17.806 – Quantitative Research Methods IV

Lecture 5: Causal Inference with Machine Learning
Part I: Variable Selection for Observed Confounders

In Song Kim

MIT

Why ML for Causal Inference?

- Traditionally, machine learning (ML) only dealt with predictive inference
  - Many models are "black boxes"
  - Some literature on "interpretable" ML ("explainability"), but no formal connection to causality

- Recently, explosion of methods for causal inference that use ML

- How can ML improve causal inference, relative to traditional design-based (Quant II) or parametric model-based (Quant III) approaches?
  1. Incorporate a large number of observed confounders
  2. Avoid strong functional form assumptions
  3. Explore heterogenous causal effects at granular levels
  4. Estimate causal effects of high-dimensional treatments and their interactions

- Caution: Better prediction $\neq$ better causal inference
  - e.g. OLS: the best linear predictor of Y vs. unbiased estimator of β under which assumptions? (Quant I)
  - Naïve regularization can lead to poor estimates of causal QoIs

Conditional Ignorability with High-Dim. Confounders

Consider causal inference under conditional ignorability:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i,$$

where

- $D_i \in \{0, 1\}$: Binary treatment
- $Y_i(d)$: Potential outcomes under treatment d
- X_i: A high-dimensional observed covariate vector, with length $p >> n$
  - Usual advice (Quant II): Control for all observed pre-treatment confounders

For simplicity, assume the DGP:

$$Y_i(d) \; = \; d\tau + X_i^\top \beta + \epsilon_i,$$

where $\mathbb{E}[\epsilon_i \mid X_i, D_i = d] = 0$ for $d \in \{0, 1\}$. τ is the causal estimand (=ATE).

- Results generalize to heterogenous effects & nonlinear models with minor modifications
- X_i could include high-order polynomials, interactions, etc. for model flexibility

Naïve approaches and why they don't work

$$Y_i(d) \;=\; d\tau + X_i^\top \beta + \epsilon_i, \text{ where } \mathbb{E}[\epsilon_i \mid X_i, d] = 0$$

- Problem: $p >> n$
  - If p is small relative to n, OLS works just fine (Quant I and II)
  - If p is large, OLS might suffer from overfitting, or even unestimable (singular model matrix)
- Key difference from non-causal ML: All we care is τ, β is just nuisance
- Possible solutions?
  1. Naïve LASSO: Regress Y_i on D_i and X_i with LASSO ($L1$) penalty on all coefficients
     $\rightarrow$ Bad idea, because we want to preserve τ
  2. LASSO (on controls only): Regress Y_i on D_i and X_i with $L1$ penalty only on $X_i^\top \beta$
  3. Post-LASSO: Use LASSO as a 1st-stage variable selection step
     1. Regress Y_i on D_i and X_i with $L1$ penalty on $X_i^\top \beta$
        $\rightarrow$ obtain estimates for β ($= \hat{\beta}^L$)
     2. OLS Y_i on D_i and the subset of X_i for which $\hat{\beta}^L \neq 0$
        $\rightarrow$ obtain $\hat{\tau}$

Regularization Bias

Unfortunately neighther LASSO nor Post-LASSO works; they produce asymptotically biased estimates of τ:

Theorem (Regularization Bias (Belloni, Chernozhukov & Hansen))

Let $\hat{\tau}$ be the estimate of τ obtained via LASSO or Post-LASSO defined above. Then:

$$\sqrt{n}|\hat{\tau} - \tau| \;\rightarrow\; \infty \quad \textit{as} \quad n \rightarrow \infty.$$

Why?

- Recall X_i is a confounder if it affects *both* Y_i *and* D_i.
- However, LASSO selects X_i to keep based on its association with Y_i only.
- Therefore, LASSO tends to miss X_i that has a moderate effect on Y_i but a strong effect on D_i.

$\rightarrow$ Regularization bias = a form of omitted variables bias!

