8. Double Machine Learning

Causal Data Science for Business Analytics

Christoph Ihl

Hamburg University of Technology

Monday, 15. June 2026

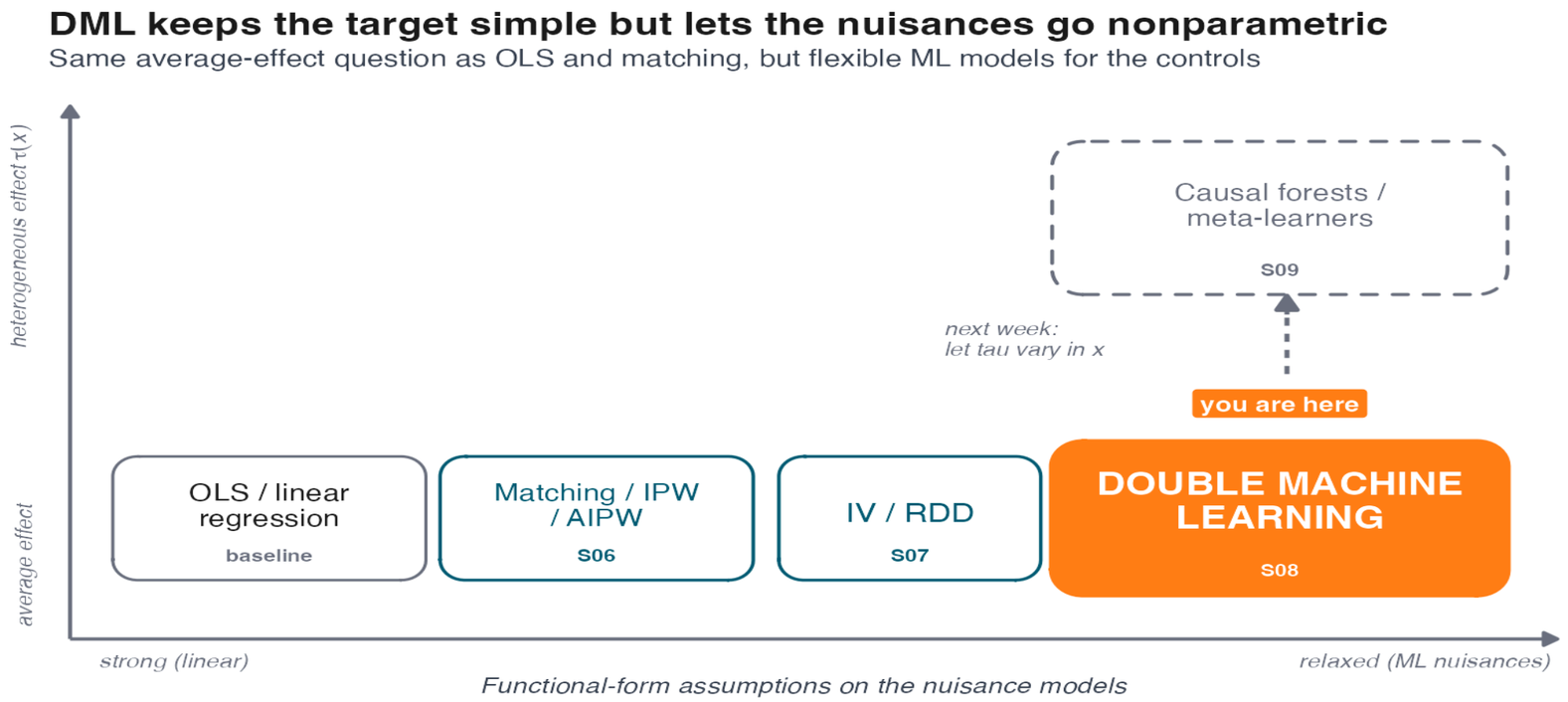


S07 went hunting for exogenous variation.

Today we have the opposite luxury: lots of covariates X, and a machine that predicts Y and D beautifully.

Can a black-box predictor hand us a causal effect — with honest error bars?

Where DML sits in the toolkit you already have



Same **average-effect** question as S06 — but the controls go **nonparametric.**
Heterogeneity $\tau(x)$ is next week.

A Machine That Predicts Is Not a Machine That Explains

What we actually want: a number AND a trustworthy interval

The deliverable is not a point estimate — it is a point **with calibrated uncertainty:**

$$\sqrt{n}\,(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, V)$$

That one line demands three things at once:

- **consistency** — $\hat{\tau} \to \tau$,
- **$\sqrt{n}$-rate** — the error shrinks fast enough to matter,
- **asymptotic normality** — so $\hat{\tau} \pm 1.96\,\widehat{\mathrm{SE}}$ has 95% coverage.

A model that only **predicts well** delivers none of these for τ — that is today's problem.

What we have: a great predictor that converges too slowly

Plug a flexible learner $\hat{g}(X)$ into a naive estimator and two biases leak into $\hat{\tau}$:

Regularization bias

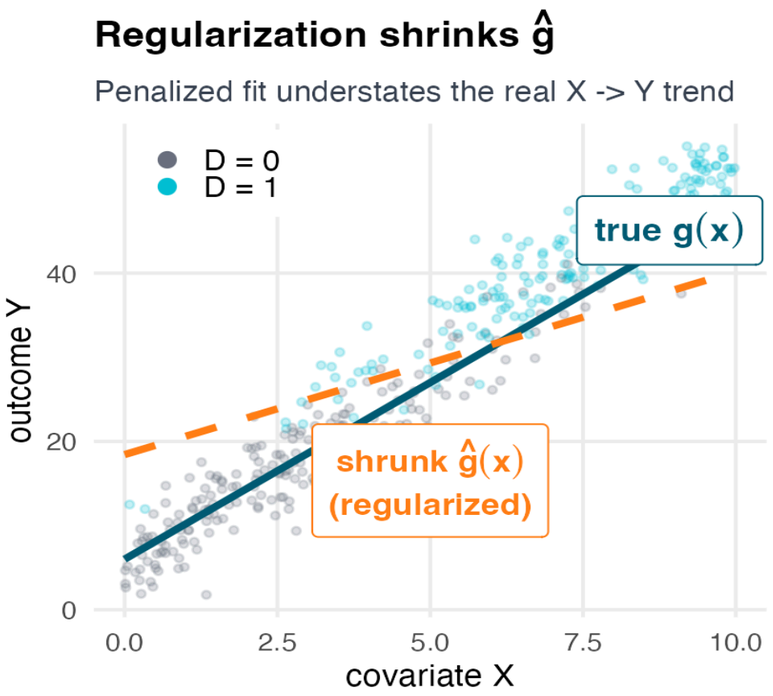
To control variance, ML shrinks (lasso penalty, shallow trees). That shrinkage is correlated with D through X – so it leaks straight into $\hat{\tau}$.

