From One LATE to Another: Machine Learning and the External Validity of IV Estimates

Alexander Kwon, Ph.D. candidate[1]; Kyungtae Lee, Ph.D. candidate[1]
[1]City University of New York, Graduate Center

Abstract

This paper examines **how machine learning methods can improve the external validity of IV estimates**. Using an empirical application on the effect of solid fuel use on cooking time across six developing countries and a series of simulation experiments, we compare the benchmark interacted two-stage least squares estimator with fixed effects (**2SLS-IF**) to a Double/Debiased Machine Learning (**DML**) approach. **The DML estimator delivers more accurate out-of-sample predictions of LATEs** when treatment effect heterogeneity and selection are driven by observable characteristics, outperforming 2SLS-IF under model misspecification. We also propose an algorithmic procedure for hyperparameter tuning (**MLtune**) that enhances the stability and generalization of DML predictions.

Conceptual Framework

Motivation
Empirical evidence from **instrumental variable (IV)** studies often guides policy decisions **beyond the original study setting**.
However, IV estimates identify **Local Average Treatment Effects (LATEs)** that apply only to **specific groups of compliers**.
When the composition of compliers differs across populations, these **LATEs may not generalize**, raising concerns about **external validity**.

Sample population → Target population

Estimate Sample LATE: $\hat{\tau}(X)$ → Predict LATE: $\hat{\tau}(X_{target})$

- Predicting target LATE using sample estimates
- Needed assumption: external unconfoundness among compliers (*Kwon and Lee, 2025*)

$$Sample\ or\ target \perp \tau | X_i, compliers$$

Methodology

2SLS-IF (benchmark):

$$y_{is} = (\alpha_0 + \alpha' \tilde{X}_i) T_i + \sum_s (\beta_s' \tilde{X}_i + \pi_s) + \varepsilon_{is}$$

estimated by IV using Z_i and $Z_i \tilde{X}$ as instruments, where i represents each observation and s represents each site in sample, $\tilde{X} = X_{is} - \bar{X}_{sample}$.
Then the predicted target LATE equals

$$\hat{\tau}_{pred}^{2SLS-IF} = \hat{\alpha}_0 + \hat{\alpha}' (\bar{X}_{target} - \bar{X}_{sample}).$$

DML estimator:
We estimate the partially linear IV model.

$$Y_i = \tau_0(X_i) T_i + f_0(X_i) + \varepsilon_i,$$

using orthogonalized moments and cross-fitting (*Chernozhukov et al., 2018*).
Machine learning (*XGBoost*) predicts nuisance functions $E[Y|X]$, $E[T|X]$, and $E[T|X, Z]$, which are used to compute debiased residuals for $\tau_0(X_i)$.
We then predict the target-site LATE a

$$\hat{\tau}_{pred}^{DML} = E[\hat{\tau}_0(X_{target})].$$

Simulation setup

Generate $X \sim N(0, \sigma_X^2)$, a binary instrument Z, and treatment T. Treatment effects $\tau(X)$ depend linearly on X_i.
We vary:

- Covariate dispersion (σ_X^2 =1,3,10)
- Instrument specification ($P(Z \mid X)$: linear vs. cubic)
- Outcome functional form (linear vs. step)
- Selection strength (θ: 0.25, 0.75) $\Longrightarrow selection\ probability = logit^{-1}(\theta X_i)$

Simulation Details

- Instrument specification: linear → $P(Z \mid X) = logit^{-1}(0.5 + 0.3X_i)$; cubic → $P(Z \mid X) = logit^{-1}(0.5 + 0.3X_i + 0.06X_i^3)$
- Outcome functional form: linear → $5X_i$; step → $5\ 1\{X_i > 0\}$

Results

Case Study
- **ML and MLtune produce more accurate LATE predictions** in Ethiopia, Honduras, Kenya, and Cambodia.
- **Nepal:** all methods fail → indicates **selection on unobservables** and **breakdown of external validity**.

Simulation Findings
- When heterogeneity & selection operate through **observables**, **DML clearly outperforms 2SLS-IF**, especially under misspecified models.
- **When covariate distributions differ sharply**, 2SLS-IF suffers **large extrapolation bias**.

Overall
•ML improves external validity when key drivers are observable.

