ECON 7377 Microeconometrics **Spring 2026**

Debiased Inference and Double Machine Learning

Contents

Part I

Regularization Bias

1 The Post-Selection Inference Problem

1.1 Invalidity of Post-Selection Inference

Consider the linear model $y = X\beta_0 + \varepsilon$ with $\varepsilon \sim (0, \sigma^2 I_n)$. Suppose we use the Lasso (or any other model selection procedure) to identify a subset $\hat{S} \subset \{1, \ldots, p\}$ of active variables, and then perform OLS on the selected model. A fundamental difficulty arises:

- The Lasso estimator is *not* asymptotically Gaussian: the event $\hat{\beta}_j = 0$ occurs with positive probability, so the distribution of $\hat{\beta}_j$ has a point mass at zero mixed with a continuous component.
- Standard confidence intervals and t-tests assume that the model was chosen *a priori*, independently of the data. When the same data are used both to select the model and to estimate coefficients, the resulting inference is invalid.

- Leeb and Pötscher (2005) established that the distribution of post-model-selection estimators can be highly non-standard: it may be bimodal, have incorrect coverage, or behave erratically across different data-generating processes. Crucially, this is not a finite-sample issue but a *fundamental impossibility*: no uniformly valid confidence interval can be obtained by naively applying classical inference after model selection.
- The key insight of Leeb and Pötscher (2005) is that even if the Lasso selects the "correct" model with high probability, there are always sequences of data-generating processes where the selection is borderline and inference breaks down. The problem persists even as $n \to \infty$.

This motivates the search for methods that can provide valid inference in high-dimensional settings *without* relying on perfect model selection.

1.2 The Post-Lasso Estimator

Before turning to debiased methods, we briefly discuss the *post-Lasso* estimator, a natural two-step procedure designed to reduce the shrinkage bias of the Lasso.

Definition 1 (Post-Lasso Estimator)**.** The post-Lasso estimator is computed in two stages:

1. Run the Lasso and let $\hat{S} = \{j : \hat{\beta}_j^L \neq 0\}$ be the set of selected variables.
2. Compute the OLS estimator using only the variables in $\hat{S}$:

$$\hat{\beta}^{PL} = \underset{\beta \in \mathbb{R}^p,\, \beta_{\hat{S}^c}=0}{\arg\min} \sum_{i=1}^{n} (y_i - x_i'\beta)^2.$$

- By refitting via OLS on the selected support, the post-Lasso removes the shrinkage bias on the nonzero coefficients. Its prediction performance is comparable to the Lasso (see Chernozhukov et al., 2015).
- However, the post-Lasso estimator remains subject to the post-selection inference problem. It is *not* asymptotically normal and standard confidence intervals are invalid.

2 Regularization Bias

2.1 Selection vs. Estimation: An Impossibility

One might hope that the Lasso can simultaneously achieve two goals: (i) correctly identify the support of β_0 (model selection), and (ii) estimate β_0 at the optimal rate (estimation). Unfortunately, these objectives are in tension.

- Yang (2005) showed that for a model selection procedure to be consistent for model identification, it must behave sub-optimally in estimating the regression function, and vice versa.
- The penalty level λ required for sign consistency (the irrepresentability condition of Tibshirani, 1996) is very different from the penalty required for optimal prediction error ($\lambda \asymp \sigma\sqrt{\log p/n}$).
- The upshot: even with the Lasso, selecting relevant covariates and accurately estimating their effects are two objectives that *cannot* be pursued simultaneously.

2.2 The Naive Plug-In Estimator

We now formalize the regularization bias problem in the context of estimating a treatment effect with high-dimensional controls.

Assumption 1 (Linear Model with Controls)**.** *Consider the i.i.d. sequence* $(Y_i, D_i, X_i)_{i=1}^n$ *satisfying*

$$Y = D\tau_0 + X'\beta_0 + \varepsilon,$$

where $\mathbb{E}[\varepsilon|D, X] = 0$*,* $D \in \{0, 1\}$ *is a binary treatment indicator, and* $X \in \mathbb{R}^p$ *is a vector of controls with* p *potentially much larger than* n*. We denote* $\mu_d := \mathbb{E}[X|D = d]$ *for* $d \in \{0, 1\}$ *and* $\pi_0 := \mathbb{E}[D_i] \neq 0$*.*

Under Assumption 1, a naive two-step procedure to estimate τ_0 proceeds as follows:

1. **Selection**: Compute the Lasso of Y on (D, X), keeping D unpenalized, and exclude variables in X with zero Lasso coefficients.
2. **Estimation**: Compute OLS of Y on D and the selected elements of X.

The resulting estimator can be written as

$$\hat{\tau} = \frac{n^{-1}\sum_{i=1}^n D_i(Y_i - X_i'\hat{\beta})}{\hat{\pi}},$$

where $\hat{\pi} = n^{-1}\sum_{i=1}^n D_i$ and $\hat{\beta}$ is the post-selection OLS estimate.

Lemma 1 (Regularization Bias, Gaillac and L'Hour (2025))**.** *Under Assumption 1, if* $\mu_1 := \mathbb{E}[X|D = 1] \neq 0$*, then*

$$\sqrt{n}|\hat{\tau} - \tau_0| \to \infty.$$

Remark 1 (Intuition for Regularization Bias)**.** Lemma 1 shows that the single-equation selection procedure creates an *omitted variable bias*. The Lasso selects variables that are strongly related to Y, but may overlook variables that have a moderate direct effect on Y yet a strong effect on D. As Belloni et al. (2014) put it: "any such variable has a moderate direct effect on the outcome, which will be incorrectly misattributed to the effect of the treatment." The root cause is that the selection uses only one equation (the outcome equation), ignoring the treatment assignment mechanism.

