UNIVERSITÉ DE GENÈVE

Weak instrumental variables due to nonlinearities in panel data: A Super Learner Control Function estimator

Monika Avila Marquez

Methods and Data Analysis, University of Geneva, Switzerland

Motivation

- A model with **endogeneity** and fixed effects for $i \in [N], t \in [T_i]$:

$$y_{it} = x_{1it}\beta_{1_o} + \alpha_{y,i} + \varepsilon_{it}, \tag{1}$$

$$x_{1it} = g_o(z_{it}) + \alpha_{x,i} + u_{it}. \tag{2}$$

Parameter of interest: β_{1_o}

Source of Identification: Exogenous variation z_{it}.

But...

- Standard: W2SLS **ignoring the nonlinearity**
- Ignoring nonlinearity may cause a problem of **weak instrumental variables**

Setup

A.1 $O_i = (y_i, x_{1i}, \tilde{X}_i, z_i)$ are $T_i \times 1$ independent. $T_i < T_{max}$, $N_T = \sum_i T_i$.

A.2 Linear structural equation:

$$\begin{gathered} y_{it} = x_{1it}\beta_{1_o} + \tilde{x}'_{it}\beta_{2_o} + \alpha_{i,y} + \varepsilon_{it}, \\ \mathbb{E}[\varepsilon_{it}|\tilde{X}_i, z_i, \alpha_{i,y}] = \mathbb{E}[\varepsilon_{it}] = 0. \end{gathered} \tag{3}$$

A.3 Nonlinear reduced form equation:

$$\begin{gathered} x_{1it} = g_o(\tilde{x}_{it}, z_{it}) + \alpha_{i,1x} + u_{it}, \\ \mathbb{E}[u_{it}|\tilde{X}_i, z_i, \alpha_{i,1x}] = 0. \end{gathered} \tag{4}$$

A.4 Linear relationship between the structural error and the reduced form error:

$$\begin{gathered} \varepsilon_{it} = \rho u_{it} + \omega_{it}, \\ \mathbb{E}[\omega_{it}|u_i] = 0. \end{gathered} \tag{5}$$

Identification: Control Function approach

- Identification of β_{1_o}

1. Transform $(\tau : \mathbb{R} \to \mathbb{R})$ model:

$$\tau y_{it} = \tau x_{1it}\beta_{1_o} + \sum_{k=2}^{K} \tau x_{kit}\beta_{2k_o} + \tau\varepsilon_{it}, \tag{6}$$

$$\tau x_{1it} = \tau g_o(\tilde{x}_{it}, z_{it}) + \tau u_{it}, i \in [N], t \in \{t_a, ..., T_i\}. \tag{7}$$

2. Augment transformed structural equation with τu_{it} :

$$\tau y_{it} = \tau x_{1it}\beta_{1_o} + \sum_{k=2}^{K} \tau x_{kit}\beta_{2k_o} + \rho_o \tau u_{it} + \tau\omega_{it}. \tag{8}$$

Population moment conditions:

$$\mathbb{E}[\phi(O_i; \theta_o, \tau g_o)] = \mathbb{E}[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i} \omega_i] = \mathbf{0}. \tag{9}$$

- τu_{it} is identified (Proposition 1)

Super Learner Control Function Estimation

Step 1:

1. Partition the set $\{1, 2, ..., N\}$ in B subsets S_1, S_2, ..., S_B with $n_{T,b} = \sum_{i \in S_b} T_i$.
2. Estimate $\tau g_o(\tilde{x}_{it}, z_{it}) = \mathbb{E}[\tau x_{1it}|I_t]$ using a Super Learner with partition $S_b^c = \{i \in S_j, j \neq b\}$, call the estimation $\widehat{\tau g_o}^{S_b^c}$.
3. Obtain the residuals $\widehat{\tau u}_{it}^{S_b} = \tau x_{1it} - \tau \widehat{g_o}^{S_b^c}(\tilde{x}_{it}, z_{it})$ for partition $S_b = \{i \in S_b\}$.
4. The estimator of θ_o for partition S_b is the solution of the sample moment conditions.

Step 2:

Average $\hat{\theta}_b$ to obtain the final estimator of θ_o as:

