Script: Double / Debiased Machine Learning for Causal Treatment Effects

Audience: MSc Statistics students with graduate-level econometrics
Duration: ∼25 minutes

Slide 1 — Title
"Hello everyone. Today I'll talk about *Double / Debiased Machine Learning for Causal Treatment Effects*.

This is based on the paper by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins. The goal is to show how to safely combine modern machine learning with classical econometric ideas to estimate causal effects."

(Briefly introduce yourself and the seminar context.)

Slide 2 — The problem: causal effects with many covariates
"Let me start with the basic problem.

We have observational data, a treatment or policy variable (D), an outcome (Y), and a big vector of potential confounders ($X \in \mathbb{R}^p$).

We're interested in a low-dimensional causal parameter θ_0: for example an average treatment effect, or a regression coefficient with a causal interpretation, or an IV effect.

In low dimensions, we'd specify a parametric model and use OLS or IV. But when the dimension of (X) is large, those parametric models are easy to misspecify.

On the other hand, modern ML methods are extremely good at predicting (Y) from ((D, X)), but they are tuned to minimize prediction loss, not to give unbiased and $\sqrt{n}$-consistent estimators of θ_0.

The central question of the talk is: **how can we combine ML with econometric identification so that we still get valid $\sqrt{n}$-inference for a causal parameter θ_0?**"

Slide 3 — Prediction vs. causal estimation
"Let me make the distinction between prediction and causal estimation explicit.

Machine learning typically tries to approximate the regression function $x \mapsto \mathbb{E}[Y \mid X = x]$, or sometimes $\mathbb{E}[Y \mid D, X]$. Algorithms like random forests, boosting and neural nets are optimized for prediction risk.

In causal inference we care about a *functional* θ_0 of the distribution of (Y, D, X). A very common way to define θ_0 is as the solution of a moment condition

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

for some score function ψ and some nuisance functions η_0.

A naive idea is: 'Fit a very flexible ML model for (Y) on ((D, X)) and read off the effect of (D).' The problem is that the regularization bias that helps ML methods predict well can destroy the $\sqrt{n}$-behavior of the estimator of θ_0.

I'll make that precise using moment conditions and a simple bias decomposition."

Slide 4 — Why econometricians love moment conditions

"In econometrics, many parameters are defined through moment conditions.

For OLS we have

$$\mathbb{E}[X(Y - X'\beta_0)] = 0,$$

for IV

$$\mathbb{E}[Z(Y - D\theta_0)] = 0,$$

and GMM stacks many such conditions.

Formally, we think of parameters as solutions of

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

That's useful here for two reasons:

1. It separates the **target** θ_0 from the **nuisance** η_0 (which we are happy to estimate with ML).
2. Concepts like orthogonal or efficient scores live naturally at the level of these moment conditions.

So in this talk we start from moment conditions for θ_0, and then ask: what happens when we plug in ML estimators for the nuisances?"

Slide 5 — Moment conditions in the partially linear model

"Let me specialize to a workhorse model: the partially linear regression (PLR).

We assume

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

Here θ_0 is the causal parameter: the effect of (D) on (Y) conditional on (X). The functions g_0 and m_0 are nuisance functions: high-dimensional and potentially nonlinear.

There are several natural moment conditions that identify θ_0, assuming g_0, m_0 are known.

1. Regression adjustment: $\mathbb{E}[(Y - D\theta_0 - g_0(X))D] = 0$.
2. Propensity-like adjustment: $\mathbb{E}[(Y - D\theta_0)(D - m_0(X))] = 0$.
3. Neyman-orthogonal (residual) score: $\mathbb{E}[(Y - D\theta_0 - g_0(X))(D - m_0(X))] = 0$.

If we had oracle access to g_0 and m_0, all three would lead to the same θ_0. The interesting part is what happens when we use ML estimates instead."

Slide 6 — Key idea (for applied work)

"Before going deeper into the math, here is the practical high-level idea.

If you want to combine ML and econometrics for causal effects, do three things:

1. Express your parameter θ_0 as the solution of a moment condition.
2. Choose a **Neyman-orthogonal** score: a score whose moment is locally insensitive to small errors in the nuisance functions.
3. Estimate the nuisances with your favourite ML methods, but use **cross-fitting** to avoid overfitting effects.

This shifts ML errors to **second order** for the parameter of interest. You keep $\sqrt{n}$-rates and asymptotic normality, so standard inference goes through.

Operationally, you can think of DML as: use ML to residualize (Y) and (D), then regress residualized (Y) on residualized (D) in a way that respects the causal assumptions.