2.3 A Favorable Case: Balanced Design

Under special conditions, the naive estimator can work. Examining when it does — and why — will motivate the general solution.

Assumption 2 (Growth Condition)**.** $s \log p / \sqrt{n} \to 0$*, where* $s = \|\beta_0\|_0$ *is the sparsity of* β_0*.*

Assumption 3 (Balanced Design)**.** *(i)* $\mu_1 = \mathbb{E}[X|D=1] = 0$ *(orthogonality).*

(ii) $\left\|n^{-1/2} \sum_{i=1}^n D_i X_i\right\|_\infty \lesssim \sqrt{\log p}$ *(concentration).*

Lemma 2 (A Favorable Case, Gaillac and L'Hour (2025))**.** *Under Assumptions 1, 2, and 3:*

$$\sqrt{n}(\hat{\tau} - \tau_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\pi_0}\right).$$

The key observation is that Assumption 3 implies

$$\mathbb{E}\left[\frac{\partial \sqrt{n}(\hat{\tau} - \tau_0)}{\partial(\hat{\beta} - \beta_0)}\right] = -\mathbb{E}\left[\frac{n^{-1/2}}{\hat{\pi}} \sum_{i=1}^n D_i X_i\right] \approx 0.$$

Under this condition, the estimator $\hat{\tau}$ is *first-order insensitive* to small deviations of $\hat{\beta}$ around β_0. This is the essential idea that the debiased methods and double machine learning will exploit in a general way: design the estimation procedure so that it is locally insensitive to errors in nuisance parameter estimation.

Part II

Debiased Lasso Inference

We now present methods that provide valid confidence intervals for individual regression coefficients in the high-dimensional linear model $y = X\beta_0 + \varepsilon$, *without* requiring that the Lasso correctly

identifies the active set. The key idea is to *debias* the Lasso estimator — correcting its shrinkage bias through a one-step Newton-type update.

1 The Zhang–Zhang (2014) Debiased Lasso

1.1 Setup and Motivation

Consider the linear model

$$y = X\beta_0 + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_n), \tag{1}$$

where $y \in \mathbb{R}^n$, $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n\times p}$ with $p > n$, and we want to construct a confidence interval for an individual coefficient $\beta_{0,j}$.

- When $p < n$, we can use the OLS estimator and form standard confidence intervals: $\hat{\beta}_j^{\text{OLS}} \pm z^{(\alpha/2)} v_j \hat{\sigma}$, where $v_j^2 = (X'X)^{-1}_{jj}$.
- When $p > n$, $X'X$ is singular, so OLS is not feasible and the Lasso estimator is biased.
- The *debiased Lasso* (also called the *low-dimensional projection estimator*, LDPE) proposed by Zhang and Zhang (2014) provides a bias correction that restores asymptotic normality.

1.2 Bias-Corrected Estimator and Score Vector

The classical OLS estimator of β_j can be written as

$$\hat{\beta}_j^{\text{OLS}} = \frac{(x_j^{\perp})' y}{(x_j^{\perp})' x_j},$$

where $X_{-j} = (x_k,\, k \neq j) \in \mathbb{R}^{n\times(p-1)}$ collects all columns of X except the j-th, and

$$x_j^{\perp} = x_j - X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}x_j = M_{-j}\, x_j, \qquad M_{-j} = I_n - X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j},$$

is the residual from the OLS projection of x_j onto the column space of X_{-j}, with M_{-j} the usual annihilator (residual-maker) matrix. Equivalently, $x_j^{\perp}$ is the component of x_j that is orthogonal to all other regressors, and $(x_j^{\perp})' x_j = \|x_j^{\perp}\|_2^2$. When $p > n$, the matrix $X'_{-j}X_{-j}$ is singular (since $\operatorname{rank}(X_{-j}) = n < p - 1$), so $x_j^{\perp}$ is not well-defined and this estimator breaks down. Zhang and Zhang (2014) propose to replace $x_j^{\perp}$ with a "relaxed" version z_j obtained via Lasso regression.

Definition 2 (Score Vector)**.** For each $j = 1, \ldots, p$, let $\hat{\gamma}_j$ be the Lasso coefficient from regressing x_j on X_{-j}:

$$\hat{\gamma}_j = \arg\min_b \left\{ \frac{\|x_j - X_{-j} b\|_2^2}{2n} + \lambda_j \|b\|_1 \right\}.$$

The *score vector* is the residual $z_j = x_j - X_{-j}\hat{\gamma}_j$.

The score vector z_j serves as a "relaxed" approximation to $x_j^{\perp}$. While z_j is not exactly orthogonal to all other columns of X, the Lasso ensures that its correlations with X_{-j} are small: $|x_k' z_j / n| \leq \lambda_j$ for $k \neq j$ (by the KKT conditions).

Definition 3 (Debiased Lasso / LDPE Estimator, Zhang and Zhang (2014))**.** The *low-dimensional projection estimator* (LDPE) of β_j is

$$\hat{\beta}_j^d = \hat{\beta}_j^{(\text{init})} + \frac{z_j'(y - X\hat{\beta}^{(\text{init})})}{z_j' x_j}, \tag{2}$$

where $\hat{\beta}^{(\text{init})}$ is the (scaled) Lasso estimator and z_j is the score vector from Definition 2.

