Statistical Inference on Predictive and Causal Effects in High-Dimensional Linear Regression Models

Gauranga Kr. Baishya, Chennai Mathematical Institute (CMI)

May 14, 2025

The Double Lasso: Frisch-Waugh-Lovell Partialling-Out

The key to inference is the application of the Frisch-Waugh-Lovell partialling-out. Consider the simple predictive model:

$$Y = \alpha D + \beta' W + \epsilon,$$

where D is the target regressor and W consists of p controls. After partialling out W, we obtain the residualized model:

$$\tilde{Y} = \alpha \tilde{D} + \epsilon, \quad \text{with } E[\epsilon \, \tilde{D}] = 0,$$

where the variables with tildes are the residuals from removing the linear effect of W (typically via linear regression).

The Double Lasso

The Double Lasso procedure is used to estimate the effect of a key variable D_i on the outcome Y_i while selecting relevant control variables W_i from high-dimensional data. It involves two main steps:

1. Running Lasso regressions to partial-out the effects of W_i from both Y_i and D_i.
2. Running an OLS regression on the residuals.

We run two separate Lasso regressions:

$$\hat{\gamma}_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i \left(Y_i - \gamma' W_i\right)^2 + \lambda_1 \sum_j \hat{\psi}_{Y,j} \left|\gamma_j\right|$$

$$\hat{\gamma}_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i \left(D_i - \gamma' W_i\right)^2 + \lambda_2 \sum_j \hat{\psi}_{D,j} \left|\gamma_j\right|$$

- $\hat{\gamma}_{YW}$ selects controls in W_i that best predict Y_i.
- $\hat{\gamma}_{DW}$ selects controls in W_i that best predict D_i.

Double Lasso: Obtaining Residuals

Using the estimates from the Lasso regressions, we obtain the residuals:

$$\check{Y}_i = Y_i - \hat{\gamma}'_{YW} W_i$$

$$\check{D}_i = D_i - \hat{\gamma}'_{DW} W_i$$

These residuals represent the parts of Y_i and D_i that are orthogonal to (or "cleaned of") the controls W_i. With the residuals $\check{Y}_i$ and $\check{D}_i$ in hand, we perform a simple OLS regression:

$$\hat{\alpha} = \arg\min_{a \in \mathbb{R}} \sum_i \left(\check{Y}_i - a\,\check{D}_i\right)^2$$

$$\hat{\alpha} = \frac{\sum_i \check{D}_i\,\check{Y}_i}{\sum_i \check{D}_i^2}$$

This final step estimates the effect α of D_i on Y_i after removing the influence of the controls.

Why this?

1. Run separate Lasso regressions of Y_i on W_i and D_i on W_i to select important controls.
2. Compute residuals $\check{Y}_i$ and $\check{D}_i$ by subtracting the predicted parts.
3. Regress $\check{Y}_i$ on $\check{D}_i$ to estimate the parameter α.

By removing (or "partialling out") the influence of W on both the outcome Y and the treatment D, we ensure that the remaining variation in D (denoted by $\tilde{D}$) is orthogonal to W. This guarantees that the estimated coefficient α in the regression

$$\tilde{Y} = \alpha \tilde{D} + \epsilon$$

is not confounded by the effects of W. In other words, it isolates the causal impact of D on Y. This method helps in obtaining a consistent estimate of α in the presence of high-dimensional control variables.

Inference with Double Lasso in High-Dimensional Regression

Under some regularity conditions,

$$\sqrt{n}(\hat{\alpha} - \alpha) \approx \frac{\sqrt{n}\,\mathbb{E}_n[\tilde{D}\,\epsilon]}{\mathbb{E}_n[\tilde{D}^2]} \overset{a}{\sim} N(0, V),$$

where

$$V = \left(E[\tilde{D}^2]\right)^{-1} E\left[\tilde{D}^2\,\epsilon^2\right] \left(E[\tilde{D}^2]\right)^{-1}.$$

Proof has been skipped.This result implies, for example, that the interval

$$\left[\hat{\alpha} \pm 1.96\sqrt{\frac{\hat{V}}{n}}\right]$$

covers α about 95% of the time.