The structure of the talk matches this:

- First, how orthogonal scores and bias decompositions work.
- Second, evidence from simulations and an application to 401(k).
- Finally, limitations and a checklist for using DML in practice.

”

Slide 7 — Naive regression adjustment: bias decomposition
“Let me start with the naive regression-adjustment moment:

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

Suppose for a moment that we do sample splitting: we use one half of the sample to train an ML estimator $\hat{g}$ of g_0, and the other half to estimate θ_0.

On the second half, the plug-in estimator is

$$\hat{\theta}_{\text{RA}} = \left(\sum D_i^2\right)^{-1} \left(\sum D_i(Y_i - \hat{g}(X_i))\right).$$

We can write

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

where

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

is well behaved and asymptotically normal by the central limit theorem.

The problem is the second term

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

In high dimensions or with complex functions, ML estimators of g_0 converge at rates $\|g_0 - \hat{g}\|_{L_2} = n^{-\varphi_g}$ with $\varphi_g < 1/2$. The term B_n is roughly of order $\sqrt{n} \cdot n^{-\varphi_g}$, which tends to infinity in general.

So even with sample splitting and quite good prediction, the regularization bias in $\hat{g}$ destroys $\sqrt{n}$-consistency of $\hat{\theta}_{\text{RA}}$.”

Slide 8 — Orthogonal (residual) score: bias decomposition
“Now let’s switch to the orthogonal score for the PLR model:

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

The corresponding estimator $\tilde{\theta}$ solves the empirical moment condition

$$\sum \psi(W_i; \tilde{\theta}, \hat{g}, \hat{m}) = 0,$$

which gives the closed-form expression

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

We again look at the scaled error:

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

The leading term is

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

which is asymptotically normal.

Crucially, the bias term now looks like

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

So instead of having a *single* estimation error, we have a *product* of two estimation errors. If the ML rates are

$$|\hat{m} - m_0| = n^{-\varphi_m}, \quad |\hat{g} - g_0| = n^{-\varphi_g},$$

then B_n^* is of order $\sqrt{n} \cdot n^{-(\varphi_m + \varphi_g)}$.

Whenever $\varphi_m + \varphi_g > 1/2$, this goes to zero. So the ML can be relatively slow, as long as the product of rates is good enough, and the bias is still second-order."

Slide 9 — Neyman orthogonality: definition & PLR example
"Why does this miracle happen? The key property is Neyman orthogonality.

General set-up: we have a score $\psi(W; \theta, \eta)$, and (θ_0, η_0) satisfies

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

Consider the Gateaux derivative of the population moment with respect to the nuisance:

$$\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta)]\big|_{\eta=\eta_0}[\eta - \eta_0].$$

We say the score is **Neyman-orthogonal** if this derivative is zero for all admissible directions $\eta - \eta_0$.

Intuitively, this means the moment condition is locally insensitive to small perturbations in the nuisance functions: plugging in slightly wrong η does not change the moment at first order.

In the PLR case,

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

and one can check that

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

That orthogonality is exactly what lets us 'kill' the first-order bias and forces estimation error into the second order."

Slide 10 — Residual interpretation & IV-style view
"An equivalent and very intuitive way to see the orthogonal score is via residuals.

Define

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

The empirical orthogonal moment condition is

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

This says: in the regression of $\tilde{Y}_i$ on $\tilde{D}_i$, the residual is uncorrelated with $\tilde{D}_i$. So

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

The solution is

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

This has an **IV-style interpretation**: $\tilde{D}_i = D_i - \hat{m}(X_i)$ behaves like an instrument for D_i after controlling flexibly for (X), and we regress the residualized outcome $(\tilde{Y})$ on the residualized treatment.

So DML is literally a bridge: ML learns the residualization, and then we plug into a very classical one-parameter IV/GMM-type estimator."

Slide 11 — Sample splitting and cross-fitting

"The last conceptual ingredient is sample splitting and cross-fitting.

The bias decomposition I showed hides a third remainder term, call it C_n^*, which involves products like

$$\sum V_i(\hat{g}(X_i) - g_0(X_i)),$$

and similar terms with m.

If we estimate $\hat{g}$ and $\hat{m}$ on the same data we use to form the score, these terms are difficult to control: the ML estimators can overfit, and the dependence structure is messy.

The solution is:

1. Split the sample into K folds.
2. On fold k, estimate $\hat{g}^{(-k)}$ and $\hat{m}^{(-k)}$ using only the other folds.
3. On fold k, compute residuals and an estimate $\hat{\theta}^{(k)}$.
4. Average over k.

