Chapter 10
Nonparametric Correlated Random-Effects Models

Daniel J. Henderson, Emma Kate Henry and Alexandra Soberon

Abstract This chapter develops methods for estimation and inference in nonparametric panel data models with correlated random-effects. Using the Mundlak specification to control for unobserved heterogeneity, this nonparametric estimation procedure can identify both the nonparametric function and a finite-dimensional parameter associated with (potentially) observed time-invariant regressors. We develop the necessary asymptotic theory for our proposed estimator. To assess the validity of our method in practice, we propose a consistent specification test for whether the model controls for the correlation between the unobserved individual effects and the regressors. Monte Carlo simulations support the asymptotic developments. We illustrate the practical utility of our approach via an empirical application.

10.1 Introduction

The analysis of panel data has a long history in econometrics/statistics (Nerlove, 2005, Chapter 2). Fixed-effects and random-effects models have served as workhorses for panel data analysis (Baltagi, 2021). While random-effects models are easy to employ, fixed-effects models are typically used in economics because the assumption of uncorrelatedness between unobserved factors and explanatory variables often proves unrealistic in practice. Correlated random-effects (CRE) models, pioneered by

Daniel J. Henderson ✉
Department of Economics, Finance and Legal Studies, University of Alabama, Tuscaloosa, AL 35487-0224, USA, e-mail: daniel.henderson@ua.edu

Emma Kate Henry
Department of Economics, Finance and Legal Studies, University of Alabama, Tuscaloosa, AL 35487-0224, USA, e-mail: ekhenry@crimson.ua.edu

Alexandra Soberon
Department of Economics, University of Cantabria, Santander, Cantabria 39005, Spain, e-mail: alexandra.soberon@unican.es

Mundlak (1978) and further developed by Chamberlain (1982), offer a middle ground that combines the simplicity of random-effects with the robustness of fixed-effects approaches (see Chapter 14).

CRE models maintain many advantages of random-effects estimation while accommodating potential correlation between individual-specific effects and explanatory variables (Wooldridge, 2019). Although the traditional CRE framework has proven fruitful, it still relies on potentially restrictive parametric assumptions about the conditional mean function that relate unobserved effects to observed covariates. This chapter proposes semiparametric methods that relax these assumptions and can accommodate non-linear relationships and interactions that standard parametric specifications might miss.

The field of nonparametric and semiparametric analysis in panel data modelling has expanded considerably. Estimators for the random (Henderson & Ullah, 2005) and fixed-effects (Henderson, Carroll & Li, 2008) settings exist and comprehensive reviews are available in the literature.[1] However, the incorporation of fixed-effects in these models presents significant computational challenges. Various estimation approaches have been proposed, ranging from iterative methodologies to profile least-squares to marginal integration techniques. These methods necessitate the estimation of fixed-effects parameters or rely on nontestable assumptions. Our semiparametric approach preserves the core insights of the CRE models and leads to straightforward estimation.[2] We both develop and demonstrate how researchers can implement these methods in practice while maintaining the interpretability and efficiency that have made CRE models valuable in applied work.

A key contribution of this chapter is the development of a specification test. This test provides researchers with a formal tool to evaluate whether the CRE specification controls for correlation between the composite error term and the explanatory variables. Our test compares the performance of our estimator, which is consistent under the null and alternative hypothesis versus the commonly used local-constant least-squares estimator, which is only consistent under the null hypothesis.

The remainder of this chapter is organized as follows. Section 10.2 introduces our semiparametric estimator and establishes its asymptotic properties. Section 10.3 develops the specification test and examines its properties under the null and alternative hypotheses. In Section 10.4, we conduct Monte Carlo simulations to investigate the finite sample performance of our estimator and test statistic. Section 10.5 demonstrates the practical utility of our methods through an empirical application that examines the relationship between R&D expenditure and industry-level regulations. Finally, Section 10.6 concludes.

[1] See the surveys of Ai and Li (2008), Henderson and Parmeter (2015), Sun, Zhang and Li (2015), Parmeter and Racine (2019), Rodriguez-Poo and Soberon (2017) and/or Su and Ullah (2011) and the references within.

[2] Bester and Hansen (2009) are able to identify average marginal effects in CRE models using sufficient statistics and index restrictions.

10.2 Estimation

Consider the standard nonparametric one-way error component model where the outcome variable, y_{it}, is related to the regressors, x_{it}, through the following regression

$$y_{it} = m(x_{it}) + \mu_i + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T \tag{10.1}$$

where $x_{it} \in \mathbb{R}^d$ is a vector of explanatory variables, μ_i is the unobserved heterogeneity that may be correlated with x_{it}, $m(\cdot)$ is an unknown smooth function to be estimated, and u_{it} is the idiosyncratic error term with $u_{it} \sim IID(0, \sigma_u^2)$. The exogeneity condition $E(u_{it}|x_{i1}, \ldots, x_{iT}, \mu_i) = 0$ for each i is assumed throughout the paper.

To relax the strict exogeneity condition and allow for correlation between the regressors of the model and the unobserved heterogeneity, we follow Mundlak (1978) and model the correlated random-effects as a linear function of all the explanatory variables averaged across time or time-invariant regressors as

$$\mu_i = \bar{x}_i^\top \psi + z_i^\top \gamma + \nu_i, \tag{10.2}$$

where $\bar{x}_i = T^{-1} \sum_{t=1}^T x_{it}$ is the temporal average of x_{it}, $z_i \in R^q$ is a vector of time-invariant regressors outside the model, and ν_i is an error term that is assumed to be independent of (x_{it}, z_i, u_{it}) and $\nu_i \sim IID(0, \sigma_\nu^2)$.

Plugging (10.2) in (10.1) gives the partially linear model:

$$y_{it} = m(x_{it}) + \omega_i^\top \theta + \varepsilon_{it}, \tag{10.3}$$

where $\omega_i^\top = (\bar{x}_i^\top, z_i^\top)$, $\theta = (\psi^\top, \gamma^\top)^\top$, and $\varepsilon_{it} = v_i + u_{it}$.

Let $\mathcal{J}_m(\mathbf{x}) = \frac{\partial m(\cdot)}{\partial \mathbf{x}}$ be the $(d \times 1)$ vector of first-order derivatives of $m(\cdot)$. Using a first-order Taylor expansion of $m(\cdot)$, the objective function (with known θ) is

$$\arg\min_{a,b} \sum_{i=1}^N \sum_{t=1}^T [y_{it} - a - (x_{it} - \mathbf{x})^\top b - \omega_i^\top \theta]^2 K_h(x_{it} - \mathbf{x}),$$

where $K_h(\cdot)$ is a kernel function. For multivariate x_{it}, we use a product kernel $K_h(v) = \prod_{l=1}^d k_h(v_l)$ with $v = (v_1, \ldots, v_d)^\top$ and $k_h(x_{it} - \mathbf{x}) = h^{-1} k((x_{it} - \mathbf{x})/h)$, where h is the smoothing (bandwidth) parameter. Let $\widehat{a}(\mathbf{x})$ and $\widehat{b}(\mathbf{x})$ be the resulting nonparametric estimators of $m(\mathbf{x})$ and $\mathcal{J}_m(\mathbf{x})$, respectively.