Double Selection

Solution: Select X_i based on both effects on Y_i and on D_i

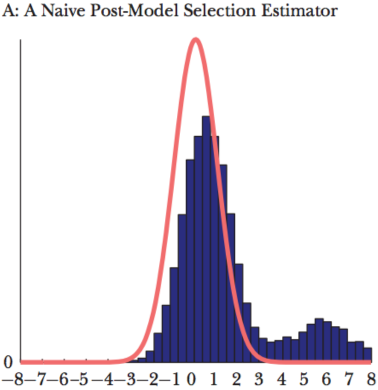
$$\begin{aligned} Y_i(d) &= d\tau + X_i^\top \beta + \epsilon_i, \text{ where } \mathbb{E}[\epsilon_i \mid X_i, d] = 0 \\ D_i &= X_i^\top \zeta + \nu_i, \text{ where } \mathbb{E}[\nu_i \mid X_i] = 0 \end{aligned}$$

The double selection method (Belloni, Chernozhukov & Hansen):

1. (Treatment selection): Regress D_i on X_i with L1 penalty $\rightarrow$ obtain $\hat{\zeta}^L$ $\rightarrow$ define $\hat{S}_D \equiv \{j = 1, ..., p : \hat{\zeta}^L \neq 0\}$.
2. (Outcome selection): Regress Y_i on X_i with L1 penalty $\rightarrow$ obtain $\hat{\beta}^L$ $\rightarrow$ define $\hat{S}_Y \equiv \{j = 1, ..., p : \hat{\beta}^L \neq 0\}$.
3. Define $\hat{S} \equiv \hat{S}_D \cup \hat{S}_Y$ (the union of the two selected covariate sets)
4. (Estimation) Regress Y_i on D_i and the subset of X_i in $\hat{S}$ via OLS $\rightarrow$ obtain $\hat{\tau}$.

Result: $\hat{\tau}$ from Step 4 is root-n consistent and asymptotically normal.

Sampling Distribution of $\hat{\tau}$ with Post-LASSO vs. Double Selection



Left panel: results based on applying LASSO only on the outcome model

Right panel: Double selection

Variance Estimation

- Note that Step 4 can also be conducted with the partialing out procedure:
  1. Regress D_i on the subset of X_i in $\hat{S}$ $\rightarrow$ save residuals $\hat{\nu}_i$
  2. Regress Y_i on the subset of X_i in $\hat{S}$ $\rightarrow$ save residuals $\hat{\epsilon}_i$
  3. Regress $\hat{\epsilon}_i$ on $\hat{\nu}_i$ $\rightarrow$ obtain $\hat{\tau}$

  This implementation is more convenient for variance estimation because of the following result:

$$\mathrm{AVar}(\hat{\tau}) \;=\; \frac{\mathbb{E}[\nu_i^2 \epsilon_i^2]}{\mathbb{E}[\nu_i^2]^2},$$

  which can be consistently estimated by a plug-in estimator:

$$\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\nu}_i^2\right)^{-2} \frac{\sum_{i=1}^{n} \hat{\nu}_i^2 \hat{\epsilon}_i^2}{n - \hat{s} - 1},$$

  where s is the number of variables in $\hat{S}$.

Example: Effect of Legalized Abortion on Crime

- Replication of Donohue and Levitt (2001), who argue that legalization of abortion reduced crime. Uses timing of legalization across states for identification.
- Original paper uses state and year fixed effects and a broad range of time varying variables
- Belloni et al use double selection to choose among 284 time-varying variables (with 600 observations):
  - Levels, differences, initial level, initial difference, and within-state average of 8 state-specific variables.
  - Initial level and initial difference of the abortion rate.
  - Quadratics in all the above variables.
  - Interactions and main effects of all the above variables with t (time trend) and t^2 (e.g., initial income squared $\times t^2$)