Conditional on the training folds, the nuisance errors and the residuals behave almost like independent objects. We can then bound terms like C_n^* with simple variance calculations.

Cross-fitting also means that in the end we're using the whole sample to estimate θ_0, so we don't lose efficiency by splitting the data once and for all."

Slide 12 — Algorithm for PLR DML (1/2)

"Let me summarize the algorithm concretely for the PLR model.

1. Choose your ML methods for $g_0(X)$ and $m_0(X)$. For example, lasso, random forests, boosting, neural nets, or some ensemble.
2. Fix the number of folds K, say $K = 5$.
3. Randomly partition the data indices $\{1, \dots, n\}$ into folds $I_1, \dots, I_K$.
4. For each fold k:
   (a) Train $\hat{g}^{(-k)}, \hat{m}^{(-k)}$ on all observations not in I_k.
   (b) For each $i \in I_k$ compute the residuals
   $$\tilde{Y}_i := Y_i - \hat{g}^{(-k)}(X_i), \quad \tilde{D}_i := D_i - \hat{m}^{(-k)}(X_i).$$

On the next slide, we use these residuals to get $\hat{\theta}_{\text{DML}}$."

Slide 13 — Algorithm for PLR DML (2/2)

"Continuing the algorithm:

5. For each fold k, on the held-out data $i \in I_k$, run a simple OLS regression of $\tilde{Y}_i$ on $\tilde{D}_i$ without an intercept:
$$\hat{\theta}^{(k)} := \frac{\sum_{i \in I_k} \tilde{D}_i \tilde{Y}_i}{\sum_{i \in I_k} \tilde{D}_i^2}.$$
6. Aggregate the fold-specific estimates:
$$\hat{\theta}_{\text{DML}} := \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)}.$$
7. Finally, estimate the asymptotic variance using the empirical influence function based on the orthogonal score, and construct confidence intervals in the usual way.

[If you add a flowchart here, you can visually show: raw data → ML fits of nuisances → residuals → regression of residuals.]"

Slide 14 — What double ML delivers (high level)
"Let me now summarize what double ML buys us, theoretically.

Under mild rate conditions on the ML estimators — essentially that the nuisance estimators are consistent and the product of their rates is fast enough — we have:

- $\hat{\theta}_{\mathrm{DML}}$ is $\sqrt{n}$-consistent and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_{\mathrm{DML}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

- We can estimate $\hat{\sigma}^2$ via the sample variance of the orthogonal score, and form standard Wald confidence intervals.

The nice feature is that this holds for a wide range of ML methods: lasso, random forests, boosting, neural nets, GAMs, and ensembles.

In the PLR model under homoscedasticity, the DML estimator actually attains the semiparametric efficiency bound. So in that sense, we lose nothing by using ML for nuisance estimation, provided we use orthogonal scores and cross-fitting."

Slide 15 — Simulation: prediction vs. causal estimation
"Chernozhukov and co-authors illustrate the difference between naive plug-in and DML using simulations.

In one design, they construct $g_0(X)$ as a function that is particularly friendly to random forests — effectively a combination of tree structures — so that random forests are almost oracle for prediction.

They then compare two estimators of θ_0:

1. Naive ML plug-in: use a random forest prediction model for (Y) on $((D, X))$, and read off the coefficient or partial dependence for (D).
2. The DML estimator based on the orthogonal score and cross-fitting.

The results are:

- The naive plug-in estimator predicts (Y) extremely well, but its sampling distribution for θ_0 is badly biased: the histogram of $\hat{\theta} - \theta_0$ is shifted away from zero.
- The DML estimator's histogram is centered at zero and well approximated by a normal curve.

The lesson is: **excellent prediction error does not imply good causal estimation**. The structure of the score — in particular, orthogonality — is what protects the causal parameter."

Slide 16 — Application: 401(k) eligibility and savings
"Let me briefly discuss an empirical illustration: 401(k) eligibility and household savings.

We consider US survey data where:

- (Y): net financial assets — including 401(k), IRAs, checking, stocks, minus non-mortgage debt.
- (D): indicator for being *eligible* for a 401(k) plan at the current job.
- (X): a rich set of covariates such as income, age, education, family size, marital status, defined benefit pension status, IRA participation, and home ownership.

Identification strategy (following the literature): conditional on these covariates, eligibility can be treated as as-good-as-random. So we aim at the average treatment effect of eligibility on net assets.

We use DML for this ATE:

- ML methods estimate outcome regressions and possibly propensity scores as functions of (X).
- We plug them into an orthogonal score and apply cross-fitting.