In this paper, we propose to use a local-constant approach to obtain the estimator of the unknown function $m(\cdot)$. Following Robinson (1988), we take the conditional expectation over x_{it} on both sides of (10.3) to obtain

$$E(y_{it}|x_{it}) = m(x_{it}) + E(\omega_i^\top|x_{it})\theta + E(\varepsilon_{it}|x_{it}), \tag{10.4}$$

and using the fact that $E(\varepsilon_{it}|x_{it}) = 0$ and subtracting (10.4) from (10.3), we obtain

$$y_{it} - E(y_{it}|x_{it}) = \{\omega_i - E(\omega_i|x_{it})\}^\top \theta + \varepsilon_{it}. \tag{10.5}$$

In order to avoid the random denominator problem which is quite common in the nonparametric kernel estimation (see Powell, Stock & Stoker, 1989), we premultiply both sides of (10.5) by the density function of x_{it}, i.e., $f(x_{it})$, to obtain

$$\widetilde{y}_{it} = \widetilde{\omega}_{it}\theta + \widetilde{\varepsilon}_{it} \tag{10.6}$$

where $\widetilde{y}_{it} = \{y_{it} - E(y_{it}|x_{it})\}f(x_{it})$, $\widetilde{\omega}_{it} = \{\omega_i - E(\omega_i|x_{it})\}f(x_{it})$, and $\widetilde{\varepsilon}_{it} = \{\varepsilon_{it} - E(\varepsilon_{it}|x_{it})\}f(x_{it})$.

Unfortunately, the resulting least-squares estimator from (10.6) is infeasible as $E(y_{it}|x_{it})f(x_{it})$, $E(\omega_i|x_{it})f(x_{it})$, and $f(x_{it})$ are unknown functions, but they can be consistently estimated using nonparametric techniques. We propose to estimate these via $\widehat{E}(y_{it}|x_{it})\widehat{f}(x_{it}) = (NT)^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T} y_{js}K_h(x_{it},x_{js})$, $\widehat{E}(\omega_i|x_{it})\widehat{f}(x_{it}) = (NT)^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}\omega_i K_h(x_{it},x_{js})$, and $\widehat{f}(x_{it}) = (NT)^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(x_{it},x_{js})$, respectively, where $K_h(x_{it},x_{js}) = h^{-d}K((x_{it}-x_{js})/h)$.

This leads to the feasible estimator of θ in (10.5) as

$$\widehat{\theta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\omega}_{it}\widetilde{\omega}_{it}^{\top}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\omega}_{it}\widetilde{y}_{it}. \tag{10.7}$$

To obtain $m(\cdot)$ in (10.1), we plug (10.7) into (10.1) to obtain

$$\ddot{y}_{it} = m(x_{it}) + e_{it}, \tag{10.8}$$

where $\ddot{y}_{it} = y_{it} - \omega_i^{\top}\widehat{\theta}$ is the new left-hand-side variable and $e_{it} = \varepsilon_{it} - \omega_i^{\top}(\widehat{\theta}-\theta)$ is the new error term.

Using a local-constant approach to estimate $m(\cdot)$ in (10.8), we propose to minimize the following objective function

$$\arg\min_{a_0}\sum_{i=1}^{N}\sum_{t=1}^{T}(\ddot{y}_{it}-a_0)^2K_h(x_{it}-\mathbf{x}),$$

and the resulting nonparametric estimator of $m(\cdot)$ is

$$\widehat{m}(\mathbf{x};h) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{y}_{it}K_h(x_{it},\mathbf{x})}{\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(x_{it},\mathbf{x})}. \tag{10.9}$$

To derive the large sample properties of the proposed estimators in (10.7) and (10.9), we first provide a definition and state some assumptions by extending what is assumed in Li and Stengos (1996) and Soberon, Rodriguez-Poo and Robinson (2021). We shall use $\mathcal{G}_{\upsilon}^{r}$ to denote the class of smooth functions such that if $g \in \mathcal{G}_{\upsilon}^{r}$, then g is bounded and υ times differentiable; g and its partial derivative functions (up to order υ) all satisfy some Lipschitz-type conditions such as $|g(\mathbf{x}) - g(\mathbf{x}')| \leq \mathcal{H}_g(\mathbf{x})\|\mathbf{x}'-\mathbf{x}\|$, where $\mathcal{H}_g(\mathbf{x})$ is a continuous function having rth moment, and where $\|\cdot\|$ denotes

the Euclidean norm, i.e., $\|\mathbf{x}\| = \sqrt{\sum_{l=1}^{d} \mathbf{x}_j^2}$ (see Definition 2 of Robinson, 1988, pp. 939). In addition, r controls the moment properties of the remainder term.

Assumption **A1:** For $t = 1, \ldots, T$, $(x_{it}, z_i, v_i, u_{it})$ are *i.i.d.* in the subscript i and strict stationarity in t for fixed i. x_{it} admits a probability density function $f \in \mathcal{G}_{\upsilon-1}^{\infty}$ (i.e., f is bounded), $m(\cdot) \in \mathcal{G}_{\upsilon}^{4}$, $E(z_i|x_{it}) \in \mathcal{G}_{\upsilon}^{4}$, and $E(\bar{x}_{iA}|x_{it}) \in \mathcal{G}_{\upsilon}^{4}$, where $\upsilon \geq 2$ is an integer. □

Assumption **A2:** $E(u_{it}|x_{it}) = 0$, $E(v_i|x_{it}) = 0$. For $\varepsilon_{it} = v_i + u_{it}$, $E(\varepsilon_{it}^2|\mathbf{x}) = \sigma_{\varepsilon}^2 \equiv \sigma_v^2 + \sigma_u^2$. $(x_{it}, z_i, v_i, u_{it})$ have finite fourth moments. □

Assumption **A3:** $K(\cdot)$ is a product kernel, the univariate kernel $k(\cdot)$ is a bounded υth order kernel, and $k(\nu) = O(1/(1+|\nu|)^{\upsilon+1})$. □

Assumption **A4:** As $N \to \infty$ for T fixed, $h \to 0$, $Nh^{2d} \to \infty$ and $Nh^{4\upsilon} \to 0$. □

Assumption **A5:** $E[(\omega_i - E(\omega_{it}|x_{it}))(\omega_i - E(\omega_{it}|x_{it}))^{\top} f^2(x_{it})]$ is a $(d+q) \times (d+1)$ non-singular matrix function. □

Assumption **A6:** For some $\delta > 0$, $E[|v_i|^{(4+\delta)}|x_{it}] < \infty$ and $E[|u_{it}|^{(4+\delta)}|x_{it}] < \infty$.□

These assumptions are fairly standard in the semiparametric literature, but some remarks are required. Assumption (A1) is stronger than what it is in Li and Stengos (1996). The density function is assumed to be bounded and at least first-order partially differentiable with a Lipschitz-continuous remainder as this is required for estimating the nonlinear portion. This stronger condition is not necessary for the asymptotic properties of $\widehat{\theta}$, only for $\widehat{m}(\cdot)$. The strict stationarity condition is imposed for the simplicity of the mathematical proofs, but can be relaxed. Assumption (A4) implies that $(2\upsilon > d)$ or $(2\upsilon \geq d+1)$ (because υ is an integer), which in turn is equivalent to $(\upsilon \geq (d+1)/2)$ as in Robinson (1988) or Li and Stengos (1996). This assumption is stronger than what is assumed in Li (1996), but it implies that a standard second-order kernel ($\upsilon = 2$) can be used if $(d \leq 3)$ and the proofs of the asymptotic properties of the proposed estimators are considerably simpler. Assumption (A6) is required to obtain the asymptotic distribution of the nonparametric estimator, $\widehat{m}(\mathbf{x}; h)$.