An example

Consider the linear regression model:

$$Y = \alpha D + \beta' W + \epsilon.$$

We want to investigate how economic growth rates (Y) are related to a country's initial wealth (D), while controlling for institutional, educational, and other characteristics (W). The coefficient α represents the "speed of convergence or divergence." :

- $\alpha < 0$ [1] implies that poor countries grow faster than rich countries (convergence).
- $\alpha > 0$ implies that poor countries grow more slowly and fall further behind rich countries (divergence).

Our focus is on determining if poor countries indeed grow faster than rich countries — is $\alpha < 0$?.

[1] corresponds to the Convergence Hypothesis predicted by the Solow growth model, Robert M. Solow, MIT.

An Example

Estimation: OLS is Expected to provide a noisy estimate of α while Double Lasso (Partialling-out) is Expected to yield a high-quality estimate.

Method	Estimate	Std. Error	95% CI
OLS	-0.009	0.032	$[-0.073,\ 0.054]$
Double Lasso	-0.045	0.018	$[-0.080,\ -0.010]$

Interpretation: OLS produces a wide confidence interval (covering both positive and negative values) and may be unreliable when p/n is not small (Here $p = 60$ and $n = 90$). However, Double Lasso provides a precise estimate: a point estimate of -4.5% with a 95% CI from -8% to -1%, supporting the conditional convergence hypothesis.

What do we infer?

Ordinary least squares (OLS) can struggle in certain settings when there are many controls relative to the number of observations, which can lead to noisy and unreliable estimates. By applying the double Lasso method, we "partial out" the influence of these controls, effectively selecting only the most relevant ones. This results in a much cleaner and more precise estimate of the true causal effect of initial wealth on growth (captured by α). If α turns out to be negative, it supports the idea that poorer countries are catching up to richer ones—a key tenet of the convergence hypothesis in economic growth theory.

Why Partialling-out Works: Neyman Orthogonality

Let $\alpha(\eta)$ be the target parameter defined implicitly by the moment condition

$$M(\alpha, \eta) := E\Big[(\tilde{Y}(\eta) - \alpha\, \tilde{D}(\eta))\, \tilde{D}(\eta)\Big] = 0,$$

where

$$\tilde{Y}(\eta) = Y - \eta_1' W, \quad \tilde{D}(\eta) = D - \eta_2' W.$$

Here, $\eta = (\eta_1, \eta_2)$ are nuisance parameters (with true value η^o).

Neyman Orthogonality means that the estimator is locally insensitive to errors in η, i.e.,

$$\partial_\eta \alpha(\eta^o) = 0.$$

Intuition Behind Neyman Orthogonality

- In high-dimensional settings, we estimate nuisance parameters η (e.g., via Lasso). These estimates are generally slightly biased.
- Neyman orthogonality ensures that small errors in η do not affect the estimate of α to the first order.
- Formally, for a small perturbation δ in η, we have:

$$\alpha(\eta^o + \delta) = \alpha(\eta^o) + \underbrace{\partial_\eta \alpha(\eta^o) \cdot \delta}_{=0} + \text{higher order terms.}$$

Proof

Since $\alpha(\eta)$ is defined implicitly by $M(\alpha, \eta) = 0$, the implicit function theorem yields:

$$\partial_\eta \alpha(\eta^o) = -\Big[\partial_a M(\alpha, \eta^o)\Big]^{-1} \partial_\eta M(\alpha, \eta^o).$$

To show $\partial_\eta \alpha(\eta^o) = 0$, it suffices to prove that:

$$\partial_\eta M(\alpha, \eta^o) = 0.$$

Recall that:

$$M(\alpha, \eta) = E\Big[(\tilde{Y}(\eta) - \alpha\, \tilde{D}(\eta))\, \tilde{D}(\eta)\Big],$$

with

$$\tilde{Y}(\eta) = Y - \eta_1' W \quad \text{and} \quad \tilde{D}(\eta) = D - \eta_2' W.$$

Proof

Derivative with respect to η_1:

$$\partial_{\eta_1} \tilde{Y}(\eta) = -W \quad \Rightarrow \quad \partial_{\eta_1} M(\alpha, \eta^o) = E\Big[-W\, \tilde{D}(\eta^o)\Big].$$

Since $\tilde{D}(\eta^o) = D - \gamma'_{D,W} W$ is the residual from regressing D on W, by construction,

$$E\Big[W\,(D - \gamma'_{D,W} W)\Big] = 0.$$

Thus, $\partial_{\eta_1} M(\alpha, \eta^o) = 0$.

