Weak instrumental variables due to ignored nonlinearities in panel data: An Orthogonal Super Learner Control Function estimator

Monika Avila Márquez * †

† University of Geneva, Methods and Data Analysis, monika.avila@unige.ch

November 15, 2026

Abstract

A triangular structural panel data model with additively separable individual-specific effects is used to model the causal effect of a covariate on an outcome variable when there are unobservable confounders with some of them time-invariant. In this setup, a linear specification for the reduced-form equation might be problematic when the conditional mean of the endogenous covariate and the instrumental variables is nonlinear in the population. The reason is that ignoring the nonlinearity could lead to weak instruments (instruments are weakly correlated with the endogenous covariate) due to misspecification as shown using a generalized concentration parameter for panel data. As a solution, we propose a triangular simultaneous equation model for panel data with additive separable individual-specific fixed effects composed of a linear structural equation with a nonlinear reduced form equation. The parameter of interest is the structural parameter of the endogenous variable. The identification of this parameter is obtained under the assumption of available exclusion restrictions and using a control function approach. We derive a Neyman orthogonal score, and we provide an estimator that we call Orthogonal Super Learner Control Function (Orthogonal SLCF) estimator. The estimation procedure is composed of two main steps and cross-fitting. First, we estimate the control function using a super learner. In the following step, we use the estimated control function to control for endogeneity in the structural equation. Cross-fitting is done across the individual dimension. The estimator is consistent and asymptotically normal achieving a parametric rate of convergence. We show that the Orthogonal SLCF estimator is equivalent to the cross-fitted plug-in IV estimator but not equivalent to the naive plug-in 2SLS estimator which is not $\sqrt{N_T}$-consistent even with cross-fitting.

Keywords— Endogeneity, Super Learner, Instrumental Variables, Identification, Control Function, Panel data, Individual Fixed Effects

*I am indebted to Prof. Rajen Shah for his invaluable feedback, as well as to Prof. Juan-Manuel Rodriguez-Poo, and Prof. Stefan Sperlich. This paper has benefited from my time at the School of Economics, University of Bristol. I thank the members of the Bristol Econometrics Group for their constructive comments. All errors are my own.

1 Introduction

A triangular structural panel data model with additive separable individual-specific effects is used to model the causal effect of a covariate on an outcome variable when there are unobservable confounders, and some of them are time-invariant. In this setup, it is common to assume that the structural equation is linear and to use a linear projection in the reduced form equation. However, a linear specification for the reduced-form equation might be problematic when the conditional mean of the endogenous covariate and the instrumental variables is nonlinear in the population. The reason is that ignoring the nonlinearity could lead to problems due to weak instruments since the strength of the linear relationship might not be high. This paper proposes to exploit the nonlinearities in the relationship between the endogenous covariate and the instrumental variables as a way to avoid problems of weak instruments due to nonlinearities [1].It is important to note that we focus on studying the issue of weak instruments due to misspecification of the reduced-form equation as well as on providing solutions to this problem. Importantly, we do not solve the problem of weak-identification since strong instruments are still required even after allowing for a flexible specification of the reduced-form equation.

An example is the relationship between air pollution and children's educational outcomes, where both variables are influenced by unobserved confounders that can be either time-varying or time-invariant. For instance, neighborhood characteristics and family income are often unobserved in publicly available data. While neighborhood characteristics typically remain constant over time, family income fluctuates. Both factors are relevant to a child's academic achievement and exposure to air pollution, leading to endogeneity in pollution exposure. To address this, fixed effects can be used to control for unobserved neighborhood characteristics, while an instrumental variable can help account for time-varying confounders. A commonly used instrumental variable is wind direction, which has a well-established nonlinear relationship with air pollution levels (Zabrocki et al., 2022).

This paper studies a triangular structural panel data model in which the reduced form equation is nonlinear and individual specific effects are present. It proposes the use of Super Learning [2] for the estimation of the nonlinearities in the model. For simplicity consider the following triangular model without covariates for $i \in \{1,...,N\}$, $t \in \{1,...,T\}$:

$$y_{it} = x_{1it}\beta_{1_o} + \alpha_{i,y} + \varepsilon_{it}, \quad with \quad \mathbb{E}[\varepsilon_{it}|x_{1i}, z_i, \alpha_{i,y}] = \mathbb{E}[\varepsilon_{it}|u_{it}], \tag{1}$$

$$x_{1it} = g_o(z_{it}) + \alpha_{i,1x} + u_{it}, \quad with \quad \mathbb{E}[u_{it}|z_i, \alpha_{i,1x}] = 0, \tag{2}$$

$$\varepsilon_{it} = \rho u_{it} + \omega_{it}, \quad with \quad \mathbb{E}[\omega_{it}|u_i] = 0, \tag{3}$$

where y_{it}, and x_{it} are continuous and bounded, $u_i' = [u_{i1}, u_{i2}, ..., u_{iT}]$, $z_i' = [z_{i1}, z_{i2}, ..., z_{iT}]$ are vectors stacking up all the observations available per individual. $\alpha_{i,y}$, and $\alpha_{i,1x}$ represent individual specific effects. [3]

The model comprises a structural equation 1 with an endogenous regressor x_{1it} that presents a non-linear relationship with the available instrumental variable z_{it} 2. Both equations present individual effects that represent time-invariant unobserved confounders. In this setup, conditioning on the individual effects allows us to control for time-invariant unobserved confounders. But, estimating all the individual effects leads to

[1] Instruments are weak when they are weakly correlated to the endogenous covariate. In a cross-sectional setup, weak IVs cause that the median of the 2SLS estimator is biased towards the OLS estimator, 2SLS t-test presents size inflation (Keane and Neal, 2024). In panel data, Within-2SLS estimator is consistent when the time dimension grows (Cai et al., 2012).

[2] It is feasible to use other machine learning methods instead of Super Learning. However, Super Learning automatically combines different methods in an optimal way achieving a convergence rate that is lower or equal to the best learner included in the library of base learners.

[3] $\alpha_{i,y}$, and $\alpha_{i,1x}$ can be equal as in the cited example but for generality we allow them to be different.

incidental parameter bias as the number of parameters to estimate grows with the sample size. Then, in order to eliminate the individual effects, we propose to transform the model (e.g. first-differencing, or a within transformation which is equivalent to obtain the deviations of the original variables from their individual mean). On the other hand, the presence of time-varying unobserved confounders is modeled through a linear relationship between the error terms in both equations (3). We deal with this issue by exploiting available exclusion restrictions (instrumental variables) along with an Orthogonal Control Function approach on the transformed model.

We derive a Neyman orthogonal score related to the augmented model, and we show that the estimator is equivalent to the IV estimator with instruments equal to transformed non-linear relationship between the instruments and the covariates. To estimate the causal parameter of interest β_{1_o}, we provide a two-step estimator. In the first step, we use a super learner (Van der Laan et al., 2007) to learn the nonlinear relationship of the transformed endogenous variable with the instrumental variables and the observable confounders. In the second step, we use the estimate the parameters of the model solving the derived Neyman orthogonal score. We show that this is equivalent to IV estimation for the β_o coefficients that uses $\widehat{\tau g_o}(\tilde{X}_i, z_i)$ as instrument. We call this estimator Orthogonal Super Learner Control Function (SLCF) estimator. The main results of the paper are that the Orthogonal Super Learner Control Function estimator of β_o is consistent and asymptotically normal with a convergence rate equal to $\sqrt{N_T}$ [4]. An extensive Monte Carlo experiment is performed to test the small sample properties of the proposed estimator. We conclude that the estimator performs well provided that we can accurately learn the nuisance parameter in the first stage. An R package is available from the author.

Our paper contributes to the panel data literature in several ways. First, this paper studies the use of machine learning to estimate panel data models with an endogenous covariate (arising from unobserved confounders or measurement error) and additive individual fixed effects using a control function approach and exploiting nonlinearities. Second, it provides general theoretical results that apply to different panel data transformations [5] while using machine learning methods and control function. Third, it shows that, under suitable conditions, one can identify the transformed nuisance parameter and as a result the transformed control function. Fourth, it provides a Neyman orthogonal score. Finally, it shows that the orthogonal control function estimator is equivalent to the plug-in IV estimation.

In contrast, the available papers using machine learning to estimate panel data models focus on exogenous covariates (time-varying confounders are observed) and the presence of individual fixed effects and focus on specific panel data transformations. For instance, Klosin and Vilgalys (2022) propose a debiased LASSO estimator to estimate heterogeneous treatment effects of a continuous treatment, which is modeled using a structural nonparametric panel data model with additive fixed effects and strictly exogenous covariates. They account for the individual specific effects by taking the first-difference of the original model. Semenova et al. (2023) focus on a structural equation model for panel data allowing for heterogeneous treatment effects with unconfoundedness on observables (sequentially exogenous covariates) and individual specific fixed effects. In order to deal with the individual fixed effects, the authors use a Mundlak approach. They provide an algorithm composed of three steps. In the first step, using cross-fitting they estimate the nuisance functions of the structural and treatment equations to obtain the residuals. In the second step, they estimate the conditional average treatment effect using lasso estimators. In the third step, they provide debiased inference.

This paper also contributes to the literature on weak instruments in panel data by being the first to focus on a setup with *short time dimension* in the presence of nonlinearities. In contrast, existing papers on weak instrumental variables in panel data focus on the asymptotic properties of the Within 2SLS

[4] $N_T = \sum_i^N T_i$.

[5] Panel data transformations are required to eliminate the individual specific effects

estimator providing conditions for consistency under weak instruments when the *time dimension grows to infinity*. Cai et al. (2012) show that the Within 2SLS is consistent when the individual and time dimension grow to infinity. They also present asymptotically pivotal tests (Anderson-Rubin test, Kleibergen test) for longitudinal data.

In Section 3, we present the baseline model. In Section 4, we describe the identification strategy, and we derive a Neyman orthogonal score. In Section 5, we present the SLCF estimator. In Section 6, we present the large sample properties of the SLCF estimator. In Section 7, we present a comparison of the proposed estimator with IV and 2SLS plug-in estimators. In Section 8, we present the simulation experiment and its results, in Section 9 we present the results of the empirical application, and in Section 10 we present the conclusions of the paper.

2 Other Related literature

This paper is related to the literature on semiparametric panel data models. Rodriguez-Poo and Soberon (2017) explain that one can either use a first-differencing approach or a profiling technique to estimate a semiparametric model with an additive separable individual specific effect and a disturbance term. Li (1996) explains that kernel estimation of the conditional expectation of the first-differenced endogenous regressor is not feasible when the dimension of the conditioning set is larger than 5. As a solution, Baltagi et al. (2002) presents an additive estimation of the first-differenced non-linear function.

In addition, this paper is related to the literature on orthogonal/debiased estimation, which focus on providing $\sqrt{N}$ consistent estimators of low-dimensional parameters in the presence of high-dimensional nuisance parameters. Chernozhukov et al. (2018) define an orthogonal score as the one that presents a vanishing Gateaux derivative with respect to the nuisance parameters when evaluated at the true finite-dimensional parameter values. The orthogonal score is closely related to the score proposed by Robinson (1988).

Finally, our work relates to the literature of mixed effects models and machine learning. Our model is related to a mixed effects model assuming that covariates are exogenous conditioning on the cluster random effects. Emmenegger and Bühlmann (2023) use double machine learning to estimate a partially linear mixed-effects model. Our setup differs from Emmenegger and Bühlmann (2023) as it allows for endogeneity, it includes a control function leading to the study of a different score function than the one presented by the authors. Young and Shah (2024) propose a sandwich boosting for accurately estimating partially linear models for grouped data.

2.1 Super Learner

The Super Learner was proposed by Van der Laan et al. (2007), it is a heterogeneous ensemble method within Supervised Learning. It consists of different base learners that are trained on the training data set, then their predictions are combined using a meta-learner by minimizing the cross-validated risk. The meta-learner can be a linear regression of the predictions of each learner on the outcome variable. Van der Laan et al. (2007) show that the Super Learner performs asymptotically as well as or better than any learner in the library of base learners.

3 The Setup

We study a model composed of a linear structural equation, the outcome variable (y_{it}) explained by an endogenous variable (x_{1it}), $K-1$ exogenous regressors ($\tilde{x}_{it}$), and an additive individual-specific effect ($\alpha_{i,y}$).

The model presents an external instrumental variable (z_{it}) for the endogenous regressor. The endogenous regressor is mean dependent on the exogenous regressors, the instrumental variable, and the individual specific effects $\alpha_{i,1x}$. We assume that:

Assumption 1 *We observe a sequence of data sets* $\{O_i = (y_i, x_{1i}, \ldots, x_{Ki}, z_i)\}_{i=1}^{N}$, *which are* $T_i \times 1$ *independent copies of the* $K+2$ *random vectors* $(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K, \mathbf{z})$ *taking values in* $[-M_y, M_y]^T \times [-M_{x1}, M_{x1}]^T \times [-M_{xK}, M_{xK}]^T \times [-M_z, M_z]^T$ *with* $M_y < \infty$, $M_{xk} < \infty$, $M_z < \infty$. *Thus, each element* t *of* $\mathbf{y}$ *is bounded and contained in* $[-M_y, M_y]$. *Similarly, each element of* x_k *is contained in* $[-M_{xk}, M_{xk}]$, *which means that* $\mathbf{x}_{kt}$ *is a continuously distributed random variable with compact support* $[-M_{xk}, M_{xk}]$. *In addition, each element of* $\mathbf{z}$ *is a continuously distributed random variable with compact support* $[-M_z, M_z]$. *Finally,* T_i *is uniformly bounded such that* $T_i < T_{max}$.

Assumption 1 states that each individual i presents T_i observations. The total number of observations is denoted by $N_T = \sum_i T_i$.

Assumption 2 *The structural equation has a linear form with an additive unobserved error term and an additive individual specific effect* $\alpha_{i,y}$

$$\begin{gathered} y_{it} = x_{1it}\beta_{1_o} + \tilde{x}'_{it}\beta_{2_o} + \alpha_{i,y} + \varepsilon_{it}, \quad i \in \{1, ..., N\}, t \in \{1, ..., T_i\}, \\ \mathbb{E}[\varepsilon_{it} | \tilde{x}_i, z_i, \alpha_{i,y}] = \mathbb{E}[\varepsilon_{it} | \tilde{x}_i, \alpha_{i,y}] = 0. \end{gathered} \tag{4}$$

where $\tilde{x}_{it} = [x_{2it}, x_{3it}, ..., x_{Kit}]'$, $z_i = [z_{i1}, \cdots, z_{iT}]'$.
The unobserved random term ε_{it} *is independent and identically distributed with a continuous cumulative distribution function. It has a compact support with zero mean conditional on* $\tilde{x}_i = [\tilde{x}_{i1}, \tilde{x}_{i2}, \ldots, \tilde{x}_{iT_i}]'$, $z_i = [z_{i1}, z_{i2}, ..., z_{iT_i}]'$, $\alpha_{i,y}$, *and* $Var(\varepsilon_{it} | \tilde{x}_i, z_i, \alpha_{i,y}) < \infty$. $\alpha_{i,y}$ *are individual specific effects with sigma-field* $\mathscr{A}_y$. *The unobserved parameter of interest is* β_{1_o}.

Under assumption 2, $\tilde{x}_{it}$ are strictly exogenous in the structural equation conditional on $\alpha_{i,y}$. In addition, $\{z_{is}, \quad s = 1, \ldots, T\}$ are uncorrelated with the structural error term ε_{it} conditional on $\alpha_{i,y}$.

