Double / Debiased Machine Learning for Causal Treatment Effects

Causal Machine Learning Seminar

Juian Cantor

Humboldt-Universität zu Berlin

December 1, 2025

The problem: causal effects with many covariates

- Observational data with many potential confounders $X \in \mathbb{R}^p$
- We want a low-dimensional causal parameter θ_0:
  - e.g. ATE, regression coefficient, IV effect
- Classical low-dimensional models: risk of misspecification when p is large
- Modern ML: excellent at predicting Y, but tuned for *prediction loss*, not for θ_0
- Question: How can we combine ML with econometric identification to get valid $\sqrt{n}$-inference for θ_0?

Prediction vs. causal estimation

- ML goal: approximate $x \mapsto \mathbb{E}[Y \mid X = x]$ (random forests, boosting, nets, ...)
- Causal goal: estimate functional $\theta_0 = \theta(P)$ defined by a *moment condition*

$$\mathbb{E}_P[\psi(W; \theta_0, \eta_0)] = 0, \qquad W = (Y, D, X).$$

- Naive idea: fit a flexible ML model for Y on (D, X) and "read off" the effect of D
- Problem (sketch): regularization bias in nuisance functions η_0 contaminates the estimator of θ_0 so that it *fails* to be $\sqrt{n}$-consistent
- We will make this precise using moment conditions and a bias decomposition.

Why econometricians love moment conditions

- Many causal/econometric parameters are defined by

$$\mathbb{E}\big[\psi(W;\theta_0,\eta_0)\big] = 0.$$

- Examples:
  - OLS: $\mathbb{E}[X(Y - X'\beta_0)] = 0$
  - IV: $\mathbb{E}[Z(Y - D\theta_0)] = 0$
  - GMM: stacked moment conditions $m(W;\theta_0) = 0$
- Advantages:
  - Clean separation between **target** θ_0 and **nuisance** η_0
  - Robustness tools (orthogonal / efficient scores) are defined at the level of moments
  - We can plug in ML estimators for η_0 and study the resulting bias in θ_0
- So we start from moment conditions for θ_0, not from a particular estimator.

Moment conditions in the partially linear model

- Partially linear regression (PLR):

$$Y = \theta_0 D + g_0(X) + U, \quad \mathbb{E}[U \mid D, X] = 0,$$
$$D = m_0(X) + V, \quad \mathbb{E}[V \mid X] = 0.$$

- Three key moment conditions for θ_0:
  1. **Regression adjustment**:
  $$\mathbb{E}\big[(Y - D\theta_0 - g_0(X))\, D\big] = 0.$$
  2. **Propensity-score-style adjustment**:
  $$\mathbb{E}\big[(Y - D\theta_0)\,(D - m_0(X))\big] = 0.$$
  3. **Neyman-orthogonal (residual) score**:
  $$\mathbb{E}\big[(Y - D\theta_0 - g_0(X))\,(D - m_0(X))\big] = 0.$$
- All three identify the same θ_0 *if* g_0, m_0 were known exactly.
- But they behave very differently once we replace g_0, m_0 by ML estimates.

Key idea (for applied work)

- **Practical takeaway:**
  - To safely combine ML with econometrics, we:
    1. Express θ_0 via a suitable moment condition,
    2. Choose a Neyman-orthogonal score (locally insensitive to nuisance errors),
    3. Estimate nuisance functions with ML + cross-fitting.
- This makes ML mistakes in η_0 second order for θ_0, so we keep $\sqrt{n}$-rates and valid asymptotic normality.
- Intuition: residualize Y and D with ML, then do a simple linear regression / IV-style step.
- Plan:
  - Methods: orthogonal score, bias decomposition, cross-fitting
  - Evidence: simulations + 401(k) application
  - Critique & checklist: when DML is (and is not) a good idea

Naive regression adjustment: bias decomposition (1/2)

- Start from regression-adjustment moment:

$$\mathbb{E}[(Y - D\theta_0 - g_0(X))D] = 0.$$

- Plug-in estimator (using sample splitting for clarity):

$$\hat{\theta}_{\mathrm{RA}} = \Big(\tfrac{1}{n}\sum D_i^2\Big)^{-1}\Big(\tfrac{1}{n}\sum D_i(Y_i - \hat{g}(X_i))\Big).$$

- Decompose

$$\sqrt{n}(\hat{\theta}_{\mathrm{RA}} - \theta_0) = A_n + B_n,$$

where

$$A_n = (E[D^2])^{-1}\frac{1}{\sqrt{n}}\sum D_i U_i,$$

$$B_n \approx (E[D^2])^{-1}\frac{1}{\sqrt{n}}\sum m_0(X_i)\big(g_0(X_i) - \hat{g}(X_i)\big).$$