Example: Effect of Legalized Abortion on Crime

Effect of Abortion on Crime

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

Assumptions for the Double Selection Method

- Double selection eliminates bias, but it requires fairly strong assumptions:
  1. Approximate sparsity: Most of the β need to be 0 in the true DGP
  2. "High quality" machine learning method for estimating β/selecting X
     1. Small asymptotic bias – bias vanishes fast enough as $n \to \infty$
        $\to$ satisfied for common methods, such as LASSO
     2. No overfitting – model complexity does not grow too fast as $n \to \infty$
        $\to$ NOT generally satisfied for most methods, such as vanilla LASSO
  3. Other (much milder) regularity conditions (see original article)
- Key: In the covariate selection context, overfitting causes bias in causal effect estimate
- Chernozhukov et al. (2018):
  - Cross-fitting to elimiate overfitting
  - Generalize method to a wider set of ML methods, including non-sparse models
  - Implementation in R and Python: `DoubleML`

Overfitting Bias

Chernozhukov et al. show the double selection estimator can be expressed as:

$$\sqrt{n}(\hat{\tau} - \tau) \ = \ a^\star + b^\star + c^\star,$$

where

- $a^\star = \frac{1}{\mathbb{E}[\nu_i^2]} \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \nu_i$
  $\xrightarrow{d}$ normal with mean 0 (per CLT)
- $b^\star = \frac{1}{\mathbb{E}[\nu_i^2]} \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^\top(\hat{\beta} - \beta) X_i^\top(\hat{\zeta} - \zeta)$
  $\rightarrow 0$ faster than $a^\star$ (captures the remaining regularization bias)
- $c^\star = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nu_i X_i(\hat{\beta} - \beta) +$ (other similar terms)
  — no guarantee to vanish because $\mathrm{Cov}(\nu_i, \hat{\beta} - \beta) \neq 0$ (overfitting bias)

Intuition: Model errors, e.g., for D_i (ν_i), are associated with estimation errors for β because observation i is used for constructing $\hat{\beta}$

Cross-fitting for Eliminating Overfitting Bias

- To eliminate the $c^\star$ term, we can split samples into K groups and, for each $j \in \{1, ..., K\}$,
  1. Use observations not in group j to estimate ζ and β
  2. Use observations in group j to construct residuals $\hat{\nu}_i$ and $\hat{\epsilon}_i$
  3. Regress $\hat{\epsilon}_i$ on $\hat{\nu}_i$ to obtain $\hat{\tau}_j$

- The $c^\star$ term for the sample splitting estimator for group j is then

$$c_j^\star = \frac{1}{\sqrt{n_j}} \sum_{i \in \text{group j}} \nu_i X_i(\hat{\beta} - \beta),$$

  where $\hat{\beta}$ is now independent of the observations used for $c_j^\star$, guaranteeing $\xrightarrow{p} 0$.

- Iterate over $j \in \{1, ..., K\}$ and combine the K estimates for full efficiency:

$$\hat{\tau}_{DML} = \frac{1}{K} \sum_{j=1}^{K} \hat{\tau}_j$$

  This is (a special case of) the cross-fitted, double (or debiased) machine learning (DML) estimator.

Example: Effect of 401(k) Pension Plan

Data: 1991 Survey of Income and Program Participation:

- Y_i: Net total financial assets
- D_i: Working at a firm that offirs a 401(k) pension plan
- X_i: Age, income, family size, education, marriage, two-earner, defined benefit pension, IRA participation, home ownership, etc.

Assumption: D_i is conditionally ignorable at the time of initial 401(k) introduction

Models used:

1. Partially linear model:

$$\begin{aligned} Y_i &= D_i\tau + g(X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i, D_i] = 0, \\ D_i &= m(X_i) + \nu_i, \quad \mathbb{E}[\nu_i | X_i] = 0 \end{aligned}$$

2. Interactive model:

$$\begin{aligned} Y_i &= g(D_i, X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i, D_i] = 0, \\ D_i &= m(X_i) + \nu_i, \quad \mathbb{E}[\nu_i | X_i] = 0, \end{aligned}$$

   where $\tau = \mathbb{E}[g(1, X_i) - g(0, X_i)]$.