Assumption 3 *The regressor* x_{1it} *is endogenous and has an unknown nonlinear reduced form equation with an additive separable individual specific effect* $\alpha_{i,1x}$:

$$\begin{gathered} x_{1it} = g_o(\tilde{x}_{it}, z_{it}) + \alpha_{i,1x} + u_{it}, \quad i \in \{1, ..., N\}, t \in \{1, ..., T_i\}, \\ \mathbb{E}[u_{it} | \tilde{x}_i, z_i, \alpha_{i,1x}] = 0, \end{gathered} \tag{5}$$

with the unknown function $g_o : [-M_x, M_x]^K \times [-M_z, M_z] \to \mathscr{S} \subset \mathbb{R}$, *with* $g_o \in \mathscr{H}(\varsigma, C)$ *where* $\mathscr{H}(\varsigma, C)$ *represents a Hölder-class with smoothness parameter equal to* $\varsigma \geq 2$[6]. $\alpha_{i,1x}$ *are individual specific effects measurable with respect to* $\mathscr{A}_y$, *therefore knowledge of* $\alpha_{i,y}$ *implies knowledge of* $\alpha_{i,1x}$. *The reduced form error term* u_{it} *is independent and identically distributed with a continuous cumulative distribution function. It has a compact support with zero mean conditional on* $\tilde{x}_i = [\tilde{x}_{i1}, \tilde{x}_{i2}, \ldots, \tilde{x}_{iT_i}]'$, $z_i = [z_{i1}, z_{i2}, ..., z_{iT_i}]'$, *and* $Var(u_{it} | \tilde{x}_i, z_i, \alpha_{i,1x}) < \infty$.

[6] If the unknown function is linear, the SLCF estimator is still consistent if instruments are not weak. We argue that if the unknown function g_o is linear, the library of the base learners includes the linear regression, and the instruments are appropriately transformed, then the SLCF estimator is equivalent to the 2SLS estimator. A proof of this statement is out of the scope of this paper.

Under assumption 3, $\tilde{x}_{it}$, z_{it} are strictly exogenous in the reduced form equation conditional on the individual specific effect $\alpha_{i,1x}$. In addition, this assumption states that the endogenous regressor has a nonlinear relationship with the exogenous regressors $\tilde{x}_{it}$, and the variable z_{it}.

As a consequence of Assumptions 2 and 3, z_{it} is an instrumental variable.
The decomposition of x_{1it} into a conditional expectation and an additive disturbance term is appropriate under the assumption that x_{1it} is continuous (Assumption 1). Assumption 3 needs to be modified when x_{1it} is a non-continuous endogenous explanatory variable.

Assumption 4 *The structural error term ε_{it} has a linear relationship with u_{it}:*

$$\begin{aligned} \varepsilon_{it} &= \rho u_{it} + \omega_{it}, \quad i \in \{1,...,N\}, t \in \{1,...,T_i\}, \\ \mathbb{E}[\omega_{it}|\tilde{x}_i, z_i, \alpha_{i,y}, u_i] &= \mathbb{E}[\omega_{it}|u_i] = 0. \end{aligned} \tag{6}$$

The error term ω_{it} is independent and identically distributed with a continuous cumulative distribution. It has a bounded compact support with zero mean conditional on $u_i = [u_{i1}, u_{i2}, \ldots, u_{iT_i}]'$, and $Var(\omega_{it}|u_{it}) = \sigma^2_\omega < \infty$.

As a consequence of Assumption 4, ω_{it} is uncorrelated with u_{it}.

Remark 1 *The assumption of a linear relationship between ε_{it} and u_{it} is a crucial assumption in panel data setups. Relaxing this assumption to allow for a nonlinear relationship with u_{it} is not trivial even when there are only additive individual fixed effects. Since one needs to eliminate fixed effects through transforming the data, one can only identify the transformed reduced form errors under assumptions 2, and 3. However, identification under a nonlinear function of u_{it} requires knowledge of the untransformed reduced form errors which is not feasible (See Section 4). For example, in the case of first-differencing, we only identify the first-differenced reduced-form errors but we would need the current and first-lag of the reduced form errors in the required information set. These reduced-form errors are not separately identifiable under assumptions 2, and 3.*

As consequence of Assumptions 3 and 4, the structural error term ε_{it} is correlated with x_{1it}. The reason is that $\mathbb{E}[x_{1it}\varepsilon_{it}] = \mathbb{E}[(g_o(\tilde{x}_{it}, z_{it}) + \alpha_{i,1x} + u_{it})\varepsilon_{it}]$ by Assumption 3. The last expression is equal to $\mathbb{E}[(g_o(\tilde{x}_{it}, z_{it}) + \alpha_{i,1x} + u_{it})(\rho_o u_{it} + \omega_{it})] = \rho_o \mathbb{E}[u^2_{it}]$ by Assumptions 3 and 4.
The error term ω_{it} has zero mean conditional on x_{1it}, $\tilde{x}_{it}$, $\alpha_{i,y}$, and u_{it} ($\mathbb{E}[\omega_{it}|x_{1it}, \tilde{x}_{it}, u_{it}, \alpha_{i,y}] = 0$) because $\mathbb{E}[\omega_{it}|x_{1i}, \tilde{x}_i, z_i, \alpha_{i,y}] = \mathbb{E}[\omega_{it}|u_i, \tilde{x}_i, z_i, \alpha_{i,y}] = \mathbb{E}[\omega_{it} \mid u_i] = 0$. The first equality is a consequence of the fact that x_{1it} is a one-to-one function with u_{it} conditional on $\tilde{x}_{it}$, z_{it}, $\alpha_{i,1x} \subseteq \alpha_{i,y}$ (Wooldridge, 2010). The second equality holds by Assumption 4. Notice that a sufficient condition for mean independence of the instruemntal variable with the structural and reduced form errors is that it is independent of both errors.

3.1 Relationship with the Potential Outcomes Framework

In this subsection, we link our setup with the potential outcomes framework. Following Holland (1988) and Guo and Small (2016), we define the potential outcome that would be observed if the treatment is continuous $x_{1,it} = x^*_{1,it}$ as $y_{it}(x^*_{1,it})$. We assume an infinite population setup over the cross-section.

Assumption 5 *Invariance to history*

$$y_{it}(x^*_{1,i}) = y_{it}(x^*_{1,it}), \tag{7}$$

with $x_{1,i} = [x_{1,i1}, \ldots, x_{1,iT}]$.

Assumption 6 *Individualistic treatment response (ITR)*

$$y_{it}(x^*_{1,t}) = y_{it}(x^*_{1,it}), \tag{8}$$

with $x_{1,t} = [x_{1,1t}, \ldots, x_{1,Nt}]$.

Assumptions 5 and 6 state that the potential outcome of individual i at period t depends only on the current level of treatment and on own treatment level.

Assumption 7 $y_{it} = y_{it}(x_{1it}) \quad when \quad x_{1it} = x^*_{1it}$

Assumption 8 *The potential outcome* $y_{it}(x^*_{it})$ *has the following linear additive form:*

$$y_{it}(x^*_{it}) = y_{it}(0) + x^*_{1it}\beta_{1_o}. \tag{9}$$

This assumption states that the potential outcome of individual i at period t if treatment $x_{1,it}$ is equal to $x^*_{1,it}$ has a linear relationship with the potential outcome without treatment and the treatment. In addition, we assume homogeneous treatment effects β_{1o}. The latter is a restrictive assumption, and relaxing it to heterogeneous treatment effects is left for future research. Notice that the instrumental variable does not have an effect on the potential outcome.

Assumption 9 *The potential outcomes without any treatment have a linear conditional mean on the pre-treatment variables* $\tilde{x}_{it}$*, and individual fixed effects:*

$$\mathbb{E}[y_{it}(0)|\tilde{x}_{it}, \alpha_{i,y}] = \tilde{x}'_{it}\beta_{2_o} + \alpha_{i,y}. \tag{10}$$

This assumption states that the potential outcome of individual i at period t without treatment presents a linear relationship with the covariates $\tilde{x}_{it}$ and the individual fixed effects. Importantly, the treatment does not have an effect on the covariates. This assumption can be relaxed to allow for a nonparametric function of the pre-treatment covariates. In this case the estimation procedure proposed is still valid after obtaining the deviations from the conditional expectation on the pre-treatment covariates (Robinson 1988, Chernozhukov et al. 2018).

Assumption 10

$$\{x_{1it}(z^*_{it}), y_{it}(x^*_{1it})\} \perp z_i \mid \tilde{x}_i, \alpha_{i,1x} \tag{11}$$

$$x_{1it} = x_{1it}(z_{it}) \quad when \quad z_{it} = z^*_{it} \tag{12}$$

$$\mathbb{E}[x_{1it}(z_{it})|\tilde{x}_{it}, \alpha_{i,1x}] = g_o(\tilde{x}_{it}, z_{it}) + \alpha_{i,1x} \tag{13}$$

This assumption states that the instrument is independent of the outcome potential outcomes, and the continuous treatment potential outcomes conditional on the covariates and the individual specific effects. Condition 11 of assumption 10 is stronger than assumption 3 as it requires conditional independence of the instrument.

Assumptions 5 to 9 , and condition 11 of assumption 10 imply Assumption 2. Assumption 10 implies assumption 3.

3.2 Estimand

The causal estimand of interest is the average treatment effect:

$$ATE = \mathbb{E}[y_{it}(x^*_{it} + 1)] - \mathbb{E}[y_{it}(x^*_{it})] = \beta_{1_o}. \tag{14}$$

4 Identification: Control Function Approach

The identification of the parameter of interest β_{1_o} relies on the presence of an instrumental variable z_{it} conditioning on the individual-specific effects (Assumptions 2 and 3). Since we run into an incidental parameter problem, we get rid of the individual-specific effects by using an appropriate transformation of the data τ. The typical transformations of the data are the first-differencing and within transformations. The first-difference transformation is equal to $y_{it} - y_{it-1}$, and the Within transformation is equal to $y_{it} - \sum_{t=1}^{T_i} y_{it}/T_i$.

$$\tau y_{it} = \tau x_{1it}\beta_{1_o} + \sum_{k=2}^{K} \tau x_{kit}\beta_{2k_o} + \tau\varepsilon_{it}, i \in \{1,...,N\}, t \in \{t_a,...,T_i\}, \tag{15}$$

$$\tau x_{1it} = \tau g_o(\tilde{x}_{it}, z_{it}) + \tau u_{it}, i \in \{1,...,N\}, t \in \{t_a,...,T_i\}. \tag{16}$$

where t_a is the first time period available after transforming the data. In the case of first-differencing $t_a = 2$, and for a Within-transformation $t_a = 1$.

Now, it is tempting to substitute equation 16 into 15 to obtain:

$$\tau y_{it} = \tau g_o(\tilde{x}_{it}, z_{it})\beta_{1_o} + \sum_{k=2}^{K} \tau x_{kit}\beta_{2k_o} + \tau u_{it}\beta_{1_o} + \tau\varepsilon_{it}, i \in \{1,...,N\}, t \in \{t_a,...,T_i\}. \tag{17}$$

In equation 17, one can believe that the estimation of the unknown parameters is straightforward since the error term $\tau u_{it}\beta_{1_o} + \tau\varepsilon_{it}$ in 17 has zero mean conditional, on an appropriate information set, as a consequence of assumptions 2, and 3. Moreover, one can think that the best estimation procedure is to estimate $\tau g_o(\tilde{x}_{it}, z_{it})$ in the first stage and in the second stage, plug it in equation 17 and perform simple OLS. But this is not possible because the regularization bias of the first stage estimator of $\tau g_o(\tilde{x}_{it}, z_{it})$ contaminates the estimation of β_{1_o} and β_{2_o} (Chernozhukov et al., 2018; Robinson, 1988; Guo et al., 2022) and produces estimates that are not $\sqrt{N_T}$-consistent (See Section 7 for more details).

In order to deal with this problem, we can use a control function approach and augment the structural equation by controlling for the unobservable τu_{it} as follows:

$$\tau y_{it} = \tau x_{1it}\beta_{1_o} + \sum_{k=2}^{K} \tau x_{kit}\beta_{2k_o} + \rho_o \tau u_{it} + \tau\omega_{it}, i \in \{1,...,N\}, t \in \{t_a,...,T_i\}. \tag{18}$$

In model 18, under assumptions 2 - 4 we can guarantee zero correlation between τx_{1it}, and $\tau\omega_{it}$ thanks to the presence of an instrumental variable in the reduced form equation that is mean independent of the error term ω_{it}.

Stacking up the time observations and letting M_{τ_i} denote the transformation operator mapping $\mathbb{R}^{T_i} \to \mathbb{R}^{T_{i,a}}$, with $T_{i,a}$ equal to T_i under the within transformation and $T_i - 1$ under the first-difference transformation, we obtain the following moment conditions:

$$\mathbb{E}[\phi(O_i; \theta_o, \tau g_o)] = \mathbf{0}, \tag{19}$$

where $\phi(O_i; \theta_o, \tau g_o)$ is a vector of score functions equal to $(M_{\tau_i}H_i)'V_i^{-1}M_{\tau_i}\omega_i$, $\theta_o = [\beta_{1_o} \quad \beta_{2_o}' \quad \rho_o]'$, $M_{\tau_i}\omega_i = M_{\tau_i}y_i - M_{\tau_i}x_{1i}\beta_{1_o} - \sum_{k=2}^{K} M_{\tau_i}x_{ki}\beta_{2k_o} - \rho_o M_{\tau_i}u_i$, $\mathbf{0} = [0 \quad \cdots \quad 0 \quad 0]'$, and $M_{\tau_i}H_i = [M_{\tau_i}x_{1i} \quad M_{\tau_i}\tilde{X}_i \quad M_{\tau_i}u_i]$, $H_i = [x_{1i} \quad \tilde{X}_i \quad u_i]$, and V_i the weighting matrix corresponding to $M_{\tau_i}M_{\tau_i}'$ under the first-difference transformation since $\Sigma_i = \mathbb{E}[(M_{\tau_i}\omega_i)(M_{\tau_i}\omega_i)'|H_i, z_i] = \sigma_\omega^2 M_{\tau_i}M_{\tau_i}' = \sigma_\omega^2 V_i$, and equal to the identity matrix I_i under the within transformation.

Remark 2 *If we use the first-difference transformation,* $M_{\tau_i} = D_i = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$*, with dimensions* $(T_i - 1) \times T_i$*. If we use the within transformation,* $M_{\tau_i} = W_i = I_i - 1/T_i \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 & 1 \end{bmatrix}$*, with dimensions* $T_i \times T_i$*,* I_i *is the identity matrix.*

Remark 3 *The moment condition is valid even under misspecification of* V_i *because we assume strict exogeneity. If we would like to relax the assumption of strict exogeneity to weak exogeneity, then it is better to use* $\mathbb{E}[(M_{\tau_i} H_i)' M_{\tau_i} \omega_i] = 0$ *as moment conditions.*

Remark 4 *If* $M_{\tau_i} = W_i$ *(the within transformation), the moment condition can be simply written as* $\mathbb{E}[(M_{\tau_i} H_i)' M_{\tau_i} \omega_i] = 0$.

Finally, we identify τu_{it} since it has zero mean conditional on an appropriate information set $A_i = \{\tilde{x}_{it-j}, z_{it-j}, L_i \leq j \leq U_i\}$ with $L_i \geq t_i - T_i$ and $U_i \leq t_i - 1$ as a consequence of Assumption 3. This identification strategy was first proposed by Avila Márquez and Krishnakumar (2023), and it was presented at the 28th International Panel Data Conference at the University of Amsterdam in 2023.

Proposition 1 (Identification θ_o**)** *If* $\{O_i = (y_i, X_i, z_i)\}_{i \in [N]}$ *satisfy assumptions 1 to 4,* $\mathbb{E}[\phi(O_i; \theta_o, \tau g_o)]$ *is continuously differentiable at* θ_o*, and* $\mathbb{E}[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i} H_i]$ *has full rank, then* θ_o *is locally identified, i.e.,* $\exists \zeta > 0$*, such that* $\mathbb{E}[\phi(O_i; \theta, \tau g_o)] \neq \boldsymbol{0}$ *for* $\theta \neq \theta_o$*,* $\theta \in \Theta_o$ *with* $\Theta_o = \{\theta : ||\theta - \theta_o||_P < \zeta\}$.

Proposition 1 states that the parameters of interest are locally identified if the vector of moment conditions is differentiable at the true parameter value, and the matrix $\mathbb{E}[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i} H_i]$ is full rank.

Proposition 2 provides conditions under which this rank condition holds.

Proposition 2 (Conditions) *If* $\{O_i = (y_i, X_i, z_i)\}_{i \in [N]}$ *satisfy assumptions 1 to 4,* $M_{\tau_i} x_{1i}$*, and* $M_{\tau_i} \tilde{X}_i$ *are not perfectly multicollinear, then*

i) $\mathbb{E}[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i} H_i]$ *has a minimum eigenvalue* $\lambda_{min} > 0$ *if and only if* $M_{\tau_i} g_o(\tilde{x}_i, z_i) \notin span(M_{\tau_i} \tilde{X}_i)$.

ii) $\mathbb{E}[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i} H_i]$ *has a minimum eigenvalue* $\lambda_{min} = 0$ *if and only if* $M_{\tau_i} g_o(\tilde{x}_i, z_i) \in span(M_{\tau_i} \tilde{X}_i)$.

