Weak Instrumental Variables Due to Nonlinearities in Panel Data

A Super Learner Control Function Estimator

Monika Avila Márquez

University of Geneva

March 2026

The Problem: Weak IVs from Nonlinearities in Panel Data

Panel data model for $i \in [N]$ and $t \in [T_i]$:

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + \alpha_{i,y} + \varepsilon_{it}, \\ x_{1it} &= g(z_{it}) + \alpha_{i,x} + u_{it}. \end{aligned} \tag{1}$$

Default in empirical work: linear first stage

$x_{1it} = z_{it}'\pi + \alpha_{i,x} + u_{it}$.

But the true relationship $g(z_{it})$ may be nonlinear.

Ignoring nonlinearity ⇒ weak instruments

When $g(\cdot)$ is nonlinear, the linear projection of x_{1it} on z_{it} may have low R^2.

Within-2SLS inconsistent, invalid inference.

The solution: Super Learner Control Function Estimator (SLCFE)

A. Augment model with control function

$\varepsilon_{it} = \rho u_{it} + \omega_{it}$

B. Transform model to eliminate individual fixed effects

C. Two-step procedure with cross-fitting

Partition individuals into B folds. For each fold b:

Step 1

Estimate $\mathbb{E}[\tau x_{1it} \mid \mathcal{J}_t]$ with a Super Learner on the complement S_b^c. Obtain residuals $\widehat{\tau u}_{it}$.

Step 2

Estimate by GLS/OLS:

$\tau y_{it} = \tau x_{1it}\beta_1 + \widehat{\tau u}_{it}\rho + \tau\omega_{it}$

Average $\widehat{\beta}_1$ across folds.

Key theoretical results

- **Neyman orthogonal score**
- **$\sqrt{NT}$ consistent and asymptotically normal**
- **Rate condition**
  $\|\widehat{\tau g} - \tau g_0\|_{P,2} = o(N^{-1/4})$
- **Cross-fitting**
  sample splitting across i

Evidence & Takeaway

Monte Carlo ($N = 1000$, $T = 2$)

- SLCFE lowest bias and RMSE across all nonlinearity levels in DGP; Within-2SLS degrades
- Coverage 95–100%

Empirical Application:
Air pollution ($PM_{2.5}$) $\rightarrow$ Educational outcome (test score) - 787 USA counties, 2009 - 2013

- **FD-SLCFE:** $\hat{\beta} = -0.979$ (s.e. 0.101) significant at 5%
- Within-2SLS: $\hat{\beta} = -0.844$ (s.e. 0.762)
- WOLS: $\hat{\beta} = -0.072$ (s.e. 0.011)

ML flexibility in the first stage · $\sqrt{NT}$-valid inference · R package available on request.