Figure1: Predicted vs. Actual LATEs by Estimation Method

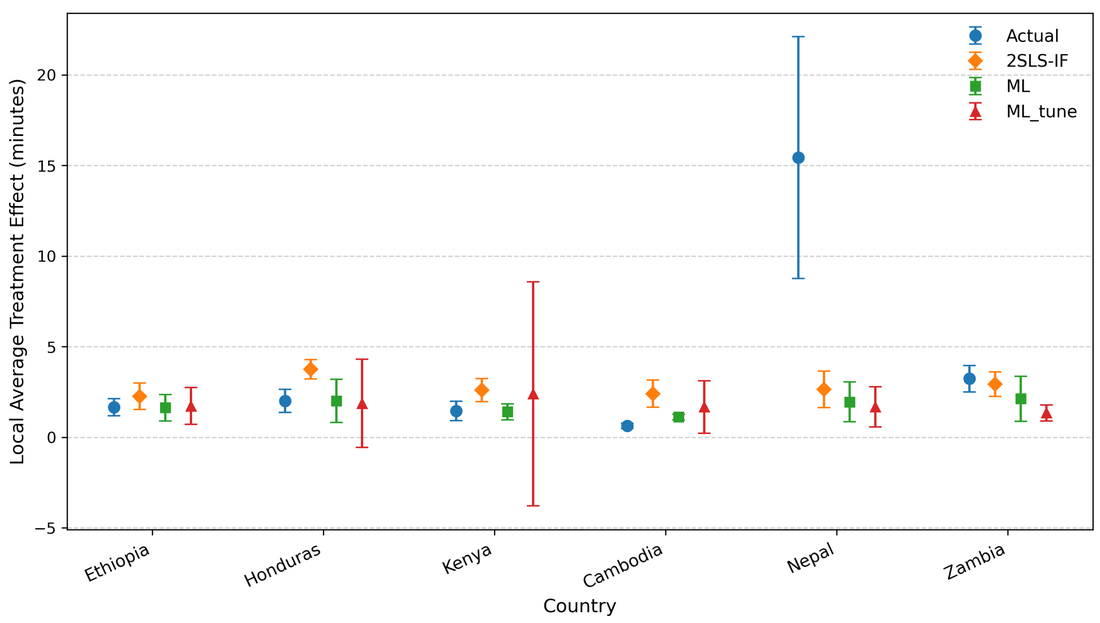
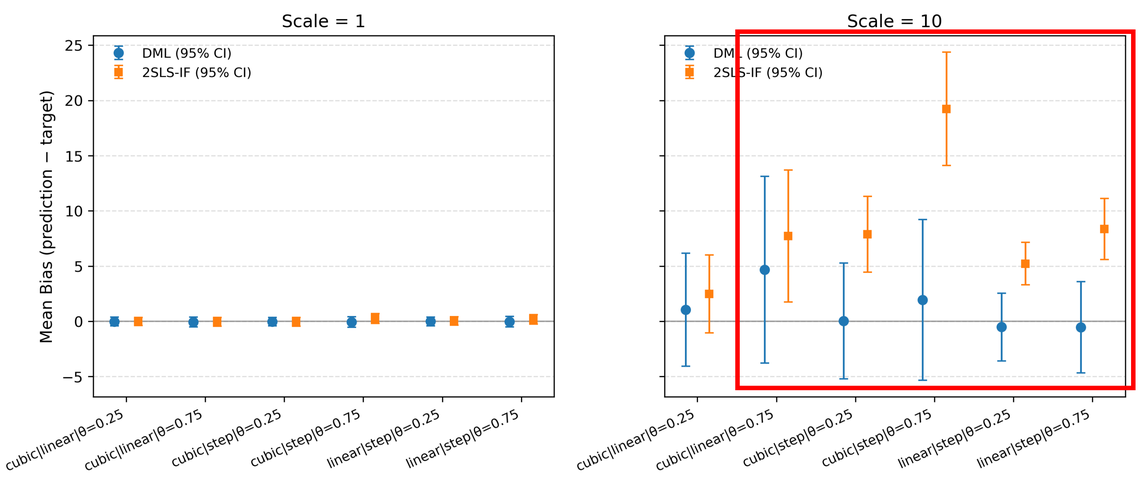


Figure2: Simulation: Mean Bias with 95% Confidence Intervals (Scale = 1 vs 10)



Discussion

With **selection on unobservables**, **both DML and 2SLS-IF fail.**

We plan to extend our study:
1. **Case where observables are correlated with unobservables.**
2. **Develop a machine learning algorithm that can improve external validity with unobservables when it is correlated with observables.**

Contact

Kyungtae Lee
CUNY Graduate Center
Email: klee5@gradcenter.cuny.edu
Website: sites.google.com/view/taelee
Phone: 608-770-1632

References

1. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj. 12097
2. Kwon, A., & Lee, K. (2025). External Validity in an Instrumental Variable Setting. Evaluation Review, 49(6), 1000-1020. https://doi.org/10.1177/0193841X251342619 (Original work published 2025)