The proof is straightforward and omitted.

This proposition states that the matrix $\mathbb{E}[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i} H_i]$ is ill-conditioned if $\tau g_o(x_{2i}, z_i)$ lies in the column span of $M_{\tau_i} \tilde{X}_i$. This could happen if the variation of the transformed nuisance parameter does not provide any different variation from the transformed covariates. This includes the degenerate situation when the transformed nuisance is equal to the null vector which intuitively means that τu_{it} approaches τx_{1it} causing high multicollinearity. Alternatively, if the functional form of $g_o(.)$ is linear, ill-conditioning happens if instruments are weakly correlated with the endogenous covariates. In those situations, the instruments do not provide additional information permitting the identification of the parameter of interest. Thus, the Gram matrix becomes ill-conditioned as instruments are weak under the conditions cited in the proposition.

Remark 5 *Notice that if z_{it} is not present in $g_o(.)$, we can still achieve identification if $\tau g_o(\tilde{x}_{it})$ does not lie in the column span of $\tau\tilde{x}_{it}$. However, identification would rely on a functional form assumption which is a strong assumption.*

Example *First-difference transformation:*

We could follow Baltagi et al. (2002) by first-differencing the structural and reduced-form equations such that we obtain:

$$\Delta y_{it} = \Delta x_{1it}\beta_{1_o} + \Delta\tilde{x}'_{it}\beta_{2_o} + \Delta\varepsilon_{it}, i \in \{1,...,N\}, t \in \{2,...,T_i\}. \tag{20}$$

$$\Delta x_{1it} = \Delta g_o(\tilde{x}_{it}, z_{it}) + \Delta u_{it}, i \in \{1,...,N\}, t \in \{2,...,T_i\}. \tag{21}$$

Now, we can use the proposed control function approach and augment the structural equation by controlling for the unobservable Δu_{it} as follows:

$$\Delta y_{it} = \Delta x_{1it}\beta_{1_o} + \Delta\tilde{x}'_{it}\beta_{2_o} + \rho_o\Delta u_{it} + \Delta\omega_{it}, i \in \{1,...,N\}, t \in \{2,...,T_i\}. \tag{22}$$

Next, we set up the following moment conditions:

$$\mathbb{E}[(D_iH_i)'\Sigma_i^{-1}D_i\omega_i] = \boldsymbol{0}, \tag{23}$$

where $D_i\omega_i = D_iy_i - D_ix_{1i}\beta_{1_o} - D_i\tilde{x}'_i\beta_{2_o} - \rho_oD_iu_i$, $\boldsymbol{0} = [0 \quad \cdots \quad 0 \quad 0]'$, $D_iH_i = [D_ix_{1i} \quad D_i\tilde{x}'_i \quad D_iu_i]'$, and $\Sigma_i = \sigma^2_\omega D_iD'_i$.

Finally, we identify Δu_{it} since it has zero mean conditional on $\tilde{x}_{it}$, $\tilde{x}_{it-1}$, z_{it}, z_{it-1} as a consequence of Assumption 3.

Proposition 3 (Identification transformed nuisance parameter) *Under assumption 3, $\Delta g(\tilde{x}_{it}, z_{it})$ is identified by $\mathbb{E}[\Delta x_{1it}|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}]$.*

Proof 1

$$\mathbb{E}[\Delta x_{1it}|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] = \mathbb{E}[\Delta g(\tilde{x}_{it}, z_{it})|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] + \mathbb{E}[\Delta u_{it}|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] \tag{24}$$

By the linearity property of the expectation operator, we have that the first term is equal to:

$$\begin{aligned}\mathbb{E}[g(\tilde{x}_{it}, z_{it})|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] - \mathbb{E}[g(\tilde{x}_{it-1}, z_{it-1})|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] = \\ g(\tilde{x}_{it}, z_{it}) - g(\tilde{x}_{it-1}, z_{it-1})\end{aligned} \tag{25}$$

Similarly for the second term, by the Law of Iterated Expectations and Assumption 3:

$$\mathbb{E}[u_{it}|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] - \mathbb{E}[u_{it-1}|\tilde{x}_{it}, z_{it}, \tilde{x}_{it-1}, z_{it-1}] = 0 \tag{26}$$

Remark 6 *A similar analysis and conclusion is obtained using a Within transformation which is equivalent to time demeaning the data. In this case, the information set $A_i = \{\tilde{x}_{it-j}, z_{it-j}, -T_i + t \leq j \leq t-1\}$.*

In Section 5, we propose an estimation method based on the Control Function approach.

4.1 Weak identification vs weak instruments due to ignored nonlinearities

In this subsection, we characterize the problem of weak instruments caused by ignored nonlinearities in the reduced-form equation. Following Mikusheva and Sun (2022) and Baltagi et al. (2012), we propose the following generalized concentration parameter:

$$\pi_o = \mathbb{E}\Big[\tau g_o(\tilde{X}_i, z_i)' \Phi^{+}_{\tau u,i} \tau g_o(\tilde{X}_i, z_i)\Big], \tag{27}$$

with $\Phi_{\tau u,i} = \mathbb{E}[M_{\tau_i} u_i u_i' M_{\tau_i}]$, and $\Phi^{+}_{\tau u,i}$ the Moore-Penrose inverse.

If π_o is bounded away from 0, we do not have an issue of weak identification. Under conditions presented in proposition 2, we guarantee that π_o is bounded away from 0. However, we do not impose a lower bound.

Proposition 4 (Concentration parameter inequality) *If $\{O_i = (y_i, X_i, z_i)\}_{i\in[N]}$ satisfy assumptions 1 to 4, $V(u_i|\tilde{X}_i, z_i) = \sigma_u^2 I_i$, and we use a linear projection in the reduced form equation, i.e. $M_{\tau_i} Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2$ after transformation, then*

$$\mathbb{E}\Big[(M_{\tau_i} Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)' \Phi^{+}_{\tau\tilde{u},i} \, (M_{\tau_i} Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)\Big] \leq \pi_o. \tag{28}$$

where a_1, a_2 represent the parameters of the linear projection, and $\Phi_{\tau\tilde{u},i} = \mathbb{E}[M_{\tau_i}\tilde{u}_i\tilde{u}_i' M_{\tau_i}]$ is the variance-covariance matrix of the transformed misspecified reduced form error term.

Proposition 4 states that misspecification of the relationship between the endogenous covariate and the instruments leads to a concentration parameter lower or equal than the concentration parameter in the population π_o. The equality holds if $\tau g_o(\tilde{x}_{it}, z_{it})$ lies in the span of $[\tau\tilde{X}_{it}, \tau Z_i]$. Thus, opening the door to a situation where instruments could be weak even if the concentration parameter in the population is bounded away from 0.

4.2 An orthogonal score

In this subsection, we derive an orthogonal score function.

Proposition 5 *The score associated to the moment condition 19 is not Neyman orthogonal.*

Proof 2 *The Gateaux derivative of the score is equal to*

$$D_h[\tau\tilde{g}] = \lim_{h\to 0} \mathbb{E}\Big[\frac{(M_{\tau_i} H_{h,i})' V_i^{-1} M_{\tau_i}(y_i - x_i\beta_o - (x_{1i} - (g_o + h\tilde{g})\rho_o)) - (M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i}(y_i - x_i\beta_o - (x_{1i} - g_o)\rho_o)}{h}\Big] \tag{29}$$

with $H_{i,h}[x_{1i}, \tilde{X}_i, u_i - h\tilde{g}]$ since $x_{1i} - (g_o + h\tilde{g}) = u_i - h\tilde{g}$, and we do not specify the arguments of g for simplicity knowing that they are equal to $\tilde{X}_i$ and z_i.

The Gateaux derivative evaluated at $h = 0$ is equal to $\mathbb{E}\Big[(M_{\tau_i} H_i)' V_i^{-1} M_{\tau_i}(\tilde{g}\rho_o))\Big]$. Note that it is not equal to 0 for all the β_o rows except for ρ_o row. More specifically, we can split $D_0[\tau\tilde{g}]$ into $D_{\beta,0}[\tau\tilde{g}]$, and $D_{\rho,0}[\tau\tilde{g}]$. Such that $D_{\beta,0}[\tau\tilde{g}] = \mathbb{E}\Big[(M_{\tau_i}[x_{1i}, \tilde{X}_i])' V_i^{-1} M_{\tau_i}(\tilde{g}\rho_o)\Big] \neq 0$, and $D_{\rho,0}[\tau\tilde{g}] = \mathbb{E}\Big[(M_{\tau_i}[u_i])' V_i^{-1} M_{\tau_i}(\tilde{g}\rho_o) - (M_{\tau_i}[\tilde{g}])' V_i^{-1} M_{\tau_i}(\omega_i)\Big] = 0$.

By the Riesz representation theorem, the Riesz representer is $\rho_o \mathbb{E}[(M_{\tau_i} H_i)|\tilde{X}_i, z_i]$. By Proposition 4 of Newey (1994), the correction term is equal to

$$\kappa = (M_{\tau_i}[g(\tilde{X}_i, z_i), \tilde{X}_i, \mathbf{0}])' V_i^{-1} M_{\tau_i}(x_{1i} - g(\tilde{X}_i, z_i))\rho. \tag{30}$$

Then, adding κ to $\phi(O_i; \theta, g)$ gives the orthogonal score

$$\psi(O_i; \theta, \tau g) = (M_{\tau_i} H_{i,g})' V_i^{-1} M_{\tau_i}(y_i - x_i\beta - \rho(x_{1i} - g(\tilde{X}_i, z_i))) + (M_{\tau_i}[g(\tilde{X}_i, z_i), \tilde{X}_i, \mathbf{0}])' V_i^{-1} M_{\tau_i}(x_{1i} - g(\tilde{X}_i, z_i))\rho, \tag{31}$$

with $H_{i,g} = [x_{1i}, \tilde{X}_i, x_{1i} - g(\tilde{X}_i, z_i)]$.

Proposition 6 (Orthogonal score) *The score 31 is Neyman Orthogonal.*

Then, the identifying moment condition is given by

$$\mathbb{E}[\psi(O_i; \theta_0, \tau g_o)] = 0. \tag{32}$$

5 Super Learner Control Function estimation

As described in the previous section, we can write the transformed structural equation for observation i at period t as follows:

$$\tau y_{it} = \tau x_{1it}\beta_{1_o} + \tau \tilde{x}'_{it}\beta_{2_o} + \rho_o \tau u_{it} + \tau \omega_{it}, i \in \{1, ..., N\}, t \in \{t_a, ..., T_i\}. \tag{33}$$

In this transformed equation, the parameters of interest are identified because τx_{1it}, $\tau \tilde{x}_{it}$ and τu_{it} are uncorrelated with error term $\tau \omega_{it}$ under assumptions 2 - 4.

In addition, it is clear that if τu_{it} is observed, we could estimate the three parameters β_{1_o}, β_{2_o}, ρ_o using the sample counterparts of the population moment conditions 32. But since τu_{it} is not observed, we propose the following estimation procedure:

First Step:

We partition the set $\{1, 2, ..., N\}$ in B subsets S_1, S_2, ..., S_B, and denote $n_{T,b} = \sum_{i \in S_b} T_i$. This sample splitting is equal to the one proposed by Emmenegger and Bühlmann (2023) and it was developed without knowledge of it. Alternatively, if the cluster structure in the data is known (individuals exhibit correlation within clusters but not across them, and both cluster membership and the number of clusters are identified) the data should be partitioned accordingly. For each $b \in \{1, \ldots, B\}$ we estimate $\tau g_o(A_i) = \mathbb{E}[\tau x_{1it}|A_i]$, with A_i the information set corresponding to the transformation τ (See Proposition 3), using a super learner with partition $S_b^c = \{i \in S_g, g \neq b\}$, and we call the estimation $\widehat{\tau g_o}^{S_b^c}$. To simplify notation, we write $\tau g_o(\tilde{x}_{it}, z_{it})$ and $\widehat{\tau g_o}^{S_b^c}(\tilde{x}_{it}, z_{it})$ for $\tau g_o(A_i)$ and $\widehat{\tau g_o}^{S_b^c}(A_i)$, where A_i is the information set of Proposition 3. Then, we 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 $S_b = \{i \in S_b\}$.

Finally, the estimator of $\theta_o = [\beta_{1_o} \quad \beta'_{2_o} \quad \rho_o]'$ for partition S_b is the solution of the sample moment conditions such that:

$$\frac{1}{n_{T,b}} \sum_{i \in S_b} \psi(O_i; \hat{\theta}_b, \widehat{\tau g_o}^{S_b^c}) = 0, \tag{34}$$

where

$$\psi(O_i; \hat{\theta}_b, \widehat{\tau g_o}^{S_b^c}) = (M_{\tau_i} H_{i,\widehat{\tau g_o}}^{S_b})' V_i^{-1}(M_{\tau_i} y_i - M_{\tau_i} X_i \hat{\beta}_b - \hat{\rho}_b(M_{\tau_i} x_{1i} - \widehat{\tau g_o}^{S_b^c}(\tilde{X}_i, z_i))) + ([\widehat{\tau g_o}^{S_b^c}(\tilde{X}_i, z_i), M_{\tau_i}\tilde{X}_i, \mathbf{0}])' V_i^{-1}(M_{\tau_i} x_{1i} - \widehat{\tau g_o}^{S_b^c}(\tilde{X}_i, z_i))\hat{\rho}_b,$$

$\hat{\beta}_b = [\hat{\beta}_{1,b} \quad \hat{\beta}'_{2,b}]'$, $M_{\tau_i}\tilde{H}^{S_b}_{i,\widehat{\tau g_o}} = [M_{\tau_i}x_{1i} \quad M_{\tau_i}\tilde{X}_i \quad \widehat{\tau u}_i^{S_b}]$, $\widehat{\tau u}_i^{S_b}$ is the vector stacking the residuals obtained in the first stage for partition S_b, $\tilde{X}_i$ a $T_i \times (K-1)$ matrix collecting all time observations of $\tilde{x}_{it}$.

This is equivalent to estimating β_o by performing instrumental variables estimation with $[\widehat{\tau g_o}^{S_b^c}, M_{\tau_i}\tilde{X}_i]$ as instruments and weighting matrix V_i^{-1} (See Proposition 7). Importantly, we can set $V_i = I_i$ which is the most efficient if $T = 2$, and ε_{it} is homoskedastic and uncorrelated across both i and t.

Second Step:

We average $\hat{\theta}_b$ to obtain the cross-fitting estimator of θ_o as:

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

Thus, the cross-fitting estimator of our parameter of interest β_{1o} is:

$$\hat{\beta}_{1o} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{1,b}.$$

In order to deal with the dependency of the estimator $\hat{\beta}_{1o}$ on the particular sample split, we follow Chernozhukov et al. (2018) by proposing to repeat the procedure SS times. The estimates are aggregated either by using the mean or the median. In addition, a correction term is added to the variance estimator of $\hat{\theta}_o$ (See Subsection 6.1.1).

Algorithm 1 Super Learner Control Function estimation

Input: Panel data sets $\{y_i, X_i, z_i\}_{i\in\{1,2,\ldots,N\}}$ from model satisfying assumptions 1 to 4, a natural number B.

Output: An estimator of the parameter of interest β_{1o}.

for $ss \in \{1,\ldots,SS\}$ **do**

  Split the individual index set $\{1,2,...,N\}$ into B sets $S_1, S_2, ..., S_B$.

  for $b \in \{1,2,...,B\}$ **do**

    Estimate the conditional expectation $\mathbb{E}[\tau x_{1it}|A_i]$ using a super learner and data corresponding to S_b^c.

    Predict $\mathbb{E}[\tau x_{1it}|A_i]$ for $i \in S_b$.

    Compute the residual $\widehat{\tau u}_{it}$ for $i \in S_b$ using the prediction of $\mathbb{E}[\tau x_{1it}|A_i]$ of previous step.