We are now ready to proceed with the asymptotic results for our parametric and nonparametric components, respectively:

Theorem 10.1 *Under assumptions (A1)-(A5), as $N \to \infty$ and T is fixed*

$$\sqrt{NT}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \Phi^{-1}\Psi\Phi^{-1}\right)$$

where $\eta_{it} = \omega_i - E(\omega_i|x_{it})$, $\Phi = E[\eta_{it}\eta_{it}^{\top} f^2(x_{it})]$ is positive definite, and $\Psi = \sigma_{\varepsilon}^2 E[\eta_{it}\eta_{it}^{\top} f^4(x_{it})]$. Moreover, $\widehat{\Phi} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\eta}_{it}\widehat{\eta}_{it}^{\top}\widehat{f}^2(x_{it})$ and $\widehat{\Psi} = (1/NT^2)\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\widehat{\eta}_{it}\widehat{\eta}_{is}^{\top}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{is}\widehat{f}(x_{it})\widehat{f}(x_{is})$ are the consistent estimators of Ψ and Φ^{-1}, respectively, where $\widehat{\eta}_{it} = \omega_i - \widehat{E}(\omega_i|x_{it})$.

The proof of Theorem 10.1 is a straightforward extension of Theorem 1 in Li and Stengos (1996). The detailed proof is available upon request.

We now turn to the asymptotic properties of the nonparametric estimator proposed for $m(\cdot)$ in (10.9). Denote $\mu_2 = \int v^2 K(v)dv$ and $\mathcal{R} = \int k^2(v)dv$, where μ_2 and $\mathcal{R}$ are scalars different from zero.

Theorem 10.2 *Under Assumptions (A1)-(A6), as $N \to \infty$ and T is fixed*

$$\sqrt{NTh^d}\left\{\widehat{m}(\mathbf{x};h) - m(\mathbf{x}) + \frac{h^2\mu_2^d}{2} tr\{\mathcal{H}_m(\mathbf{x})\} + o_p(h^2)\right\} \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2 \mathcal{R}^d}{f(\mathbf{x})}\right),$$

where $\mathcal{H}_m(\mathbf{x}) = \partial m(\mathbf{x})/\partial\mathbf{x}\partial\mathbf{x}^\top$ is the Hessian matrix of $m(\cdot)$.

The proof of Theorem 10.2 is obtained via a similar scheme as in Fan and Gijbels (1995) or Soberon et al. (2021), among others. The detailed proof is available upon request.

10.3 Inference

In this section, we discuss how to test for whether the estimation procedure proposed in this chapter accounts for correlation between the individual effects v_i and the regressor vector x_{it}. The null and alternative hypothesis can be written as

$$H_0 : Pr[E(v_i|x_{it}) = 0] = 1 \text{ for all } i$$
$$H_1 : Pr[E(v_i|x_{it}) \neq 0] > 0 \text{ for some } i.$$

The commonly used local-constant estimator (i.e., Nadaraya-Watson estimator) is consistent when $E(v_i|x_{it}) = 0$, but is inconsistent when $E(v_i|x_{it}) \neq 0$. Our nonparametric correlated random-effects estimator is consistent in both cases.

Motivated by Sun, Carroll and Li (2009) in a different context, we propose a test statistic based on

$$I = \int [\widehat{m}(\mathbf{x};h) - \widetilde{m}(\mathbf{x};h)]^2 \, d\mathbf{x}, \tag{10.10}$$

where $\widehat{m}(\mathbf{x})$ is our nonparametric correlated random-effects local-constant estimator and $\widetilde{m}(\mathbf{x})$ is the Nadaraya-Watson estimator of the form

$$\widetilde{m}(\mathbf{x};h) = \frac{\sum_{i=1}^N \sum_{t=1}^T y_{it} K_h(x_{it},\mathbf{x})}{\sum_{i=1}^N \sum_{t=1}^T K_h(x_{it},\mathbf{x})}.$$

Using $\sum_{i=1}^N \sum_{t=1}^T K_h(x_{it},\mathbf{x})$ to remove the random denominators and by defining $\widehat{\epsilon}_{it} = y_{it} - \omega_i^\top \widehat{\theta} - \widetilde{m}(\mathbf{x};h)$, the test statistic becomes

$$I_n = \frac{1}{N^2T^2h^d}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\int \widehat{\epsilon}_{it}\widehat{\epsilon}_{js}K_h(x_{it},\mathbf{x})K_h(x_{js},\mathbf{x})d\mathbf{x}, \tag{10.11}$$

because by rearranging terms, we can write

$$\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(x_{it},\mathbf{x})\left[\widehat{m}(\mathbf{x};h)-\widetilde{m}(\mathbf{x};h)\right] = \sum_{i=1}^{N}\sum_{t=1}^{T}K_h(x_{it},\mathbf{x})\widehat{\epsilon}_{it}.$$

To simplify the integration in (10.11), we use the fact that the typical element of the integration term can be written as a twofold convolution kernel that acts as a weighting function to select the observations such that only those (x_{it},x_{js}) close to each other are used. Finally, to avoid the asymptotic bias term of this type of double summation test, we remove the $i=j$ terms of the kernel matrix of I_n, and obtain the following test statistic

$$\widehat{I}_n = \frac{1}{N^2T^2h^d}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\widehat{\epsilon}_{it}\widehat{\epsilon}_{js}K_h(x_{it},x_{js}). \tag{10.12}$$

To derive the asymptotic properties of the test statistic proposed in (10.12), the following assumptions are required:

***Assumption* A7:** For all $t\neq s$, (x_{it},x_{is}) has a joint probability density function, $f_{ts}(\mathbf{x}_1,\mathbf{x}_2)$, that is continuously differentiable and $\sup_{t\neq s}\int f_{ts}(\mathbf{x},\mathbf{x})d\mathbf{x} < M < \infty$. $E(x_{l,it}^{\upsilon}|x_{is})$, $f(\mathbf{x})$, $f_{ts}(\mathbf{x}_1,\mathbf{x}_2)$, and their first-order partial derivatives are all uniformly bounded for $l=1,\ldots,d$, $\upsilon=1,\ldots,4$, and for all i and $t\neq s$. □

***Assumption* A8:** As $N\to\infty$ for T fixed, $h\to 0$, and $Nh^d\to\infty$. □

The asymptotic distribution under the null and the power of our test are given by the following theorems. They are obtained following similar reasoning as to that in Zheng (1996) and/or Henderson and Soberon (2024), among others. The detailed proofs are available upon request.

