Causal Inference: Part II

Sukjin Han

University of Bristol

June 2024

Center for Data Science, Zhejiang University

Causal Inference: Roadmap for Part II

Machine learning (ML) for causal inference

- Issues of naive ML approach
- Double/debiased ML approach
- Neyman orthogonality
- Sample splitting

1. Example 1: Average treatment effects
3. Example 2: Partially linear models

Example 1: Estimating Average Treatment Effects

- Assume $Y_d \perp D|X$ for $d \in \{0,1\}$
  - Conditional independence
  - X is potentially high-dimensional
- Suppose $\theta_0 = E[Y_1 - Y_0]$
- By conditional independence,

$$\begin{aligned}\theta_0 &= E[E[Y|D=1,X] - E[Y|D=0,X]] \\ &= E[g_0(1,X) - g_0(0,X)]\end{aligned}$$

where $g_0(D,X) \equiv E[Y|D,X]$

Naive Approach: Plug-In

- Naive approach for estimation:
- Use ML to learn $g_0(1, X)$ and $g_0(0, X)$
- i.e., obtain $\hat{g}(1, X)$ and $\hat{g}(0, X)$
  - e.g., lasso, random forest, neural network
- Then, use a plug-in estimator:

$$\hat{\theta}_{plug} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{g}(1, X_i) - \hat{g}(0, X_i)\}$$

- The plug-in estimator is biased, inconsistent and not asymptotically normal
  - Even if predictive performance of $\hat{g}$ is superb!

Naive Plug-In Estimator

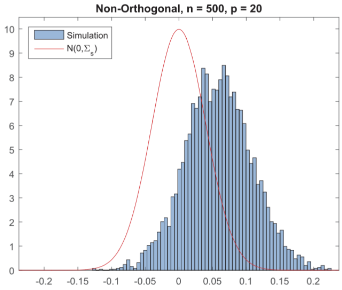

Figure: Bias of Plug-In Estimator of θ_0

Naive Plug-In Estimator

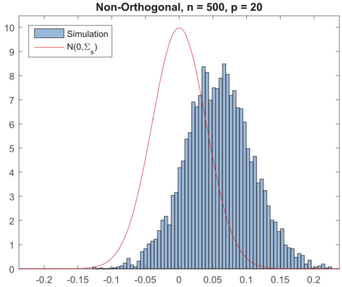

Figure: Bias of Plug-In Estimator of θ_0

- This is because bias and error in estimating g_0 influence $\hat{\theta}$
  - e.g., regularization bias
- Q: How to guarantee $\hat{\theta}$ is $\sqrt{n}$-asymptotically normal?
- Q: How to make estimation of θ_0 insensitive to variations in g?

Double ML Approach

- Suppose $g_0 \in G$ where G is space of sq-integrable functions
- Suppose $\exists \alpha_0 \in G$ such that

$$E[\alpha_0(D,X)g(D,X)] = E[g(1,X) - g(0,X)] \quad \forall g \in G \quad (1)$$

  - Existence of α_0 by Riesz representation theorem
  - α_0 is the Riesz representer
- Then, by (1)

$$\begin{aligned}
\theta_0 &= E[g_0(1,X) - g_0(0,X)] \\
&= E[\alpha_0(D,X)g_0(D,X)] \\
&= E[\alpha_0(D,X)E[Y|D,X]] \\
&= E[E[\alpha_0(D,X)Y|D,X]] \\
&= E[\alpha_0(D,X)Y]
\end{aligned}$$

  - Three different representations of θ_0 —(∗)

Double ML Approach

- What is α_0 in this case? With $P(X) \equiv P[D = 1|X]$,

$$\alpha_0(D, X) = \frac{D}{P(X)} - \frac{1 - D}{1 - P(X)}$$

  - Inverse probability weighting (IPW)

- This is because, e.g.,

$$\begin{aligned} E\left[\frac{D}{P(X)} g(D, X)\right] &= E\left[E\left[\frac{D}{P(X)} g(D, X)\middle| D\right]\right] \\ &= E\left[E\left[\frac{D}{P(X)} g(1, X)\middle| D\right]\right] \\ &= E\left[\frac{D}{P(X)} g(1, X)\right] \\ &= E\left[\frac{E[D|X]}{P(X)} g(1, X)\right] = E[g(1, X)] \end{aligned}$$

Double ML Approach

- Motivated from $(*)$, let

$$\begin{aligned}\theta_0 &= M(g_0, \alpha_0) \\ &\equiv E[g_0(1, X) - g_0(0, X)] + E[\alpha_0(D, X)(Y - g_0(D, X))]\end{aligned}$$

  - g_0 and α_0 are nuisance functions
  - $\alpha_0(D, X)(Y - g_0(D, X))$ is influence function adjustment