    Estimate $\theta_o = [\beta_{1o}, \beta'_{2o}, \rho_o]'$ solving the sample moment conditions 34 for $i \in S_b$.

  end for

  Compute $\hat{\theta}_{ss} = \frac{\sum_1^B \widehat{\theta_{b,ss}}}{B}$.

end for

Compute $\hat{\theta}_o = \frac{\sum_1^{SS} \widehat{\theta_{ss}}}{SS}$, or the median.

Example *First-difference transformation (continued):*

As described in the previous section, we can write the first-differenced structural equation for observation i at period t as follows:

$$\Delta y_{it} = \Delta x_{1it}\beta_{1_o} + \Delta\tilde{x}'_{it}\beta_{2_o} + \rho_o\Delta u_{it} + \Delta\omega_{it}, i \in \{1,...,N\}, t \in \{2,...,T_i\}. \tag{35}$$

First, we partition the set $\{1,2,...,N\}$ in B subsets S_1, S_2, ..., S_B, and for each $b \in \{1,\ldots,B\}$ we estimate $\mathbb{E}[\Delta x_{1it}|\tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}]$ using a super learner with partition S_b^c. Then, we get the residuals $\widehat{\Delta u}_{it}^{S_b} =$

$\Delta x_{1it} - \widehat{\Delta g_o}^{S_b^c}(\tilde{x}_{it}, z_{it})$ *for partition* S_b. *We obtain the estimator of* θ_o *for partition* S_b *as the solution of the sample moment conditions such that:*

$$\frac{1}{n_{T,b}} \sum_{i \in S_b} \psi(O_i; \hat{\theta}_b, \widehat{\Delta g_o}^{S_b^c}) = 0, \tag{36}$$

where

$$\psi(O_i; \hat{\theta}_b, \widehat{\Delta g_o}^{S_b^c}) = (D_i H_{i,\widehat{\tau g_o}}^{S_b^c})' V_i^{-1} (D_i y_i - D_i X_i \hat{\beta}_b - \hat{\rho}_b (D_i x_{1i} - \widehat{\Delta g_o}^{S_b^c}(\tilde{X}_i, z_i))) + ([\widehat{\Delta g_o}^{S_b^c}(\tilde{X}_i, z_i), D_i \tilde{X}_i, \boldsymbol{0}])' V_i^{-1} (D_i x_{1i} - \widehat{\Delta g_o}^{S_b^c}(\tilde{X}_i, z_i)) \hat{\rho}_b,$$

$\hat{\beta}_b = [\hat{\beta}_{1,b} \quad \hat{\beta}'_{2,b}]'$, $D_i \tilde{H}_{i,\widehat{\tau g_o}}^{S_b} = [D_i x_{1i} \quad D_i \tilde{X}_i \quad \widehat{\Delta u}_i^{S_b}]$, *and* $\widehat{\Delta u}_i^{S_b}$ *are the residuals obtained in the first stage for partition* S_b.

Finally, since $\hat{\beta}_{1o,b} \in \hat{\beta}_{o,b}$ *we estimate our parameter of interest* β_{1o} *using:*

$$\hat{\beta}_{1o} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_{1,b}.$$

Algorithm 2 Super Learner Control Function Estimation: First-difference transformation

Input: Panel data sets $\{y_i, X_i, z_i\}_{i \in \{1,2,...,N\}}$ from model satisfying assumptions 1 to 4, a natural number B.
Output: An estimator of the parameter of interest β_{1o}.
for $ss \in \{1, 2, ..., SS\}$ **do**
  Split the individual index set $\{1, 2, \ldots, N\}$ into B sets $S_1, S_2, ..., S_B$.
  for $b \in \{1, 2, ..., B\}$ **do**
    Estimate the conditional expectation $\mathbb{E}[\Delta x_{1it} | \tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}]$ using a super learner and data corresponding to S_b^c.
    Predict $\mathbb{E}[\Delta x_{1it} | \tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}]$ for $i \in S_b$.
    Compute the residual $\widehat{\Delta u}_{it}$ for $i \in S_b$ using the prediction of $\mathbb{E}[\Delta x_{1it} | \tilde{x}_{it}, \tilde{x}_{it-1}, z_{it}, z_{it-1}]$ of previous step.
    Estimate $\theta_o = [\beta_{1o}, \beta'_{2o}, \rho_o]'$ solving the sample moment conditions 36 for $i \in S_b$.
  end for
  Compute $\hat{\theta}_{ss} = \frac{\sum_1^B \widehat{\theta_{b,ss}}}{B}$.
end for
Compute $\hat{\theta}_o = \frac{\sum_1^{SS} \widehat{\theta_{ss}}}{SS}$, or the median.

6 Large sample properties of the Orthogonal SLCF estimator

Theorem 1 *Consistency of the Orthogonal SLCF estimator*
If $\{O_i = (y_i, X_i, z_i)\}_{i \in [N]}$ *satisfy assumptions 1 to 4, assumptions of subsection 11.3 in the Annex hold, i)* Ω_i *is a positive semi-definite matrix and* $\Omega_i \mathbb{E}[\psi(O_i; \theta, \tau g_o)] = 0$ *only if* $\theta = \theta_o$, *ii)* $\theta \in \Theta$ *with* Θ *compact, iii)* $\psi(O_i; \theta, \tau g_o)$ *is continuous at each* $\theta \in \Theta$, *iv) the function class* $\mathscr{F} = \{\psi(O_i; \theta, \tau g_o)\}$ *satisfies a uniform integrability condition* $\mathbb{E}[\sup_{\in \Theta} \|\psi(O_i; \theta, \tau g_o)\|] < \infty$, *then* $\hat{\theta}_o \xrightarrow{p} \theta_o$.

The proof is presented in the appendix (Subsection 11.6).

Theorem 2 *Asymptotic normality of the Orthogonal SLCF estimator of β_o*
If $\{O_i = (y_i, X_i, z_i)\}_{i\in[N]}$ satisfy assumptions 1 to 4, assumptions of subsection 11.3 in the Annex hold, then,

$$\sqrt{N_T}(\hat{\beta}_o - \beta_o) \xrightarrow{d} \mathcal{N}(0, \Sigma_o), \tag{37}$$

with $\Sigma_o = \mathbb{E}_P[\tilde{Z}_i' V_i^{-1} M_{\tau_i} X_i]^{-1} \mathbb{E}_P[\tilde{Z}_i' V_i^{-1} M_{\tau_i} \varepsilon_i \varepsilon_i' M_{\tau_i}' V_i^{-1} \tilde{Z}_i] \mathbb{E}_P[\tilde{Z}_i' V_i^{-1} M_{\tau_i} X_i]^{-1}$, *and* $\tilde{Z} = [\tau g_o(\tilde{X}_i, z_i) \quad M_{\tau_i}\tilde{X}_i]$.

The proof is presented in the appendix (Subsection 11.7).

Remark 7 *The asymptotic normality of the orthogonal SLCF estimator of the parameter ρ_o are left for later study. The reason is that the ρ_o has no correction in the orthogonal score derivation.*

6.1 Estimator of the variance-covariance matrix

The estimator of the variance-covariance matrix of the SLCF estimator of β_o is given by:

$$\hat{\Sigma}_o = \hat{J}_o^{-1} \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_{n,b}[\widehat{\tilde{Z}}_i'^{S_b} V_i^{-1} (M_{\tau_i} y_i^{S_b} - M_{\tau_i} X_i^{S_b} \hat{\beta}_o)(M_{\tau_i} y_i^{S_b} - M_{\tau_i} X_i^{S_b} \hat{\beta}_o)' V_i^{-1} \hat{\tilde{Z}}_i^{S_b}] (\hat{J}_o^{-1})', \tag{38}$$

where:

$$\hat{J}_o = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_{n,b}[\widehat{\tilde{Z}}_i'^{S_b} V_i^{-1} X_i^{S_b}]. \tag{39}$$

Remark 8 *In the case of Within transformation and ε_{it} are identical and independently distributed, there is need of adjusting the variance-covariance matrix by multiplying by the factor $\frac{NT-K}{NT-N-K}$. The reason is that this provides valid statistical inference after whiping out the individual fixed effects (Ding 2021). Otherwise, the cluster-robust variance-covariance matrix of the Within estimator does not need any correction (Ding 2021, Colin Cameron and Miller 2015).*

6.1.1 Including sample splitting uncertainty

In order to take into account the sample splitting uncertainty, we can repeat the cross-fitting process several times as explained in Section 5 and algorithm 1. We also need to add correction term to the variance-covariance estimator $\hat{\Sigma}_o$ given by:

$$C = \frac{1}{SS} \sum_{ss=1}^{SS} (\hat{\beta}_{ss} - \hat{\beta}_o)(\hat{\beta}_{ss} - \hat{\beta}_o)', \tag{40}$$

with $\hat{\beta}_o = \frac{1}{SS} \sum_{ss=1}^{SS} \hat{\beta}_{ss}$.
It is possible to use the median instead of the mean (Chernozhukov et al. 2018).

Algorithm 3 Estimator of the variance-covariance of the Super Learner Control Function estimator

Input: Panel data sets $\{y_i, X_i, z_i\}_{i\in\{1,2,...,N\}}$ from model satisfying assumptions 1 to 4, a natural number B. The selected index set $\{1,2,...,N\}$ for each set $S_1, S_2, ..., S_B$ across all samples splits (SS) obtained in algorithm 1. The estimated $\hat{\theta}_{ss}$ obtained in algorithm 1.
Output: An estimator of the variance-covariance matrix of β_{1o}.
for $ss \in \{1,2,...,SS\}$ **do**
  for $b \in \{1,2,...,B\}$ **do**
    Estimate $\hat{J}_{o,b}^{-1} = \mathbb{E}_{n,b}[\tilde{Z}'^{S_b}_i V_i^{-1} X_i^{S_b}]$.
    Estimate $\mathbb{E}_{n,b}[\tilde{Z}'^{S_b}_i V_i^{-1}(M_{\tau_i} y_i^{S_b} - M_{\tau_i} X_i^{S_b}\hat{\beta}_o)(M_{\tau_i} y_i^{S_b} - M_{\tau_i} X_i^{S_b}\hat{\beta}_o)' V_i^{-1} \tilde{Z}_i^{S_b}]$.
  end for
  Compute $\hat{\Sigma}_s$.
end for
Compute $C = \frac{1}{SS}\sum_{ss=1}^{SS}(\hat{\beta}_{ss} - \hat{\beta}_o)(\hat{\beta}_{ss} - \hat{\beta}_o)'$.
Compute $\hat{\Sigma}_o = \frac{1}{SS}\sum_{ss=1}^{SS}(\hat{\Sigma}_s) + C$.

7 Comparison of the Orthogonal SLCFE with plug-in IV estimator

Proposition 7 *For each fold b, let* $\widehat{\tilde{Z}} := [\widehat{\tau g_o}^{S_b^c}(\tilde{X}_i, z_i),\ M_{\tau_i}\tilde{X}_i]$ *and* $\hat{u}_i := \widehat{\tau u_i}^{S_b}$ *for* $i \in S_b$*, and suppose* $\sum_{i\in S_b} \widehat{\tilde{Z}}' V_i^{-1} M_{\tau_i} X_i$ *is nonsingular and* $\sum_{i\in S_b} \hat{u}_i' V_i^{-1} \hat{u}_i > 0$*. Then* $\hat{\theta}_b = (\hat{\beta}_b', \hat{\rho}_b)'$ *solves the sample moment conditions (34) if and only if*

$$\hat{\beta}_b = \Big(\sum_{i\in S_b} \widehat{\tilde{Z}}' V_i^{-1} M_{\tau_i} X_i\Big)^{-1} \sum_{i\in S_b} \widehat{\tilde{Z}}' V_i^{-1} M_{\tau_i} y_i, \tag{41}$$

$$\hat{\rho}_b = \frac{\sum_{i\in S_b} \hat{u}_i' V_i^{-1}(M_{\tau_i} y_i - M_{\tau_i} X_i \hat{\beta}_b)}{\sum_{i\in S_b} \hat{u}_i' V_i^{-1} \hat{u}_i}. \tag{42}$$

That is, the orthogonalized control function estimator of β_o *is the instrumental variables estimator with the learned instrument* $\widehat{\tau g_o}^{S_b^c}$ *(and* $M_{\tau_i}\tilde{X}_i$ *as its own instrument) and weighting matrix* V_i^{-1}*.*

Proof 3 *Write* $r_i(\beta) := M_{\tau_i} y_i - M_{\tau_i} X_i \beta$ *and* $\sum$ *for* $\sum_{i\in S_b}$*. By definition of the residual,* $M_{\tau_i} x_{1i} - \widehat{\tau g_o}^{S_b^c}(\tilde{X}_i, z_i) = \hat{u}_i$*, hence*

$$M_{\tau_i} X_i - \widehat{\tilde{Z}} = [\hat{u}_i,\ 0] = \hat{u}_i e_1', \qquad e_1 = (1,0,\ldots,0)'. \tag{43}$$

Summing the score over $i \in S_b$*, the moment conditions (34) are*

$$\sum (M_{\tau_i} X_i)' V_i^{-1}(r_i(\beta) - \rho\hat{u}_i) + \rho \sum \widehat{\tilde{Z}}' V_i^{-1} \hat{u}_i = 0 \qquad (\beta \text{ rows}), \tag{44}$$

$$\sum \hat{u}_i' V_i^{-1}(r_i(\beta) - \rho\hat{u}_i) = 0 \qquad (\rho \text{ row}). \tag{45}$$

Collecting the terms in ρ *in (44) and using (43),*

$$-\rho \sum (M_{\tau_i} X_i - \widehat{\tilde{Z}})' V_i^{-1} \hat{u}_i = -\rho\, e_1 \sum \hat{u}_i' V_i^{-1} \hat{u}_i,$$

so (44) is equivalent to

$$\sum (M_{\tau_i} X_i)' V_i^{-1} r_i(\beta) = e_1 \rho \sum \hat{u}_i' V_i^{-1} \hat{u}_i. \tag{46}$$

(For the rows of β_2 the two terms in ρ cancel exactly; for the row of β_1 they reduce to $\rho \sum \hat{u}_i' V_i^{-1} \hat{u}_i$.) By (45), $\rho \sum \hat{u}_i' V_i^{-1} \hat{u}_i = \sum \hat{u}_i' V_i^{-1} r_i(\beta)$. Substituting in (46) and using (43) again,

$$\sum (M_{\tau_i} X_i - \hat{u}_i e_1')' V_i^{-1} r_i(\beta) = \sum \widehat{\tilde{Z}}' V_i^{-1} (M_{\tau_i} y_i - M_{\tau_i} X_i \beta) = 0,$$

which, under the nonsingularity condition, is equivalent to (41); and (45) is (42). Each step is an equivalence (adding zero, regrouping, and replacing one side of (45) by the other), so the converse also holds.