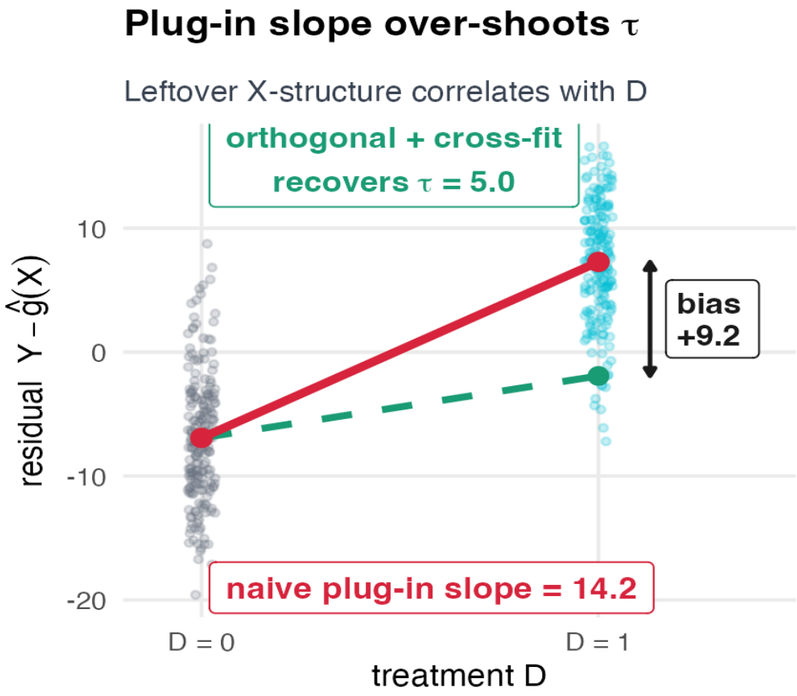
Overfitting bias

Evaluate the nuisance on the same rows used to fit it and the model has partly memorised its own noise – biasing $\hat{\tau}$ and shrinking its SE.

ML nuisances converge at a slow $n^{-1/4}$ rate; a naive plug-in inherits that rate for τ — far too slow for a $\sqrt{n}$ CI.

Why a naive ML plug-in is biased — see it once





A shrunk $\hat{g}(X)$ leaves structure in the residuals that is **correlated with** D — and that leftover slope masquerades as a treatment effect. Orthogonalizing + cross-fitting Y and D removes it.

LearnLoop: does the recommender actually lift exam scores?

Symbol	LearnLoop	Role
D	recommended (0/1)	treatment
Y	exam_score	outcome
X	age, region, device, prior courses, baseline score, …	observed covariates

A naive comparison says **+14.03 points**. But the recommender targets stronger learners — they would have scored higher **anyway**. The lift is confounded by X, and the confounding runs through a *nonlinear* interaction (baseline score × prior courses) that a linear control won't catch.

The whole session in one recipe

The DML recipe

1. Write a Neyman-orthogonal score for τ (insensitive to small nuisance errors).
2. Estimate the nuisances g, m with any ML learner, using cross-fitting.
3. Solve the empirical moment for $\hat{\tau}$.
4. Read the SE off the influence function.

Two ingredients do the real work: orthogonality (kills regularization bias) and cross-fitting (kills overfitting bias). The rest is plumbing.

Partialling-Out: From FWL to Robinson

You already know the trick: Frisch–Waugh–Lovell

To get the coefficient on D in $Y \sim D + X$, you do **not** need the full regression. FWL says: it equals the slope of a **residual-on-residual** regression.

- Regress Y on X, keep residuals $\tilde{Y} = Y - \hat{E}[Y \mid X]$.
- Regress D on X, keep residuals $\tilde{D} = D - \hat{E}[D \mid X]$.
- Regress $\tilde{Y}$ on $\tilde{D}$ — that slope **is** $\hat{\tau}$, numerically identical to the full OLS.

"Partial out X, then read the slope." We met this in S06 (matching) and S07 (2SLS purges D). Today we **swap OLS for ML** in the two partialling-out steps. (Frisch and Waugh 1933; Lovell 1963)

Two kinds of parameter: target vs nuisance

Don't interpret every coefficient

In $Y = \tau D + g(X) + \varepsilon$, only τ is the target. The function $g(X)$ is a nuisance — needed for estimation, never interpreted.

This split is liberating: we no longer need a *correct, interpretable* model of g. We only need to **predict** it well enough to partial it out.

That is exactly the job ML is good at — and exactly why we can hand g to a black box without guilt.

The partially linear model (Robinson, 1988)

Treatment enters **linearly**; the controls enter **arbitrarily** — that's what *partially linear* means. (Robinson 1988)

$$Y = \tau D + g(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid D, X] = 0$$

$$D = m(X) + \nu, \qquad \mathbb{E}[\nu \mid X] = 0$$

- τ — the **target** (constant treatment effect, the ATE).
- $g(X)$ — the **structural** outcome function.

We partial out the **two nuisances**:

- $\ell(X) = \mathbb{E}[Y \mid X] = \tau\, m + g$ — the reduced-form outcome (**not** g).
- $m(X) = \mathbb{E}[D \mid X]$ — the **propensity**.

Partialling-out, annotated

$$\hat{\tau} = \frac{\widehat{\mathrm{Cov}}(\tilde{Y},\, \tilde{D})}{\widehat{\mathrm{Var}}(\tilde{D})}, \qquad \tilde{Y} = Y - \hat{\ell}(X), \quad \tilde{D} = D - \hat{m}(X)$$

- $\tilde{Y}$ — the part of the exam score **not predicted** by who the learner is.
- $\tilde{D}$ — the part of the recommendation **not predicted** by who the learner is (the as-good-as-random slice).
- The slope of $\tilde{Y}$ on $\tilde{D}$ is the treatment effect, purged of X.

Here $\hat{\ell}(X) = \hat{\mathbb{E}}[Y \mid X]$ is the **reduced-form** outcome predictor — $\ell = \tau m + g$, folding in the indirect path of X through D. It is DoubleML's ml_l (*not* the structural g).