Naive regression adjustment: bias decomposition (2/2)

- Recall the decomposition:

$$\sqrt{n}(\hat{\theta}_{\text{RA}} - \theta_0) = A_n + B_n,$$

- $A_n = (E[D^2])^{-1} \frac{1}{\sqrt{n}} \sum D_i U_i$ is well-behaved:
  - CLT applies $\Rightarrow$ asymptotically normal.
- $B_n \approx (E[D^2])^{-1} \frac{1}{\sqrt{n}} \sum m_0(X_i)\big(g_0(X_i) - \hat{g}(X_i)\big)$ is the regularization bias term:
  - With ML, $\|g_0 - \hat{g}\|_{L_2} = n^{-\varphi_g}$ with $\varphi_g < 1/2$ typically.
  - So B_n does *not* vanish at $\sqrt{n}$-scale.
- Conclusion: $\hat{\theta}_{\text{RA}}$ is generally not $\sqrt{n}$-consistent.
- Key insight: the bias enters linearly through $g_0 - \hat{g}$, making it first-order.

Orthogonal (residual) score: bias decomposition (1/2)

- Orthogonal (Neyman) score for PLR:

$$\psi(W;\theta,g,m) = (Y - \theta D - g(X))\,(D - m(X)).$$

- Define DML estimator $\tilde{\theta}$ as the solution of

$$\frac{1}{n}\sum \psi(W_i;\tilde{\theta},\hat{g},\hat{m}) = 0 \quad \Rightarrow \quad \tilde{\theta} = \frac{\frac{1}{n}\sum (D_i - \hat{m}(X_i))(Y_i - \hat{g}(X_i))}{\frac{1}{n}\sum (D_i - \hat{m}(X_i))^2}.$$

- Decompose

$$\sqrt{n}(\tilde{\theta} - \theta_0) = A_n^* + B_n^* + C_n^*.$$

- Leading term:

$$A_n^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum V_i U_i \quad \Rightarrow \quad A_n^* \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

Orthogonal (residual) score: bias decomposition (2/2)

- Recall the decomposition:

$$\sqrt{n}(\tilde{\theta} - \theta_0) = A_n^* + B_n^* + C_n^*.$$

- A_n^* is the well-behaved leading term (asymptotically normal).
- B_n^* is the bias term, but now it involves a *product* of nuisance errors:

$$B_n^* = (E[V^2])^{-1} \frac{1}{\sqrt{n}} \sum (\hat{m}(X_i) - m_0(X_i))(\hat{g}(X_i) - g_0(X_i)).$$

- If $\|\hat{m} - m_0\| = n^{-\varphi_m}$ and $\|\hat{g} - g_0\| = n^{-\varphi_g}$, then $B_n^* = O_p\big(\sqrt{n}\, n^{-(\varphi_m + \varphi_g)}\big)$.
- As soon as $\varphi_m + \varphi_g > 1/2$, we have $B_n^* = o_p(1)$.
- Key insight: ML can be slow (e.g., $\varphi_g = \varphi_m = 1/4$), yet bias is still negligible at $\sqrt{n}$-scale.
- This is the power of orthogonal scores: bias is now second-order.

Neyman orthogonality: definition (1/2)

- General setup: parameter θ_0 and nuisance η_0 solve

$$\mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0.$$

- **Neyman orthogonality**:

$$\partial_\eta \, \mathbb{E}[\psi(W; \theta_0, \eta)]\big|_{\eta=\eta_0}[\eta - \eta_0] = 0 \quad \text{for all admissible directions } \eta - \eta_0.$$

  (Gateaux derivative of the moment w.r.t. η vanishes at η_0.)
- Interpretation:
  - Moment condition is locally insensitive to small errors in η.
  - First-order bias from estimating η_0 cancels; remaining bias is second-order.

Neyman orthogonality: PLR example (2/2)

- Recall the orthogonal score for PLR:

$$\psi(W; \theta, g, m) = (Y - \theta D - g(X))(D - m(X)).$$

- At the true parameter values (θ_0, g_0, m_0), we have:

$$\mathbb{E}[\psi(W; \theta_0, g_0, m_0)] = 0.$$

- One can verify that this score satisfies Neyman orthogonality:

$$\partial_{(g,m)} \mathbb{E}[\psi(W; \theta_0, g, m)]\big|_{(g,m)=(g_0,m_0)} = 0.$$

- This is the key structural reason why DML works with high-dimensional ML.
- The orthogonality property ensures that errors in estimating (g_0, m_0) only affect the bias at second order.