Remark 9 *The proposition is exact algebra in every sample and for any first-stage estimator; it uses neither conditional expectations nor asymptotics. Cross-fitting is not needed for the equivalence; it is needed for the asymptotic properties of the estimator (Theorems 1, 2). The learned-instrument interpretation connects the estimator with the literature on estimated optimal instruments (Newey, 1990; Belloni et al., 2012) and with machine-learning first stages in instrumental variables (Chen et al., 2020).*

7.1 Comparison of the Naive Plug-in 2SLS estimator and Plug-in IV estimation are not equivalent to Naive SLCFE

The naive Super Learner Control function estimator based on the moment condition 19 is not numerical equivalent to a 2SLS estimator that plugs in the estimated transformed nuisance parameter $\tau g_o(\tilde{x}_{it}, z_{it})$. Without loss of generality, we assume that exogenous covariates are not present (model 1 to 3), and $V_i = I_i$. As described in Section 4, one could be tempted to estimate the parameter of interest using the naive 2SLS estimator:

$$\hat{\beta}_{2SLS,1_o} = \sum_b^B \frac{1}{B} \Big(\sum_{i \in S_b} \widehat{\tau g_o}^{S_b^c}(z_i)' \widehat{\tau g_o}^{S_b^c}(z_i) \Big)^{-1} \Big(\sum_{i \in S_b} \widehat{\tau g_o}^{S_b^c}(z_i)' M_{\tau_i} y_i \Big). \tag{47}$$

Remark 10 *If $\{O_i = (y_i, X_i, z_i)\}_{i \in [N]}$ satisfy assumptions 1 to 4, assumptions 11 - 13 of subsection 11.3 in the Annex hold, Naive plug-in 2SLS estimation is not numerically equivalent to the naive SLCFE when $g_o(.)$ is nonlinear*

Assuming that $B = 1$, we can write the estimators as:

$$\hat{\beta}_{2SLS,1_o} = \Big(\sum_i \widehat{\tau g_o}(z_i)' \widehat{\tau g_o}(z_i) \Big)^{-1} \Big(\sum_i \widehat{\tau g_o}(z_i)' M_{\tau_i} y_i \Big), \tag{48}$$

$$\hat{\beta}_{Naive_SLCFE,1_o} = \Big(\sum_i (M_{\tau_i} x_{1i})' \hat{A}_{\tau u,i} M_{\tau_i} x_{1i} \Big)^{-1} \Big(\sum_i (M_{\tau_i} x_{1i})' \hat{A}_{\tau u,i} M_{\tau_i} y_i \Big), \tag{49}$$

where $\hat{A}_{\tau u,i} = I_{T_{ai}} - \widehat{\tau u_i} (\widehat{\tau u_i}' \widehat{\tau u_i})^{-1} \widehat{\tau u_i}'$.
Because $\hat{A}_{\tau u,i} M_{\tau_i} x_{1i} \neq \widehat{\tau g_o}(z_i)$, the two estimators are not equivalent. They are numerically equivalent if $\widehat{\tau g_o(z)} \in col(\tau z)$.

Remark 11 *If $\{O_i = (y_i, X_i, z_i)\}_{i \in [N]}$ satisfy assumptions 1 to 4, assumptions 11 - 13 of subsection 11.3 in the Annex hold, the Naive plug-in 2SLS estimator is not $\sqrt{N_T}$-consistent.*

The naive plug-in 2SLS estimator with cross-fitting is given by:

$$\hat{\beta}_{2SLS,1_o} = \frac{1}{B} \sum_b^B \Big(\sum_{i \in S^b} \widehat{\tau g_o}^{S_b^c}(z_i)' \widehat{\tau g_o}^{S_b^c}(z_i) \Big)^{-1} \Big(\sum_{i \in S^b} \widehat{\tau g_o}^{S_b^c}(z_i)' M_{\tau_i} y_i \Big). \tag{50}$$

Replacing $M_{\tau_i} y_i$ by $\tau g_o(z_i)\beta_{1_o} + M_{\tau_i} u_i \beta_{1_o} + M_{\tau_i}\varepsilon_i$, adding and subtracting $\widehat{\tau g_o}^{S_b^c}(z_i)\beta_{1_o}$, it can be written as:

$$\hat{\beta}_{2SLS,1_o} = \beta_{1_o} + \frac{1}{B}\sum_b^B \Big(\sum_{i\in S^b} \widehat{\tau g_o}^{S_b^c}(z_i)' \widehat{\tau g_o}^{S_b^c}(z_i) \Big)^{-1} \Big(\sum_{i\in S^b} \widehat{\tau g_o}^{S_b^c}(z_i)' (\xi_i^{S_b^c}\beta_{1,o} + M_{\tau_i} u_i \beta_{1_o} + M_{\tau_i}\varepsilon_i) \Big), \tag{51}$$

with $\xi_i^{S_b^c} = \tau g_o(z_i) - \widehat{\tau g_o}(z_i)^{S_b^c}$.
Now, adding and subtracting $\tau g_o(z_i)$ to $\widehat{\tau g_o}^{S_b^c}(z_i)$ we obtain that:

$$\hat{\beta}_{2SLS,1_o} = \beta_{1_o} + \frac{1}{B}\sum_b^B \Big(\sum_{i\in S^b} \widehat{\tau g_o}^{S_b^c}(z_i)' \widehat{\tau g_o}^{S_b^c}(z_i) \Big)^{-1} \Big(\sum_{i\in S^b} (\xi_i^{S_b^c} + \tau g_o(z_i))' (\xi_i^{S_b^c}\beta_{1,o} + M_{\tau_i} u_i \beta_{1,o} + M_{\tau_i}\varepsilon_i) \Big), \tag{52}$$

Heuristically, the second term does not converge to zero at rate faster than $\sqrt{N_T}$ because $\tau g_o(z_i)$, unlike $M_{\tau_i} u_i$ and $M_{\tau_i}\varepsilon_i$, does not have conditional mean zero given z_i, so the term $\sum_i \tau g_o(z_i)'\xi_i$ has non-zero mean and, after scaling by $\sqrt{N_T}$, diverges (Chernozhukov et al., 2018, Section 1.1) Thus, the naive plug-in estimator presents a first order bias term that does not vanish using sample-splitting after scaling by $\sqrt{N_T}$. Thus, the naive plug-in 2SLS estimator is not $\sqrt{N_T}$-consistent.

The IV estimator is given by:

$$\hat{\beta}_{IV,1_o} = \sum_b^B \frac{1}{B} \Big(\sum_{i\in S_b} \widehat{\tau g_o}^{S_b^c}(z_i)' M_{\tau_i} x_{1i} \Big)^{-1} \Big(\sum_{i\in S_b} \widehat{\tau g_o}^{S_b^c}(z_i)' M_{\tau_i} y_i \Big). \tag{53}$$

Remark 12 *If $\{O_i = (y_i, X_i, z_i)\}_{i\in[N]}$ satisfy assumptions 1 to 4, assumptions of subsection 11.3 in the Annex hold, plug-in IV estimation is not numerically equivalent to the naive SLCFE when $g_o(.)$ is nonlinear*

Assuming that $B = 1$, we can write the estimators as:

$$\hat{\beta}_{IV,1_o} = \Big(\sum_i \widehat{\tau g_o}(z_i)' M_{\tau_i} x_{1i} \Big)^{-1} \Big(\sum_i \widehat{\tau g_o}(z_i)' M_{\tau_i} y_i \Big), \tag{54}$$

$$\hat{\beta}_{Naive_SLCFE,1_o} = \Big(\sum_i (M_{\tau_i} x_{1i})' \hat{A}_{\tau u,i} M_{\tau_i} x_{1i} \Big)^{-1} \Big(\sum_i (M_{\tau_i} x_{1i})' \hat{A}_{\tau u,i} M_{\tau_i} y_i \Big), \tag{55}$$

where $\hat{A}_{\tau u,i} = I_{T_{ai}} - \widehat{\tau u_i}(\widehat{\tau u_i}'\widehat{\tau u_i})^{-1}\widehat{\tau u_i}'$.
Because $\hat{A}_{\tau u,i} M_{\tau_i} x_{1i} \neq \widehat{\tau g_o}(z_i)$, the two estimators are not equivalent. They are numerically equivalent if $\widehat{\tau g_o(z)} \in col(\tau z)$.

8 Monte Carlo simulations

8.1 Design

We generate a balanced panel with $N = 2500$ individuals and $T = 2$ periods. For $i = 1,\ldots,N$ and $t = 1,2$,

$$\begin{aligned} x_{1it} &= 0.5\,\alpha_i + g_o(\tilde{x}_{it}, z_{it}) + u_{it}, \\ y_{it} &= 0.5\,\alpha_i + \tilde{x}_{it}\beta_{2_o} + x_{1it}\beta_{1_o} + \varepsilon_{it}, \qquad \varepsilon_{it} = \rho_o u_{it} + \omega_{it}, \\ g_o(\tilde{x}_{it}, z_{it}) &= a\,|z_{it}| - 2\tanh(\tilde{x}_{it}) + z_{it}/a, \end{aligned}$$

with $\alpha_i \sim U(-1,1)$, $\tilde{x}_{it} = \alpha_i + U(-2,2)$, $z_{it} \sim U(-2,2)$, $u_{it} \sim U(-2,2)$ and $\omega_{it} \sim U(-1,1)$, all independent, $\beta_{1_o} = \beta_{2_o} = 1$ and $\rho_o = 0.9$. The parameter $a \in \{1,\dots,10\}$ moves the first stage from a design with a strong linear component ($a = 1$) to one in which the dependence of x_1 on z is almost entirely nonlinear ($a = 10$): the linear part z/a shrinks and the nonlinear part $a|z|$ grows. Linear instruments therefore become weak as a increases, while g_o remains a strong instrument. We use the first-difference transformation; with $T = 2$ it gives the same estimator as the within transformation. Each design is replicated $R = 500$ times. Appendix 12 reports a design with heterogeneous ρ_i.

8.2 Estimators

All estimators are computed on first-differenced data with regressors $(\Delta x_{1it}, \Delta\tilde{x}_{it})$. We report results for β_{1_o}.

- **FD-OLS**: OLS of Δy on $(\Delta x_1, \Delta\tilde{x})$.
- **FD-2SLS**: 2SLS with instruments $(\Delta z, \Delta\tilde{x})$.
- **FD-2SLS poly**: 2SLS with instruments (orthogonal polynomials of degree 5 in Δz, $\Delta\tilde{x}$).
- **Oracle IV**: IV with instruments $(\Delta g_o, \Delta\tilde{x})$, using the true g_o. This estimator is infeasible and serves as a benchmark.
- **Orthogonal SLCF**: the estimator of Algorithm 2. By Proposition 7, it equals the cross-fitted IV estimator with instruments $(\widehat{\Delta g}, \Delta\tilde{x})$.
- **Naive SLCF**: the control function estimator without the correction term, that is, OLS of Δy on $(\Delta x_1, \Delta\tilde{x}, \widehat{\Delta u})$ with the same cross-fitted residuals and the usual OLS sandwich variance.
- **Plug-in 2SLS**: OLS of Δy on $(\widehat{\Delta g}, \Delta\tilde{x})$, cross-fitted.
- **IV no CF**: IV with instruments $(\widehat{\Delta g}, \Delta\tilde{x})$, where $\widehat{\Delta g}$ is fitted on the full sample, without cross-fitting.

The last three estimators isolate the ingredients of the Orthogonal SLCF estimator. Naive SLCF is cross-fitted but not orthogonal. IV no CF is orthogonal but not cross-fitted. Plug-in 2SLS is cross-fitted but uses $\widehat{\Delta g}$ as a regressor instead of as an instrument.

8.3 Implementation

The super learner combines the sample mean, linear regression, a neural network with one hidden layer of 10 units and linear output, and gradient boosting (`xgboost`). Its weights are chosen by 5-fold cross-validation with folds formed by individual. Its features are $(z_{it}, z_{it-1}, \tilde{x}_{it}, \tilde{x}_{it-1}, \Delta\tilde{x}_{it}, \Delta z_{it})$ for all a. Cross-fitting uses $B = 5$ folds formed by individual. Within each sample split, the estimator is computed fold by fold and averaged over folds. We repeat this for $SS = 3$ sample splits and take the median $\hat{\beta}_1 = \text{median}_{ss}\,\hat{\beta}_{1,ss}$, with variance $\text{median}_{ss}\{\hat{V}_{ss} + (\hat{\beta}_{1,ss} - \hat{\beta}_1)^2\}$ (Subsection 6.1.1). All standard errors are cluster-robust at the individual level. With $R = 500$, the Monte Carlo standard error of a coverage rate is at most 0.022, and 0.010 at 0.95.

8.4 Results

Table 1 reports the median bias, the mean squared error (MSE) and the coverage of nominal 95% confidence intervals. Figure 1 plots the bias and the coverage against a. First-stage diagnostics, interquartile ranges, confidence-interval lengths and further robustness checks are in Appendix 12.

The first stage behaves as intended. The cluster-robust F statistic of Δz falls from 1840.5 at $a = 1$ to 0.9 at $a = 10$ and is below 10 for $a \geq 6$, while the out-of-fold R^2 of the super learner increases from 0.70 to 0.95 (Table 3).

FD-OLS has median bias between 0.383 and 0.035 and zero coverage; its MSE is nevertheless smaller than that of FD-2SLS for $a \geq 5$. FD-2SLS has median bias of at most 0.023 in absolute value, but its dispersion grows with a: its interquartile range increases from 0.026 at $a = 1$ to 0.256 at $a = 10$, and its coverage is at least 0.99 for $a \geq 7$. For $a \geq 7$, its MSE is dominated by a few replications with very large errors, as expected for an exactly identified IV estimator with a weak instrument. FD-2SLS poly has smaller dispersion than FD-2SLS for large a, but median bias between 0.034 and 0.059 for $a \geq 5$, the bias towards OLS expected with several weak instruments. Polynomials of Δz do not help because, in first differences, they are correlated with Δg_o only through the linear term (Appendix 12).

The Orthogonal SLCF estimator has median bias of at most 0.002 in absolute value for all a and coverage between 0.93 and 0.97, within two Monte Carlo standard errors of 0.95. Its interquartile range and confidence-interval length are within 0.003 of those of the oracle IV estimator, and equal to three decimals for $a \geq 4$. Relative to FD-2SLS, its interquartile range is about the same at $a = 1$ (0.024 against 0.026) and about 50 times smaller at $a = 10$ (0.005 against 0.256). Its MSE exceeds that of the oracle by at most 9% (at $a = 1$) and by at most 6% for $a \geq 2$. It is smaller than the MSE of FD-2SLS for every a, by a factor of 1.2 at $a = 1$ and of more than 200 at $a = 5$, and smaller than that of FD-2SLS poly by a factor of more than 100 for $a \geq 5$. In this design, the Orthogonal SLCF estimator thus attains the precision of the infeasible estimator that uses the true g_o.

The three alternative estimators fail in different ways.

- **Naive SLCF** has median bias between -0.014 and -0.036, increasing in absolute value with a, and coverage that falls from 0.72 at $a = 1$ to 0.18 at $a = 10$. Its confidence intervals are too short: the OLS sandwich treats $\widehat{\Delta u}$ as data and ignores the effect of the first-stage estimation on the limit distribution, which is not zero because this score is not orthogonal.

- **Plug-in 2SLS** has median bias between 0.017 and 0.050, generally increasing with a, and coverage that falls from 0.95 at $a = 1$ to 0.25 at $a = 10$. The estimation error $\widehat{\Delta g} - \Delta g_o$ enters the regressor, so the estimator is consistent only if $\widehat{\Delta g}$ is consistent for Δg_o. In the IV estimator, $\widehat{\Delta g}$ enters only as an instrument, and cross-fitting makes it uncorrelated with the structural error of the same individual, whether or not it is close to Δg_o.

- **IV no CF** has median bias of 0.072 at $a = 1$, decreasing to 0.010 at $a = 10$, and coverage between 0.06 and 0.34. When $\widehat{\Delta g}$ is fitted on the full sample, the prediction for individual i depends on Δx_{1i}, hence on Δu_i, and is therefore correlated with $\Delta \varepsilon_i = \rho_o \Delta u_i + \Delta \omega_i$.

Their MSEs are between 7 and 393 times that of SLCF: CF uncorr. between 2.33×10^{-3} and 4.37×10^{-3}, Plug-in 2SLS between 3.30×10^{-3} and 6.66×10^{-3}, and IV no CF between 0.126×10^{-3} and 7.91×10^{-3}. Cross-fitting and the orthogonal (IV) form are therefore both needed. Dropping either gives bias of the order of the standard deviation of the estimator or larger, and coverage well below the nominal level.

Identification through functional form. In this design, g_o is nonlinear in the exogenous regressor $\tilde{x}$ as well as in z. The learned instrument can therefore also use the nonlinearity in $\tilde{x}$, which is excluded from

the outcome equation only through the assumption that $\tilde{x}$ enters it linearly (Angrist and Frandsen, 2022; Chen, Chen and Lewis, 2020).