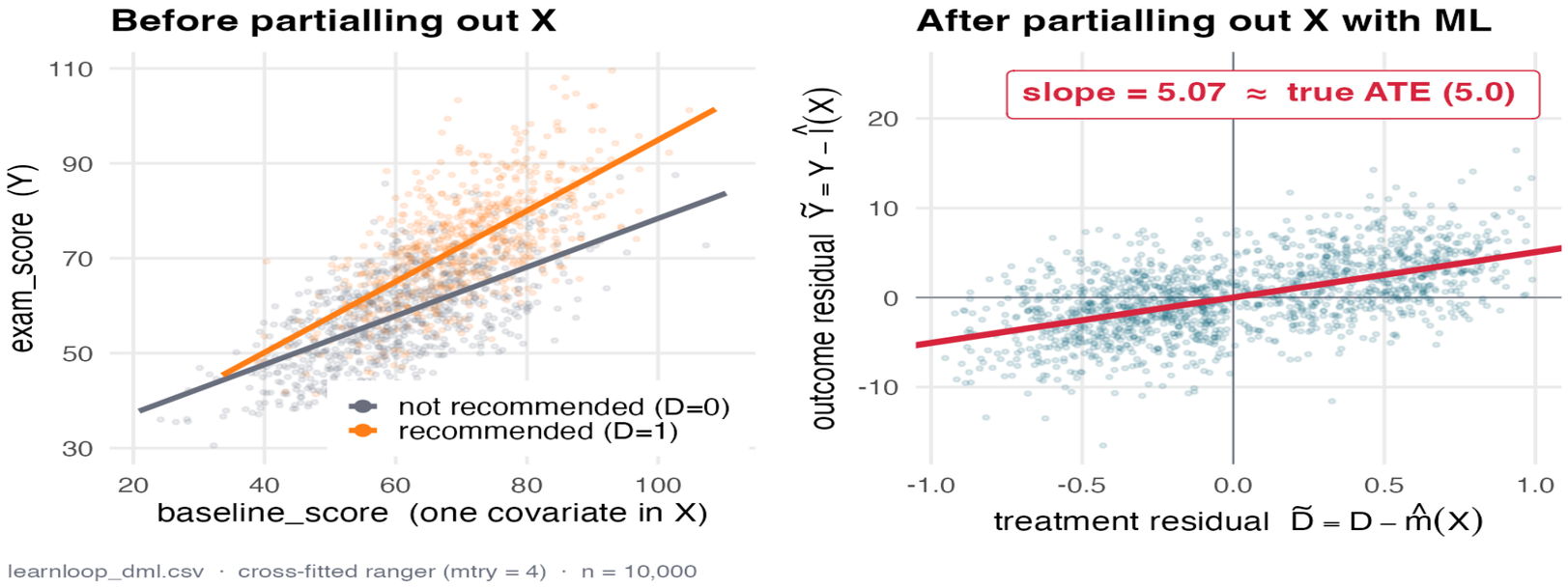
The leap: swap OLS for any ML learner

OLS partials out X by a **linear** projection. Nothing requires that.

Residualize with…	$\hat{\mathbb{E}}[Y \mid X]$, $\hat{\mathbb{E}}[D \mid X]$
Linear regression	only linear/additive structure
Random forest, lasso, boosting, neural net	**arbitrary nonlinearity + interactions**

In LearnLoop the confounding hides in a **baseline-score × prior-courses interaction.** A linear control misses it; a random forest catches it — and the residual-on-residual slope snaps onto the truth.

See it on LearnLoop: let the machine remove X, then read the slope



learnloop_dml.csv · cross-fitted ranger (mtry = 4) · n = 10,000

Left: treated learners sit higher partly because of X. Right: after ML residualization, the slope of $\tilde{Y}$ on $\tilde{D}$ is ≈ 5.0 — the partialling-out estimate.

Two Ingredients: Orthogonality + Cross-Fitting

What is a score? An estimating equation for τ

Score (moment function)

A score $\psi(W;\tau,\eta)$ ties together the data $W=(Y,D,X)$, the target τ, and the nuisances $\eta=(\ell,m)$. Its moment – the population expectation – is zero at the truth: $\mathbb{E}[\psi(W;\tau_0,\eta_0)]=0$, which pins down τ_0.

Every score here is **linear in τ**: $\psi=\psi_a(W;\eta)\,\tau+\psi_b(W;\eta)$, so the root is closed-form, $\hat{\tau}=-\mathbb{E}_n[\psi_b]/\mathbb{E}_n[\psi_a]$ (ML $\hat{\eta}$ plugged in).

Partialling-out is an exact example of this:

$$\psi=(\tilde{Y}-\tau\tilde{D})\,\tilde{D}=\underbrace{(-\tilde{D}^2)}_{\psi_a}\tau+\underbrace{\tilde{Y}\tilde{D}}_{\psi_b}\;\Rightarrow\;\tau=\frac{\mathrm{Cov}(\tilde{Y},\tilde{D})}{\mathrm{Var}(\tilde{D})}$$

The two ingredients: **orthogonality** (a property of this score) and **cross-fitting** (how $\hat{\eta}$ enters it).

Step 1 — an orthogonal score (Neyman orthogonality)

Orthogonality

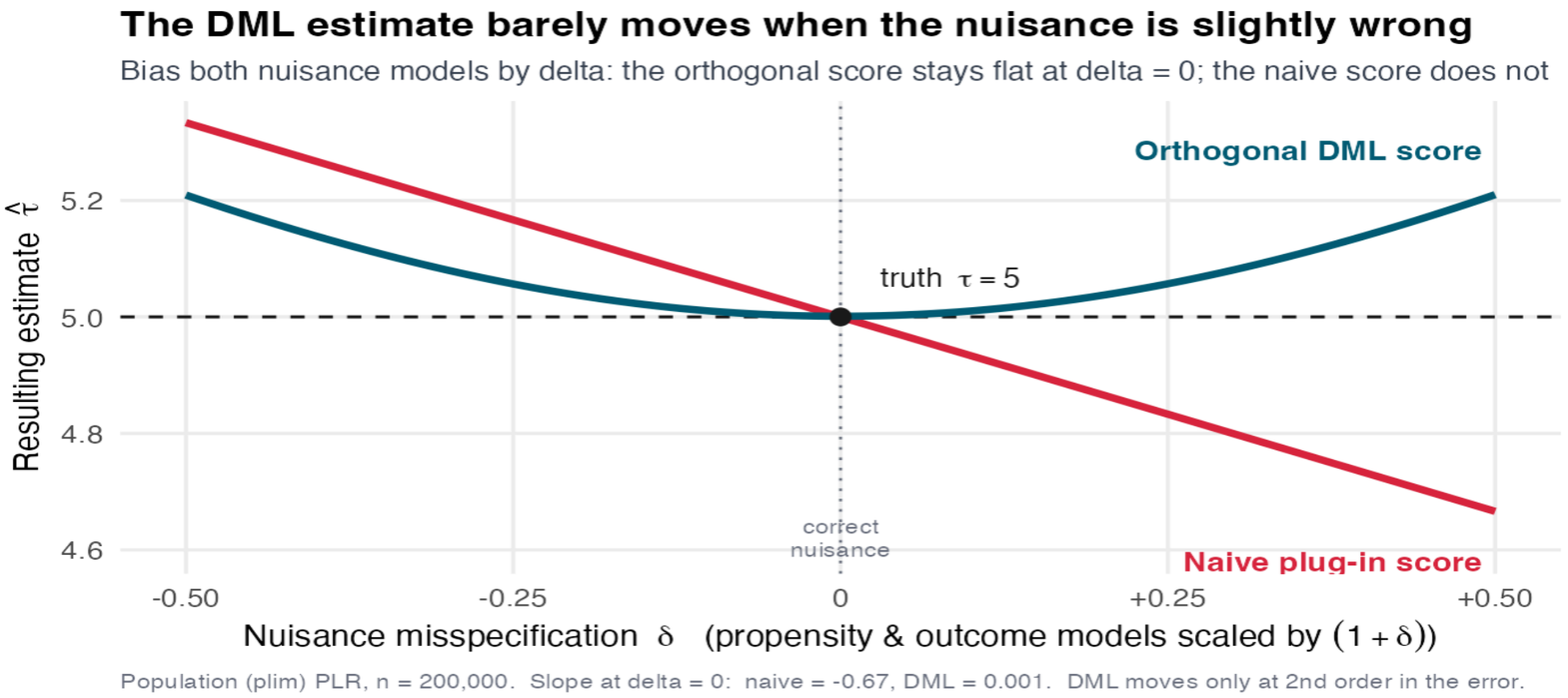
A score is Neyman-orthogonal if slightly-wrong nuisance models barely move the estimate: its sensitivity to nuisance error is zero to first order – the linear (δ) term cancels, leaving only a second-order (δ^2) effect.

Formally, the moment $\psi(W;\tau,\eta)$ has zero **Gateaux derivative** in the nuisance $\eta = (\ell, m)$ at the truth:

$$\partial_\eta \, \mathbb{E}\big[\psi(W;\tau_0,\eta)\big]\big|_{\eta=\eta_0} = 0$$

Consequence: the bias in $\hat{\tau}$ depends on the **product** of the two nuisance errors, not their sum. Two slow $n^{-1/4}$ errors multiply to $n^{-1/2}$ — fast enough for a valid CI.