To understand why this works, decompose the estimation error:

$$\hat{\beta}_j^d - \beta_{0,j} = \underbrace{\frac{z_j'\varepsilon}{z_j' x_j}}_{\text{noise}} + \underbrace{\sum_{k \neq j} \frac{z_j' x_k}{z_j' x_j} (\beta_{0,k} - \hat{\beta}_k^{(\text{init})})}_{\text{approximation error (bias)}}. \tag{3}$$

- The first term is approximately $N(0, \sigma^2 \|z_j\|_2^2 / (z_j' x_j)^2)$ — this is the "noise factor."
- The second term (bias) is controlled by two quantities:
  - $\eta_j = \max_{k \neq j} |z_j' x_k| / \|z_j\|_2$ — the *bias factor*, measuring how well z_j approximates orthogonality.
  - $\|\hat{\beta}^{(\text{init})} - \beta_0\|_1$ — the ℓ_1-error of the initial estimator.
- When η_j is small (ensured by the Lasso) and the initial estimator is consistent, the bias term vanishes and the debiased estimator is asymptotically normal.

The procedure is summarized as follows.

Algorithm 1: Debiased Lasso (Zhang and Zhang, 2014)

Input: Data (y, X), coordinate j, significance level α

Step 1: Compute an initial Lasso estimate $\hat{\beta}^{(\text{init})}$ and a noise level estimate $\hat{\sigma}$;

Step 2: Regress x_j on X_{-j} via Lasso to obtain the score vector z_j;

Step 3: Compute the LDPE: $\hat{\beta}_j^d = \hat{\beta}_j^{(\text{init})} + z_j'(y - X\hat{\beta}^{(\text{init})})/(z_j' x_j)$;

Step 4: Compute the noise factor $\tau_j = \|z_j\|_2 / |z_j' x_j|$;

Output: Point estimate $\hat{\beta}_j^d$ and confidence interval $[\hat{\beta}_j^d \pm \Phi^{-1}(1-\alpha/2)\,\hat{\sigma}\,\tau_j]$

1.3 Confidence Intervals

Under regularity conditions, the LDPE achieves asymptotic normality:

Theorem 1 (Asymptotic Normality of the LDPE, Zhang and Zhang (2014))**.** *Suppose* $\|\beta_0\|_0 \le s$ *and* $s\log(p)/\sqrt{n} \to 0$*, and assume appropriate regularity conditions on the design matrix* X*. Then:*

$$\frac{\hat{\beta}_j^d - \beta_{0,j}}{\hat{\sigma}\,\tau_j} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\tau_j = \|z_j\|_2 / |z_j' x_j|$.

- An approximate $(1-\alpha)$ confidence interval for $\beta_{0,j}$ is

$$\hat{\beta}_j^d \pm \Phi^{-1}(1-\alpha/2)\,\hat{\sigma}\,\tau_j.$$

- The result does *not* require the uniform signal strength condition

$$\min_{j:\,\beta_{0,j}\neq 0} |\beta_{0,j}| \ge C\sigma\sqrt{\log(p)/n}$$

  that is needed for variable selection consistency. This is the key advantage over selection-based approaches.

- The dimension condition $s\log(p)/\sqrt{n} \to 0$ is mild: it allows p to grow exponentially in n as long as the model is sufficiently sparse.

Remark 2. When $p < n$ and X has full column rank, the debiased Lasso reduces to the OLS estimator and the confidence intervals coincide with the classical ones. The debiased Lasso is therefore a genuine generalization of classical inference to the high-dimensional regime.

2 General Debiased Lasso Framework

2.1 Approximate Inverse Covariance Approach

The Zhang–Zhang procedure is a special case of a more general debiasing framework presented in Hastie et al. (2015). Let $\hat{\beta}_\lambda$ denote the Lasso estimator at regularization level λ and $\hat{\Sigma} = X'X/n$.

Proposition 1 (Debiased Lasso via Approximate Inverse)**.** Define the *debiased Lasso* estimator as

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{n}\Theta X'(y - X\hat{\beta}_\lambda), \tag{4}$$

where Θ is an approximate inverse of $\hat{\Sigma}$ (i.e., $\|\Theta\hat{\Sigma} - I_p\|_\infty$ is small). Then:

$$\hat{\beta}^d = \beta_0 + \frac{1}{n}\Theta X'\varepsilon + \underbrace{(I_p - \Theta\hat{\Sigma})(\hat{\beta}_\lambda - \beta_0)}_{\hat{\Delta}\text{ (remainder)}}. \tag{5}$$

- The first term $\Theta X'\varepsilon/n$ is a sum of independent random variables — it is asymptotically $N(0, \sigma^2\Theta\hat{\Sigma}\Theta'/n)$ by the CLT.
- The remainder $\hat{\Delta}$ is controlled by $\|\Theta\hat{\Sigma} - I_p\|_\infty \cdot \|\hat{\beta}_\lambda - \beta_0\|_1$, which vanishes under sparsity.
- The key challenge: constructing Θ so that $\|\Theta\hat{\Sigma} - I\|_\infty$ is small, even when $\hat{\Sigma}$ is singular ($p > n$).

2.2 Methods for Constructing Θ

Several methods have been proposed for constructing the approximate inverse Θ:

1. **Nodewise regression** (van de Geer et al., 2014): For each j, regress x_j on X_{-j} by Lasso to obtain residuals z_j. This constructs Θ row by row. It can be interpreted as estimating the *precision matrix* Σ^{-1} via neighborhood-based sparse graph estimation. This is precisely the approach of Zhang and Zhang (2014) applied coordinate-wise.

2. **Convex program** (Javanmard and Montanari, 2014): For each j, define the j-th row m_j of Θ as the solution to

$$\begin{aligned} \min_{m \in \mathbb{R}^p} \quad & m'\hat{\Sigma}m \\ \text{subject to} \quad & \|\hat{\Sigma}m - e_j\|_\infty \le \gamma, \end{aligned}$$

   where e_j is the j-th unit vector. This directly minimizes the variance of $\hat{\beta}_j^d$ subject to the constraint that the bias is controlled.

