(1) - Y_i(0)|X_i \in l(x|\Pi)\right]$$

$$= \mu(1, x|\Pi) - \mu(0, x|\Pi)$$

Model and Estimation

- The objective now is to estimate the sample average treatment effect within leaf $\hat{tau}$
- a Simple tree

$$MSE_0 = \frac{1}{N}\sum(Y_i - \bar{Y})^2 \qquad \text{All observations}$$

$$MSE_1 = \frac{1}{N}\sum(Y_i - \bar{Y}_{j:x_j \in l(x_i|\Pi)})^2 \quad X_i < c_1 \quad X_i \geq c_2$$

- Partition $\Pi \in P$

$$\{l_1 = \{x_i : x_i < c_1\}, l_2 = \{x_i : x_i \geq c_2\}\} \tag{10}$$

- Prediction is

$$\hat{\mu}(x) = \bar{Y}_{j:x_j \in l(x_i|\Pi)} \tag{11}$$

Model and Estimation

- Given a partition Π the MSE

$$MSE_{\mu}(S^{te}, S^{est}, \Pi) = \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ \left(Y_i - \hat{\mu}(X_i, S^{est}, \Pi) \right)^2 - Y_i^2 \right\} \tag{12}$$

- The target is

$$max\, Q^C(\pi) = -E_{S^{te}, S^{tr}} \left[MSE_{\mu}(S^{te}, S^{tr}, \pi(S^{tr}) \right] \tag{13}$$

Causal Trees

- The MSE for treatment effects:

$$MSE_\tau(S^{te}, S^{est}, \Pi) = \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ \left(\tau_i - \hat{\tau}(X_i | S^{est}, \Pi\right)^2 - \tau_i^2 \right\}$$

- In the paper they show that you can estimate this

Honest Inference for Treatment Effects

- To ensure valid estimates of the treatment effect within each subgroup, Athey and Imbens propose a sample-splitting approach that they refer to as **honesty**
  - a method is honest if it uses one subset of the data to estimate the model parameters, and
  - a different subset to produce estimates given these estimated parameters.
- honesty implies that the asymptotic properties of treatment effect estimates within leaves are the same as if the tree partition had been exogenously given,
- It is one of the assumptions required to produce unbiased and asymptotically normal estimates of the treatment effect.

Example: Green and Kern, 2012