

On the use of Machine Learning methods to estimate Triangular Two-level Panel Data Models with Individual Fixed Effects

Authors: Monika Avila Márquez, Jaya Krishnakumar

July 1, 2023

MAD, University of Geneva

GSEM, University of Geneva

Motivation

Consider the linear structural equation for individual i :

$$y_{it} = x_{it1}\beta_{1o} + \tilde{x}_{it}'\beta_{2o} + \alpha_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T, \quad (1)$$

where y_{it} outcome variable, x_{it1} endogenous explanatory variable (EEV), $\tilde{x}_{it}$ vector of exogenous regressors, z_{it} instrumental variables, α_i individual fixed effect.

EEV has a nonlinear relationship with the instrumental variables and the exogenous regressors:

$$x_{it1} = g_o(\tilde{x}_{it}, z_{it}) + \alpha_i + u_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N, \quad (2)$$

with $E[u_{it} | \tilde{x}_{i1}, z_{i1}, \dots, \tilde{x}_{iT}, z_{iT}, \alpha_i] = 0$.

Parameter of interest: β_{1o} .

Source of Identification: Exogenous variation (Instrumental variables).

But..

- Should we perform 2SLS on the transformed model ignoring the nonlinearity in the reduced form equation?
- Would be beneficial to take into account the nonlinearity in the reduced form equation?
- Wooldridge (2015), CF is more efficient than 2SLS when the nonlinear reduced form equation is correctly specified but it is non robust to misspecification of the nonlinear form.

Control function approach using random forest regression to estimate the nonlinear relationship.

Main contributions and results

1. Linear structural equation:

- 1.1 Control function approach.
- 1.2 Two estimation methods (CF, CF-OS) using random forest regression to estimate the nonlinear relationship.
- 1.3 Using a Monte Carlo experiment, we conclude that CF-OS has lower bias and RMSE than 2SLS on the first-differenced model in almost all scenarios.

2. Partial linear structural equation:

- 2.1 Control function approach.
- 2.2 Three methods (CF, CF-OS, CF-IT) using random forest regression to estimate the nonlinear relationship.
- 2.3 Using a Monte Carlo experiment, we conclude that CF-IT with data splitting has lower bias and RMSE than 2SLS on the first-differenced model.

Agenda

1. Literature review
2. The setting: linear structural equation.
 - 2.1 Identification
 - 2.2 Estimation
 - 2.3 Monte Carlo experiment
3. Relaxing the linearity assumption in the structural equation.
 - 3.1 Identification
 - 3.2 Estimation
 - 3.3 Monte Carlo experiment
4. Conclusions

1. Control function approach:

- 1.1 Wooldridge (2015), CF is more efficient than 2SLS when the nonlinear reduced form equation is correctly specified.

2. Semi-parametric linear panel data model with EEV.

- 2.1 RE: Li and Stengos (1996), IV using kernel estimator on the model in levels; FE: Qiang and Wang (2012), IV technique using kernel estimator on the first-differenced model; Baltagi (2002), IV using series estimator for the first-differenced model.

The set up: Linear structural equation

The set up

A.1 Observed data: $w_i = (y_i, x_{i1}, \tilde{x}_i, z_i)$ are $T \times 1$ identical and independent copies of the vector random variables $(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K, \mathbf{z})$. T fixed and small.

A.2a Linear structural equation:

$$y_{it} = x_{it1}\beta_{1o} + \tilde{x}_{it}'\beta_{2o} + \alpha_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N. \quad (3)$$

A.3 α_i individual fixed effect.

A.4 $\tilde{x}_{it}$ is strictly exogenous

$$E[\varepsilon_{it} | \tilde{x}_{i1}, \tilde{x}_{i2}, \dots, \tilde{x}_{iT}, \alpha_i] = 0. \quad (4)$$

Linear structural equation

A.5 x_{it1} is an endogenous explanatory variable (EEV). It is nonlinearly related to $\tilde{x}_{it}$, and an instrumental variable z_{it} .

$$x_{it1} = g(\tilde{x}_{it}, z_{it}) + \alpha_i + u_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N, \quad (5)$$

with $E[u_{it} | \tilde{x}_{i1}, \tilde{x}_{i2}, \dots, \tilde{x}_{iT}, z_{i1}, z_{i2}, \dots, z_{iT}, \alpha_i] = 0$, and

$$E[\varepsilon_{it} | z_{i1}, z_{i2}, \dots, z_{iT}, \alpha_i] = 0.$$

A.6 The structural and reduced-form errors are linearly dependent

$$\varepsilon_{it} = \rho_o u_{it} + \omega_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N, \quad (6)$$

with $E[\omega_{it} | u_{i1}, u_{i2}, \dots, u_{iT}] = 0$.