Remark 3. Both methods produce Θ such that $\hat{\Sigma}\Theta \approx I$ in the ℓ_∞ sense. The nodewise regression approach is more common in the econometrics literature due to its simplicity and interpretability. The convex program approach of Javanmard and Montanari (2014) can yield somewhat smaller confidence intervals by explicitly minimizing the variance.

2.3 Asymptotic Normality

Using either method for constructing Θ, the resulting debiased estimator $\hat{\beta}^d$ from (4) satisfies:

Theorem 2 (Asymptotic Normality, General Debiased Lasso)**.** *Suppose $\|\beta_0\|_0 \leq s$ with $s\log(p)/\sqrt{n} \to 0$, and assume appropriate restricted eigenvalue conditions on X. Let Θ be constructed by either nodewise regression or the convex program above, so that $\|\hat{\Delta}\|_\infty = \|(I_p - \Theta\hat{\Sigma})(\hat{\beta}_\lambda - \beta_0)\|_\infty \xrightarrow{p} 0$. Then, for each $j = 1, \ldots, p$:*

$$\frac{\hat{\beta}_j^d - \beta_{0,j}}{\sigma\sqrt{(\Theta\hat{\Sigma}\Theta')_{jj}/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

In particular, an approximate $(1-\alpha)$ confidence interval for $\beta_{0,j}$ is

$$\hat{\beta}_j^d \pm \Phi^{-1}(1-\alpha/2)\,\hat{\sigma}\sqrt{(\Theta\hat{\Sigma}\Theta')_{jj}/n},$$

where $\hat{\sigma}$ is a consistent estimator of σ.

Part III

Post-Double-Selection

1 The Partially Linear Model

1.1 Model Setup

Assumption 4 (Partially Linear Model with High-Dimensional Controls)**.** *The data $(y_i, d_i, z_i)_{i=1}^n$ are i.i.d. and satisfy:*

$$y_i = d_i\alpha_0 + g(z_i) + \zeta_i, \quad \mathbb{E}[\zeta_i | z_i, d_i] = 0, \tag{6}$$

$$d_i = m(z_i) + v_i, \quad \mathbb{E}[v_i | z_i] = 0, \tag{7}$$

where y_i is the outcome, d_i is the treatment/policy variable, z_i is a vector of confounding factors, α_0 is the treatment effect of interest, $g(\cdot)$ and $m(\cdot)$ are unknown functions, and ζ_i, v_i are disturbances.

- We use linear combinations of p potential regressors $x_i = P(z_i)$ to approximate $g(z_i) \approx x_i'\beta_{g0}$ and $m(z_i) \approx x_i'\beta_{m0}$, where p may be much larger than n.
- The partially linear model was introduced by Robinson (1988). In classical settings with $p \ll n$, estimation of α_0 proceeds via the Frisch-Waugh-Lovell (FWL) theorem. The challenge here is that $p \gg n$.
- Under approximate sparsity, the functions g and m can each be well-approximated by a small number of terms: $\|\beta_{g0}\|_0 \leq s$ and $\|\beta_{m0}\|_0 \leq s$ with $s \ll n$.

1.2 Why Single Selection Fails

Example 1 (Single Selection Bias, $p = 1$)**.** To illustrate the problem, consider the simplest case with a single control x:

$$y_i = \alpha_0 d_i + \beta_g x_i + \zeta_i,$$
$$d_i = \beta_m x_i + v_i.$$

A single-equation model selection procedure (Lasso on the outcome equation only) will drop x_i whenever $|\beta_g|$ is small enough relative to the noise level. But even when β_g is small, the omitted variable bias is

$$\text{bias} \propto \frac{\beta_g \cdot \beta_m \cdot \sigma_x^2}{\sigma_d^2},$$

which can be large if β_m is large (i.e., x is strongly related to treatment).

Under sequences where $\beta_g \to 0$ slowly, the single-selection estimator is *not* root-n consistent. Its distribution is bimodal: sometimes x is selected (correct behavior) and sometimes it is dropped (omitted variable bias). See the left panel of Figure 1.

Remark 4. The double-selection estimator resolves this by selecting controls from *both* equations. The variable x is only omitted when its coefficients in *both* equations are small, which ensures that the omitted variable bias is negligible. See the right panel of Figure 1.

2 The Post-Double-Selection Procedure

2.1 The Algorithm

Algorithm 2: Post-Double-Selection (Belloni et al., 2014)

Input: Data $(y_i, d_i, x_i)_{i=1}^n$

Step 1 (Selection on treatment): Run Lasso of d on x, obtain $\hat{S}_D = \{j : \hat{\delta}_j^L \neq 0\}$;

Step 2 (Selection on outcome): Run Lasso of y on x, obtain $\hat{S}_Y = \{j : \hat{\beta}_j^L \neq 0\}$;

Step 3 (Estimation): Run OLS of y on d and the controls in $\hat{S}_D \cup \hat{S}_Y$;

Output: Treatment effect estimate $\check{\alpha}$, standard errors, confidence intervals

- The two selection steps can use either Lasso or post-Lasso. The analyst may also include additional controls $\hat{S}_3$ (an "amelioration set") based on domain knowledge.
- The final OLS regression uses d_i and $x_{i,\hat{S}}$ where $\hat{S} = \hat{S}_D \cup \hat{S}_Y \cup \hat{S}_3$.
- Inference on α_0 is performed using conventional OLS standard errors from the final regression.