9 Empirical application: the price elasticity of electricity demand

9.1 Design

Burke and Abayasekara (2018) estimate the short-run price elasticity of electricity demand with a panel of US states s, sectors g (residential, commercial, industrial) and years t. In levels, the model is

$$\ln E_{sgt} = \beta_1 \ln P_{sgt} + \alpha_{sg} + \lambda_{sg} t + \mu_{gt} + \varepsilon_{sgt}, \tag{56}$$

$$\ln P_{sgt} = g(\ln P^{\text{other}}_{sgt}, g, t) + \alpha^{x}_{sg} + u_{sgt}, \qquad \varepsilon_{sgt} = \rho\, u_{sgt} + \omega_{sgt}, \tag{57}$$

where E_{sgt} is electricity sales, P_{sgt} is the average price (revenue divided by sales), α_{sg} is a state-by-sector effect, $\lambda_{sg} t$ a state-by-sector linear trend, and μ_{gt} a sector-by-year effect. The price is endogenous: it is set jointly with demand, and an average price falls mechanically when sales rise if part of the bill is fixed. The instrument is $\ln P^{\text{other}}_{sgt}$, the log average price paid by the other sectors of the same state in the same year (for the residential sector: the commercial, industrial and transportation sectors). It moves with the costs of supplying electricity in the state but is not directly affected by demand in sector g. Burke and Abayasekara (2018) specify the first stage as linear, $g(z,g,t) = \pi z + \kappa_{gt}$; we leave g unrestricted.

First differences remove α_{sg} and α^{x}_{sg} and give the equation that Burke and Abayasekara (2018) estimate (their Table 5, column 6):

$$\Delta \ln E_{sgt} = \beta_1 \Delta \ln P_{sgt} + \lambda_{sg} + \Delta\mu_{gt} + \Delta\varepsilon_{sgt},$$

in which the trends become state-by-sector effects and $\Delta\mu_{gt}$ are sector-by-year effects. Observations are weighted by the square root of the number of customers, and standard errors are clustered by state and sector. They report $\hat{\beta}_1 = -0.06$ (0.02), with a first-stage coefficient of 0.93 and $F = 2{,}421$, on 1,728 observations (48 states, 3 sectors, 12 years).

In the notation of Section 2, the unit is $i = (s,g)$, with $N = 144$ units, and t is the year, with $T = 12$ differences, so that $x_{1it} = \ln P_{sgt}$ and $z_{it} = \ln P^{\text{other}}_{sgt}$. The regressor x_{1it} is endogenous, z_{it} is the instrument, and $\tilde{x}$ contains the sector-by-year effects and the state-by-sector trends of (56). We keep the outcome, the regressor, the instrument, the fixed effects and the weights, and change only the first stage. We report first-differenced OLS, first-differenced 2SLS with their linear instrument, and FD orthogonal SLCF, the estimator of Algorithm 2 with the first-difference transformation and $V_i = I$, which by Proposition 7 is IV with the learned instrument.

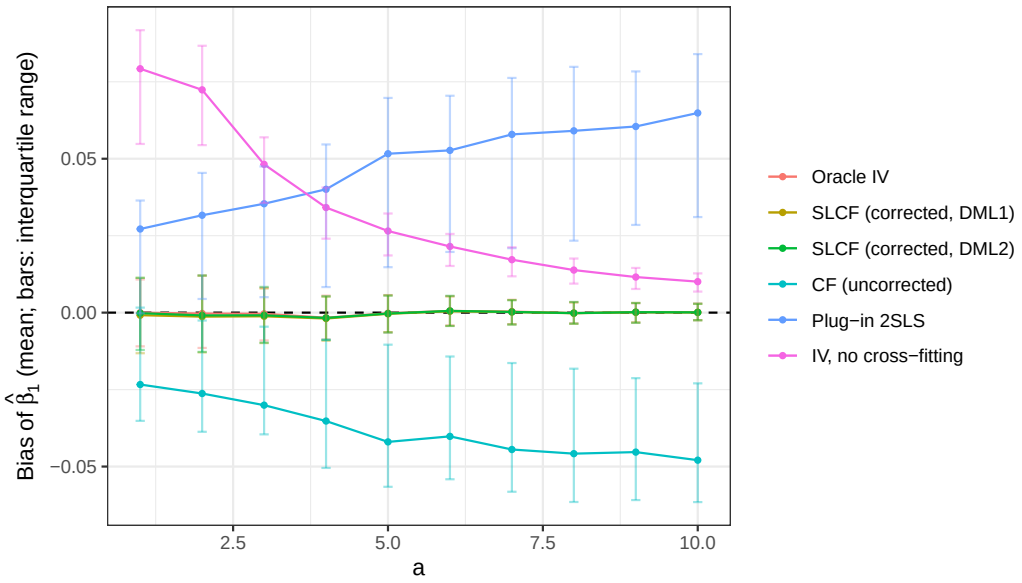
First stage. With first differences, the first stage is $\Delta \ln P_{sgt} = g(z_{sgt}, g, t) - g(z_{sg,t-1}, g, t-1) + \Delta u_{sgt}$. The super learner therefore estimates $\mathbb{E}[\Delta \ln P_{sgt} \mid z_{sgt}, z_{sg,t-1}, g, t]$. Its library contains the mean, linear regression and gradient boosting (learning rate 0.1, depth 6, 300 rounds), with weights chosen by non-negative least squares on 3-fold cross-validated predictions. Functions of the sector and year alone are absorbed by the sector-by-year effects, so the learned instrument varies only through the instrument and its interactions with sector and year.

Cross-fitting and inference. The instrument of one sector is built from the prices of the other sectors in the same state, so the three sectors of a state are dependent. We therefore form the $B = 5$ cross-fitting

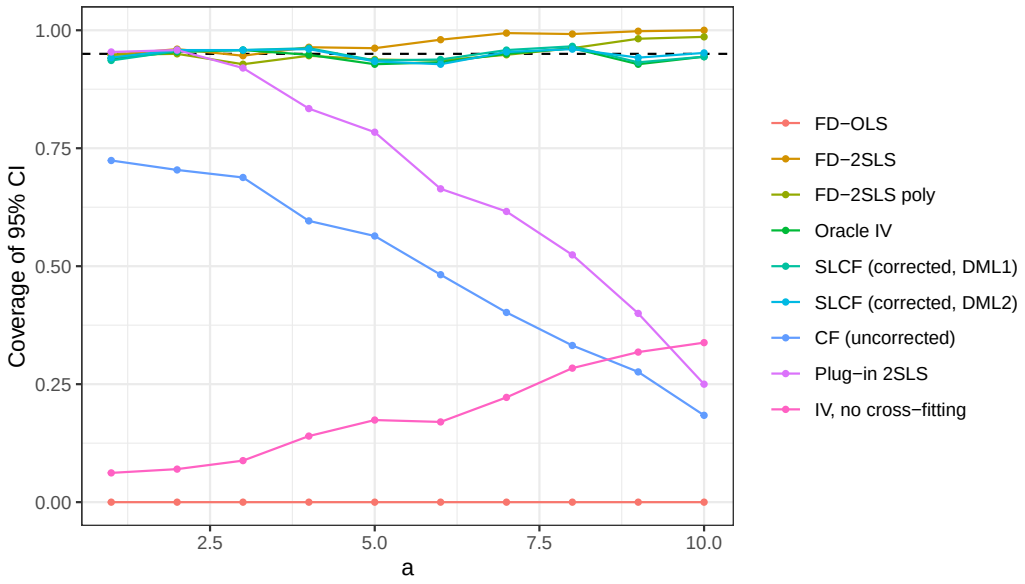
Table 1: Median bias, mean squared error and coverage of $\hat{\beta}_1$

a	FD-OLS	FD-2SLS	FD-2SLS poly	Oracle IV	Orthogonal SLCF	Naive SLCF	Plug-in 2SLS	IV no CF
A. Median bias								
1	0.383	0.001	0.003	0.001	0.000	−0.014	0.017	0.072
2	0.383	0.002	0.007	0.000	−0.002	−0.017	0.021	0.071
3	0.261	0.003	0.014	−0.000	−0.001	−0.019	0.021	0.047
4	0.174	0.003	0.022	−0.001	−0.002	−0.021	0.022	0.033
5	0.122	0.004	0.034	0.000	0.000	−0.026	0.030	0.026
6	0.090	0.001	0.039	0.001	0.001	−0.029	0.038	0.021
7	0.067	0.022	0.059	0.000	−0.000	−0.031	0.042	0.016
8	0.053	0.009	0.046	−0.000	−0.000	−0.033	0.041	0.013
9	0.043	0.007	0.047	0.000	0.000	−0.036	0.048	0.011
10	0.035	0.023	0.036	0.000	−0.000	−0.036	0.050	0.010
B. Mean squared error $\times 1000$								
1	146	0.438	0.443	0.328	0.358	3.13	4.65	7.91
2	146	1.68	1.68	0.317	0.335	2.33	3.30	6.30
3	67.8	4.30	4.16	0.158	0.167	2.65	3.89	2.82
4	30.2	7.95	6.64	0.106	0.107	3.07	4.41	1.47
5	14.9	18.5	10.7	0.076	0.077	4.37	6.66	0.890
6	8.04	51.2	16.7	0.049	0.051	3.24	5.38	0.564
7	4.57	2,963	16.6	0.035	0.036	3.71	6.39	0.370
8	2.78	6,134	20.4	0.025	0.025	3.62	6.01	0.236
9	1.81	207,650	17.2	0.022	0.022	3.19	5.65	0.168
10	1.22	2,296	18.4	0.016	0.017	3.59	6.52	0.126
C. Coverage of nominal 95% confidence intervals								
1	0.00	0.95	0.95	0.94	0.94	0.72	0.95	0.06
2	0.00	0.96	0.95	0.95	0.96	0.70	0.96	0.07
3	0.00	0.95	0.93	0.96	0.96	0.69	0.92	0.09
4	0.00	0.96	0.95	0.95	0.96	0.60	0.83	0.14
5	0.00	0.96	0.94	0.93	0.94	0.56	0.78	0.17
6	0.00	0.98	0.94	0.93	0.94	0.48	0.66	0.17
7	0.00	0.99	0.95	0.95	0.96	0.40	0.62	0.22
8	0.00	0.99	0.96	0.96	0.97	0.33	0.52	0.28
9	0.00	1.00	0.98	0.93	0.93	0.28	0.40	0.32
10	0.00	1.00	0.99	0.94	0.94	0.18	0.25	0.34

500 replications, $N = 2500$, $T = 2$, first differences, $\beta_{1_o} = 1$. Panel A: median of $\hat{\beta}_1 - \beta_{1_o}$. Panel B: mean of $(\hat{\beta}_1 - \beta_{1_o})^2$, multiplied by 1000. FD-2SLS is exactly identified; for $a \geq 7$ its Monte Carlo mean squared error is driven by a few replications with very large $|\hat{\beta}_1 - \beta_{1_o}|$ (up to 256 at $a = 9$) and should be read together with the median bias and the interquartile range (Table 4). Panel C: Monte Carlo standard error at most 0.022, and 0.010 at 0.95.



(a) Mean bias of $\hat{\beta}_1$; bars are interquartile ranges of $\hat{\beta}_1 - \beta_{1_o}$. FD-OLS, FD-2SLS and FD-2SLS poly are omitted for scale.



(b) Coverage of nominal 95% confidence intervals. The dashed line is 0.95.

Figure 1: Bias and coverage of $\hat{\beta}_1$ by a. 500 replications, $N = 2500$, $T = 2$, first differences.

folds by state: the learner never predicts a state it was trained on. As in Burke and Abayasekara (2018), standard errors are clustered by state and sector; we also report standard errors clustered by state. We solve the moment conditions once over all folds and repeat the procedure for $SS = 10$ sample splits; we report the median, with variance $\text{median}_{ss}\{\hat{V}_{ss} + (\hat{\beta}_{1,ss} - \hat{\beta}_1)^2\}$ (Subsection 6.1.1).

Data. Sales, revenue and customers by state and sector come from the historical EIA-861 files of the U.S. Energy Information Administration (U.S. Energy Information Administration, 2026), the source used by Burke and Abayasekara (2018). The price is revenue divided by sales; the price paid by the other sectors is total revenue minus the sector's revenue, divided by total sales minus the sector's sales, where totals include the transportation sector. Alaska, Hawaii and the District of Columbia are excluded. The sample is 2004–2015 (differences of 2003–2015 levels): 1,728 observations in 144 units, as in the original study.

9.2 Results

Table 2: Short-run price elasticity of electricity demand, US states, 2004–2015

	FD-OLS	FD-2SLS linear	FD orth. SLCF	Published 2SLS
$\hat{\beta}_1$	−0.099	−0.045	−0.046	−0.06
SE, clustered by state × sector	(0.022)	(0.027)	(0.027)	(0.02)
SE, clustered by state		(0.023)	(0.023)	
$\hat{\rho}$			−0.131	
			(0.050)	
AR 95% set			[−0.101, 0.011]	
First-stage partial R^2		0.669	0.667	
First-stage F		861.2	732.3	2,420.9
Out-of-fold R^2			0.649	
Observations	1,728	1,728	1,728	1,728

Outcome: Δln electricity sales; endogenous regressor: Δln average price. Units: 48 states × 3 sectors; first differences over 2004–2015. All columns include sector-by-year and state-by-sector effects and are weighted by the square root of the number of customers. Instrument: Δln average price paid by the other sectors of the same state. FD orthogonal SLCF: cross-fitting folds by state, median over 10 sample splits; super learner weights (mean, linear, boosting) 0.00/0.74/0.26. AR: Anderson–Rubin set with the learned instrument averaged over splits. Hansen J test with the learned and the linear instrument: $p = 0.75$. Published: Burke and Abayasekara (2018), Table 5, column 6, with state-level controls and real prices.

Replication. OLS gives −0.099 (0.022), the same as the −0.10 (0.03) of their estimate without instruments and controls (their Table 5, column 1). Our FD-2SLS estimate is −0.045 (0.027), against their −0.06 (0.02); the difference is within one standard error, and our specification does not yet include their controls. The first stage is strong: the partial R^2 is 0.67 and the cluster-robust F statistic is 861; they report $F = 2{,}421$ with their controls and possibly a different variance estimator.

First stage. The super learner puts a weight of 0.74 on linear regression and 0.26 on boosting, but the learned instrument has the same partial R^2 as the linear one (0.667 against 0.669). In first differences of

logs, the change in a sector's price is close to linear in the change in the price paid by the other sectors of the same state.

Estimates. FD orthogonal SLCF gives -0.046 (0.027), and FD-2SLS -0.045 (0.027). The estimators agree because the learned instrument is, up to estimation error, the linear one, and the Hansen J test with both instruments does not reject ($p = 0.75$). With standard errors clustered by state, which also allow for dependence between the sectors of a state, both standard errors are 0.023. The Anderson–Rubin 95% set, $[-0.101, 0.011]$, just includes zero. The control function coefficient $\hat{\rho} = -0.131$ (0.050) is negative: the part of the price change not explained by the instrument is negatively correlated with the demand shock. This is what an average price implies: when demand rises, fixed charges are spread over more kilowatt-hours and the average price falls, so OLS is biased downward, as the gap between OLS (-0.099) and IV (-0.045) shows.

Summary. In this published panel design the linear instrument is strong and the first stage is close to linear, so 2SLS is a reliable benchmark. The super learner gives most weight to the linear model, and FD orthogonal SLCF reproduces the FD-2SLS estimate and its standard error, which are close to the published estimate. The application shows that the estimator and its inference work in real panel data, and that learning the first stage costs nothing when there is little nonlinearity to learn. The gains from learning a nonlinear first stage, for which real data provide no benchmark, are shown in Section 8.

10 Conclusions

This paper proposes a triangular simultaneous equation model for panel data with additive separable individual-specific fixed effects composed of a linear structural equation with a nonlinear reduced form equation to deal with the issue of weak instrumental variables due to nonlinearities in panel data settings. The parameter of interest is the structural parameter of the endogenous variable. The identification of this parameter is obtained under the assumption of available exclusion restrictions and using a control function approach. Estimating the parameter of interest is done using an estimator that we call the orthogonal Super Learner Control Function (Orthogonal SLCF) estimator. The estimation procedure is composed of two steps and cross-fitting. We estimate the control function with a super learner using sample splitting. In the following step, we use the estimated control function to control for endogeneity in the structural equation. Cross-fitting is done across the individual dimension. The estimator is consistent and asymptotically normal achieving a parametric rate of convergence. We show that the estimator of the β_o coefficients is numerically equivalent to the IV estimator using as instruments the transformed nuisance parameter and the covariates. We perform a Monte Carlo simulation to test the performance of the estimators proposed. We conclude that the estimator performs well, provided that we can accurately learn the nuisance parameter in the first stage.