Identification: Control Function approach

1. Augment the model using $\varepsilon_{it} = \rho u_{it} + \omega_{it}$, and obtain:

$$y_{it} = x_{it1}\beta_{1o} + \tilde{x}_{it}'\beta_{2o} + \alpha_i + \rho_o u_{it} + \omega_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N. \quad (7)$$

2. First-difference the model to eliminate the individual fixed effect:

$$\Delta y_{it} = \Delta x_{it1}\beta_{1o} + \Delta \tilde{x}_{it}'\beta_{2o} + \rho_o \Delta u_{it} + \Delta \omega_{it}. \quad (8)$$

The population moment conditions are:

$$\mathbb{E}[\Delta \omega_{it} \Delta H_{it}] = 0 \quad \forall t, \quad (9)$$

with $\Delta H_{it} = [\Delta x_{it1} \quad \Delta \tilde{x}_{it}' \quad \Delta u_{it}]'$.

Estimation: Method 1

Step 1: Estimate $\Delta g(\tilde{x}_{it}, z_{it})$ by running a random forest regression of Δx_{it1} on $\tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}$, and obtain $\widehat{\Delta u_{it}} = \Delta x_{1it} - \widehat{\Delta g(\tilde{x}_{it}, z_{it})}$.

Step 2: Estimate the parameters of the augmented model using the sample moment conditions:

$$\frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=2}^T \Delta \omega_{it} \widehat{\Delta H_{it}} = 0. \quad (10)$$

Estimation: Method 2

Step 1: Estimate $\Delta g(\tilde{x}_{it}, z_{it})$ by running a random forest regression of $\Delta x_{1,it}$ on $\tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}$, and obtain $\widehat{\Delta u_{it}} = \Delta x_{1,it} - \widehat{\Delta g(\tilde{x}_{it}, z_{it})}$.

Step 2:

2.1 Use the estimated residuals $\widehat{\Delta u_{it}}$ in place of Δu_{it} in the first-differenced augmented model:

$$\Delta y_{it} = \Delta x_{1,it} \beta_{1o} + \Delta \tilde{x}_{it}' \beta_{2o} + \rho_o \widehat{\Delta u_{it}} + \Delta \omega_{it}. \quad (11)$$

2.2 Partial out the exogenous regressors in 11, written in matrix form:

$$M_{\Delta \tilde{x}} \Delta y_{it} = M_{\Delta \tilde{x}} \Delta x_{1,it} \beta_{1o} + \rho_o M_{\Delta \tilde{x}} \widehat{\Delta u_{it}} + M_{\Delta \tilde{x}} \Delta \omega_{it}. \quad (12)$$

Step 2:

2.3 Estimate ρ_o by partialing out $M_{\Delta\tilde{x}}\Delta x_{1,it}$.

2.4 Obtain $\widehat{\Delta y_{it}^*} = M_{\Delta\tilde{x}}\Delta y_{it} - \hat{\rho}_o M_{\Delta\tilde{x}}\widehat{\Delta u_{it}}$.

2.5 Estimate the parameter of interest by regressing: $\widehat{\Delta y_{it}^*}$ on $M_{\Delta\tilde{x}}\Delta x_{1,it}$.

