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 $i \in [N], t \in [T_i]$ :

$$y_{it} = x_{1it}\beta_{1_o} + \alpha_i + \varepsilon_{it}, \quad (1)$$

$$x_{1it} = g_o(z_{it}) + \gamma_i + u_{it}. \quad (2)$$

Parameter of interest: $\beta_{1_o}$

**Source of Identification:** Exogenous variation $z_{it}$.

**But...**

- W2SLS **ignoring the nonlinearity**?
- Ignoring nonlinearity leads to 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:

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

A.3 Nonlinear reduced form equation:

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

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

$$\varepsilon_{it} = \rho u_{it} + \omega_{it}, \quad \mathbb{E}[\omega_{it}|u_i, \tilde{X}_i, z_i, \alpha_{i,y}] = 0. \quad (5)$$

## Identification: Control Function approach

- Identification of $\beta_{1_o}$
  1. Transform ($\tau : \mathbb{R} \to \mathbb{R}$) model:

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

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

  2. Augment transformed structural equation with $\tau u_{it}$ :

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

Population moment conditions:

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

with $\theta_o = [\beta_{1_o} \quad \beta_{2_o}' \quad \rho_o]'$, $M_{\tau_i}H_i = [M_{\tau_i}x_{i1} \quad M_{\tau_i}\tilde{X}_i \quad M_{\tau_i}u_i]$, $V_i = \mathbb{E}[(M_{\tau_i}\omega_i)(M_{\tau_i}\omega_i)'|H_i, z_i] = \sigma_\omega^2 M_{\tau_i}M_{\tau_i}'$, and $M_{\tau_i}$ transformation matrix.

- $\tau u_{it}$ 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 $\widehat{\tau u}_{it}^{S_b} = \tau x_{1it} - \widehat{\tau g_o^{S_b^c}(\tilde{x}_{it}, z_{it})}$ for partition $S_b = \{i \in S_b\}$.
4. The estimator of $\theta_o$ for partition $S_b$ is the solution of the sample moment conditions.

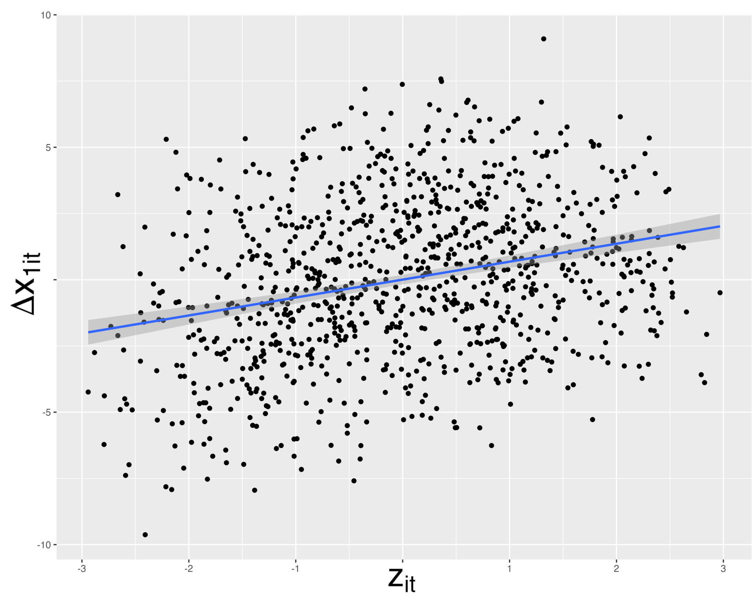
**Step 2**:

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

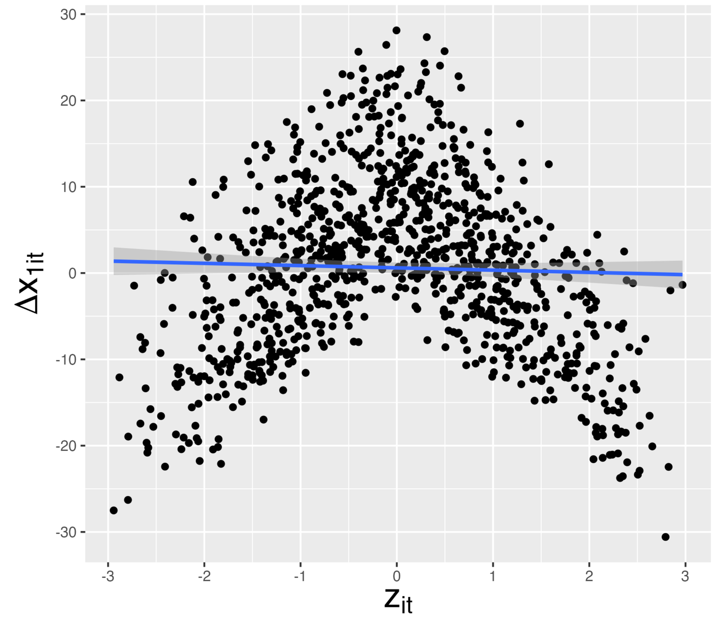
$$\hat{\theta}_o = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b.$$