Derivative with respect to η_2: Differentiating $\tilde{D}(\eta) = D - \eta'_2 W$ with respect to η_2 gives:

$$\partial_{\eta_2} \tilde{D}(\eta) = -W.$$

Applying the product rule to $M(\alpha, \eta)$, we get:

$$\partial_{\eta_2} M(\alpha, \eta^o) = -E\Big[W\, \tilde{Y}(\eta^o)\Big] + 2\alpha\, E\Big[W\, \tilde{D}(\eta^o)\Big].$$

Proof

Since $\tilde{Y}(\eta^o) = Y - \gamma'_{Y,W} W$ is the residual from regressing Y on W, we have:

$$E\Big[W\,\tilde{Y}(\eta^o)\Big] = 0,$$

and as before, $E\Big[W\,\tilde{D}(\eta^o)\Big] = 0$. Therefore,

$$\partial_{\eta_2} M(\alpha, \eta^o) = 0.$$

Since both derivatives with respect to η_1 and η_2 are zero, we conclude:

$$\partial_{\eta} M(\alpha, \eta^o) = 0.$$

Therefore, by the implicit function theorem:

$$\partial_{\eta}\alpha(\eta^o) = -\Big[\partial_{a} M(\alpha, \eta^o)\Big]^{-1} \cdot 0 = 0.$$

This completes the proof of Neyman orthogonality: small perturbations in the nuisance parameters η have no first-order impact on the target parameter α.

What happens when there is no Neyman's Orthogonality: Single Selection

In high-dimensional linear models, a seemingly intuitive approach could be to:

1. Apply Lasso regression of Y_i on D_i and W_i that selects the relevant covariates W_Y (plus the covariate of interest D_i)

$$Y_i = \alpha D_i + \beta' W_i + \epsilon_i$$

2. Refit the model by least squares of Y_i on D_i and the selected covariates W_Y:

$$Y_i = \alpha D_i + \beta_Y' W_{Y,i} + \epsilon_i$$

3. Carry out inference for the target parameter α using standard inference based on this regression.

What happens when there is no Neyman's Orthogonality: Single Selection

This method is acceptable if the goal is solely prediction, but it can lead to misleading conclusions when inferring α. This naive approach relies on the moment condition:

$$M(a, b) = E\Big[(Y - aD - b'W)D\Big] = 0.$$

When $b = \beta$ (the true coefficient vector), the true value $a = \alpha$ satisfies:

$$M(\alpha, \beta) = E\Big[(Y - \alpha D - \beta' W)D\Big] = 0.$$

This is the classical moment condition in OLS, which makes the prediction errors orthogonal to each predictor. Mathematically, Neyman Orthogonality requires:

$$\frac{\partial}{\partial b} M(\alpha, \beta) = E[D\, W] = 0.$$

However, in many practical cases, D is not orthogonal to W, so $E[D\, W] \neq 0$.

What happens when there is no Neyman's Orthogonality: Single Selection

In high dimensions, the Lasso estimator for β converges at a slower-than-parametric rate, roughly

$$\sqrt{\frac{s \log(max(p, n))}{n}},$$

where s is the sparsity level and p is the number of controls.

Without Neyman orthogonality, any error in estimating the nuisance parameter β transmits directly to the estimation of α. The moment condition is sensitive to these errors. Consequently, the estimator for α converges at a rate slower than $\sqrt{n}$. [2]

[2] In "pure" RCTs where treatment is assigned independently of everything, $D's$ are orthogonal to $W's$, after de-meaning D, so Neyman orthogonality automatically holds in this setting.

What happens when there is no Neyman's orthogonality: Single Selection

The reason that the naive estimator does not perform well is that it only selects controls that are strong predictors of the outcome, thereby omitting weak predictors of the outcome. However, weak predictors of the outcome could still be strong predictors of D, in which case dropping these controls results in a strong omitted variable bias.

In contrast, the orthogonal approach solves two prediction problems – one to predict Y and another to predict D – and finds controls that are relevant for either. The resulting residuals are therefore approximately "de-confounded".