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \alpha_i + 0.2 \varepsilon_{it}^2, \quad (13)$$

where:

$$x_{1it} = g(x_{it2}, z_{it}) + \alpha_i + 0.2 u_{it}^2, \quad (14)$$

$$g(x_{it2}, z_{it}) = x_{it2} + z_{it} + 2 \exp(-16 x_{it2}^2 - 16 z_{it2}^2)$$

, (15)

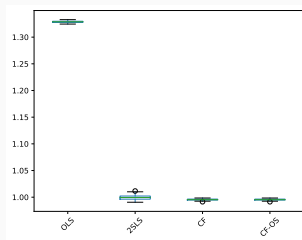
$$x_{2it} = \alpha_i + \zeta_{it}, \quad \text{with} \quad \zeta_{it} \sim U(-2, 2), \quad (16)$$

$$z_{2it} = \alpha_i + \nu_{it}, \quad \text{with} \quad \nu_{it} \sim U(-2, 2), \quad (17)$$

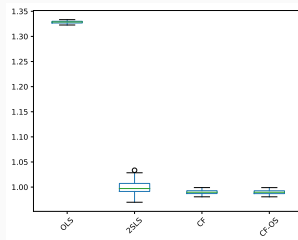
$$\begin{pmatrix} u_{it} \\ \varepsilon_{it} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right), \quad (18)$$

$$\alpha_i \sim N(0, 1). \quad (19)$$

Simulation results



(a) $N=5000$, $T=20$



(b) $N=5000$, $T=3$

Figure 1: Linear structural equation and nonlinear reduced-form equation: DGP 2-Scenario 1. lvs: x_{it2} , z_{it} , x_{it-12} , z_{it-1}

Note: The true parameter value is 1 ($\beta_1 = 1$).

Relaxing the linearity assumption

A.2b The structural equation:

$$y_{it} = x_{it1}\beta_{1o} + f(\tilde{x}_{it}) + \alpha_i + \varepsilon_{it}. \quad (20)$$

Identification

1. Augment the model using $\varepsilon_{it} = \rho u_{it} + \omega_{it}$ to obtain:

$$y_{it} = x_{it1}\beta_{1o} + f(\tilde{x}_{it}) + \alpha_i + \rho_o u_{it} + \omega_{it}. \quad (21)$$

2. First-difference the model to eliminate the individual fixed effect:

$$\Delta y_{it} = \Delta x_{it1}\beta_{1o} + \Delta f(\tilde{x}_{it}) + \rho_o \Delta u_{it} + \Delta \omega_{it}. \quad (22)$$

3. Obtain the conditional expectation of Δy_{it} on $\tilde{x}_{it}, \tilde{x}_{it-1}$:

$$\begin{aligned} E[\Delta y_{it} | \tilde{x}_{it}, \tilde{x}_{it-1}] &= E[\Delta x_{1,it} | \tilde{x}_{it}, \tilde{x}_{it-1}] \beta_{1o} + \Delta f(\tilde{x}_{it}) \\ &+ E[\Delta u_{it} | \tilde{x}_{it}, \tilde{x}_{it-1}] + E[\Delta \omega_{it} | \tilde{x}_{it}, \tilde{x}_{it-1}] \end{aligned} \quad (23)$$

under A.7 $E[\Delta f(\tilde{x}_{it}) | \tilde{x}_{it}, \tilde{x}_{it-1}] = \Delta f(\tilde{x}_{it})$.

5. Obtain the deviation of Δy_{it} from the conditional expectation of Δy_{it} on $\tilde{x}_{it}, \tilde{x}_{it-1}$:

$$\Delta v_{it} = \widetilde{\Delta x_{1,it}} \beta_{1o} + \rho_o \widetilde{\Delta u_{it}} + \Delta \omega_{it}. \quad (24)$$

6. Population moment conditions:

$$E[\Delta \omega_{it} \widetilde{\Delta x_{1,it}}] = 0 \text{ under A.5 and A.6.}$$

$$E[\Delta \omega_{it} \widetilde{\Delta u_{it}}] = 0 \text{ under A.6}$$

Estimation: CF

The estimation procedure has three steps:

Step 1: Following Robinson1988E, estimate the reduced form residuals $\widehat{\Delta u_{it}} = \Delta x_{it1} - \overline{E[\Delta x_{it1} | \tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}]}$ using random forest regression.