11 Appendix

11.1 Notation

A set $\{1, 2, ..., N\}$ is denoted as $[N]$. $|\cdot|$ refers to the cardinality of a set. $\|\cdot\|_{L^p}$ is the L^p norm defined on the space $L^p(\Omega, P)$. $\|\cdot\|_p$ is the p-norm.

11.2 Definitions

We follow Emmenegger and Bühlmann (2023) and provide the following definitions. Each individual i presents T_i observations. We assume that T_i is uniformly bounded such that $T_i < T_{max}$. The total number of observations is denoted by $N_T = \sum_i T_i$. The number of sample splits B is not random and independent of N. The data is defined as $\{O_i = (X_i, z_i, y_i)\}_{i\in[N]}$ with $X_i = [x_{1i}, \tilde{X}_i]$, and denote the probability distribution P of the grouped data O_i. $\{\mathscr{P}_N\}_{N\geq 1}$ is a sequence of probability distributions. A partition of the set $[N]$ is $S_1, S_2, ..., S_B$. The total number of observations belonging to a partition $S_b \in [N]$ is equal to $n_{T,b} = \sum_{i\in S_b} T_i$. We assume that the subsets S_b have similar size such that $Bn_{T,b} = N_T + o(1), \quad \forall b \in [B]$. We denote $S_b^c = \{O_i = (y_i, X_i, z_i)\}_{i\in S_b^c}$, and the estimator of the nuisance parameter $\widehat{\tau g_o}^{S_b^c}$ with data from S_b^c. The empirical mean $\mathbb{E}_{n_{T,b}} = \frac{1}{n_{T_b}} \sum_{i\in S_b} \phi(O_i; \theta, \tau g)$. We define $\{\delta_N\}_{N\geq B}$ as a sequence of positive numbers that converge to 0 as $N \to \infty$, and $\{\delta_N\} \geq N^{-1/4}$.

11.2.1 Population Loss Function

We define the following population loss function:

$$Q(\theta, \tau g_o) = \mathbb{E}[(M_{\tau_i}(y_i - H_i\theta))' V_i^{-1} M_{\tau_i}(y_i - H_i\theta)],$$

where V_i is equal to the identity matrix I_i under the within transformation, and $\Sigma = \mathbb{E}[(M_{\tau_i}\omega_i)(M_{\tau_i}\omega_i)'|H_i]$ under first-difference transformation, also notice that V_i is a scalar if $T_i = 2, \forall i \in [N]$, $M_{\tau_i} = D_i$, and the error term ω_{it} is homoskedastic and uncorrelated across i and t. Finally, $\theta = [\beta_1, \quad \beta_2', \quad \rho]'$

11.2.2 Empirical Loss Function

We define the empirical loss function:

$$L_N(\theta, \tau g_o) = \sum_i^N \frac{1}{N} [(M_{\tau_i} y_i - M_{\tau_i} H_i \theta)' V_i^{-1} (M_{\tau_i} y_i - M_{\tau_i} H_i \theta)].$$

11.2.3 Population moment condition

We define the population moment condition:

$$m(O_i; \theta, \tau g_o) = \mathbb{E}[(M_{\tau_i} y_i - M_{\tau_i} H_i \theta)' V_i^{-1} M_{\tau_i} H_i]$$

with $m(O_i; \theta_o, \tau g_o) = 0$.

11.2.4 Non Neyman orthogonal score function

The score function is $\phi(O_i; \theta, \tau g_o) = (M_{\tau_i} y_i - M_{\tau_i} H_i \theta)' V_i^{-1} M_{\tau_i} H_i$.

11.2.5 Sample moment condition

We define the sample moment condition:

$$\hat{m}(O_i; \theta, \tau g_o) = \frac{1}{N} \sum_i (M_{\tau_i} y_i - M_{\tau_i} H_i \theta)' V_i^{-1} M_{\tau_i} H_i.$$

11.3 Assumptions

Assumption 11 *$\theta \in \Theta$, the set Θ is bounded, contains θ_o.*

This assumption is similar to assumption 4.1 of Emmenegger and Bühlmann (2023).

Assumption 12 *The set $\mathscr{G}$ consists P-integrable functions τg with pth moment that exists and it contains τg_o,*

$$\|\tau g - \tau g_o\|_{P,2} \leq \delta_N$$

,

$$\|\tau g - \tau g_o\|_{P,2}^2 \leq N_T^{-1/2}\delta_N$$

.

Assumption 13 *For all N, all $i \in [N]$, all $P \in \mathscr{P}_N$, all $b \in [B]$, and $p > 8$ we have:*

13.1 At the true θ_0 and the true g_0, the data $\{O_i = (X_i, Z_i, Y_i)\}_{i\in[N]}$ satisfies the identifiability condition.

$$\mathbb{E}_P\Big[\mathbb{E}_{n_{T,b}}[\phi(O_i;\theta_o,\tau g_o)]\Big] = 0.$$

13.2 $\mathbb{E}_P\Big[\mathbb{E}_{n_{T,b}}[\phi(O_i;\theta_o,\tau g_o)\phi(O_i;\theta_o,\tau g_o)']\Big] = \tilde{\Omega}_i$ with $\tilde{\Omega}_i$ a positive semi-definite matrix, with eigenvalues bounded away from 0 and finite, with the minimum eigenvalue $\lambda_T > 0$, and the maximum eigenvalue $\lambda_1 < \infty$.

13.3 $\mathbb{E}_P\Big[\mathbb{E}_{n_{T,b}}[\psi(O_i;\theta_o,\tau g_o)\phi(O_i;\theta_o,\tau g_o)']\Big] = \Omega_i$ with Ω_i a positive semi-definite matrix, with eigenvalues bounded away from 0 and finite, with the minimum eigenvalue $\tilde{\lambda}_T > 0$, and the maximum eigenvalue $\tilde{\lambda}_1 < \infty$.

13.4 V_i is a positive semi-definite matrix, with eigenvalues bounded away from 0 and finite, with the minimum eigenvalue $\overline{\lambda}_T > 0$, and the maximum eigenvalue $\overline{\lambda}_1 < \infty$.

13.5 There exists finite real constants C_1,C_2, C_3, C_4 satisfying $||y_i||_{P,p} \leq C_1$, $||z_i||_{P,p} \leq C_2$, $||x_{1,i}||_{P,p} \leq C_3$, $||\tilde{x}_i||_{P,p} \leq C_4$ for all $k > 2$.

Assumption 14 *$\mathbf{E}[\tau x_{k,i}] = 0$ for all $i \in [N]$, and all $k \in [K]$.*

Assumption 13.1 states that the true parameter vector is identifiable. Assumption 13.5 states that the data has finite moments. This assumption is equivalent to Assumption 5.1 of Emmenegger and Bühlmann (2023), and Newey and McFadden (1994).

Assumption 14 states that the unconditional expectation of the transformed regressors is equal to 0. This is plausible if the covariates are stationary.

11.4 Proof of Proposition Identification

For the proof, we follow Chen et al. (2014).

The first-derivative of $\mathbb{E}[\phi(O_i;\theta,\tau g_o)]$ with respect θ_o is $\mathbb{E}[\phi'(O_i;\theta_o,\tau g_o)] = -\mathbb{E}[(M_{\tau_i}H_i)'V_i^{-1}M_{\tau_i}H_i]$. Since the first-derivative of the moment $\mathbb{E}[\phi(O_i;\theta,\tau g_o)]$ has full rank, then $\mathbb{E}[\phi'(O_i;\theta_o,\tau g_o)'\phi'(O_i;\theta_o,\tau g_o)]$ has smallest eigenvalue λ_{min} is positive. Also, $|\mathbb{E}[\phi'(O_i;\theta_o,\tau g_o)]'\ell| \geq \lambda_{min}\ell$. Then, $\exists\zeta$ such that

$$\frac{|\mathbb{E}[\phi(O_i;\theta,\tau g_o)] - \mathbb{E}[\phi(O_i;\theta_o,\tau g_o)] - \mathbb{E}[\phi'(O_i;\theta_o,\tau g_o)(\theta-\theta_o)]|}{|\mathbb{E}[\phi'(O_i;\theta_o,\tau g_o)](\theta-\theta_o)|} < 1 \tag{58}$$

Thus, $\mathbb{E}[\phi(O_i;\theta,\tau g_o)] \neq 0$ for any $\theta \neq \theta_o$.

11.5 Proof of Proposition Concentration parameter inequality

Step 1 By the Projection Theorem, we can re-write the function $\tau g_o(\tilde{x}_{it}, z_{it})$ as $[\tau\tilde{x}_{it} \quad \tau z_{it}]'a + \tau\tilde{g}_o(\tilde{x}_{it}, z_{it})$.

Step 2 By the Pythagorean Theorem, we have that:

$$||[\tau\tilde{x}_{it} \quad \tau z_{it}]'a||_{P,2} \leq ||\tau g_o(\tilde{x}_{it}, z_{it})||_{P,2} \tag{59}$$

Step 3 Define the error term $\tau\tilde{u}_{it} = \tau x_{1it} - [\tau\tilde{x}_{it} \quad \tau z_{it}]'a$.
Notice that

$$\begin{aligned} \tau\tilde{u}_{it} &= \tau x_{1it} - [\tau\tilde{x}_{it} \quad \tau z_{it}]'a \\ &= \tau g_o(x_{it}, z_{it}) + \tau u_{it} - [\tau\tilde{x}_{it} \quad \tau z_{it}]'a \\ &= \tau\tilde{g}_o(x_{it}, z_{it}) + \tau u_{it} \end{aligned} \tag{60}$$

Thus, the variance of the error term $\tau\tilde{u}_{it}$ is larger or equal than the variance of $\tau u_{it} = \tau x_{1it} - \tau g_o(x_{it}, z_{it})$ since $\mathbb{E}[\tau\tilde{g}_o(x_{it}, z_{it})\tau u_{it}] = 0$ by L.I.E. and Assumption 3. Stacking up the time observations, we have that:

$$\begin{aligned} \mathbb{E}[(\tau\tilde{g}_o(\tilde{X}_i, z_i) + \tau u_i)(\tau\tilde{g}_o(\tilde{X}_i, z_i) + \tau u_i)'] = \\ \mathbb{E}[\tau\tilde{g}_o(\tilde{X}_i, z_i)\tau\tilde{g}_o(\tilde{X}_i, z_i)' + \tau u_i) + 2\tau\tilde{g}_o(\tilde{X}_i, z_i)\tau u_i' + \tau u_i\tau u_i'] = \\ \mathbb{E}[\tau\tilde{g}_o(\tilde{X}_i, z_i)\tau\tilde{g}_o(\tilde{X}_i, z_i)'] + \mathbb{E}[\tau u_i\tau u_i'], \end{aligned} \tag{61}$$

since the $\mathbb{E}[\tau\tilde{g}_o(\tilde{X}_i, z_i)\tau u_i'] = 0$ by Assumption 3.
Then, $\forall \mathbf{l} \in \mathbb{R}^{T_{ai}}$,

$$\mathbf{l}'\Big[\mathbb{E}[\tau\tilde{g}_o(\tilde{X}_i, z_i)\tau\tilde{g}_o(\tilde{X}_i, z_i)' + \tau u_i\tau u_i'] - \mathbb{E}[\tau u_i\tau u_i']\Big]\mathbf{l} \geq 0 \tag{62}$$

Step 4 By step 3, we have that

$$\mathbb{E}\Big[(M_{\tau_i}Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)'\Phi^{+}_{\tau\tilde{u},i}(M_{\tau_i}Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)\Big] \leq \mathbb{E}\Big[(M_{\tau_i}Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)'\Phi^{+}_{\tau u,i}(M_{\tau_i}Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)\Big]. \tag{63}$$

And by Step 2 we have that

$$\mathbb{E}\Big[(M_{\tau_i}Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)'\Phi^{+}_{\tau\tilde{u},i}(M_{\tau_i}Z_i a_1 + M_{\tau_i}\tilde{X}_i a_2)\Big] \leq \pi_o. \tag{64}$$

The equality holds when $g_o(\tilde{x}_{it}, z_{it}) \in span(\tilde{x}_{it}, z_{it})$.

11.6 Consistency

The proof of consistency of $\hat{\theta}_o$ follows Emmenegger and Bühlmann (2023), and Newey and McFadden (1994).

Proof 4 *Proof of Theorem 1*
Condition i) is satisfied by Assumptions 13.1, and 13.3, condition ii) is satisfied by Assumption 11, condition iii) and iv) are satisfied by lemma 12, lemma 15 of Emmenegger and Bühlmann (2023), and lemma 1. This implies that the estimator $\hat{\theta}$ is consistent (Emmenegger and Bühlmann, 2023).

Lemma 1 *We prove that lemma 13 of Emmenegger and Bühlmann (2023) holds in our set up.*

Proof 5

$$\begin{aligned}\psi(O_i;\theta,\tau g)-\psi(O_i;\theta,\tau g_o) &= (M_{\tau_i}H_{i,g})'V_i^{-1}M_{\tau_i}(y_i-x_i\beta-\rho(x_{1i}-g(\tilde{X}_i,z_i)))\\ &+(M_{\tau_i}[g(\tilde{X}_i,z_i),\tilde{X}_i,\boldsymbol{0}])'V_i^{-1}M_{\tau_i}(x_{1i}-g(\tilde{X}_i,z_i))\rho\\ &-(M_{\tau_i}H_{i,g_o})'V_i^{-1}M_{\tau_i}(y_i-x_i\beta-\rho(x_{1i}-g_o(\tilde{X}_i,z_i)))\\ &-(M_{\tau_i}[g_o(\tilde{X}_i,z_i),\tilde{X}_i,\boldsymbol{0}])'V_i^{-1}M_{\tau_i}(x_{1i}-g_o(\tilde{X}_i,z_i))\rho\end{aligned} \tag{65}$$

Since $g(\tilde{X}_i,z_i)=g_o(\tilde{X}_i,z_i)+\xi_i$, we are left with:

$$\psi(O_i;\theta,\tau g)-\psi(O_i;\theta,\tau g_o) = (M_{\tau_i}u_i)'V_i^{-1}M_{\tau_i}\xi_i\rho+(M_{\tau_i}\omega_i)'V_i^{-1}M_{\tau_i}\xi_i \tag{66}$$

With $P\in\{\mathscr{P}_N\}_{N\geq 1}$, then:

$$\begin{aligned}\sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\left\|\mathbb{E}_P[\mathbb{E}_{n_{T,b}}[(M_{\tau_i}(u_i\rho+\omega_i))'V_i^{-1}M_{\tau_i}\xi_i]]\right\| &\leq \sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\left\|\mathbb{E}_P[(M_{\tau_i}u_i)'V_i^{-1}M_{\tau_i}\xi_i\rho]\right\|+\sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\left\|\mathbb{E}_P[(M_{\tau_i}\omega_i)'V_i^{-1}M_{\tau_i}\xi_i]\right\|\\ &\leq \sup_{\in\Theta,\tau g\in\mathscr{G}}\left\|\mathbb{E}_P[\sum_t\lambda_t(q_t'M_{\tau_i}u_i)(q_t'M_{\tau_i}\xi_i)\rho]\right\|\\ &+\sup_{\in\Theta,\tau g\in\mathscr{G}}\left\|\mathbb{E}_P[\sum_t\lambda_t(q_t'M_{\tau_i}\omega_i')(q_t'M_{\tau_i}\xi_i)]\right\|\\ &\leq \sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\left\|\mathbb{E}_P[\lambda_1\sum_t(q_t'M_{\tau_i}\omega_i)(q_t'M_{\tau_i}\xi_i)]\right\|+\\ &\sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\left\|\lambda_1\mathbb{E}_P\sum_t(q_t'M_{\tau_i}\xi_i(q_t'M_{\tau_i}u_i)\rho]\right\|\\ &\leq \sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\lambda_1\left\|\|M_{\tau_i}\omega_i\|_{P,2}\|M_{\tau_i}\xi_i\|_{P,2}\right\|+\\ &\sup_{\theta\in\Theta,\tau g\in\mathscr{G}}\lambda_1\rho\left\|\|M_{\tau_i}u_i\|_{P,2}\|M_{\tau_i}\xi_i\|_{P,2}\right\|\end{aligned} \tag{67}$$

Since Θ is bounded, $\|\tau g-\tau g_o\|_{P,2}\leq\delta_N$, the last expression is bounded by δ_N by assumptions 11, and 12.