2.2 Asymptotic Normality

Theorem 3 (Estimation and Inference on Treatment Effects, Belloni et al. (2014))**.** *Under approximate sparsity conditions on $g(\cdot)$ and $m(\cdot)$, sparse eigenvalue conditions on the Gram matrix, and structural moment conditions, the post-double-selection estimator $\check{\alpha}$ satisfies:*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \xrightarrow{d} N(0,1),$$

where $\sigma_n^2 = [\mathbb{E}[v_i^2]]^{-1}\mathbb{E}[v_i^2\zeta_i^2][\mathbb{E}[v_i^2]]^{-1}$. Under homoscedasticity ($\mathbb{E}[\zeta_i^2|z_i] = \sigma_\zeta^2$), the asymptotic variance reduces to $\sigma_\zeta^2(\sigma_v^2)^{-1}$, which is the semiparametric efficiency bound of Robinson (1988).

A consistent estimator of the asymptotic variance is

$$\hat{\sigma}_\alpha^2 = \left[\frac{1}{n}\sum_{i=1}^n \hat{v}_i^2\right]^{-2} \frac{1}{n - \hat{s} - 1}\sum_{i=1}^n \hat{v}_i^2\hat{\zeta}_i^2,$$

where $\hat{\zeta}_i = y_i - \check{\alpha}d_i - x_i'\hat{\beta}$, $\hat{v}_i = d_i - x_i'\hat{\delta}$, and $\hat{s} = |\hat{S}_D \cup \hat{S}_Y|$. A $(1-\alpha)$-confidence interval is $\check{\alpha} \pm \Phi^{-1}(1-\alpha/2)\,\hat{\sigma}_\alpha/\sqrt{n}$.

2.3 Connection to Orthogonal Moments and FWL

The success of the double-selection procedure can be understood through the lens of *Neyman orthogonality*, which will be central to the DML framework in Part IV.

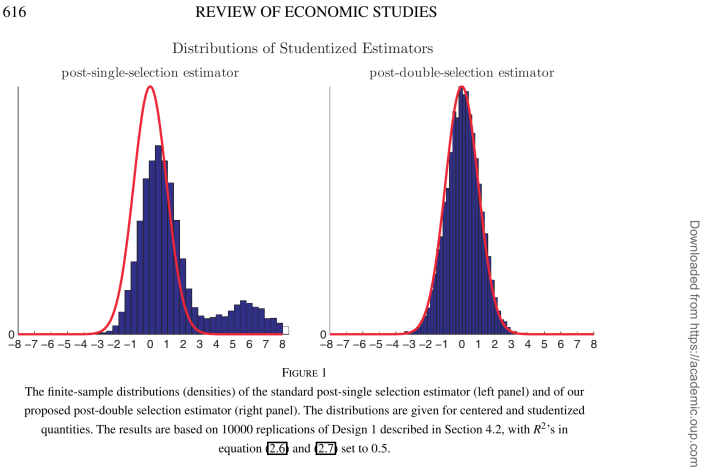

Figure 1: Finite-sample distributions of the treatment effect estimator under single selection (left, bimodal) and post-double-selection (right, approximately normal). The single-selection estimator can exhibit severe omitted variable bias. Source: Belloni et al. (2014, Figure 1).

Consider the moment condition underlying the treatment effect:

$$\mathbb{E}[\psi(Z, \alpha_0, \eta_0)] := \mathbb{E}\left[(Y - D\alpha_0 - X'\beta_0)\,(D - X'\delta_0)\right] = 0, \tag{8}$$

where $\eta_0 = (\beta_0', \delta_0')'$ collects the nuisance parameters. This is the Frisch-Waugh-Lovell moment condition: α_0 is the coefficient from regressing outcome residuals on treatment residuals.

Proposition 2 (Neyman Orthogonality of the Double-Selection Moment)**.** Under Assumption 4, the moment function $\psi(Z, \alpha, \eta)$ satisfies

$$\mathbb{E}\left[\partial_\eta \psi(Z, \alpha_0, \eta_0)\right] = 0.$$

That is, the moment condition (8) is first-order insensitive to errors in estimating the nuisance parameters (β_0, δ_0).

Proof. The Jacobian with respect to $\eta = (\beta, \delta)$ is

$$\partial_\eta \psi = \begin{pmatrix} -(D - X'\delta)X \\ -(Y - D\alpha - X'\beta)X \end{pmatrix}.$$

At the true values, $\mathbb{E}[\partial_\beta \psi] = -\mathbb{E}[\xi X] = 0$ by the treatment equation, and $\mathbb{E}[\partial_\delta \psi] = -\mathbb{E}[\varepsilon X] = 0$ by the outcome equation. □

- The orthogonal moment (8) is the FWL representation:

$$\alpha_0 = \frac{\text{Cov}[D - X'\delta_0, Y - X'\beta_0]}{\mathbb{E}[(D - X'\delta_0)^2]}.$$

  By partialling out X from *both* Y and D, we use only the variation in D and Y that is *orthogonal* to X. Small errors in estimating $g(\cdot)$ and $m(\cdot)$ have only higher order effects.

- This orthogonality is exactly what the balanced design assumption (Assumption 3) provided in the favorable case of Part I — but now it is *built into the estimator* rather than assumed as a property of the data.

- The double-selection procedure creates the necessary orthogonality by performing variable selection in *both* equations, ensuring that the relevant confounders are controlled for regardless of which equation they appear in.

- This perspective leads naturally to the *double machine learning* framework (Part IV), which generalizes this idea beyond Lasso to arbitrary machine learning methods.

Part IV

Double Machine Learning

The post-double-selection method of Part III uses the Lasso for variable selection in a *linear* partially linear model. *Double/debiased machine learning* (DML), introduced by Chernozhukov et al. (2018), generalizes this to allow *any* machine learning method for estimating nuisance functions, including random forests, neural networks, and boosting. The two key ingredients are *Neyman orthogonality* and *cross-fitting*.