Theorem 10.3 *Under Assumptions (A1)-(A8) and assuming $f(x)>0$ for each x in the support of x_{it}, under H_0, as $N\to\infty$ and T is fixed,*

$$J_n = NTh^{d/2}\frac{\widehat{I}_n}{\sqrt{\widehat{\Sigma}}} \xrightarrow{d} N(0,1)$$

where $\widehat{\Sigma} = \frac{1}{N^2T^2h^d}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\widehat{\epsilon}_{it}^2\widehat{\epsilon}_{js}^2K_h^2(x_{it},x_{js})$ is a consistent estimator of the asymptotic variance of $NTh^{d/2}\widehat{I}_{NT}$ where

$$\Sigma = \frac{2\sigma_{\varepsilon}^2\mathcal{R}^d(K)}{T^2}E[f(x_{it})].$$

Theorem 10.4 *Under Assumptions (A1)-(A8) and assuming $f(x) > 0$ for each x in the support of x_{it}, under H_1, we have $Pr(J_n \geq c) \to 1$ as $N \to \infty$, where (c) is any positive constant.*

In practice, the asymptotic distribution of the test statistic is not useful for finite samples. As is standard in nonparametric kernel based tests, we resort to a bootstrap procedure. The steps for the wild bootstrap are as follows:

1. Compute the test statistic $\widehat{J}_n$ for the original sample of data $\{ y_{it}, x_{it}, \omega_i \}$.
2. For each observation i, for each time period t, draw a wild residual bootstrap v_{it}^* and construct the bootstrapped left-hand-side variable as $y_{it}^* = \widehat{m}(x_{it}) + \omega_i^\top \widehat{\theta} + v_{it}^*$ and call $\{ y_{it}^*, x_{it}, \omega_i \}$ the bootstrap sample.[3]
3. Calculate $\widehat{T}_n^*$ where $\widehat{T}_n^*$ is calculated the same way as $\widehat{T}_n$ except that y_{it} is replaced by y_{it}^*.
4. Repeat steps 2 and 3 a large number (B) of times and then construct the sampling distribution of the bootstrapped test statistics. Reject the null if the estimated statistic $\widehat{T}_n$ is greater than the upper $\alpha-$percentile of the bootstrapped test statistics.

10.4 Simulations

We investigate the finite sample performance of our proposed estimators and test via Monte Carlo simulations. We begin with Equation (10.1) and consider various forms for $m(\cdot)$ and μ_i. Our study considers two distinct functional forms for $m(\cdot)$:

$$m(x_{it}) = x_{it}^\top \beta$$
$$m(x_{it}) = \sin(x_{it}),$$

where x_{it} is generated as a standard random normal variable. We examine three specifications for the unobserved heterogeneity, μ_i:

$$\mu_i = \bar{x}_i^\top \psi + v_i$$
$$\mu_i = z_i^\top \gamma + v_i$$
$$\mu_i = z_i^\top \gamma + \bar{x}_i^\top \psi + v_i,$$

and when we study the size of our test, we set $\mu_i = v_i$.

Both error terms (v_i and ε_{it}) are generated as i.i.d. random variables, normally distributed with zero mean and variance $1/2$. Our simulation design varies both the cross-sectional dimension $N \in \{100, 200, 400\}$ and the time dimension $T \in \{3, 5\}$. For each configuration, we conduct 999 simulations, and use Gaussian kernel functions with bandwidths selected according to Silverman (1986). For our testing procedure, we use 399 bootstrap replications for each simulation.

[3] $v_{it}^* = \widehat{\epsilon}_{it} * \varphi_i^b$, where $\varphi_i^b \overset{i.i.d.}{\sim} N(0,1)$. It is easy to see that $E[\widehat{\epsilon}_{it}\varphi_i^b] = 0$, $Cov[\widehat{\epsilon}_{it}\varphi_i^b, \widehat{\epsilon}_{js}\varphi_j^b] = Cov[\widehat{\epsilon}_{it}, \widehat{\epsilon}_{js}]$, for all $i, j = 1, \ldots, N, t, s = 1, \ldots, T$.

10.4.1 Estimation

To assess estimation accuracy, we compute the average mean squared error (AMSE) for the nonparametric function $m(\cdot)$

$$AMSE[\widehat{m}(\cdot)] = \frac{1}{999}\sum_{j=1}^{999}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{m}_j(x_{it}) - m(x_{it}))^2\right],$$

and for specifications involving γ, we also analyze its behavior via $MSE(\widehat{\gamma}) = \frac{1}{999}\sum_{j=1}^{999}(\widehat{\gamma}_j - \gamma)^2$.

The results of the simulations are summarized in Figures 10.1 and 10.2.[4] The results reveal several key patterns. For the linear specification $m(\cdot) = x_{it}^\top\beta$, as expected, the least-squares estimator consistently outperforms the nonparametric approach across all specifications, exhibiting lower AMSE values and reduced variance. This advantage is particularly pronounced for smaller sample sizes. However, the performance gap narrows as the sample size increases, with $N = 400$ showing markedly improved precision for both methods.

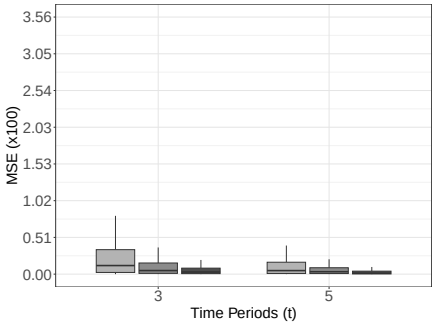
The nonparametric estimation of the nonlinear function $m(\cdot) = \sin(x_{it})$ demonstrates consistent patterns across all three μ_i specifications. The AMSE values decrease with larger sample sizes, and estimation precision improves when $T = 5$ compared to $T = 3$. The consistency properties shown in the theorems appear to be validated here.

The estimation results appear relatively robust across the different specifications of unobserved heterogeneity, though the more complex specification $\mu_i = z_i^\top\gamma + \bar{x}_i^\top\psi + v_i$ shows slightly higher MSE values, particularly in smaller samples. This suggests that the additional complexity of controlling for both time-invariant regressors and time averages of x_{it} introduces modest efficiency costs in finite samples.

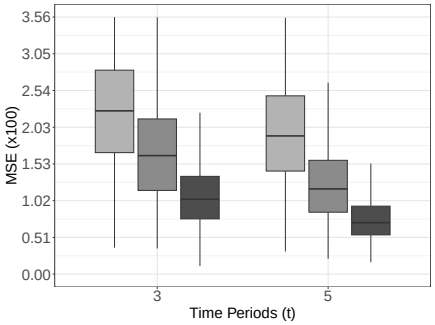
The results for the estimation of the parameter γ can be found in Figures 10.3 and 10.4. The estimates here come from the same simulations runs in panels (c-f) in Figure 10.1. What is obvious from the comparison between the two figures is that in Figure 10.3, we can see that the parametric components from the semiparametric procedure are estimated as precisely as those from the least-squares estimator. As expected, the results for Figure 10.4 for the semiparametric estimator look nearly identical to those from the previous figure. Again, the least-squares estimates are not given in Figure 10.4 as the estimates of $m(\cdot)$ are inconsistent.