Orthogonality, drawn



Perturb the nuisances by δ. The **naive** score cuts steeply through the truth (slope $\neq 0$). The **orthogonal** score is flat at $\delta = 0$ — it only feels the error at *second* order.

Step 2 — cross-fitting kills the overfitting bias

The danger: if the **same rows** train the nuisance and enter the moment for τ, the model has seen its own noise.

The cross-fitting mantra

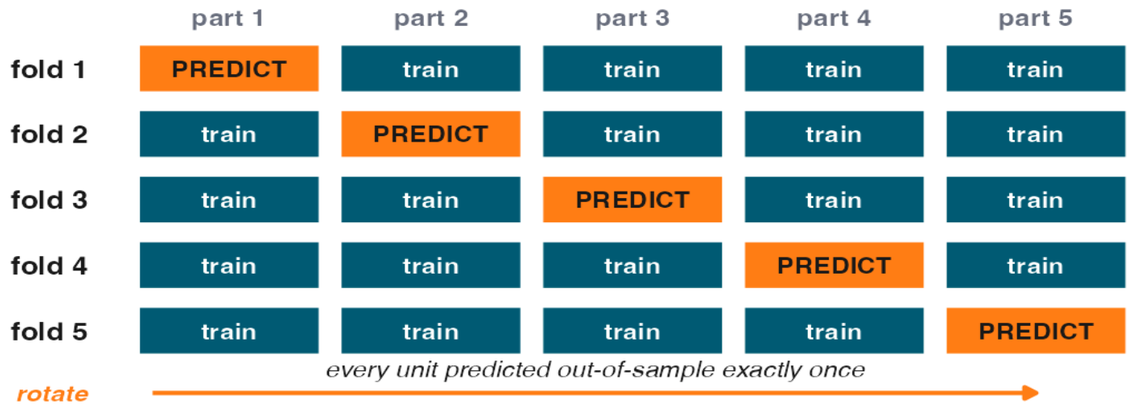
A unit must never be predicted by a model that was trained on it.

So we split the sample: fit $\hat{g}, \hat{m}$ on **other** folds, predict the **held-out** fold, and rotate until every unit has an out-of-fold prediction. Out-of-fold predictions are independent of each unit's own noise — the overfitting bias vanishes.

Cross-fitting, drawn

A unit is never predicted by a model trained on it

K = 5-fold cross-fitting: hold out one fold, fit nuisances on the rest, predict the held-out fold, rotate



Flow: train nuisances on K-1 folds -> predict the held-out fold -> rotate until every unit is predicted out-of-sample once.
Orange = held-out (predict) block; petrol = training blocks.

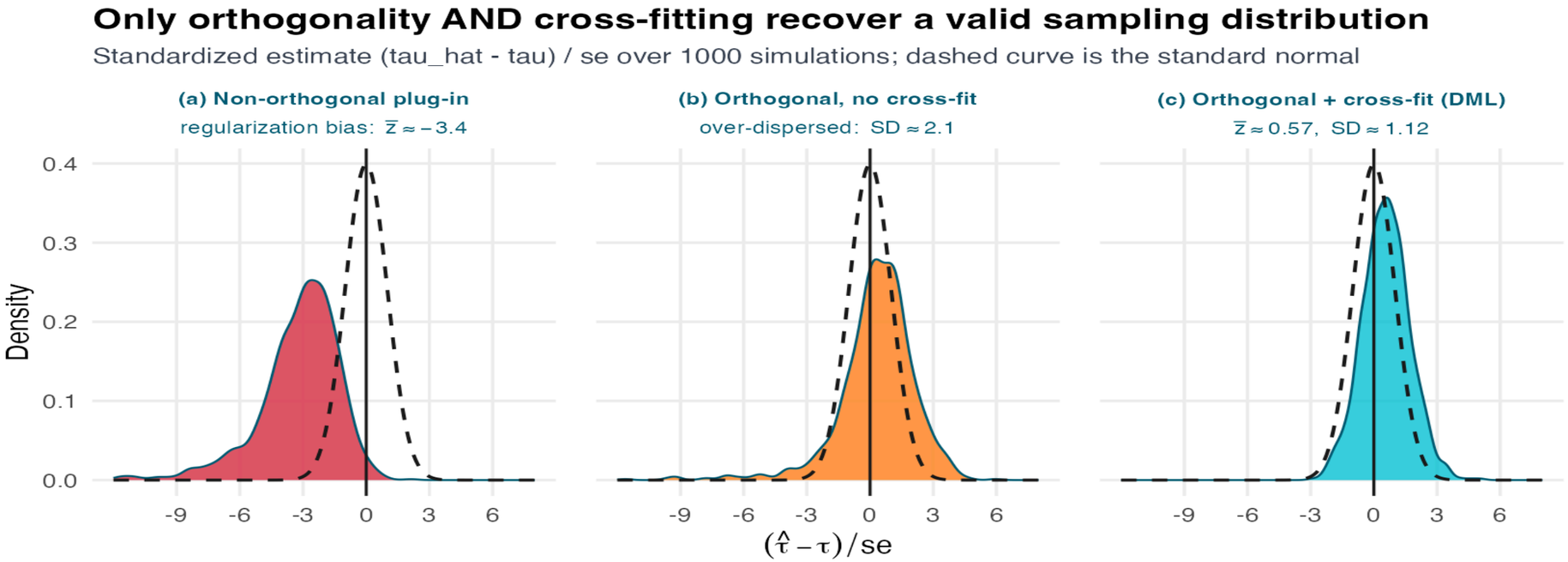
Five folds: each part is predicted by a model trained on the **other four** — so every unit is predicted out-of-sample exactly once, never by a model that saw it.

How many folds — and why repeat?

- **Five folds** is the workhorse default (2 is fine for huge n; more folds = more training data per fold, more compute).
- The fold split is **random** — a single split adds noise. Repeat the whole cross-fitting n_{rep} times and take the **median** estimate / aggregate the SEs.
- Use a genuinely flexible learner (random forest, boosting); don't pre-commit to one — **stack / super-learn** if unsure. (Ahrens et al. 2025)

In LearnLoop the key knob is the forest's mtry: too small and it can't see the interaction — the estimate drifts. Tuning matters because the nuisance quality is what orthogonality leans on.

The payoff: only orthogonality AND cross-fitting work



(a) non-orthogonal → shifted (regularization bias). **(b)** orthogonal, no cross-fit → over-dispersed (SE too small). **(c)** orthogonal + cross-fit → centered, correctly scaled $\mathcal{N}(0,1)$.

Step 3 — average the per-unit scores into $\hat{\tau}$

Each unit has a per-unit score $\psi_i = \psi(W_i; \tau, \hat{\eta}_{-k(i)})$ — its own out-of-fold nuisances — and it is **linear in** τ: $\psi_i = \psi_{a,i}\,\tau + \psi_{b,i}$.