Empirically, across a range of learners (lasso, trees, forests, boosting, nets, ensembles), the estimated ATE is substantial and positive, on the order of \$7–\$9k in extra net financial assets for being eligible.

The key point here is not the exact number, but that we are able to flexibly control for a high-dimensional (X) with ML **and** still get standard errors and confidence intervals for the causal effect."

Slide 17 — Beyond PLR: ATE, ATTE, PLIV, LATE

“The PLR model is just the starting point.

The same DML blueprint extends to several other core causal estimands.

1. **ATE / ATTE under unconfoundedness.** We observe (Y, D, X) with binary (D), and assume that conditional on (X), treatment is as-good-as-random. Orthogonal scores combine outcome regression $g_0(d, X)$ and the propensity score $m_0(X) = \mathbb{E}[D \mid X]$. Again, we estimate nuisances with ML and plug them into the orthogonal score with cross-fitting.

2. **Partially linear IV (PLIV).** Here we have an instrument (Z) and model

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

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

   The exclusion restriction is that, given (X), (Z) affects (Y) only through (D). DML uses an orthogonal score that involves the residualized outcome and a residualized version of the instrument.

3. **LATE with binary (D) and (Z).** In the local average treatment effect setting, the parameter is a ratio of two expectations (a Wald estimand). DML constructs orthogonal scores for numerator and denominator, with ML for nuisance functions such as $\mu_0(z, x) = \mathbb{E}[Y \mid Z = z, X = x]$, $m_0(z, x) = \mathbb{E}[D \mid Z = z, X = x]$, and the instrument propensity.

In all these settings, the pattern is the same: low-dimensional causal parameter, high-dimensional nuisance functions, orthogonal score, ML, cross-fitting.”

Slide 18 — What DML does *not* fix

“So far I’ve emphasized what DML does well. Let me be explicit about what it does *not* solve.

DML relaxes *functional-form* assumptions by allowing ML to estimate complex nuisance relationships. But it does not relax the underlying **causal identification assumptions**:

- In PLR or ATE settings, we still need that all confounders are observed and appropriately controlled for in (X).
- In IV or LATE settings, we still need valid instruments (Z): relevance, exclusion, and monotonicity or other structural conditions where required.
- We still need sufficient overlap or positivity: we shouldn’t have many units with propensity scores extremely close to 0 or 1.
- We must avoid ‘bad controls’ — variables that are colliders or mediators in the causal graph.

If important confounders are unobserved, DML cannot magically remove the resulting bias.

So ML plus DML is not a substitute for good research design. You still have to think carefully about the causal structure and variable choice.”

Slide 19 — Evidence from method evaluation studies

“There are also some lessons from method evaluation studies that are useful for practice.

Simulation papers that compare OLS, naive ML, and DML across different designs typically find:

- When confounding is truly linear and the parametric model is correctly specified, OLS and DML both perform well. In such cases, DML doesn’t buy you much beyond robustness to mild misspecification.
- When confounding is nonlinear — say interactions, U-shaped effects, or thresholds — DML shines **if** you use flexible learners for the nuisance functions.

  If you run DML with a very rigid learner, such as plain lasso on raw covariates and no transformations, you can get OLS-like bias because the nuisance fits are poor.

  If you instead use flexible methods like random forests, boosting, GAMs, or neural nets, DML can substantially reduce bias while keeping variance under control.

The main message: the ‘double’ in double ML does not guarantee good performance by itself. You still need to choose ML methods that are appropriate for the structure of your problem.”

Slide 20 — Checklist & take-home message
“Let me finish with a checklist and a summary you can take into your own projects.

When is DML a good fit?

- You have many covariates and believe the nuisance structure is complex but learnable.
- You care about a low-dimensional parameter: ATE, ATTE, a regression coefficient in a PLR model, an IV effect, a LATE.
- You have enough data to fit flexible ML models reasonably well.

What should you actually do?

1. Write down the causal estimand and the moment condition that identifies it.
2. Derive or look up an orthogonal score for that estimand.
3. Choose ML algorithms capable of capturing the nonlinearities you care about.
4. Use cross-fitting and examine sensitivity to ML choices and to the number of folds (K).
5. Always think about the identification assumptions in parallel — ML cannot fix a bad instrument or missing confounder.

If you remember just one sentence, let it be this:

> Use ML where it’s strong — in estimating high-dimensional nuisance functions — but protect the causal parameter with orthogonal moments and cross-fitting, so that ML errors only show up as second-order bias.

Thanks, and I’m happy to discuss how this applies to your own empirical settings.”