[4] We show the results for both the least-squares and semiparametric estimates for the linear specification, but only the semiparametric results for the nonlinear specification as our least-squares estimators are inconsistent in this setting.

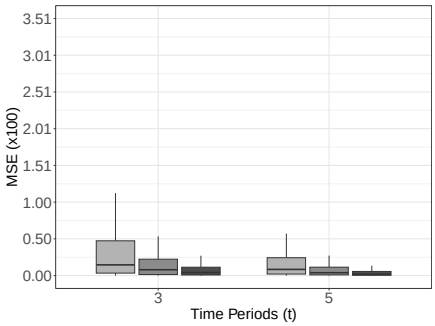
Fig. 10.1: AMSE for the linear specification of $m(x) = x_{it}\beta$ for least-squares and semiparametric estimators: $N \in \{100, 200, 400\}$ and $T \in \{3, 5\}$



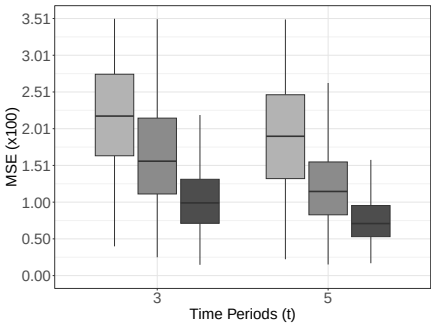
(a) $\mu_i = \bar{x}_i^\top \psi + v_i$: Least-Squares



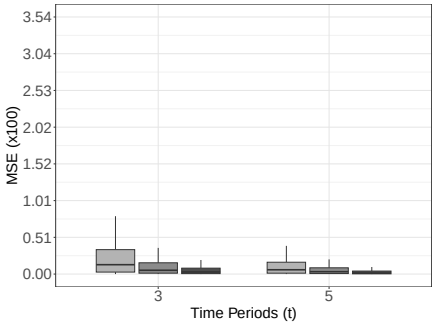
(b) $\mu_i = \bar{x}_i^\top \psi + v_i$: Semiparametric



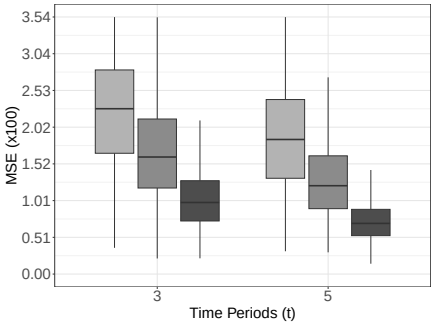
(c) $\mu_i = Z_i \gamma + v_i$: Least-Squares



(d) $\mu_i = z_i^\top \gamma + v_i$: Semiparametric

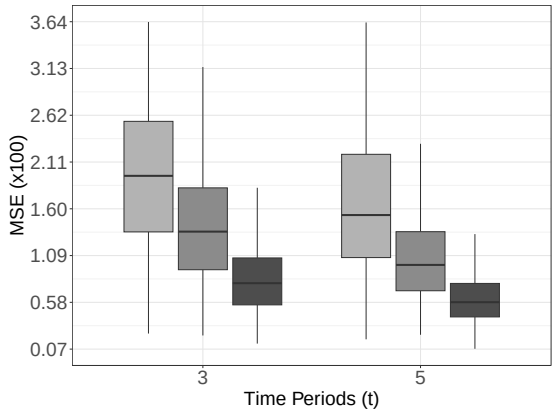


(e) $\mu_i = z_i^\top \gamma + \bar{x}_i^\top \psi + v_i$: Least-Squares

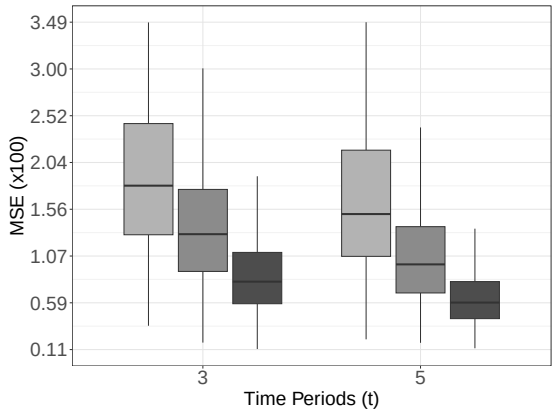


(f) $\mu_i = z_i^\top \gamma + \bar{x}_i^\top \psi + v_i$: Semiparametric

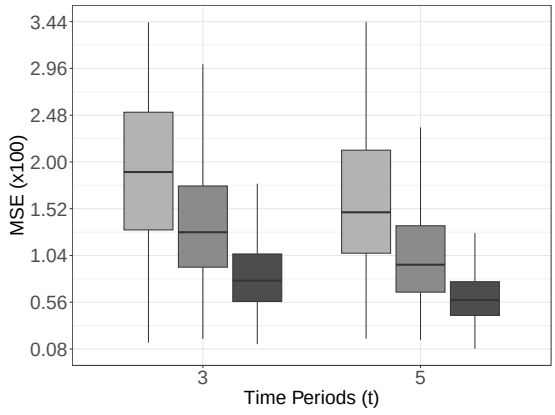
Fig. 10.2: AMSE for the nonlinear specification of $m(x) = \sin(x_{it})$ for the semiparametric estimator: $N \in \{100, 200, 400\}$ and $T \in \{3, 5\}$



(a) $\mu_i = \bar{x}_i^\top \psi + v_i$

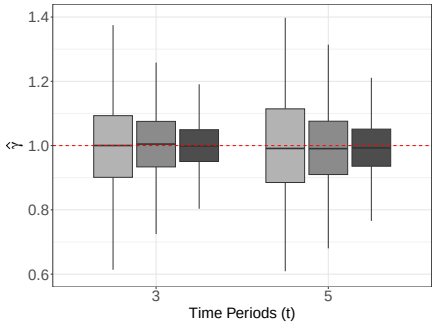


(b) $\mu_i = z_i^\top \gamma + v_i$

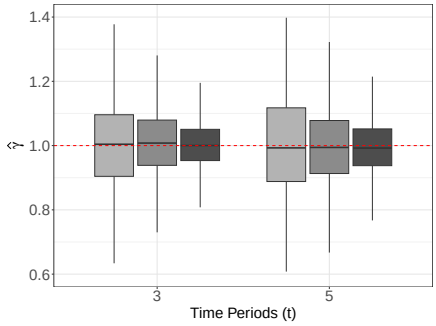


(c) $\mu_i = z_i^\top \gamma + \bar{x}_i^\top \psi + v_i$

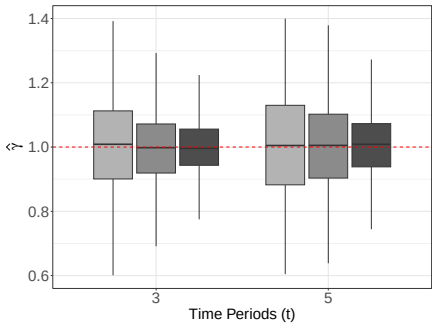
Fig. 10.3: Estimates of $\widehat{\gamma}$ for the linear specification of $m(x) = x_{it}^{\top}\beta$ for least-squares and semiparametric estimators: $N \in \{100, 200, 400\}$ and $T \in \{3, 5\}$