Result: Effect of 401(k) Pension Plan

Table 2. Estimated effect of 401(k) eligibility on net financial assets.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
Panel A: interactive regression model							
ATE	6830	7713	7770	7806	7764	7702	7546
(twofold)	[1282]	[1208]	[1276]	[1159]	[1328]	[1149]	[1360]
	(1530)	(1271)	(1363)	(1202)	(1468)	(1170)	(1533)
ATE	7170	7993	8105	7713	7788	7839	7753
(fivefold)	[1201]	[1198]	[1242]	[1155]	[1238]	[1134]	[1237]
	(1398)	(1236)	(1299)	(1177)	(1293)	(1148)	(1294)
Panel B: partially linear regression model							
ATE	7717	8709	9116	8759	8950	9010	9125
(twofold)	[1346]	[1363]	[1302]	[1339]	[1335]	[1309]	[1304]
	(1749)	(1427)	(1377)	(1382)	(1408)	(1344)	(1357)
ATE	8187	8871	9247	9110	9038	9166	9215
(fivefold)	[1298]	[1358]	[1295]	[1314]	[1322]	[1299]	[1294]
	(1558)	(1418)	(1328)	(1328)	(1355)	(1310)	(1312)

- Naïve diff in means: \$19,559 with SE $= 1413$
- Results largely robust across different ML methods

Inference with Selection among Many instruments

- The general results of Chernozhukov et al. hold for a broad class of estimating equations for β
- An important special case is instrumental variables design:

$$
\begin{aligned}
Y_i(d) &= d\tau + X_i^\top \beta + \epsilon_i, \text{ where } \mathbb{E}[\epsilon_i \mid X_i, Z_i] = 0 \\
D_i &= Z_i^\top \delta + X_i^\top \zeta + \nu_i, \text{ where } \mathbb{E}[\nu_i \mid X_i, Z_i] = 0 \\
Z_i &= X_i^\top \xi + \eta_i, \text{ where } \mathbb{E}[\eta_i \mid X_i] = 0
\end{aligned}
$$

A DML estimator for τ is, for each $j \in \{1, ..., K\}$,

1. Regress D_i on X_i and Z_i with L1 penalty $\rightarrow$ obtain $\hat{\zeta}$ and $\hat{\delta}$ $\rightarrow$ retain fitted values $\hat{D}_i = X_i^\top \hat{\zeta} + Z_i^\top \hat{\delta}$
2. Regress Y_i on X_i with L1 penalty $\rightarrow$ obtain $\hat{\beta}$
3. Regress $\hat{D}_i$ on X_i with L1 penalty $\rightarrow$ obtain coefficients $\hat{\phi}$
4. Obtain residuals from <u>non-j</u> observations for each step: $\hat{\nu}_i = D_i - \hat{D}_i$, $\hat{\theta}_i = Y_i - X_i^\top \hat{\beta}$, and $\hat{\omega}_i = \hat{D}_i - X_i^\top \hat{\phi}$.
5. Run 2SLS with $\hat{\theta}_i$ as the outcome, $\hat{\nu}_i$ as the treatment, $\hat{\omega}_i$ as instrument.

Example: Effect of 401(k) Pension Plan

Table 3. Estimated effect of 401(k) participation on net financial assets.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
LATE	8978	11073	11384	11329	11094	11119	10952
(twofold)	[2192]	[1749]	[1832]	[1666]	[1903]	[1653]	[1657]
	(3014)	(1849)	(1993)	(1718)	(2098)	(1689)	(1699)
LATE	8944	11459	11764	11133	11186	11173	11113
(fivefold)	[2259]	[1717]	[1788]	[1661]	[1795]	[1641]	[1645]
	(3307)	(1786)	(1893)	(1710)	(1890)	(1678)	(1675)

- Original estimate with linear IV (Poterba et al. 1994): \$13,102 with SE $= 1922$