1 The General DML Framework

1.1 Semiparametric Setup and Regularization Bias

Consider the general semiparametric estimation problem where we observe an i.i.d. sample $\{W_i\}_{i=1}^n$ and want to estimate a low-dimensional target parameter $\theta_0 \in \mathbb{R}^{d_\theta}$ that is identified by a moment condition:

$$\mathbb{E}[m(W; \theta_0, \eta_0)] = 0, \tag{9}$$

where $m(\cdot;\theta,\eta)$ is a known score (moment) function and η_0 is a nuisance parameter that may be high-dimensional or infinite-dimensional (e.g., a conditional expectation function).

- In the partially linear model, θ_0 is the treatment effect, $\eta_0 = (g_0, m_0)$ are the conditional expectation functions, and $W = (Y, D, X)$.
- In the ATE setting, $\theta_0 = \mathbb{E}[Y(1) - Y(0)]$, $\eta_0 = (e_0, \mu_0)$ are the propensity score and outcome regression.
- The nuisance parameter η_0 is typically estimated using ML methods.

A natural *plug-in* estimator replaces η_0 with an ML estimate $\hat{\eta}$ and solves the sample analog:

$$\frac{1}{n}\sum_{i=1}^{n} m(W_i;\hat{\theta},\hat{\eta}) = 0.$$

However, this plug-in estimator generally fails to be $\sqrt{n}$-consistent due to two biases (Chernozhukov et al., 2018; Ahrens et al., 2025):

1. **Regularization bias**: ML estimators introduce bias through regularization (e.g., shrinkage in Lasso, pruning in trees). Since the score m is generally sensitive to perturbations in η, this bias propagates into $\hat{\theta}$.

   Formally, a Taylor expansion gives:

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \underbrace{-J_{\theta\theta}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(W_i;\theta_0,\eta_0)}_{\text{CLT term}} + \underbrace{J_{\theta\theta}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}m(W_i;\theta_0,\eta_0)(\hat{\eta}-\eta_0)}_{(\star)\text{: first-order impact of nuisance estimation}}, \tag{10}$$

   where $J_{\theta\theta} = \mathbb{E}[\partial_\theta m(W;\theta_0,\eta_0)]$. The term $(\star)$ generally does not vanish when $\partial_\eta m \neq 0$.

2. **Overfitting bias**: When $\hat{\eta}$ is estimated on the same data used to evaluate the moment condition, the estimation errors $\hat{\eta} - \eta_0$ are correlated with the scores $m(W_i;\theta_0,\eta_0)$, creating a bias even when $\mathbb{E}[\partial_\eta m] = 0$.

1.2 Neyman Orthogonality

The first key ingredient of DML addresses regularization bias.

Definition 4 (Neyman Orthogonality)**.** The score function $m(W;\theta,\eta)$ satisfies *Neyman orthogonality* at (θ_0,η_0) with respect to a nuisance realization set $\mathcal{T}_N \subset \mathcal{T}$ if

$$\frac{\partial}{\partial r}\mathbb{E}\left[m(W;\theta_0,\eta_0 + r(\eta-\eta_0))\right]\Big|_{r=0} = 0, \quad \forall\eta\in\mathcal{T}_N. \tag{11}$$

- Intuitively, condition (11) means that the moment condition is *locally insensitive* to perturbations of the nuisance parameter around its true value. Small errors in estimating η_0 have only a *second-order* effect on the moment condition.

- Returning to the expansion (10): Neyman orthogonality ensures that $\mathbb{E}[\partial_\eta m(W;\theta_0,\eta_0)] = 0$, so the problematic term $(\star)$ becomes

$$(\star) = J_{\theta\theta}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial_\eta m(W_i;\theta_0,\eta_0)(\hat{\eta}-\eta_0),$$

  which is $\sqrt{n}$ times a sample average of mean-zero terms (times $\hat{\eta}-\eta_0$). Under appropriate rate conditions, this vanishes.

- **Rate requirement**: The nuisance estimators need to satisfy a product rate condition: $\sqrt{n}\|\hat{\eta}-\eta_0\|^2 \to 0$, or equivalently, the mean-squared convergence rate faster than $n^{-1/4}$. This is achievable by many ML methods under structural assumptions (sparsity, smoothness, etc.).

Remark 5 (Why Neyman Orthogonality Matters)**.** Without orthogonality, the bias term $(\star)$ in (10) involves $\sqrt{n}$ times $\mathbb{E}[\partial_\eta m]\cdot(\hat{\eta}-\eta_0)$. Even if $\hat{\eta}$ converges at rate $n^{-1/4}$ (typical for nonparametric ML), this gives $\sqrt{n}\cdot n^{-1/4} = n^{1/4} \to \infty$. With orthogonality, the bias depends on $\sqrt{n}\|\hat{\eta}-\eta_0\|^2 \lesssim \sqrt{n}\cdot n^{-1/2} \to 0$. Orthogonality converts a *first-order* bias into a *second-order* one.

1.3 Cross-Fitting

The second key ingredient of DML addresses overfitting bias.

Definition 5 (K-Fold Cross-Fitting)**.** Randomly partition $\{1,\dots,n\}$ into K folds $I_1,\dots,I_K$ of approximately equal size. For each fold $k = 1,\dots,K$:

(i) Estimate $\hat{\eta}_k$ using data $\{W_i : i \notin I_k\}$ (all observations *except* fold k).

(ii) Evaluate the score $m(W_i;\theta,\hat{\eta}_k)$ for observations $i \in I_k$.

The cross-fitted DML estimator $\check{\theta}$ solves:

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k} m(W_i;\check{\theta},\hat{\eta}_k) = 0. \tag{12}$$

- Since $\hat{\eta}_k$ is estimated on data *excluding* fold k, the estimation errors $\hat{\eta}_k - \eta_0$ are *independent* of the observations $\{W_i : i \in I_k\}$ used in the moment condition. This breaks the dependence that causes overfitting bias.