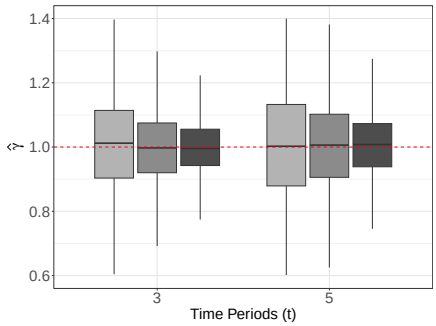
(a) $\mu_i = Z_i\gamma + v_i$: Least-Squares



(b) $\mu_i = z_i^{\top}\gamma + v_i$: Semiparametric



(c) $\mu_i = z_i^{\top}\gamma + \bar{x}_i^{\top}\psi + v_i$: Least-Squares



(d) $\mu_i = z_i^{\top}\gamma + \bar{x}_i^{\top}\psi + v_i$: Semiparametric

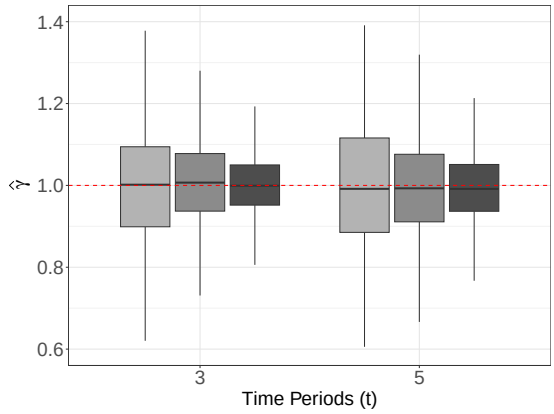
10.4.2 Inference

To study the performance of our testing procedure, we first study its finite sample size properties. The results are similar for the linear and nonlinear functions of $m(\cdot)$ and so we only consider the linear function. Everything remains the same as in the previous subsection except to study size, we set $\mu_i = v_i$. 399 bootstrap replications for each of our 999 simulation.

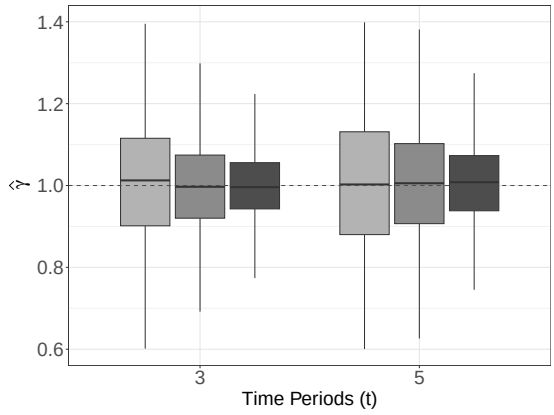
The results for size can be found in Figure 10.5. We plot the actual level versus the nominal level. Each of the actual levels for the tests appear to be near the nominal levels, but note that the performance improves with N, as expected. These simulations appear to support the theory developed in Section 10.3.

The results for the power of our test can be found in Figure 10.6. We plot the actual level versus the nominal level. The power of the test improves with N, as expected, for each specification of μ_i. That being said, we see that the probability of correctly rejecting the null is higher when controlling for $\bar{x}_i$ as compared to z_i, and is even

Fig. 10.4: Estimates of $\widehat{\gamma}$ for the nonlinear specification of $m(x) = \sin(x_{it})$ for the semiparametric estimator: $N \in \{100, 200, 400\}$ and $T \in \{3, 5\}$



(a) $\mu_i = z_i^\top \gamma + v_i$



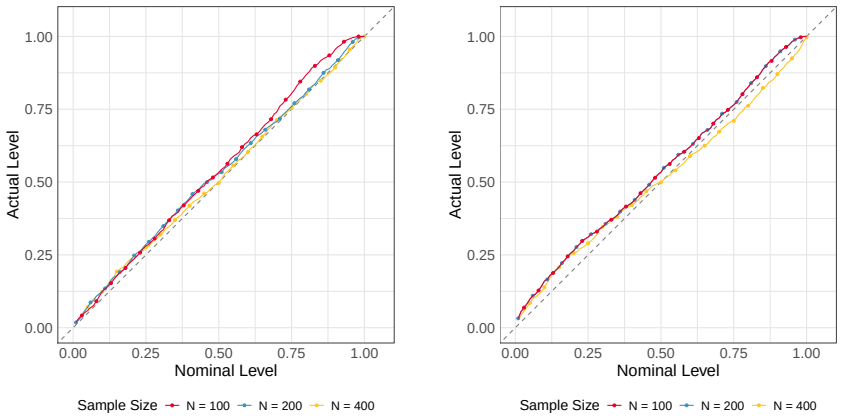
(b) $\mu_i = z_i^\top \gamma + \bar{x}_i^\top \psi + v_i$

higher when controlling for both. These simulations appear to support the theory developed in Section 10.3.

10.5 Empirical Illustration

To demonstrate the empirical relevance of our methodological innovations, we analyze the relationship between firms' research and development expenditures, current assets, and regulatory restrictions across different industries. Our data combines firm-level

Fig. 10.5: Nominal size (vs actual size) of our specification test: 999 simulations each with 399 bootstrap replications



(a) $\mu_i = v_i, T = 3$ **(b)** $\mu_i = v_i, T = 5$

financial information from Compustat (Wharton Research Data Services, 2024) with industry-level regulatory data from RegData U.S. 4.1 (QuantGov, 2024).

RegData U.S. 4.1, released in March 2022, provides comprehensive measurements of federal regulations and their industry-specific impacts. The dataset quantifies regulatory restrictions through algorithmic identification of prohibited or required activities in the Code of Federal Regulations, offering a systematic measure of regulatory burden at the industry level.

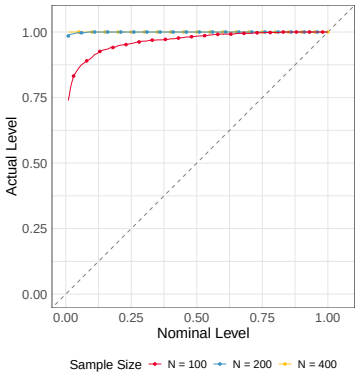
Instead of a traditional panel data model (firms measured over time), in our application, we consider a repeated measure problem. Our unit of observation will be the industry and the repeated measure will be observing different firms in each industry. We will repeat this analysis over three different time periods: 2019, 2020 and 2021.