Inference on Many Coefficients: An Intuitive Explanation

We extend our model to allow for multiple target coefficients:

$$Y = \sum_{\ell=1}^{p_1} \alpha_\ell D_\ell + \sum_{j=1}^{p_2} \beta_j \bar{W}_j + \epsilon,$$

where: D_ℓ ($\ell = 1, \ldots, p_1$) are the predictors (or policies) of interest & $\bar{W}_j$ ($j = 1, \ldots, p_2$) are additional controls. Why do we need it?

– **Multiple Policies:** We might be interested in the effects of several different policies simultaneously.
– **Heterogeneous Effects:** The effect of a treatment may vary across subgroups. For example, a job training program might work differently for young vs. older workers.
– **Nonlinear Effects:** The impact of a continuous policy variable (e.g., price) might be nonlinear. For example, the effect of raising a price from \$1 to \$2 might differ from that of raising it from \$2 to \$3.

Example

Example: Suppose we study how a subsidy (our base treatment D_0) affects firm performance. We suspect the effect might differ by industry (heterogeneous effects) and might not be linear (nonlinear effects). We can generate multiple predictors as:

$$D_1 = D_0, \quad D_2 = D_0 \times \text{Industry Indicator}, \quad D_3 = D_0^2, \quad \text{etc.}$$

Double Lasso is applied *one-by-one* to each D_ℓ to obtain valid inference on each α_ℓ, even when both p_1 and p_2 are large relative to the sample size.

Other Approaches That Have the Neyman Orthogonality Property: Double Selection

Double Selection is a procedure designed to accurately estimate the causal effect of a treatment variable D on an outcome Y when there are many confounding variables W. In this approach, Lasso is applied in two separate steps. First, Lasso is used to select those control variables that are strong predictors of the outcome Y. Second, Lasso is applied to select those controls that are strong predictors of the treatment D. The union of these two sets of controls is then formed, and an ordinary least squares (OLS) regression of Y on D and the combined controls is run:

- **Step 1:** Use Lasso to select controls that best predict the outcome Y; denote this set as W_Y.
- **Step 2:** Use Lasso to select controls that best predict the treatment D; denote this set as W_D.
- **Step 3:** Regress Y on D and the union of controls $W_Y \cup W_D$, and proceed with standard inference.

Comparison with Double Lasso

This process ensures that any variable that is important for predicting either Y or D is included, thereby reducing the risk of omitted variable bias.

By effectively "partialling out" the influence of W from both the outcome and the treatment, the method produces a robust estimate of the causal effect that is less sensitive to errors in variable selection.

The Double Lasso approach directly constructs residuals by removing the influence of W from both Y and D using regularized estimators and then estimates the treatment effect via a moment condition that enforces orthogonality. In contrast, Double Selection explicitly selects a union of controls through separate Lasso regressions for Y and D before running a final OLS regression. Double Selection tends to be more conservative, ensuring that no important confounders are omitted.

Other Approaches That Have the Neyman Orthogonality Property: Debiased/Desparsified Lasso

Desparsified Lasso corrects the bias introduced by the standard Lasso's shrinkage. While Lasso is great for handling many predictors by selecting a subset and shrinking coefficients toward zero, this shrinkage systematically biases the estimates downward. Desparsified Lasso "undoes" this bias, yielding estimates that are closer to the true values and asymptotically normal, which enables constructing confidence intervals and hypothesis tests.
This debiasing step adjusts for the shrinkage bias in the Lasso estimates, providing an estimator of α that is asymptotically normal and suitable for inference.

Debiased Lasso: procedure

1. Run a Lasso regression of Y on D and W (with an appropriate penalty λ) to obtain $\hat{\beta}$.
2. Run a Lasso regression of D on W (with a suitable λ) to obtain $\hat{\gamma}$.
3. Construct the residualized treatment:

$$\tilde{D}(\hat{\gamma}) = D - \hat{\gamma}' W.$$

4. Obtain the estimator $\hat{\alpha}$ as the solution to the moment condition:

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} D_i - \hat{\beta}' W_i \right) \tilde{D}_i(\hat{\gamma}) = 0.$$

This yields the explicit formula:

$$\hat{\alpha} = \left(\frac{1}{n} \sum_{i=1}^{n} D_i \, \tilde{D}_i(\hat{\gamma}) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}' W_i \right) \tilde{D}_i(\hat{\gamma}) \right).$$