Residual interpretation & IV-style view (1/2)

- Define residuals using ML nuisances:

$$\tilde{Y}_i := Y_i - \hat{g}(X_i), \qquad \tilde{D}_i := D_i - \hat{m}(X_i).$$

- Orthogonal moment is

$$\frac{1}{n}\sum \tilde{D}_i(\tilde{Y}_i - \theta\,\tilde{D}_i) = 0.$$

  Equivalently:

$$\tilde{Y}_i = \theta\,\tilde{D}_i + \text{error}, \qquad \mathbb{E}[\tilde{D}_i \cdot \text{error}] = 0.$$

- Estimator:

$$\tilde{\theta} = \frac{\sum \tilde{D}_i \tilde{Y}_i}{\sum \tilde{D}_i^2}.$$

- This is simply OLS of residualized outcome on residualized treatment.

Residual interpretation & IV-style view (2/2)

- Recall the DML estimator:

$$\tilde{\theta} = \frac{\sum \tilde{D}_i \tilde{Y}_i}{\sum \tilde{D}_i^2}, \quad \tilde{Y}_i = Y_i - \hat{g}(X_i), \quad \tilde{D}_i = D_i - \hat{m}(X_i).$$

- **IV-style intuition**:
  - Think of $\tilde{D}_i = D_i - \hat{m}(X_i)$ as an "optimal" instrument for D.
  - We regress the residualized outcome $\tilde{Y}$ on $\tilde{D}$ using this instrument.
  - The variation in $\tilde{D}_i$ is orthogonal to X by construction.
- This bridges two worlds:
  - **ML**: flexible, high-dimensional residualization via $\hat{g}$ and $\hat{m}$
  - **Econometrics**: IV / GMM inference on a low-dimensional parameter θ_0
- Key insight: ML handles the complex confounding structure; econometric theory protects the causal parameter.

Sample splitting and cross-fitting

- Remaining issue: if we estimate $\hat{g}$, $\hat{m}$ and $\tilde{\theta}$ on the *same* data, overfitting can create extra terms like
$$\frac{1}{\sqrt{n}} \sum V_i(\hat{g}(X_i) - g_0(X_i)),$$
which may not vanish.
- Solution: **sample splitting + cross-fitting**.
  - Split data into K folds.
  - On each fold k, estimate $(\hat{g}^{(-k)}, \hat{m}^{(-k)})$ using all other folds.
  - Compute residuals and $\tilde{\theta}^{(k)}$ on fold k only.
  - Average over k.
- Conditional on the training folds, nuisance errors and score residuals are nearly independent; remainder terms are controlled by simple variance bounds.
- Cross-fitting also recovers efficiency (we use the whole sample for θ_0).

Algorithm for PLR DML (1/2)

1. Choose ML methods for $g_0(X)$ and $m_0(X)$:
   - e.g. lasso, random forest, boosting, neural nets, ensembles
2. Fix number of folds K (e.g. $K = 5$).
3. Randomly split $\{1, \ldots, n\}$ into folds $I_1, \ldots, I_K$ of (roughly) equal size.
4. For each fold $k = 1, \ldots, K$:
   4.1 Train $\hat{g}^{(-k)}, \hat{m}^{(-k)}$ on data $\{i \notin I_k\}$.
   4.2 For each $i \in I_k$, compute

$$\tilde{Y}_i := Y_i - \hat{g}^{(-k)}(X_i), \quad \tilde{D}_i := D_i - \hat{m}^{(-k)}(X_i).$$

Algorithm for PLR DML (2/2)

4. For each fold k:
   4.1 On $i \in I_k$, run OLS of $\tilde{Y}_i$ on $\tilde{D}_i$ (no intercept):

$$\hat{\theta}^{(k)} := \frac{\sum_{i \in I_k} \tilde{D}_i \tilde{Y}_i}{\sum_{i \in I_k} \tilde{D}_i^2}.$$

5. Aggregate:

$$\hat{\theta}_{\text{DML}} := \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)}.$$

6. Estimate asymptotic variance using the empirical influence function, and form CIs in the usual way.

What double ML delivers (high level)