More formally, we examine the relationship between research and development expenditures (XRD) as our left-hand-side variable, current assets (ACT) as our firm-varying explanatory variable, and industry-specific regulatory restrictions (RESTRICTIONS) as our firm-invariant variable. Our specification follows the form:

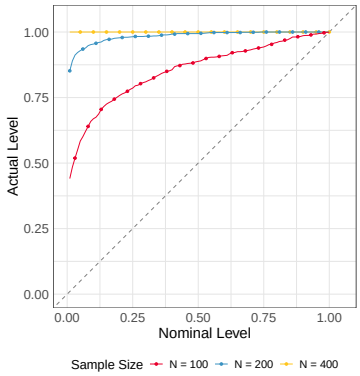
$$y_{ij} = x_{ij}^\top \beta + \bar{x}_i^\top \psi + z_i^\top \gamma + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, J, \tag{10.13}$$

where y_{ij} represents (the log of) research and development expenditures of firm j in industry i, x_{ij} denotes current (the log of) assets of firm j in industry i, $\bar{x}_i$ is the average value of x_{ij} over the firms (j) in industry i, and z_i captures regulatory restrictions in industry i. To ensure sufficient within-industry variation, we select the top five firms ($J = 5$) from each industry and exclude industries with fewer than five

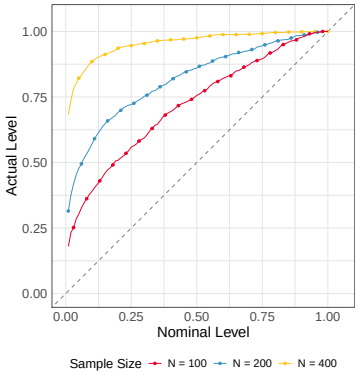
Fig. 10.6: Power of our specification test: 999 simulations each with 399 bootstrap replications



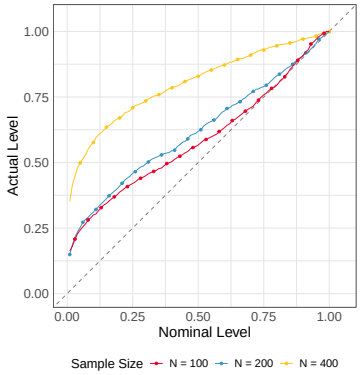
(a) $\mu_i = \overline{x}_i^\top \psi + v_i$, $T = 3$



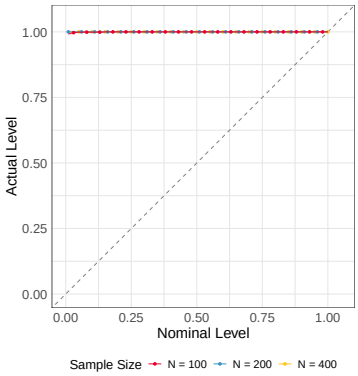
(b) $\mu_i = \overline{x}_i^\top \psi + v_i$, $T = 5$



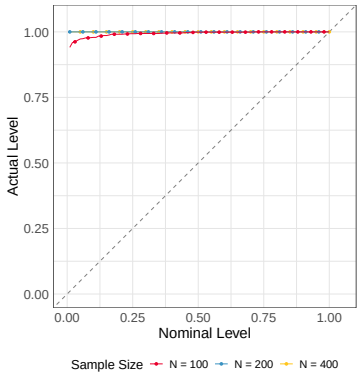
(c) $\mu_i = z_i^\top \gamma + v_i$, $T = 3$



(d) $\mu_i = z_i^\top \gamma + v_i$, $T = 5$



(e) $\mu_i = \overline{x}_i^\top \psi + z_i^\top \gamma + v_i$, $T = 3$



(f) $\mu_i = \overline{x}_i^\top \psi + z_i^\top \gamma + v_i$, $T = 5$

firms. This filtering process yields $N = 47$ industries in 2019, 48 industries in 2020, and 50 industries in 2021.

The results for both the parametric and semiparametric estimation of Equation 10.13 can be found in Figures 10.7 and 10.8. The former (Figure 10.7) displays scatterplots of $\ddot{y}_{ij} = y_{ij} - \omega_i^\top \widehat{\theta}$ versus x_{ij} and lays the parametric ($x_{ij}^\top \widehat{\beta}$) and nonparametric fits ($\widehat{m}(x_{ij})$) on top of the scatterplots for each year (rows) in the first and second columns, respectively. The parametric and nonparametric fits are difficult to distinguish from one another. Any functional form test for these sample sizes would likely lead to a failure to reject the parametric model. That being said, the semiparametric fit appears to go through the center of the points whereas the parametric fit appears to be impacted by a few large (outlier) values for $\ddot{y}_{ij}$ for smaller values of x_{ij}. The semiparametric (local-estimator) is less impacted by these values. The slightly better fit is summarized via the relative values of pseudo-R^2 (squared correlation between y and the fitted value of y).[5] Figure 10.8 gives the point estimates of γ, for each method, for each year, for each estimation method (ψ is a nuisance parameter and is not reported). We can see that the point estimates for γ are negative for each method in each year.[6] If we are to take these estimates literally, it suggests that increased regulation leads to less expenditure on research and development. That being said, these models are very simple and the estimates of $\widehat{\gamma}$ are insignificant (for each estimator) in each time period.

10.6 Conclusion

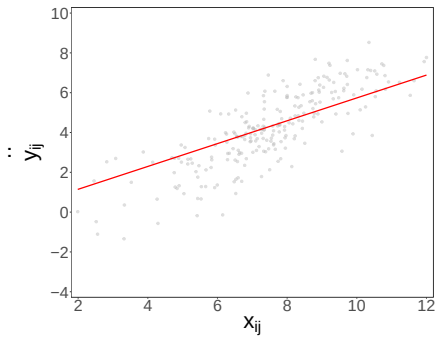
In this chapter, we proposed a semiparametric procedure for estimating CRE models. Our estimators result in closed-form solutions and achieve the optimal rates of convergence for the nonparametric and parametric components of our models. We further develop a test to check if the CRE specification captures the correlation between the unobserved effects and the regressors. Our finite sample simulations support our asymptotic theory. Finally, we provided an empirical illustration to show how the methods work with real data.

Acknowledgements The authors thank the editors, Badi Baltagi and László Mátyás, as well as Soroush Ghazi and Chris Parmeter for useful comments and suggestions. Henderson and Soberon acknowledge financial support from the I+D+i project Ref. TED2021-131763A-I00 financed by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

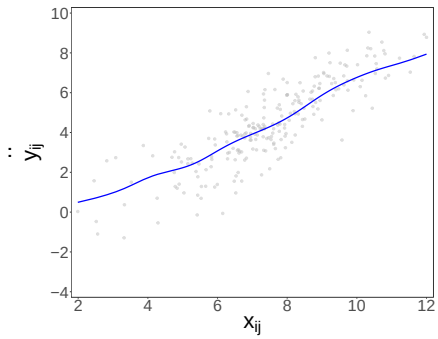
[5] The pseudo-R^2 values are; for 2019: least-squares = 0.7450, semiparametric = 0.7880; 2020: least-squares = 0.7342, semiparametric = 0.7893; and for 2021: least-squares = 0.7195, semiparametric = 0.7847.