Step 2: Using the residuals obtained in previous step, estimate the residuals $\widehat{\Delta v_{it}} = \Delta y_{it} - \overline{E[\Delta y_{it} | \tilde{x}_{it}, \tilde{x}_{it-1}]}$, $\widehat{\Delta x_{1it}} = \Delta x_{it1} - \overline{E[\Delta x_{it1} | \tilde{x}_{it}, \tilde{x}_{it-1}]}$, and $\widehat{\Delta u_{it}} = \widehat{\Delta u_{it}} - \overline{E[\widehat{\Delta u_{it}} | \tilde{x}_{it}, \tilde{x}_{it-1}]}$ learning the conditional expectations using random forests.

Step 3: Use the estimated residuals to estimate the parameter of interest.

We call this estimator CF in the simulation study.

The estimation procedure has five steps:

Step 1 and **Step 2** equal to CF.

Step 3: Use the residuals estimate ρ_o using Frisch-Waugh-Lovell theorem.

Step 4: Obtain $\Delta \tilde{v}_{it} = \widehat{\Delta v_{it}} - \widehat{\Delta u_{it} \rho_o}$

Step 5: Use the new residuals to estimate the parameter of interest.

We call this estimator CF-OS in the simulation study.

The estimation procedure has four steps:

Step 1 as in CF.

Step 2: Set up starting values for $\rho_o^{(0)}$, $\beta_{1_o}^{(0)}$.

Step 3: Obtain $\Delta y_{it} - \Delta x_{it1} \beta_{1_o}^{(0)} - \rho_o^{(0)} \widehat{\Delta u}_{it} = \Delta f_o(\tilde{x}_{it}) + \Delta \omega_{it}$, and estimate $\Delta f(\tilde{x}_{it})$.

Step 4: Obtain $\Delta y_{it} - \widehat{\Delta f_o(\tilde{x}_{it})}$, and estimate β_{1_o} , ρ_o .

We call this estimator CF-IT in the simulation study.

$$y_{it} = \beta_1 x_{it1} + f(x_{it2}) + \alpha_i + 0.8u_{it} + \omega_{it}, \quad (25)$$

$$x_{1it} = g(x_{it2}, z_{it}) + \alpha_i/2 + u_{it}, \quad \text{with } u_{it} \sim N(0, 1), \quad (26)$$

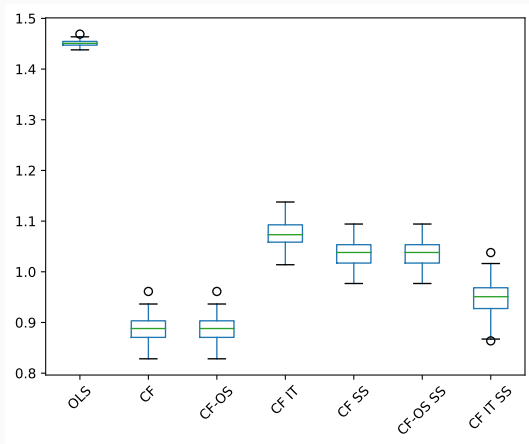
with: $f(x_{it2}) = \sin(2 * x_{it2}) + \exp(-16 * x_{it2}^2)$,
 $g(x_{it2}) = \log(0.1 * x_{it2}^2 + z_{it}^2)$.

$$x_{2it} = \alpha_i/2 + \zeta_{it}, \quad \text{with } \zeta_{it} \sim U(-2, 2), \quad (27)$$

$$z_{it} = \alpha_i/2 + \nu_{it}, \quad \text{with } \nu_{it} \sim U(-2, 2), \quad (28)$$

$$u_{it} \sim U(-2, 2), \omega_{it} \sim U(-1, 1), \alpha_i \sim N(0, 1).$$

Simulation results



(a) $N=5000$, $T=3$

Figure 2: Linear structural equation and nonlinear reduced-form equation: DGP 5-Scenario 1. lvs: x_{it2} , z_{it} , x_{it-12} , z_{it-1}

1. Linear structural equation:

- 1.1 Using a Monte Carlo experiment we conclude that CF-OS has lower bias and RMSE than 2SLS on the first-differenced model in almost all scenarios.

2. Partial linear structural equation:

- 2.1 Using a Monte Carlo experiment we conclude that CF-IT with data splitting has lower bias and RMSE than 2SLS on the first-differenced model.