- Then,

$$\begin{aligned}\theta_0 &= M(g_0, \alpha_0) \\ &= M(g, \alpha_0) \quad \forall g \in G \\ &= M(g_0, \alpha) \quad \forall \alpha \in G\end{aligned}$$

Double ML Approach

- Also, when taking directional derivative w.r.t. nuisance functions in any direction $\nu \in G$,

  $$\frac{\partial}{\partial t} M(g_0 + t\nu, \alpha_0)|_{t=0}$$
  $$= E[\nu(1, X) - \nu(0, X)] - E[\alpha_0(D, X)\nu(D, X)] = 0$$

  by (1) and

  $$\frac{\partial}{\partial t} M(g_0, \alpha_0 + t\nu)|_{t=0}$$
  $$= E[\nu(D, X)(Y - g_0(D, X))]$$
  $$= E[E[\nu(D, X)(Y - g_0(D, X))|D, X]] = 0$$

  - Neyman orthogonality
- M is locally insensitive to either g or α
  - Double robustness

DML Estimation

- Using the DML formula,

$$\hat{\theta}_{DML} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{g}(1, X_i) - \hat{g}(0, X_i) + \hat{\alpha}(D_i, X_i)(Y_i - \hat{g}(D_i, X_i))\}$$

- Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{DML} - \theta_0) \rightsquigarrow N(0, \sigma^2)$$

- Bias in estimation of g and α does not transmit to estimation of θ (at least to the first order)
- Rate of convergence of $\hat{\alpha}$ and $\hat{g}$ only needs to be faster than $n^{-1/4}$ (more later)
  - This holds for most "simple" ML

Sample Splitting

- It is advised to split the sample
  1. Calculate $\hat{g}$ and $\hat{\alpha}$ using one sample
  2. Calculate $\hat{\theta}_{DML}$ using another sample
- This removes dependence between $(\hat{g}, \hat{\alpha})$ and $\hat{\theta}_{DML}$
  - Asymptotic normality is guaranteed under weaker conditions
  - i.e., remove bias induced by overfitting
- More generally, cross validation can be used to improve efficiency

Double ML Estimator

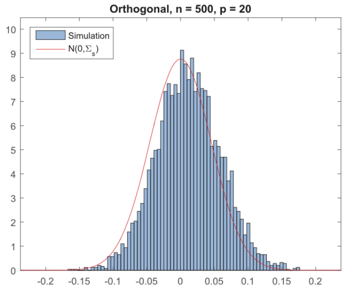

Figure: Bias of Plug-In Estimator of θ_0

More General Framework

- In general, suppose $g_0(X) \equiv E[Y|X]$ and

$$\theta_0 = E[m(Z; g_0)]$$

  - e.g., ATE (above) and average derivate ($\theta_0 = E[\partial g(D, X)/\partial D]$ with continuous D)

- Then

$$\theta_0 = M(g_0, \alpha_0) \equiv E[m(Z; g_0) + \alpha_0(X)(Y - g_0(X))]$$

  where $\alpha_0 \in G$ is Riesz representer s.t.

$$E[m(Z; g)] = E[\alpha_0(X)g(X)] \quad \forall g \in G$$

- Then

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\{m(Z_i; \hat{g}) + \hat{\alpha}(X_i)(Y_i - \hat{g}(X_i))\}$$

Example 2: Partially Linear Models

- Partially linear model with continuous D

$$Y = D\theta_0 + g_0(Z) + U, \quad E[U|Z, D] = 0$$

  - D: treatment; θ is parameter of interest
  - Z: high-dim covariates (i.e., "controls" or "measured confounders")

- Z are confounders in the sense that

$$D = m_0(Z) + V, \quad E[V|Z] = 0$$

Naive Approach: Prediction-Based ML

- Predict Y using D and Z and obtain

$$D\hat{\theta} + \hat{g}(Z)$$

  - e.g., Estimation using alternating minimization:
    1. Choose initial guess $\hat{\theta}$
    2. Run random forest of $Y - D\hat{\theta}$ on Z to fit $\hat{g}(Z)$
    3. Run OLS on $Y - \hat{g}(Z)$ on D to fit $\hat{\theta}$
    4. Repeat until convergence

- Again, excellent prediction performance but $\hat{\theta}$ is biased and not asymptotically normal

Double ML Approach

1. Predict Y and D using Z by $\widehat{E[Y|Z]}$ and $\widehat{E[D|Z]}$
2. Residualize $\hat{W} = Y - \widehat{E[Y|Z]}$ and $\hat{V} = D - \widehat{E[D|Z]}$
3. Regress $\hat{W}$ on $\hat{V}$ to get $\hat{\theta}_{DML}$