Set the **average** to zero and solve:

$$\mathbb{E}_n[\psi] = 0 \quad \Longrightarrow \quad \hat{\tau} = -\frac{\mathbb{E}_n[\psi_b]}{\mathbb{E}_n[\psi_a]} \qquad \text{(average the pieces, then divide).}$$

Robinson PLR — $\psi_{a,i} = -\tilde{D}_i^2,\ \psi_{b,i} = \tilde{Y}_i\tilde{D}_i$: $\quad \hat{\tau} = \dfrac{\overline{\tilde{Y}\tilde{D}}}{\overline{\tilde{D}^2}} = \dfrac{\mathrm{Cov}(\tilde{Y}, \tilde{D})}{\mathrm{Var}(\tilde{D})}.$

Same crank everywhere — only ψ changes: **AIPW** ($\psi_a = -1$) makes $\hat{\tau} = \overline{\psi_b}$ a *plain average* of scores; **IV** is a *ratio* of two averaged scores (Wald).

Step 4 — the SE comes from the influence function

A good estimator is **asymptotically linear:** $\hat{\tau} - \tau \approx \frac{1}{n}\sum_i \phi(W_i)$. The **influence function** ϕ_i measures how much unit i tips the estimate; for DML it is the **normalized score:**

$$\phi_i = -\frac{\psi_i}{\mathbb{E}[\psi_a]} \ \text{(mean 0)} \quad \Rightarrow \quad \sqrt{n}\,(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\big(0,\ \mathbb{E}[\phi^2]\big), \qquad \widehat{\mathrm{SE}} = \sqrt{\frac{\mathbb{E}_n[\psi^2]}{\mathbb{E}_n[\psi_a]^2\, n}}$$

Robinson: $\phi_i = \dfrac{(\tilde{Y}_i - \tau\tilde{D}_i)\,\tilde{D}_i}{\mathrm{Var}(\tilde{D})}$ — exactly the robust SE of the residual-on-residual regression. **Same formula for PLR, AIPW, IV;** only ψ, ψ_a change (AIPW $\psi_a = -1 \Rightarrow \widehat{\mathrm{SE}} = \sqrt{\widehat{\mathrm{Var}}(\psi_i)/n}$).

The recipe, as a four-step checklist

DML in four steps

1. **Orthogonal score** – choose a Neyman-orthogonal $\psi(W;\tau,\eta)$, linear in τ: $\psi = \psi_a(W;\eta)\,\tau + \psi_b(W;\eta)$.
2. **Cross-fitted nuisances** – fit $\hat{\eta}$ on held-out folds; every unit predicted out-of-fold.
3. **Average & solve** – $\mathbb{E}_n[\psi] = 0 \;\Rightarrow\; \hat{\tau} = -\,\mathbb{E}_n[\psi_b] \,/\, \mathbb{E}_n[\psi_a]$.
4. **Influence-function SE** – $\phi_i = -\,\psi_i / \mathbb{E}_n[\psi_a]$, so $\widehat{\mathrm{SE}} = \sqrt{\mathbb{E}_n[\psi^2] \,/\, \mathbb{E}_n[\psi_a]^2 \,/\, n}$.

Memorise this: only the **score** ψ (step 1) changes across models (PLR, AIPW, IV) — steps 2–4 are identical. (Chernozhukov et al. 2018)

Throughout the course the effect is τ; the DML papers and the DoubleML package write it θ (reported as coef) — same object.

AIPW: the Doubly-Robust Score

When the effect isn't a single number: the interactive model

Same two equations as PLR — but **drop the linearity in D**:

$$Y = \mu(D, X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid D, X] = 0$$

$$D = e(X) + \nu, \qquad \mathbb{E}[\nu \mid X] = 0$$

- $\mu(D, X) = \mathbb{E}[Y \mid D, X]$ — the **outcome surface** (arms $\mu_1(X) = \mu(1, X)$, $\mu_0(X) = \mu(0, X)$).
- $e(X) = \mathbb{E}[D \mid X]$ — the **propensity** (PLR's m).
- No constant τ: the effect $\mu_1(X) - \mu_0(X)$ may **vary with** X.
  - we target the ATE $= \mathbb{E}[\mu_1(X) - \mu_0(X)]$.

Plug-in alone ($\overline{\hat{\mu}_1 - \hat{\mu}_0}$) inherits the regularization bias → need an orthogonal score: **AIPW**.

AIPW = Direct estimate + Residual correction

$$\psi_i = \underbrace{\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right]}_{\text{direct (outcome regression)}} + \underbrace{\frac{D_i\,(Y_i - \hat{\mu}_1)}{\hat{e}(X_i)} - \frac{(1 - D_i)(Y_i - \hat{\mu}_0)}{1 - \hat{e}(X_i)}}_{\text{bias correction (propensity-weighted residual)}}$$

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_i \psi_i$$

The outcome model gives a first guess; the propensity-weighted residual **repairs** whatever the outcome model got wrong. (Robins, Rotnitzky, and Zhao 1994; Bang and Robins 2005)

AIPW slots into the same four steps

The per-unit AIPW signal ψ_i (last slide) is the score's ψ_b; subtract τ for the orthogonal score:

$$\psi = \psi_i - \tau \qquad \Longrightarrow \qquad \psi_b = \psi_i, \quad \psi_a = -1$$

- **Step 2 — nuisances** $\eta = (\mu_1, \mu_0, e)$: now **three** (PLR's m → two arms $+$ propensity), cross-fitted.
- **Step 3 — average & solve** $\hat{\tau} = -\mathbb{E}_n[\psi_b]/\mathbb{E}_n[\psi_a] = \mathbb{E}_n[\psi_i]$: a **plain average** of the per-unit signals.
- **Step 4 — SE** $\widehat{\mathrm{SE}} = \sqrt{\widehat{\mathrm{Var}}(\psi_i)/n}$ (influence function $= \psi_i - \bar{\psi}$).

Only the **score** changed — orthogonality (= double robustness, next) and cross-fitting carry over verbatim.

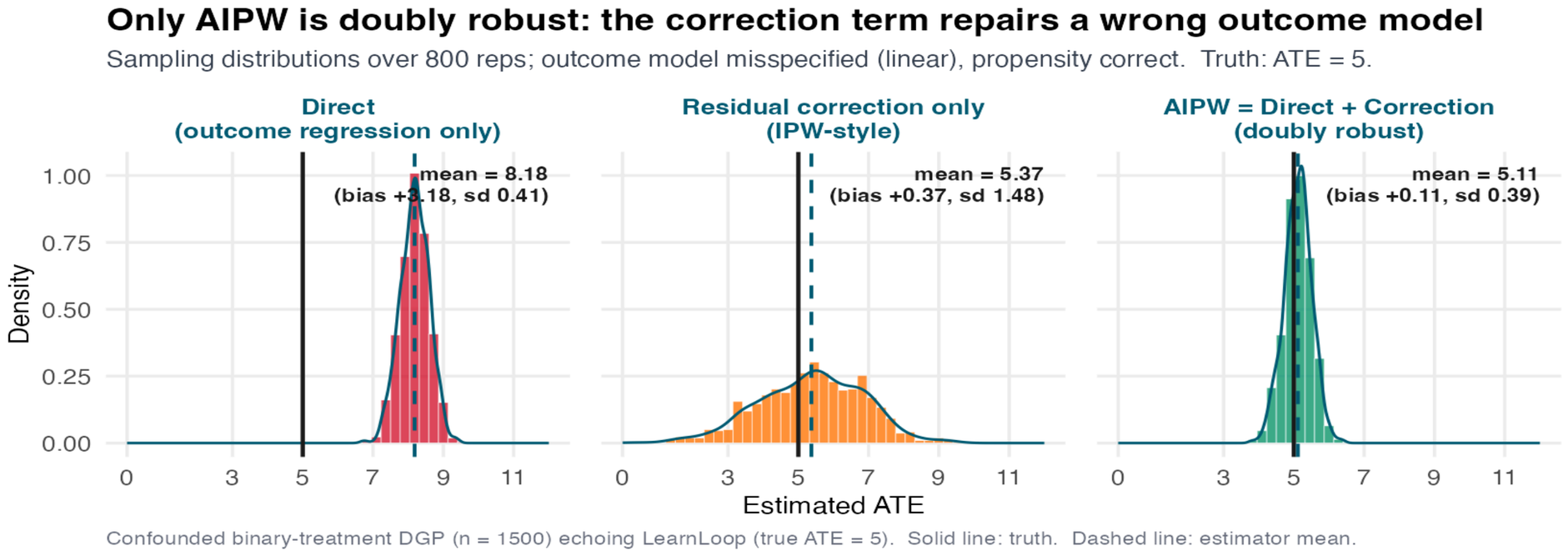
Double robustness: right if EITHER nuisance is right