- By rotating the roles of folds, cross-fitting ensures that *all* observations contribute to both nuisance estimation and parameter estimation, preserving efficiency.
- Cross-fitting is analogous to cross-validation, but serves a different purpose: it is not for tuning parameter selection, but for ensuring valid inference.
- Typically $K = 5$ or $K = 10$ is used in practice.

2 DML for the Partially Linear Model

2.1 Setup and Two Score Functions

Consider the partially linear regression (PLR) model as in (6)–(7):

$$Y = D\theta_0 + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0, \tag{13}$$

$$D = m_0(X) + V, \quad \mathbb{E}[V|X] = 0, \tag{14}$$

where θ_0 is the target parameter, $\eta_0 = (\ell_0, r_0) = (\mathbb{E}[Y|X], \mathbb{E}[D|X])$ are the nuisance functions, and $g_0(X) = \ell_0(X) - \theta_0 r_0(X)$, $U = Y - \ell_0(X)$, $V = D - r_0(X)$ (Robinson, 1988).

Two natural score functions identify θ_0:

1. **Naive score (not orthogonal):**

$$m_{\text{naive}}(W; \theta, \eta) = (Y - g(X) - \theta D)D. \tag{15}$$

   This corresponds to regressing $Y - g(X)$ on D. The nuisance is $\eta(X) = g(X)$.

2. **Orthogonal score (FWL / Robinson):**

$$m_{\text{PLR}}(W; \theta, \eta) = [(Y - \ell(X)) - \theta(D - r(X))]\,(D - r(X)). \tag{16}$$

   This corresponds to the FWL approach: partial out X from *both* Y and D and regress the residuals. The nuisances are $\eta(X) = (\ell(X), r(X))$.

Remark 6. While both scores identify θ_0 (i.e., $\mathbb{E}[m(\cdot; \theta_0, \eta_0)] = 0$), only the orthogonal score m_{PLR} is suitable for DML. The naive score is sensitive to errors in estimating $g(X)$: any mistake in $\hat{g}$ that is correlated with D directly biases $\hat{\theta}$.

2.2 Verifying Neyman Orthogonality

Proposition 3 (Neyman Orthogonality of the PLR Score)**.** The score m_{PLR} in (16) satisfies Neyman orthogonality at (θ_0, η_0), while the naive score m_{naive} in (15) does not.

Proof. **Naive score is not orthogonal:** Let $\Delta g(X) = g(X) - g_0(X)$. Then

$$\frac{\partial}{\partial r}\mathbb{E}[m_{\text{naive}}(W; \theta_0, g_0 + r\Delta g)]\Big|_{r=0} = \mathbb{E}[-\Delta g(X) \cdot D],$$

which is generally nonzero when D and X are correlated.

Orthogonal score: Let $\eta(X) = (\ell(X), r(X))$ with perturbation $\Delta\eta = (\Delta\ell, \Delta r)$ where $\Delta\ell = \ell - \ell_0$, $\Delta r = r - r_0$. Then:

$$\begin{aligned}
&\frac{\partial}{\partial\lambda}\mathbb{E}[m_{\text{PLR}}(W; \theta_0, \eta_0 + \lambda\Delta\eta)]\Big|_{\lambda=0} \\
&= \mathbb{E}[-\Delta\ell(X)(D - r_0(X)) - \Delta r(X)(Y - \ell_0(X)) + 2\theta_0\Delta r(X)(D - r_0(X))] \\
&= -\mathbb{E}[\Delta\ell(X)V] - \mathbb{E}[\Delta r(X)U] + 2\theta_0\mathbb{E}[\Delta r(X)V] = 0,
\end{aligned}$$

where the last equality uses $\mathbb{E}[V|X] = 0$, $\mathbb{E}[U|X] = 0$, and $\ell_0(X) = \mathbb{E}[Y|X]$, $r_0(X) = \mathbb{E}[D|X]$. □

2.3 The DML Estimator and Asymptotic Normality

Using the orthogonal score m_{PLR} from (16) with cross-fitting, the DML estimator for the PLR model is:

$$\check{\theta} = \left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\hat{V}_{i,k}D_i\right)^{-1}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\hat{V}_{i,k}(Y_i - \hat{\ell}_k(X_i))\right), \tag{17}$$

where $\hat{V}_{i,k} = D_i - \hat{r}_k(X_i)$ are the treatment residuals and $\hat{\ell}_k, \hat{r}_k$ are nuisance function estimates obtained from fold-k out-of-sample predictions.

Theorem 4 (Asymptotic Normality of DML, Chernozhukov et al. (2018))**.** *Under Neyman orthogonality of the score, K-fold cross-fitting, and the rate condition*

$$\|\hat{\ell} - \ell_0\|_2 \cdot \|\hat{r} - r_0\|_2 = o_P(n^{-1/2}),$$

the DML estimator $\check{\theta}$ satisfies:

$$\sqrt{n}(\check{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \mathbb{E}[V^2U^2]/(\mathbb{E}[V^2])^2$. Under homoscedasticity, $\sigma^2 = \sigma_U^2/\sigma_V^2$, the semiparametric efficiency bound.

Remark 7 (The Product Rate Condition)**.** The condition $\|\hat{\ell} - \ell_0\|_2 \cdot \|\hat{r} - r_0\|_2 = o_P(n^{-1/2})$ is the key requirement. It allows each nuisance function to converge at a rate *slower* than $n^{-1/4}$ — for example, $n^{-1/3}$ for one and $n^{-1/6}$ for the other. Many ML methods (Lasso, random forests, neural networks, boosting) achieve such rates under structural assumptions (Ahrens et al., 2025). This product rate condition is a direct consequence of Neyman orthogonality: without it, a single rate of $n^{-1/2}$ would be required, which is unattainable for most nonparametric estimators.