- Split sample between Step 1 and Step 2

- Then

$$\sqrt{n}(\hat{\theta}_{DML} - \theta_0) \rightsquigarrow N(0, \Sigma)$$

Moment Conditions

- Two approaches rely on different moment conditions:

$$E[(Y - D\theta_0 - g_0(Z))D] = 0 \quad (2)$$
$$E[(Y - D\theta_0)(D - E[D|Z])] = 0 \quad (3)$$
$$E\left[\{(Y - E[Y|Z]) - (D - E[D|Z])\theta_0\}(D - E[D|Z])\right] = 0 \quad (4)$$

  - (2): Regression adjustment
  - (3): Propensity score adjustment
  - (4): Neyman-orthogonal

- Both approaches generate estimators of θ_0 that solve the empirical analog of the moment conditions above...
  - after plugging in ML-based estimators for

$$g_0(Z), \quad m_0(Z) \equiv E[D|Z], \quad \ell_0(Z) \equiv E[Y|Z]$$

    using set-aside sample

Naive Approach from (2): Prediction-Based ML

- Suppose we use (2) with an estimator $\hat{g}(Z)$ to estimate θ_0:

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} D_i \left(Y_i - \hat{g}(Z_i) \right)$$

- Then

$$\sqrt{n}(\hat{\theta} - \theta_0) = A + B$$

where

$$A \equiv \left(\frac{1}{n} \sum_{i=1}^{n} D_i^2 \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i U_i$$

$$B \equiv \left(\frac{1}{n} \sum_{i=1}^{n} D_i^2 \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \left(g_0(Z_i) - \hat{g}(Z_i) \right)$$

- $A \rightsquigarrow N(0, \tilde{\Sigma})$ under standard conditions

Naive Approach from (2): Prediction-Based ML

- Generally, $B \to \infty$:

$$B \approx \left(ED^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(Z_i)\left(g_0(Z_i) - \hat{g}(Z_i)\right)$$

  - $g_0(Z_i) - \hat{g}(Z_i)$ is the error in estimating g_0
- Heuristics:
  - In nonparametric setting, the error is of order $n^{-\varphi}$ for $0 < \varphi < 1/2$
  - Then B will then look like $\sqrt{n}n^{-\varphi} \to \infty$
- Therefore, $\hat{\theta}$ is not $\sqrt{n}$-consistent
- Similar heuristics apply to estimation with (3)

Double ML Approach from (4)

- Suppose we use (4) to estimate θ_0:

$$\hat{\theta}_{DML} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{V}_i^2\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{V}_i\hat{W}_i$$

where $\hat{V} = D - \hat{m}(Z)$ and $\hat{W} = Y - \hat{\ell}(Z)$

- Under mild conditions, can write

$$\sqrt{n}(\hat{\theta} - \theta_0) = A^* + B^* + C^*$$

where $C^* = o_p(1)$ and

$$A^* \equiv \left(\frac{1}{n}\sum_{i=1}^{n}V_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_iU_i$$

$$B^* \equiv \left(\frac{1}{n}\sum_{i=1}^{n}V_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m_0(Z_i) - \hat{m}(Z_i)\right)\left(g_0(Z_i) - \hat{g}(Z_i)\right)$$

Double ML Approach from (4)

- $A^* \rightsquigarrow N(0, \Sigma)$ under standard conditions
- B^* now depends on product of estimation errors in both nuisance functions
- Then B^* will look like $\sqrt{n}n^{-(\varphi_m + \varphi_\ell)}$ where φ_m and φ_ℓ are convergence rates of $\hat{m}(Z)$ and $\hat{\ell}(Z)$, resp.
  - $o(n^{-1/4})$ is often attainable rate for ML estimators
- C^* contains terms like

$$\left(\frac{1}{n}\sum_{i=1}^{n} V_i^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \left(m_0(Z_i) - \hat{m}(Z_i)\right)$$

  - With sample splitting, easy to control and claim $o_p(1)$
  - Without sample splitting, hard to control and claim $o_p(1)$

Neyman Orthogonality of (4)

- Key difference between (2) and (4) is that (4) satisfies Neyman orthogonality condition:
- Let

$$\eta_0 \equiv (\ell_0, m_0) \equiv (E[Y|Z], E[D|Z]), \qquad \eta \equiv (\ell, m)$$

- The Gateaux derivative of (4) w.r.t. η vanishes:

$$\partial_\eta E\left[\{(Y - \ell(Z)) - (D - m(Z))\theta_0\}(D - m(Z))\right]\Big|_{\eta=\eta_0} = 0$$

  - i.e., the moment condition remains "valid" under "local" mistakes in the nuisance functions
- This property generally does not hold with (2) for nuisance function g

References

- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. rem, 12(1), 338.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. Econometrics Journal. 21(1):C1-C68.

Thank You! ☺