Two shots on goal

The AIPW estimate is consistent if either the outcome models μ_1, μ_0 or the propensity e is correctly specified — you don't need both.

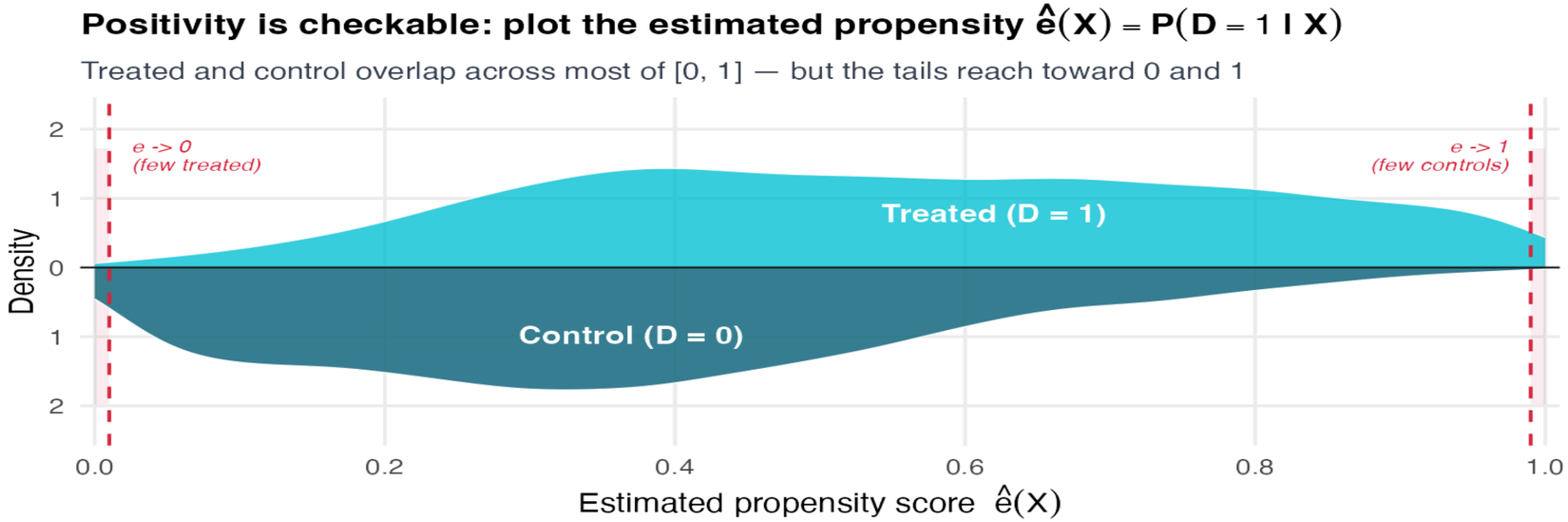
Orthogonality is **why**: the score's first-order sensitivity to each nuisance is zero, so an error in one is absorbed as long as the other holds. Double robustness is orthogonality, viewed from the modelling side.

AIPW's bias–variance payoff



Outcome model **misspecified** on purpose: **Direct** is biased (≈ 8.2), **IPW-only** is centered but noisy, **AIPW** is centered on 5 and tightest.

Before you trust the weights: check overlap



AIPW **divides by** $\hat{e}$ and $1 - \hat{e}$ — as $\hat{e} \to 0, 1$ the weights, and the variance, blow up. Always plot this first; then the question is **calibrate or trim** (next).

Trim, clip, or calibrate? Change the question vs. fix the weights

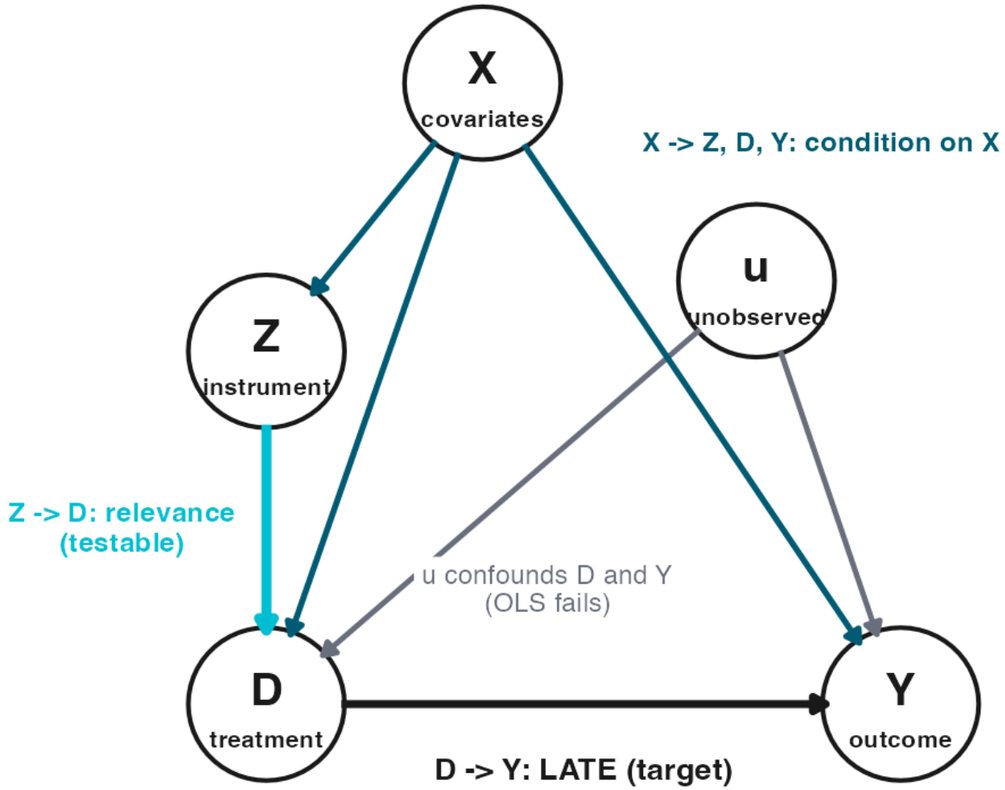
Three responses when AIPW weights $1/\hat{e}$ explode as $\hat{e} \to 0, 1$:

	what it does to $\hat{e}$	**estimand**	**use when**
Trim	**drops** units with $\hat{e} \notin [\alpha, 1-\alpha]$	**overlap-ATE** $\neq$ ATE	genuine non-overlap
Clip	**caps** $\hat{e}$ at $[\alpha, 1-\alpha]$, keeps units	full ATE, capped units **biased**	quick blunt guard
Calibrate	**rescales** $\hat{e} \to \tilde{e}$ so $\mathbb{E}[D \mid \tilde{e}] = \tilde{e}$	full ATE, **honest** weights	ML over-confidence