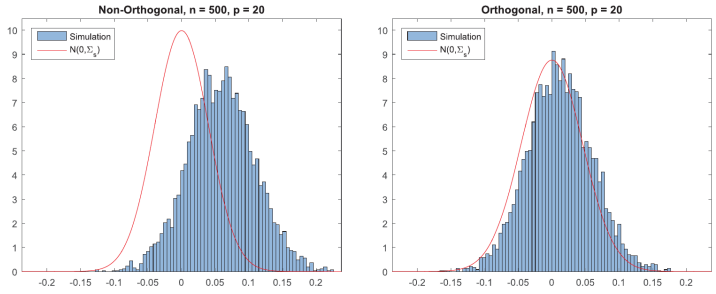

Figure 1. Comparison of the conventional and double ML estimators. [Colour figure can be viewed at wileyonlinelibrary.com]

splitting will play a key role in allowing us to guarantee that $c^* = o_P(1)$ under weak conditions as outlined below and discussed in detail in Section 3.

Figure 1 provides a numerical illustration of the negative impact of regularization bias and the benefit of orthogonalization. The left panel shows the behaviour of a conventional (non-

Figure 2: Comparison of conventional (non-orthogonal, left) and DML (orthogonal, right) estimators in the PLR model with random forest nuisance estimators. The conventional estimator is badly biased; the DML estimator is approximately normal and centered at the truth. Source: Chernozhukov et al. (2018, Figure 1).

2.4 DML Algorithm

The full DML procedure for the PLR model is summarized below.

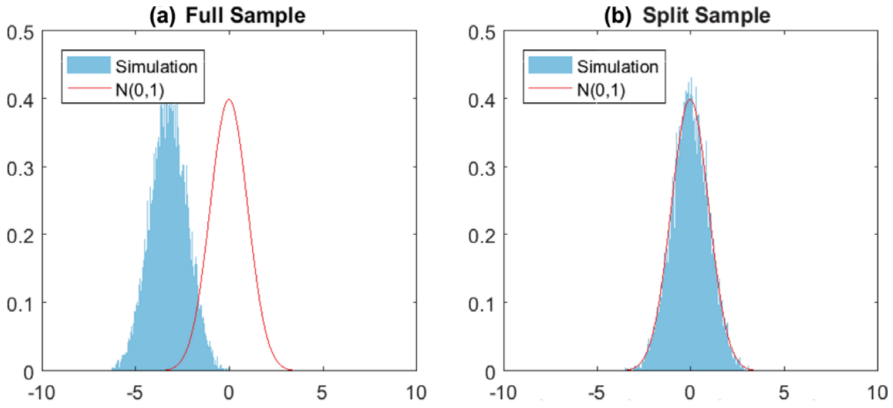

Figure 2. Comparison of full-sample and cross-fitting procedures. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 3: Comparison of full-sample (left) and cross-fitted (right) DML estimators. Without cross-fitting, overfitting bias distorts the distribution even with orthogonal scores. Cross-fitting removes this bias at no cost to efficiency. Source: Chernozhukov et al. (2018, Figure 2).

Algorithm 3: DML2 for the Partially Linear Model (Chernozhukov et al., 2018)

Input: Data $(Y_i, D_i, X_i)_{i=1}^n$, number of folds K, ML method $\mathcal{M}$

Step 1: Randomly partition $\{1, \ldots, n\}$ into K folds $I_1, \ldots, I_K$ of approximately equal size;

for $k = 1, \ldots, K$ **do**

  Step 2a: Estimate $\hat{\ell}_k(\cdot)$ by regressing Y on X using $\mathcal{M}$ on data $\{i \notin I_k\}$;

  Step 2b: Estimate $\hat{r}_k(\cdot)$ by regressing D on X using $\mathcal{M}$ on data $\{i \notin I_k\}$;

  Step 2c: Compute residuals $\hat{U}_{i,k} = Y_i - \hat{\ell}_k(X_i)$ and $\hat{V}_{i,k} = D_i - \hat{r}_k(X_i)$ for $i \in I_k$;

end

Step 3: Compute $\check{\theta} = \left(\sum_k \sum_{i \in I_k} \hat{V}_{i,k} D_i\right)^{-1} \left(\sum_k \sum_{i \in I_k} \hat{V}_{i,k} \hat{U}_{i,k}\right)$;

Step 4: Estimate variance: $\hat{\sigma}^2 = \left[n^{-1} \sum_i \hat{V}_i^2\right]^{-2} n^{-1} \sum_i \hat{V}_i^2 (Y_i - D_i \check{\theta} - \hat{\ell}(X_i))^2$;

Output: $\check{\theta}$, standard error $\hat{\sigma}/\sqrt{n}$, CI: $\check{\theta} \pm z_{\alpha/2}\, \hat{\sigma}/\sqrt{n}$

Remark 8 (Software)**.** DML is implemented in several software packages. The `DoubleML` package is available in R and Python. The `hdm` R package implements the post-double-selection method. See Ahrens et al. (2025) for a comprehensive review of available tools and practical recommendations.

`Zhan Gao, April 8, 2026`

References

Ahrens, A., V. Chernozhukov, C. Hansen, D. Kozbur, M. Schaffer, and T. Wiemann (2025). An introduction to double/debiased machine learning. *arXiv preprint arXiv:2504.08324*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics 7*(1), 649–688.

Gaillac, C. and J. L'Hour (2025). *Machine Learning for Econometrics*. Oxford University Press.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*(1), 2869–2909.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika 92*(4), 937–950.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(1), 217–242.