Example: Acemoglu, Johnson and Robinson

Table 4. Estimated effect of institutions on output.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
Twofold	0.85	0.81	0.84	0.77	0.94	0.80	0.83
	[0.28]	[0.42]	[0.38]	[0.33]	[0.32]	[0.35]	[0.34]
	(0.22)	(0.29)	(0.30)	(0.27)	(0.28)	(0.30)	(0.29)
Fivefold	0.77	0.95	0.90	0.73	1.00	0.83	0.88
	[0.24]	[0.46]	[0.41]	[0.33]	[0.33]	[0.37]	[0.41]
	(0.17)	(0.45)	(0.40)	(0.27)	(0.30)	(0.34)	(0.39)

- Original estimate with linear IV: 1.10 with SE $= 0.46$
- Only $N = 64$, so take with a grain of salt

Efficient Doubly-Robust Estimators

In Quant II, we learned the "doubly-robust" estimator (Robins *et al.*)

$$\hat{\tau}_{DR} \equiv \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}(1, X_i) + \frac{1}{N} \sum_{i=1}^{N} \frac{T_i(Y_i - \hat{\mu}(1, X_i))}{\hat{\pi}(X_i)} \right\} - \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}(0, X_i) + \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - T_i)(Y_i - \hat{\mu}(0, X_i))}{1 - \hat{\pi}(X_i)} \right\}$$

- Consistent if either the propensity score model or the outcome model is correct (semi-parametrically efficient when the propensity score model is correct)
- In high-dimension settings, we can estimate β and $\pi(X_i)$ using LASSO or post-LASSO instead of OLS
- However, the assumption of sparse propensity score model might be too strong (i.e., if treatment assignment is a complex function of confounders)
- Note: very unstable with extreme values of $\hat{\pi}$

Difficulty in Estimating Propensity Scores

- Note that we need to weight observations with the inverse of $\hat{\pi}$: unstable with extrememe values
- Estimating the "right" propensity score $\pi(X_i)$ is difficult especially in high-dimension (and $\hat{\pi}(X_i)$ is often ≈ 0)
- Residual balancing methods directly calculates $\gamma = \frac{1}{\hat{\pi}(X_i)}$
  - Intuition is the same as Horvitz-Thompson estimator (1952. *J. Am. Stat. Assoc.*): weight each observation in the control group such that it looks like the treatment group (i.e., good covariate balance)
  - relaxes the assumption of sparsity of propensity score model (but still maintains the sparsity of the outcome model)

"Residual Balancing" (Athey, Imbens and Wager, 2018)

1. Compute positive approximately balancing weights γ:

$$\gamma \quad = \quad \arg\min_{\tilde{\gamma}} \left\{ (1-\zeta)||\tilde{\gamma}||_2^2 + \zeta||\overline{X_t} - \mathbf{X}_c^\top \tilde{\gamma}||_\infty^2 \right\}$$

   directly finding weights γ (rather than estimating propensity scores and using $1/\hat{\pi}(X_i)$) such that it balances treated and control group ($\zeta = 0.5$ by default). This is a quadratic programming problem! (L-infinity-norm is defined as $||\mathbf{x}||_\infty = \max_k ||x_k||$)

2. Estimate β_c using a LASSO (or Ridge or Elastic net) on the control group ($\alpha = 0.9$ by default)

$$\hat{\beta}_c = \arg\min_\beta \left\{ \sum_{\{i:T_i=0\}} \left(Y_i^{\mathrm{obs}} - X_i\beta \right)^2 + \lambda \left((1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \right) \right\} \qquad (1)$$

3. Estimate the conditional ATT (adjusting with the weighted average of the residuals):

$$\hat{\tau} = \overline{Y}_t - \left(\overbrace{\overline{X}_t \cdot \hat{\beta}_c}^{\text{counterfactual prediction}} + \underbrace{\sum_{\{i:T_i=0\}} \gamma_i \left(Y_i^{\mathrm{obs}} - X_i \cdot \hat{\beta}_c \right)}_{\text{weighted average of the residuals}} \right)$$

You can implement this with balanceHD package in **R**