- Under mild rate conditions on ML nuisances (e.g. $\|\hat{g} - g_0\|, \|\hat{m} - m_0\| = o_p(1)$ and $\varphi_g + \varphi_m > 1/2$):
  - $\hat{\theta}_{\text{DML}}$ is $\sqrt{n}$-consistent and asymptotically normal
  - Standard Wald CIs are valid
- Works with a wide range of ML:
  - Lasso, random forests, boosting, neural nets, generalized additive models, ensembles
- In PLR and several other settings, the DML estimator is semiparametrically efficient (under homoscedasticity).
- The same pattern generalizes:
  - Define an orthogonal score for your parameter
  - Estimate nuisances with ML + cross-fitting
  - Solve empirical orthogonal moment for θ_0

Simulation: prediction vs. causal estimation

- Design (Chernozhukov et al.):
  - $g_0(X)$ built from trees $\Rightarrow$ random forest is (nearly) oracle for prediction
  - Treatment equation $D = m_0(X) + V$ with nontrivial confounding
- Compare two estimators of θ_0:
  1. Naive ML regression-adjustment (plug-in $\hat{g}$)
  2. DML estimator based on orthogonal score + cross-fitting
- Findings:
  - Naive plug-in: excellent out-of-sample prediction of Y, but biased for θ_0, distribution shifted away from the truth
  - DML: distribution of $\hat{\theta}_{\text{DML}}$ is centered at θ_0 and close to normal
- Moral: good prediction error is *not* enough; the score must be orthogonal to ML errors.

Application: 401(k) eligibility and savings

- Data: US households, outcome Y = net financial assets; treatment D = 401(k) eligibility.
- High-dimensional covariates X (income, age, education, family status, etc.).
- Identification: treat eligibility as (conditionally) exogenous given X.
- DML for the ATE of eligibility:
  - $g_0(1, X), g_0(0, X)$ estimated with flexible ML
  - Orthogonal score for ATE, ML nuisances cross-fitted
- Result (roughly): substantial positive effect ($\sim$ \$7–9k), robust across many ML learners.
- Demonstrates: we can combine rich ML adjustment for X with standard errors for the causal effect.

Beyond PLR: ATE, ATTE, PLIV, LATE

- **ATE / ATTE under unconfoundedness**:
  - Orthogonal scores combining outcome regression and propensity score
  - Same pattern: residuals + cross-fitting
- **Partially linear IV model**:

$$Y = \theta_0 D + g_0(X) + U, \quad \mathbb{E}[U \mid X, Z] = 0,$$
$$D = m_0(X, Z) + V, \quad \mathbb{E}[V \mid X, Z] = 0.$$

  - Z affects Y only via D given X (exclusion)
  - DML uses an orthogonal score involving $(Y - \theta D - g_0(X))(Z - m_0(X))$
- **LATE with binary** D, Z:
  - Orthogonal scores for Wald-type estimands, with ML nuisance functions for m_0, p_0, μ_0.
- In all cases: low-dimensional θ_0, high-dimensional η_0, orthogonal score + ML + cross-fitting.

What DML does *not* fix

- DML relaxes *functional-form* assumptions but assumes:
  - Correct causal graph / identification:
    - Unconfoundedness with respect to observed X, or
    - Valid instruments Z with exclusion restrictions
  - Overlap / positivity (no extreme propensity scores)
  - No bad controls (no colliders or mediators in X)
- If key confounders are unobserved, DML cannot repair the resulting bias.
- ML cannot substitute for research design: careful variable selection, timing, instruments, etc. still crucial.

Evidence from method evaluation studies

- Method evaluations (e.g. simulation studies comparing OLS, naive ML, and DML) highlight:
  - With **linear** confounding, OLS and DML often perform similarly.
  - With **nonlinear** confounding (interactions, thresholds):
    - DML with very rigid learners (e.g. plain lasso on raw X) can inherit OLS-like bias.
    - DML with flexible learners (RF, boosting, GAMs, nets) can drastically reduce bias.
  - ML choice inside DML matters a lot for finite-sample bias–variance trade-offs.
- Practical message: the "D" in DML does *not* automatically guarantee good performance; you still need sensible ML.

Checklist & take-home message

- **When is DML a good fit?**
  - Many covariates, complex but plausibly learnable nuisance structure
  - Target is low-dimensional (ATE, PLR coefficient, IV effect, LATE)
  - Sample size large enough to support flexible ML
- **What you should actually do in applied work:**
  1. Write down the causal estimand and its moment condition.
  2. Derive (or look up) an orthogonal score for that estimand.
  3. Choose ML methods rich enough to capture plausible nonlinearities.
  4. Use cross-fitting; report sensitivity to choice of learner and K.
- One-sentence takeaway:
  - Use ML to learn the nuisance parts, but protect the causal parameter with orthogonal moments and cross-fitting so that ML errors only show up as second-order bias.