11.7 Asymptotic Normality

Proof 6 *In order to prove the asymptotic normality of the SLCF estimator of β_{1o}, we write the estimator as:*

$$\hat{\beta}_{1o}=\frac{1}{B}\sum_{b=1}^{B}\Big(\sum_{i\in S_b}(\widehat{\tilde{Z}}'_iV_i^{-1}M_{\tau_i}X_i)'\Big)^{-1}\Big(\sum_{i\in S_b}(\widehat{\tilde{Z}}'V_i^{-1}M_{\tau_i}y_i)'\Big), \tag{68}$$

with $\widehat{\tilde{Z}}=[\widehat{g}(\tilde{X}_i,z_i),\tilde{X}_i]$.

This is equal to:

$$\sqrt{N_T}(\hat{\beta}_o-\beta_o)=\sum_{b=1}^{B}\Big(\frac{1}{n_{T,b}}\sum_{i\in S^b}\widehat{\tilde{Z}}'V_i^{-1}M_{\tau_i}X_i)\Big)^{-1}\Big(\frac{\sqrt{N_T}}{n_{T,b}}\sum_{i\in S^b}\widehat{\tilde{Z}}'V_i^{-1}M_{\tau_i}\varepsilon_i\Big). \tag{69}$$

with

$$\widehat{\tilde{Z}}'V_i^{-1}M_{\tau_i}\varepsilon_i=\tilde{Z}_i'V_i^{-1}M_{\tau_i}\varepsilon_i+e_1\,\xi_i'V_i^{-1}M_{\tau_i}\varepsilon_i,\qquad \xi_i:=\widehat{\tau g}_i^{S_b^c}-\tau g_{o,i},\quad e_1=(1,0,\ldots,0)'.$$

The remainder $R_b := (NT)^{-1/2}\sum_{i\in S_b}\xi_i' V_i^{-1} M_{\tau_i}\varepsilon_i$ *is* $o_p(1)$ *by sample splitting: conditional on the data in* S_b^c, ξ_i *is a fixed function of* $(\tilde{X}_i, z_i)$, *so the summands are independent with mean zero (because* $\mathbb{E}[M_{\tau_i}\varepsilon_i \mid \tilde{X}_i, z_i] = 0$*) and* $\mathbb{E}[R_b^2 \mid S_b^c] \le C\,\|\xi\|_{P,2}^2 \to_p 0$. *Hence* $R_b = o_p(1)$ *by Chernozhukov et al. (2018, Lemma 6.1).*

Then, we have that:

$$\frac{\sqrt{N_T}}{n_{T,b}}\sum_{i\in S^b}\widehat{\tilde{Z}}' V_i^{-1} M_{\tau_i}\varepsilon_i) = \frac{\sqrt{N_T}}{n_{T,b}}\sum_{i\in S^b}\tilde{Z}_i' V_i^{-1} M_{\tau_i}\varepsilon_i + o_p(1) \tag{70}$$

We define:

$$\Sigma_o = \mathbb{E}[\tilde{Z}_i' V_i^{-1} M_{\tau_i}\varepsilon_i\varepsilon_i' M_{\tau_i}' V_i^{-1}\tilde{Z}_i], \tag{71}$$

Then, by the Lindeberg-Feller Central Limit Theorem we have that:

$$\frac{1}{\sqrt{N_T}}\sum_i \tilde{Z}_i V_i^{-1}\varepsilon_i \xrightarrow{d} \mathcal{N}(0,\Sigma_o). \tag{72}$$

Thus,

$$\sqrt{N_T}(\hat{\beta}_o - \beta_o) \xrightarrow{d} \mathcal{N}(0,\Sigma_o), \tag{73}$$

with $\Sigma_o = \mathbb{E}_P[\tilde{Z}_i' V_i^{-1} M_{\tau_i} X_i]^{-1}\mathbb{E}_P[\tilde{Z}_i' V_i^{-1} M_{\tau_i}\varepsilon_i\varepsilon_i' M_{\tau_i}' V_i^{-1}\tilde{Z}_i](\mathbb{E}_P[\tilde{Z}_i' V_i^{-1} M_{\tau_i} X_i]^{-1})'$.

12 Additional simulation results

This appendix complements Section 8. All results use the design and implementation of that section: $N = 2500$, $T = 2$, first differences, $R = 500$ replications.

First-stage diagnostics. Table 3 reports, for each a, the median over replications of the cluster-robust first-stage F statistics (Wald statistics divided by the number of excluded instruments) of Δz, of a degree-5 polynomial in Δz and of $\widehat{\Delta g}$, the out-of-fold R^2 of the super learner and $\hat{\rho}$. For $\widehat{\Delta g}$, the F statistic treats $\widehat{\Delta g}$ as a fixed regressor and is reported only as a descriptive measure of instrument strength. The F statistic of Δz is below 10 for $a \ge 6$, that of the polynomial for $a \ge 4$, and that of $\widehat{\Delta g}$ is at least 2933 and increases with a.

Why polynomials of Δz do not help. The F statistic of the polynomial is smaller than that of Δz for every a. Given $\Delta z_{it} = d$, the map $(z_{it}, z_{it-1}) \mapsto (-z_{it-1}, -z_{it})$ leaves Δz_{it} and the joint distribution unchanged, because z_{it} and z_{it-1} are i.i.d. and symmetric around zero, and it changes the sign of $|z_{it}| - |z_{it-1}|$. Hence $\mathbb{E}[a(|z_{it}| - |z_{it-1}|) \mid \Delta z_{it}] = 0$, and any function of Δz_{it} is correlated with Δg_o only through the linear term $\Delta z_{it}/a$. The higher-order terms add instruments without adding information. The nonlinear part of Δg_o is a function of (z_{it}, z_{it-1}), which the super learner uses as features.

FD-OLS. The bias of FD-OLS falls with a because the variance of Δx_1 explained by g_o grows with a, while the variance of Δu does not change. It is the same at $a = 1$ and $a = 2$ because $\mathrm{Var}(a|z_{it}| + z_{it}/a) = a^2/3 + 4/(3a^2)$ takes the value $5/3$ at both.

Dispersion and interval length. Tables 4 and 5 report the interquartile range of $\hat{\beta}_1$ and the median length of the 95% confidence intervals. For CF uncorr., the intervals are too short relative to the sampling dispersion. Under normality, an interquartile range of 0.039 (at $a = 10$) corresponds to a standard deviation of $0.039/1.349 \approx 0.029$ and to a 95% interval length of about 0.113, whereas the median reported length is 0.033.

Estimate of ρ_o. The median of $\hat{\rho}$ decreases from 0.834 at $a = 1$ to 0.683 at $a = 10$, against the true value 0.9. The control is $\widehat{\Delta u} = \Delta u + (\Delta g_o - \widehat{\Delta g})$, so the first-stage error acts as measurement error in the control and attenuates $\hat{\rho}$. The scale of g_o, and with it the scale of the estimation error, increases with a, which is consistent with the attenuation increasing with a. Because the score of β is orthogonal, this error does not affect $\hat{\beta}_1$ to first order.

Aggregation over folds and sample splits. Table 6 compares averaging over folds (DML1) with pooling over folds (DML2), and the median with the mean over sample splits. The three versions have median bias within 0.001 of each other and coverage within 0.01 of each other for every a.

Heterogeneous ρ_i. We also consider $\rho_i = 0.9 + 0.5\,\alpha_i$. Since ρ_i depends only on α_i, which is independent of u_i, this design satisfies the condition $\mathbb{E}[\rho_i M u_i \mid \tilde{x}_i, z_i] = 0$. Table 7 reports the results. At $a = 1$, they are the same as with homogeneous ρ_o up to Monte Carlo error: SLCF has median bias -0.000 and coverage 0.94, CF uncorr. has coverage 0.75 and IV no CF has coverage 0.08.

Table 3: First-stage diagnostics

a	F: Δz	F: poly(Δz)	F: $\widehat{\Delta g}$	Out-of-fold R^2	$\hat{\rho}$
1	1840.5	379.2	2937.4	0.702	0.834
2	285.1	64.6	2933.1	0.702	0.837
3	77.3	20.7	5278.5	0.766	0.820
4	28.9	9.3	8808.6	0.823	0.803
5	12.2	5.0	13066.3	0.865	0.780
6	6.3	3.0	18439.7	0.895	0.758
7	3.6	2.2	24430.8	0.917	0.740
8	2.0	1.7	30680.7	0.931	0.720
9	1.5	1.4	38395.6	0.943	0.703
10	0.9	1.2	45198.4	0.951	0.683

Medians over 500 replications, $N = 2500$, $T = 2$, first differences. F: cluster-robust first-stage Wald statistic divided by the number of excluded instruments. Out-of-fold R^2 of the super learner for Δx_1. True $\rho_o = 0.9$.

References

Avila Márquez, M. and J. Krishnakumar (2023, July). On the use of random forests to estimate triangular two-level panel data models with individual fixed effects. In *Proceedings of the 28th International Panel Data Conference*, Amsterdam, Netherlands. Paper presented at the University of Amsterdam. 9

Baltagi, B. H., C. Kao, and L. Liu (2012). On the estimation and testing of fixed effects panel data models with weak instruments. 11

Table 4: Interquartile range of $\hat{\beta}_1$

a	FD-OLS	FD-2SLS	FD-2SLS poly	Oracle IV	SLCF	CF uncorr.	Plug-in 2SLS	IV no CF
1	0.014	0.026	0.027	0.022	0.024	0.037	0.037	0.037
2	0.014	0.056	0.056	0.024	0.025	0.036	0.041	0.032
3	0.012	0.089	0.088	0.017	0.018	0.035	0.042	0.021
4	0.010	0.114	0.105	0.014	0.014	0.042	0.046	0.017
5	0.010	0.151	0.124	0.012	0.012	0.046	0.055	0.014
6	0.009	0.180	0.139	0.010	0.010	0.040	0.051	0.010
7	0.007	0.196	0.146	0.008	0.008	0.042	0.055	0.009
8	0.006	0.246	0.154	0.007	0.007	0.043	0.056	0.008
9	0.006	0.251	0.134	0.006	0.006	0.040	0.050	0.007
10	0.005	0.256	0.136	0.005	0.005	0.039	0.053	0.006

Interquartile range of $\hat{\beta}_1$ over 500 replications, $N = 2500$, $T = 2$, first differences.

Table 5: Median length of 95% confidence intervals

a	FD-OLS	FD-2SLS	FD-2SLS poly	Oracle IV	SLCF	CF uncorr.	Plug-in 2SLS	IV no CF
1	0.040	0.080	0.080	0.069	0.072	0.048	0.146	0.061
2	0.040	0.161	0.159	0.069	0.072	0.048	0.147	0.062
3	0.035	0.243	0.237	0.052	0.053	0.043	0.112	0.048
4	0.031	0.323	0.305	0.040	0.040	0.040	0.090	0.037
5	0.027	0.411	0.367	0.032	0.032	0.045	0.081	0.031
6	0.023	0.503	0.416	0.027	0.027	0.044	0.076	0.026
7	0.021	0.596	0.446	0.023	0.023	0.039	0.072	0.022
8	0.019	0.749	0.459	0.020	0.020	0.038	0.067	0.020
9	0.017	0.825	0.448	0.018	0.018	0.036	0.066	0.018
10	0.015	0.931	0.436	0.016	0.016	0.033	0.056	0.016

Median over 500 replications, $N = 2500$, $T = 2$, first differences.

Table 6: SLCF: aggregation over folds and sample splits

	Median bias			Coverage		
a	DML1, median	DML2, median	DML1, mean	DML1, median	DML2, median	DML1, mean
1	0.000	0.001	0.001	0.94	0.94	0.95
2	−0.002	−0.001	−0.002	0.96	0.96	0.96
3	−0.001	−0.001	−0.001	0.96	0.96	0.96
4	−0.002	−0.002	−0.002	0.96	0.96	0.96
5	0.000	0.000	0.000	0.94	0.93	0.94
6	0.001	0.001	0.001	0.94	0.93	0.93
7	−0.000	−0.000	0.000	0.96	0.95	0.95
8	−0.000	−0.000	−0.000	0.97	0.96	0.97
9	0.000	0.000	0.000	0.93	0.94	0.94
10	−0.000	−0.000	−0.000	0.94	0.95	0.95

DML1: fold-by-fold estimates averaged over folds. DML2: one estimate pooled over folds. Median or mean over the 3 sample splits. The first column is the SLCF estimator of Table 1. 500 replications, $N = 2500$, $T = 2$, first differences.

Table 7: Heterogeneous $\rho_i = 0.9 + 0.5\,\alpha_i$: median bias and coverage

a	FD-OLS	FD-2SLS	FD-2SLS poly	Oracle IV	SLCF	CF uncorr.	Plug-in 2SLS	IV no CF
Median bias								
1	0.382	0.002	0.003	0.001	−0.000	−0.016	0.017	0.072
Coverage								
1	0.00	0.93	0.93	0.93	0.94	0.75	0.95	0.08

500 replications, $N = 2500$, $T = 2$, first differences. Median $\hat{\rho} = 0.833$; out-of-fold $R^2 = 0.702$.

Baltagi, B. H., D. Li, et al. (2002). Series estimation of partially linear panel data models with fixed effects. *Annals of economics and finance 3*(1), 103–116. 4, 10

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429. 17

Burke, P. J. and A. Abayasekara (2018). The price elasticity of electricity demand in the United States: A three-dimensional analysis. *The Energy Journal 39*(2), 123–146. 21, 24

Cai, Z., Y. Fang, and H. Li (2012). Weak instrumental variables models for longitudinal data. *Econometric Reviews 31*(4), 361–389. 2, 4

Chen, J., D. L. Chen, and G. Lewis (2020). Mostly harmless machine learning: Learning optimal instruments in linear IV models. 17

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014). Local identification of nonparametric and semiparametric models. *Econometrica 82*(2), 785–809. 27

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68. 4, 7, 8, 13, 15, 18

Colin Cameron, A. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources 50*(2), 317–372. 15

Ding, P. (2021). The frisch–waugh–lovell theorem for standard errors. *Statistics & Probability Letters 168*, 108945. 15

Emmenegger, C. and P. Bühlmann (2023). Plug-in machine learning for partially linear mixed-effects models with repeated measurements. *Scandinavian Journal of Statistics 50*(4), 1553–1567. 4, 12, 26, 27, 28

Guo, Z. and D. S. Small (2016). Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research 17*(100), 1–35. 6

Guo, Z., M. Zheng, and P. Bühlmann (2022). Robustness against weak or invalid instruments: Exploring nonlinear treatment models with machine learning. *arXiv preprint arXiv:2203.12808*. 8

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series 1988*(1), i–50. 6

Keane, M. P. and T. Neal (2024). A practical guide to weak instruments. *Annual Review of Economics 16*. 2

Klosin, S. and M. Vilgalys (2022). Estimating continuous treatment effects in panel data using machine learning with a climate application. *arXiv preprint arXiv:2207.08789*. 3

Li, Q. (1996). On the root-n-consistent semiparametric estimation of partially linear models. *Economics Letters 51*(3), 277–285. 4

Mikusheva, A. and L. Sun (2022). Inference with many weak instruments. *The Review of Economic Studies 89*(5), 2663–2686. 11

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*(4), 809–837. 17

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382. 12

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier. 27, 28

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954. 4, 7, 8

Rodriguez-Poo, J. M. and A. Soberon (2017). Nonparametric and semiparametric panel data models: Recent developments. *Journal of Economic Surveys 31*(4), 923–960. 4

Semenova, V., M. Goldman, V. Chernozhukov, and M. Taddy (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics 14*(2), 471–510. 3

U.S. Energy Information Administration (2026). Electricity: Detailed state data. historical EIA-861 sales, revenue, customers and price by state and sector. https://www.eia.gov/electricity/data/state/. 24

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical applications in genetics and molecular biology 6*(1). 3, 4

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press. 6

Young, E. H. and R. D. Shah (2024). Sandwich boosting for accurate estimation in partially linear models for grouped data. *Journal of the Royal Statistical Society Series B: Statistical Methodology 86*(5), 1286–1311. 4

Zabrocki, L., A. Alari, and T. Benmarhnia (2022). Improving the design stage of air pollution studies based on wind patterns. *Scientific Reports 12*(1), 7917. 2