Only **trim** changes the *question*; clip & calibrate keep the full ATE — clip patches the *symptom*, calibrate fixes the *cause*. → **Calibrate first.** (Ballinari and Bearth 2024; Crump et al. 2009)

When Unconfoundedness Fails: DML-IV

If a confounder is unobserved, bring back the instrument



Partially linear IV — the model and the score

D is **endogenous** (u leaks in) — residualize on X and **instrument with** Z (exogeneity on Z, not D):

$$Y = \tau D + g(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z, X] = 0$$

$$Z = r(X) + \nu, \qquad \mathbb{E}[\nu \mid X] = 0$$

- **Three** nuisances: $\ell = \mathbb{E}[Y \mid X]$, $m = \mathbb{E}[D \mid X]$, $r = \mathbb{E}[Z \mid X]$ — residuals $\tilde{Y}, \tilde{D}, \tilde{Z}$.
- The orthogonal score uses $\tilde{Z}$ as the instrument:

$$\psi = (\tilde{Y} - \tau\tilde{D})\,\tilde{Z} = \underbrace{(-\tilde{D}\tilde{Z})}_{\psi_a}\tau + \underbrace{\tilde{Y}\tilde{Z}}_{\psi_b}$$

PLR with one change — $\psi_a = -\tilde{D}\tilde{Z}$ instead of $-\tilde{D}^2$: **instrument the residual.** (Robinson 1988)

Partially linear IV — the estimate and its influence function

Step 3 — solve $\mathbb{E}_n[\psi] = 0$:

$$\hat{\tau} = -\frac{\mathbb{E}_n[\psi_b]}{\mathbb{E}_n[\psi_a]} = \frac{\mathrm{Cov}(\tilde{Y}, \tilde{Z})}{\mathrm{Cov}(\tilde{D}, \tilde{Z})} \qquad \text{(residualized 2SLS / Wald)}$$

Step 4 — influence function $\phi_i = -\psi_i / \mathbb{E}_n[\psi_a]$:

$$\phi_i = \frac{(\tilde{Y}_i - \tau \tilde{D}_i)\,\tilde{Z}_i}{\mathrm{Cov}(\tilde{D}, \tilde{Z})}, \qquad \widehat{\mathrm{SE}} = \sqrt{\frac{\mathbb{E}_n[\psi^2]}{\mathbb{E}_n[\psi_a]^2\, n}}$$

Exactly the **IV-robust SE** — the same Step 4 as PLR/AIPW; only ψ, ψ_a change.

Interactive IV (IIVM) — the model and the score

Drop the constant τ — let D enter g arbitrarily; keep the instrument equation:

$$Y = g(D, X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z, X] = 0$$

$$Z = r(X) + \nu, \qquad \mathbb{E}[\nu \mid X] = 0$$

- Nuisances: $\mu(Z, X) = \mathbb{E}[Y \mid Z, X]$, $p(Z, X) = \mathbb{E}[D \mid Z, X]$, $r(X) = \mathbb{E}[Z \mid X]$.
- Score = numerator $- \tau \times$ denominator, each an **AIPW signal:**

$$\psi = \underbrace{\Gamma^Y}_{\psi_b} + \underbrace{(-\Gamma^D)}_{\psi_a}\tau, \qquad \Gamma^Y = \mathrm{AIPW}(Z \to Y), \;\; \Gamma^D = \mathrm{AIPW}(Z \to D)$$

AIPW with a **ratio:** $\psi_a = -\Gamma^D$ (AIPW had -1). (Chernozhukov et al. 2018)

Interactive IV (IIVM) — LATE & influence function

Each signal is an AIPW pseudo-outcome (weighted by the instrument propensity r):

$$\Gamma^Y = \big[\mu(1,X) - \mu(0,X)\big] + \frac{Z\,(Y - \mu(1,X))}{r} - \frac{(1-Z)(Y - \mu(0,X))}{1-r}$$

$$\Gamma^D = \big[p(1,X) - p(0,X)\big] + \frac{Z\,(D - p(1,X))}{r} - \frac{(1-Z)(D - p(0,X))}{1-r}$$

Step 3 — solve: $\hat{\tau}_{\mathrm{LATE}} = -\dfrac{\mathbb{E}_n[\psi_b]}{\mathbb{E}_n[\psi_a]} = \dfrac{\mathbb{E}_n[\Gamma^Y]}{\mathbb{E}_n[\Gamma^D]}$ — ITT ÷ complier share.

Step 4 — influence function: $\phi_i = \dfrac{\Gamma_i^Y - \tau\,\Gamma_i^D}{\mathbb{E}_n[\Gamma^D]}, \qquad \widehat{\mathrm{SE}} = \sqrt{\dfrac{\mathbb{E}_n[\psi^2]}{\mathbb{E}_n[\psi_a]^2\, n}}$

Same Step 4 machinery; identification (relevance, exclusion, monotonicity) is inherited from S07.

How Wrong Could I Be?

DML assumes X is enough — it never **proves** it (orthogonality won't save you from an *unmeasured* confounder). So, exactly as in S06, we ask: **how strong would an unobserved confounder have to be to overturn the finding?** (Cinelli and Hazlett 2020)

The Riesz representer: the score's treatment side

Omitting a confounder U biases the estimate by the **product of two errors** — and which treatment object α appears depends on the **score**:

$$\text{bias} = \mathbb{E}\big[\ \underbrace{(g_0 - \hat{g})}_{\text{outcome-reg. error}}\ \underbrace{(\alpha_0 - \hat{\alpha})}_{\text{treatment-side error}}\ \big], \qquad g = \mathbb{E}[Y \mid D, X]$$

	treatment-side α	**so cf_d is…**
DML-PLR / -PLIV	residual $\tilde{D}$ ($\propto$ the coefficient)	a plain partial R^2 — **just sensemakr**
DML-AIPW / IRM / IIVM	**Riesz representer** $D/e - (1-D)/(1-e)$	the **Riesz-norm** increase — new

cf_y is always a genuine outcome R^2; only the **treatment dial** changes with the score.