[6] The small values for $\widehat{\gamma}$ should not be surprising as a one-unit increase in regulation should lead to only small changes in the log of research and development expenditure.

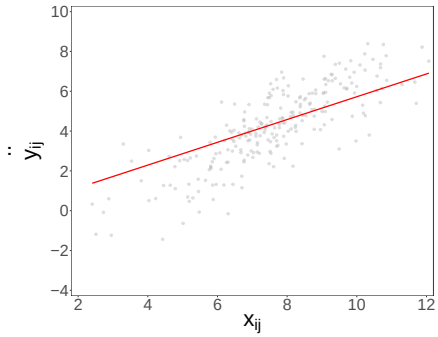
Fig. 10.7: Application: $x_{ij}^{\top}\widehat{\beta}$ and $\widehat{m}(x_{ij})$ versus x_{ij} for the least-squares and semiparametric estimators



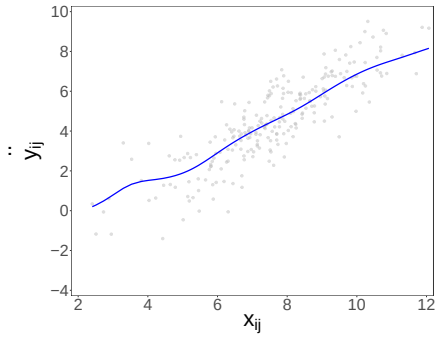
(a) 2019: Least-Squares



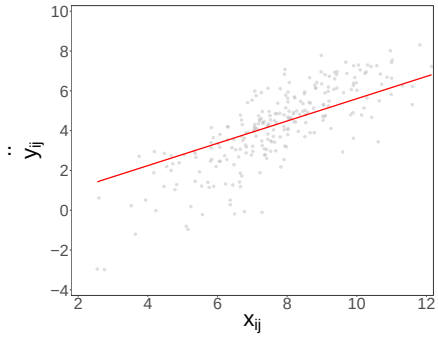
(b) 2019: Semiparametric



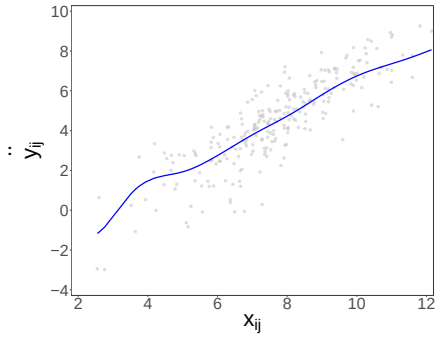
(c) 2020: Least-Squares



(d) 2020: Semiparametric

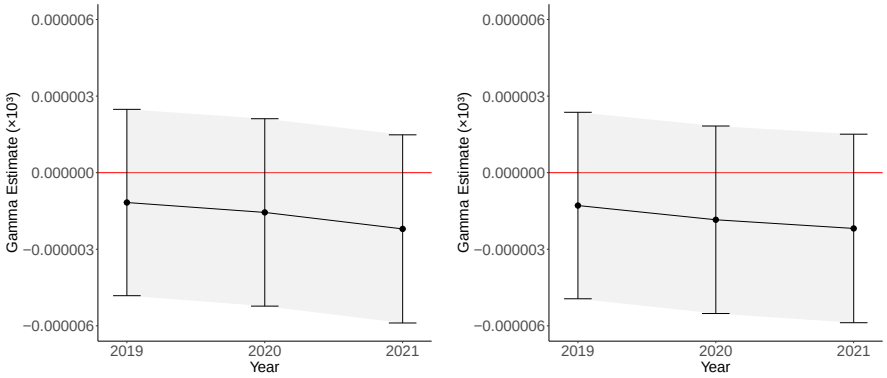


(e) 2021: Least-Squares



(f) 2021: Semiparametric

Fig. 10.8: Application: point estimates of γ and 90% confidence bounds for the years 2019, 2020, and 2021 for both the least-squares and semiparametric estimators



(a) Least-Squares **(b)** Semiparametric

References

Ai, C. & Li, Q. (2008). Semi-parametric and non-parametric methods in panel data models. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice* (pp. 451–478). Springer.

Baltagi, B. (2021). *Econometric Analysis of Panel Data*. Springer.

Bester, C. A. & Hansen, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics*, *27*(2), 235–250.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, *18*(1), 5–46.

Fan, J. & Gijbels, I. (1995). *Local Polynomial Modelling and its Applications*. Chapman & Hall.

Henderson, D. J., Carroll, R. J. & Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, *144*(1), 257–275.

Henderson, D. J. & Parmeter, C. F. (2015). *Applied Nonparametric Econometrics*. Cambridge University Press.

Henderson, D. J. & Soberon, A. (2024). Nonparametric models with fixed effects. In L. Matyas (Ed.), *The Econometrics of Multi-dimensional Panels, Advanced Studies in Theoretical and Applied Econometrics 54* (pp. 285–323). Springer Nature Switzerland AG.

Henderson, D. J. & Ullah, A. (2005). A nonparametric random effects estimator. *Economics Letters*, *88*(3), 403–407.

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

Li, Q. & Stengos, T. (1996). Semiparametric estimation of partially linear panel data models. *Journal of Econometrics*, *71*(1), 389–397.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69–85.

Nerlove, M. (2005). *Essays in Panel Data Econometrics*. Cambridge University Press.

Parmeter, C. F. & Racine, J. S. (2019). Nonparametric estimation and inference for panel data models. *Panel Data Econometrics*, 97–129.

Powell, J. L., Stock, J. H. & Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, *57*(6), 1403–1430.

QuantGov. (2024). *RegData United States 4-1.* https://www.quantgov.org/csv-download. (Accessed: 12-2024)

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954.

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Routledge.

Soberon, A., Rodriguez-Poo, J. M. & Robinson, P. M. (2021). Nonparametric panel data regression with parametric cross-sectional dependence. *Econometrics Journal*, *25*(1), 114–133.

Su, L. & Ullah, A. (2011). Nonparametric and semiparametric panel econometric models: estimation and testing. *Handbook of Empirical Economics and Finance*, 455–497.

Sun, Y., Carroll, R. J. & Li, D. (2009). Semiparametric estimation of fixed effects panel data varying coefficient models. *Advances in Econometrics*, *25*, 101–130.

Sun, Y., Zhang, Y. Y. & Li, Q. (2015). Nonparametric panel data regression models. In B. Baltagi (Ed.), *The Oxford Handbook of Panel Data* (pp. 285–324). Oxford University Press.

Wharton Research Data Services. (2024). *Compustat Annual Fundamentals Database.* https://wrds-www.wharton.upenn.edu/pages/grid-items/compustat-annual-updates-fundamentals-annual-demo/. (Accessed: 12-2024)

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics*, *211*(1), 137–150.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*(2), 263–289.