$$\hat{\theta}_o = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b.$$

Scan for paper & references

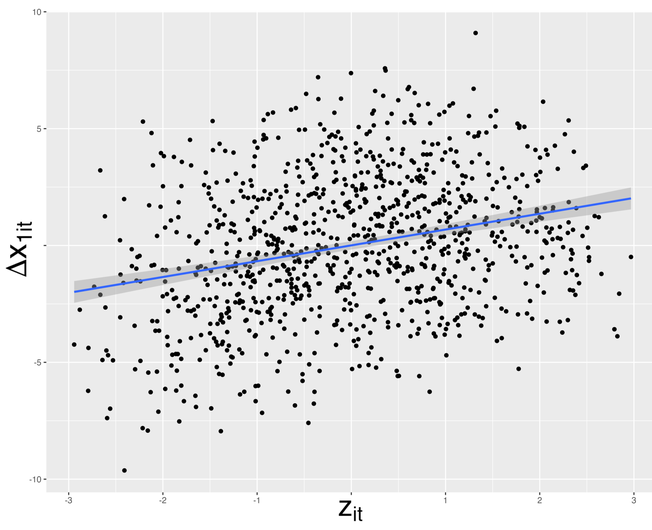
An R package is available on request

monika.avila@unige.ch

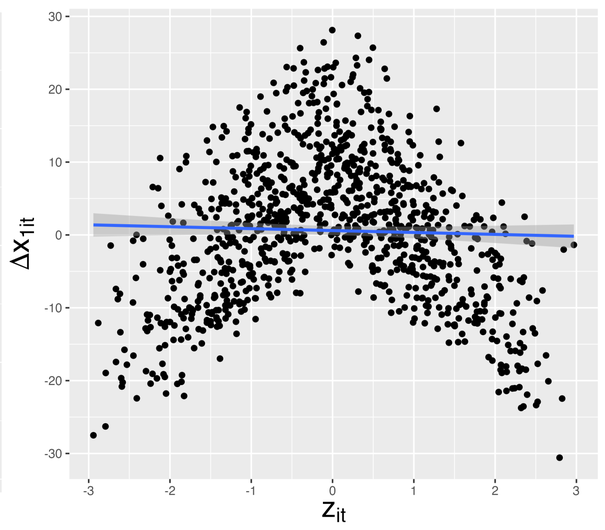
monikaavila.com

Results

- Score $\phi(O_i; \theta_o, \tau g_o)$ is an Orthogonal score (Proposition 2).
- **Large sample properties**
  SLCFE is consistent and asymptotically normal with parametric convergence rate $\sqrt{N_T}$ (Theorems 1 and 2).
- **Monte Carlo simulation**
  The design: Relationship between the EEV and the instrument

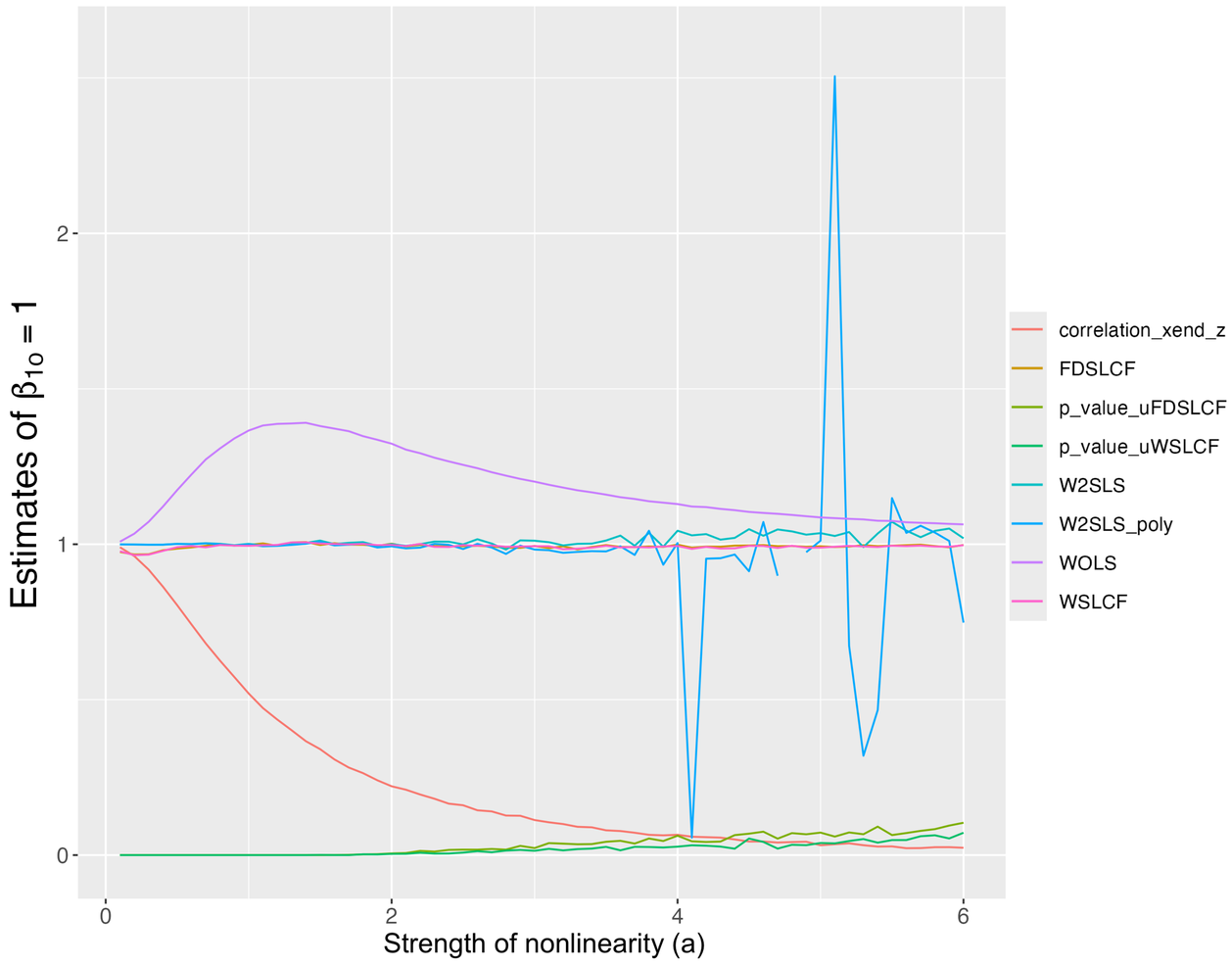


a = 1



a = 6

The results: Average estimated β_{1_o}



Average estimates of β_{1_o} for 100 samples simulated with different values of a, $N = 1000$, $T = 2$

Empirical Application

Causal effect of air pollution on educational outcomes:

Table 1: Estimated effect of air pollution (PM 2.5 concentration) on student performance

WOLS	W2SLS	FDSLCF	WSLCF
-0.072	-0.844	-0.979	-0.859
(0.011)	(0.762)	(0.101)	(0.085)

Note: US counties, N = 787, 2009-2013

Conclusion

- τu_{it} is identified (Proposition 1)
- Score is an orthogonal score (Proposition 2)
- SLCFE is $\sqrt{N_T}$-consistent (Theorems 1 and 2)