Auto-DML: for a general target, *learn* the Riesz representer α from data — the orthogonal correction follows, no hand-derived propensity inversion. (Chernozhukov, Newey, and Singh 2022; Chernozhukov et al. 2021)

Three dials set the bias: two strengths and their alignment ρ

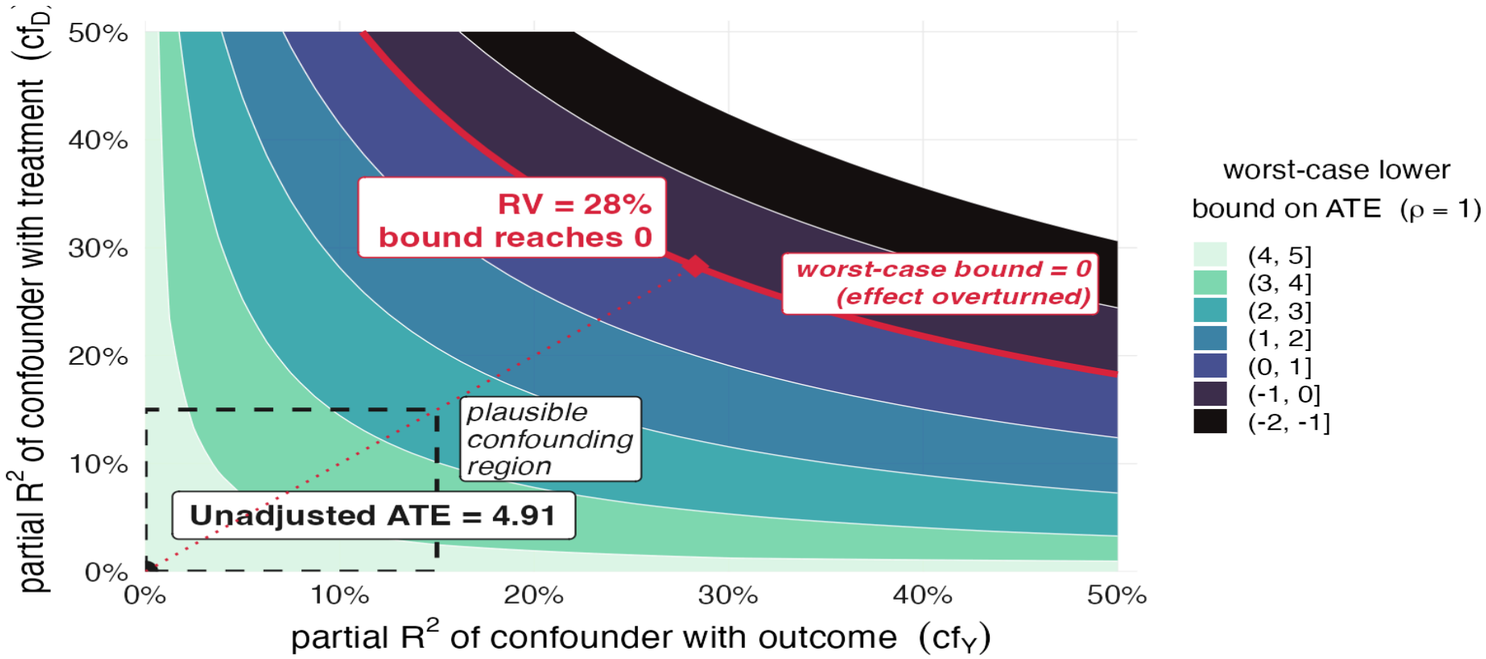
Cauchy–Schwarz turns that product into **magnitudes × correlation**:

$$|\text{bias}| = \underbrace{|\rho|}_{\text{alignment}} \cdot \underbrace{\sqrt{cf_y}}_{\text{outcome}} \cdot \underbrace{\sqrt{\frac{cf_d}{1-cf_d}}}_{\text{treatment}} \cdot S$$

- cf_y — partial R^2 of U with the **outcome** (its share of $g_0 - \hat{g}$),
- cf_d — partial R^2 of U with the **Riesz representer** (its share of $\alpha_0 - \hat{\alpha}$),
- $\rho = \text{Corr}\big(g_0 - \hat{g},\ \alpha_0 - \hat{\alpha}\big)$ — the **alignment** of U's two pulls.

$\rho = 0$ → **no bias** even if U is strong; $\rho = \pm 1$ → **maximum** bias. We report the worst case $\rho = 1$, so we argue only about *strength* — and the **robustness value (RV)** is the strength (at $\rho = 1$) that drives the bound to 0. (Chernozhukov et al. 2026)

How strong would hidden confounding have to be?



Python doubleml DoubleMLIRM (AIPW; RandomForest, max_features = 4, 5-fold) on learnloop_dml; sensitivity_analysis(rho = 1). Unadjusted ATE = 4.91, RV = 28%, RVa = 26%.

Across the plausible region the bound stays **positive** — overturning the effect needs a confounder about twice as strong as everything we measured.

The **AIPW/IRM** ATE (DoubleMLIRM): cf_d is the Riesz-norm dial. A PLR/PLIV effect would give the plain-sensemakr version.

Recap: one recipe, two ingredients, honest intervals

- **FWL with a machine**: partial X out of Y and D with any ML learner, then read the slope.
- **Orthogonality** kills regularization bias; **cross-fitting** kills overfitting bias — together they restore the $\sqrt{n}$, normal CI.
- One **score** per estimand: PLR (constant effect), AIPW/IRM (doubly-robust ATE), IIVM (LATE).
- DML solves **estimation, not identification** — so always **stress-test** unconfoundedness.

Next week (S09): stop averaging the score ψ_i and start **projecting** it on X — heterogeneous effects $\tau(x)$, meta-learners, causal forests. DML also powers DiD in Part V. (Chang 2020)

References

Ahrens, Achim, Christian B. Hansen, Mark E. Schaffer, and Thomas Wiemann. 2025. "Model Averaging and Double Machine Learning." *Journal of Applied Econometrics* 40 (3): 249–69. https://doi.org/10.1002/jae.3103.

Ballinari, Daniele, and Nora Bearth. 2024. "Improving the Finite Sample Estimation of Average Treatment Effects Using Double/Debiased Machine Learning with Propensity Score Calibration." https://arxiv.org/abs/2409.04874.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73. https://doi.org/10.1111/j.1541-0420.2005.00377.x.

Chang, Neng-Chieh. 2020. "Double/Debiased Machine Learning for Difference-in-Differences Models." *The Econometrics Journal* 23 (2): 177–91. https://doi.org/10.1093/ectj/utaa001.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68. https://doi.org/10.1111/ectj.12097.

Chernozhukov, Victor, Carlos Cinelli, Whitney K. Newey, Amit Sharma, and Vasilis Syrgkanis. 2026. "Long Story Short: Omitted Variable Bias in Causal Machine Learning." *Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest.a.1705.

Chernozhukov, Victor, Whitney K. Newey, Víctor Quintas-Martínez, and Vasilis Syrgkanis. 2021. "Automatic Debiased Machine Learning via Riesz Regression." https://arxiv.org/abs/2104.14737.

Chernozhukov, Victor, Whitney K. Newey, and Rahul Singh. 2022. "Automatic Debiased Machine Learning of Causal and Structural Effects." *Econometrica* 90 (3): 967–1027. https://doi.org/10.3982/ECTA18515.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society: Series B* 82 (1): 39–67. https://doi.org/10.1111/rssb.12348.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96 (1): 187–99. https://doi.org/10.1093/biomet/asn055.

Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1 (4): 387–401. https://doi.org/10.2307/1907330.

Lovell, Michael C. 1963. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association* 58 (304): 993–1010. https://doi.org/10.1080/01621459.1963.10480682.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66. https://doi.org/10.1080/01621459.1994.10476818.

Robinson, Peter M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–54. https://doi.org/10.2307/1912705.

Thank you for your attention!

TUHH
Institute of
Entrepreneurship
STARTUPENGINEER.io



www.startupengineer.io/team/ihl/

christoph-ihl

christophihl