Improving Two-Stage Least Squares via Calibrated Machine Learning First Stage

Gaurav Sood

Introduction

Two-stage least squares (2SLS) is a standard estimator for identifying causal effects when a treatment D is endogenous but an instrument Z and covariates X satisfy exclusion and relevance. The classical procedure fits

$$D_i = \gamma_0 + \pi Z_i + \delta^\top X_i + u_i$$

in the *first stage* and then estimates

$$Y_i = \alpha + \beta \widehat{D}_i + \theta^\top X_i + v_i$$

in the *second stage*, where $\widehat{D}_i$ are fitted values from the first stage. Modern machine learning (ML) can flexibly approximate $E[D \mid Z, X]$, but raw ML predictions often suffer from miscalibration or violate known monotonicity (stronger Z implies no lower D probability). We show that applying *isotonic calibration* to ML-first-stage predictions yields *calibrated* fitted values $\widetilde{D}_i$ that improve first-stage fit, strengthen the instrument, and reduce bias and variance in the 2SLS estimator.

Revised 2SLS Procedure

1. First Stage with Calibration

1. Fit an ML model to predict D using (Z, X):

$$\hat{D}_i^{\mathrm{ML}} = f(Z_i, X_i).$$

2. Apply isotonic regression to enforce monotonicity and recalibrate probabilities:

$$\widetilde{D}_i = \arg\min_{g \in \mathcal{M}} \sum_i \big(D_i - g(\hat{D}_i^{\mathrm{ML}})\big)^2,$$

   where $\mathcal{M}$ is the set of nondecreasing functions. The result $\widetilde{D}_i$ are the calibrated first-stage fitted values.

2. Second Stage

Regress the outcome on the calibrated first-stage predictions and covariates:

$$Y_i = \alpha + \beta\, \widetilde{D}_i + \theta^\top X_i + v_i.$$

In fact, this is the standard 2SLS approach using Z as the instrument: the second-stage regression automatically employs the fitted values from regressing D on (Z, X).

Key Benefits and Theoretical Insights

- **Reduced First-Stage MSE:** Isotonic regression projects raw ML predictions onto the monotonic cone, ensuring
$$\sum_i (D_i - \widetilde{D}_i)^2 \leq \sum_i (D_i - \hat{D}_i^{\text{ML}})^2.$$
- **Stronger Instrument:** A better first-stage fit raises both the partial R^2 of $\tilde{D}$ on Z (controlling for X) and the corresponding F-statistic for testing Z's relevance, mitigating weak-instrument bias.
- **Lower 2SLS Variance:** Asymptotically,
$$(\hat{\beta}_{2SLS}) \approx \frac{\sigma_v^2}{n\, R^2_{D\sim Z|X}},$$
so increasing the first-stage partial R^2 via calibration reduces the estimator's variance.
- **No Additional Endogeneity:** Calibration is a monotonic transformation of predictions based only on (Z, X), preserving the exclusion restriction and not introducing bias.

Simulation Evidence

We simulate data with $n = 1000$ and 200 replications:

$$Z_i \sim \text{Bernoulli}(0.5), \quad X_i \sim \mathcal{U}(-1, 1),$$
$$D_i \sim \text{Bernoulli}(\text{logit}^{-1}(0.5Z_i + 0.5X_i)), \quad Y_i = 2D_i + X_i + \varepsilon_i,\ \varepsilon_i \sim N(0, 1).$$

We compare:

- *Raw ML 2SLS*: second stage uses $\hat{D}_i^{\text{ML}}$.
- *Calibrated 2SLS*: second stage uses $\widetilde{D}_i$ after isotonic calibration.

Table 1: Simulation Results: Raw vs Calibrated ML First Stages

Method	First-Stage F	Partial R^2	MSE($\hat{\beta}$)
Raw ML 2SLS	8.2	0.043	0.39
Calibrated 2SLS	12.9	0.063	0.05

Calibration substantially boosts first-stage strength and sharply reduces estimation error in β.

Regarding Propensity Score ATE and Calibration

Calibration also improves Average Treatment Effect (ATE) estimation via propensity-score methods, though the mechanism differs from 2SLS.

Why Calibration Helps PSA

- *Weighting Efficiency:* In inverse-probability weighting (IPW), weights

- *Matching Quality:* Nearest-neighbor matching on calibrated scores $\tilde{e}(X_i)$ yields better covariate balance, since matched units have more accurate treatment-probability alignment, reducing bias from imperfect matches.
- *Overlap Assessment:* Well-calibrated scores correctly identify regions of common support, ensuring that treated and control groups overlap in propensity score space.

Theoretical Argument Under unconfoundedness, the IPW representation of the ATE is

$$\tau = E[Y(1) - Y(0)] = E\Big[\frac{Y\,D}{e(X)} - \frac{Y\,(1-D)}{1-e(X)}\Big].$$

When $\hat{e}(X)$ is miscalibrated, the sample analog uses $\hat{e}(X_i)$, introducing both bias (from systematic deviations) and variance (from extreme weights). Applying isotonic calibration:

$$\tilde{e} = rg\min_{g\in\mathcal{M}} \sum_i (D_i - g(\hat{e}(X_i)))^2,$$

ensures $\tilde{e}(X)$ is a monotonic, well-calibrated estimate of the true $e(X)$, which:

- Reduces bias from misaligned weights.
- Lowers variance by avoiding extreme weights.
- Improves the performance of matching and weighting estimators in finite samples.

Conclusion

Calibrating ML-based first-stage predictions via isotonic regression improves the quality of variation extracted from instruments, yielding stronger first-stage statistics and more precise 2SLS estimates. This simple adjustment reconciles flexible ML modeling with econometric identification assumptions, enabling more reliable causal inference.