## 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**

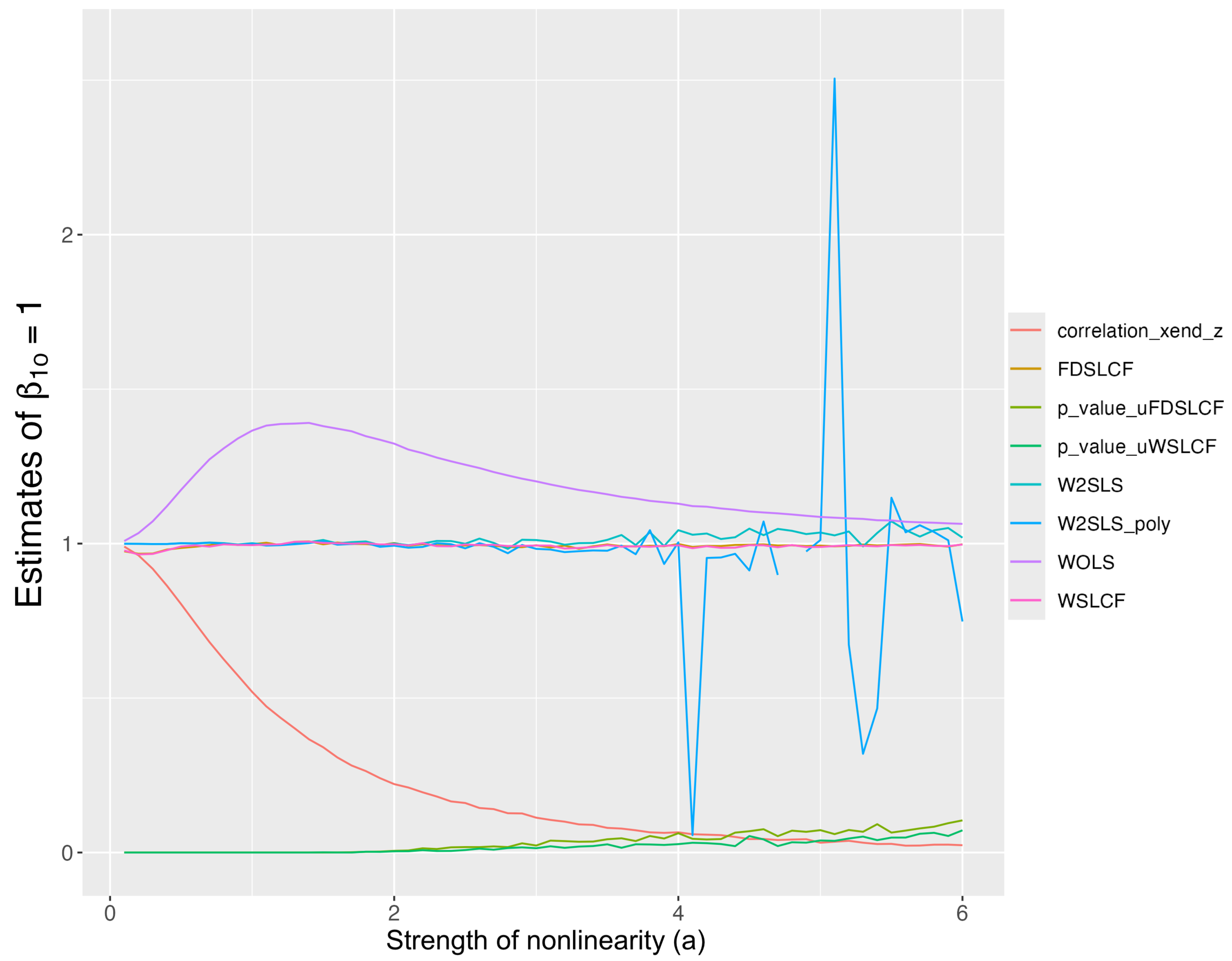



a = 1




a = 6

**The results: Average estimated $\beta_{1_o}$**




Average estimates of $\beta_{1_o}$ 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

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

## Contact me

Scan for paper & references

An R package is available on request

monika.avila@unige.ch

monikaavila.com