SEEDS
Sustainability Environmental Economics
and Dynamics Studies

Working Paper Series

Zero-inflated regression for unobserved effects panel data models and difference-in-differences estimation

by

Hervé Cardot, Antonio Musolesi

11/2021

SEEDS Working Paper 11/2021
December 2021
By Hervé Cardot, Antonio Musolesi

The opinions expressed in this working paper do not necessarily reflect the position of SEEDS as a whole.

Zero-inflated regression for unobserved effects panel data models and difference-in-differences estimation

Hervé CARDOT
Institut de Mathématiques de Bourgogne, UMR CNRS 5584,
Université de Bourgogne Franche-Comté
Antonio MUSOLESI
Department of Economics and Management (DEM),
University of Ferrara and SEEDS

November 30, 2021

Abstract

We introduce a statistical model combining a continuous response regression model, which can take either positive or negative values, and a mass at zero. The proposed zero-inflated regression model may be appropriate in many empirical circumstances such as unobserved effects panel data models, difference-in-differences treatment effect estimation and, more generally, when the dependent variable is expressed in terms of variation over time. We provide a mathematical formalization by means of conditional mixtures, and we first show that in this context the classical ordinary least squares estimator is generally biased. We then propose a subset estimator based on the subsample of units for which the dependent variable has non-null values and derive its asymptotic properties under a conditional independence assumption. Such an estimator can be used, along with a binary response model for the conditional probability of facing a mass at zero, to compute the partial effects arising from zero-inflated regression models. We prove the asymptotic normality of the estimator as well as consistency of the empirical bootstrap. Then, we focus on unobserved effects panel data models and on difference-in-differences estimation under zero inflation and propose an estimator of the average treatment effect that is proven to be consistent. We finally provide a Monte Carlo simulation study as well as empirical illustrations showing the usefulness of the proposed approach and bringing new insights on the size of the bias in commonly used regression models, which are based on the assumption that the response variable is continuous.

JEL classification: C21; C23; C25.

Keywords: Mixture of Distributions; Zero Inflation; Bootstrap; Panel Data; Policy Evaluation.

1 Introduction

In econometric specifications, the dependent variable is often expressed in terms of variation over time. A prime example includes commonly adopted unobserved effects panel data models, where the typical approach to estimating the parameters of interest consists of adopting a transformation, such as individual differencing over time or within transformation, to eliminate the unobserved component and then applying ordinary least squares (OLS) (see, for example, Wooldridge, 2010). A similar strategy is adopted in program evaluation within a difference-in-differences (DID) framework, where for identification purposes and to address the issue of selection on unobservables, it is commonly assumed that the conditional independence assumption holds for the difference in the outcome before and after the beginning of the policy and then a before–after approach is adopted (Heckman and Hotz, 1989; Lechner, 2011, 2015). Another relevant example is provided by cross-sectional data models when the interest lies in directly modeling outcome variation over time, such as when studying economic growth or employment dynamics as a function of some explanatory variables observed at a given point in time (Sala-i Martin, 1997).

However, while most of the economic variables such as employment, wages, production, investments, consumption, etc. take non-negative values, a crucial consequence of modeling the individual deviations of the outcome variable over time is that these deviations can take either positive or negative values. Importantly, it may also be—especially at a micro-data level—that for a non-negligible fraction of the statistical units under investigation the variable of interest does not vary over time, so that we have to face a peculiar kind of zero-inflated phenomenon.

As an empirical illustration of this phenomenon, in Figure 1 we illustrate the estimated distribution of the variation of employment in time $\mathrm{EMP}_{it} - \mathrm{EMP}_{it-1}$, EMP_{it} being the employment level in French municipalities for $t = 1994$. This distribution can be approximated by a mixture of a mass at 0 and a continuous density function, which is defined over both positive and negative values.

===== Figure 1 =====

With this scenario, common zero-inflated approaches, which are based on negative binomial or Poisson distributions and can only deal with non-negative count data, are not appropriate. The model under study is also different from the corner solution model, which arises when the response variable has a continuous distribution over strictly positive values and there is a mass at zero with non-null probability. This is a corner solution outcome with a corner at zero and,

and in this case a tobit model can be used for estimation. Unlike the corner solution model, the proposed model also allows for negative values.

This paper aims to provide a theoretical formalization of such a zero-inflated model and bring new evidence based both on simulated and real data. We propose a statistical model based on a conditional mixture of a continuous linear regression model and a mass at zero. We first show that in this context, the classical ordinary least squares estimator is biased unless the probability of observing zero does not depend on the explanatory variables of the regression function. We then propose a subset estimator based on the subsample of units for which the dependent variable has non-null values and derive its asymptotic properties. We prove that under a specific conditional independence assumption, which is closely related to the well-known missin- at-random (MAR) condition in missing data problems (see Little and Rubin, 2002) or the classical unconfoundedness assumption in program evaluation (see Imbens and Wooldridge, 2009), the proposed estimator is consistent and asymptotically Gaussian. This subset estimator can be used, along with a binary response model for the conditional probability of facing a mass at zero, to compute the partial effects arising from such a peculiar zero-inflated regression model. We also prove that empirical paired bootstrap procedures can be employed to obtain consistent approximations of the distribution of the unknown parameters and to build confidence intervals for prediction with a given asymptotic confidence level when the conditional probability of observing zero can be expressed as a probit or logit model.

After formalizing the model, we focus our attention on commonly used panel data approaches. In particular, we first consider standard unobserved effects panel data models, for which the proposed estimator, the computation of partial effects, and the bootstrap procedure can be quite straightforwardly adopted. We also extend the theoretical work by focusing on DID estimation under zero inflation and propose an estimator of the average treatment effect (ATE) that is proven to be consistent.

We then provide a simulated example that aims to illustrate the effect of zero inflation on the expected value of the response variable and to check the ability of paired bootstrap procedures to produce reliable confidence intervals. It is shown that the zero-inflated phenomenon can produce very different functional relations depending on the underlying parameters and that the linear model provides misleading results. In particular, when the underlying relation is non-monotonic it clearly provides a senseless fit. In contrast, the proposed estimator, which handles the zero-inflation, provides a very faithful description of the underlying DGP. Our simulation also offers evidence of the validity of the non-parametric bootstrap in the proposed zero-inflated framework, even in the case of small samples.

Finally, the usefulness of our methodology is illustrated on real data examples, bringing new insight into the size of the bias of commonly used regression models, which are based on the assumption that the response variable is continuous. We first revisit a classical Mincer wage equation with zero-inflated data and exploit the panel data of Baltagi and Khanti-Akom (1990). We also consider the problem of estimating the ATE of two distinct public policies that were devoted to boosting rural development in France and have been recently investigated in Cardot and Musolesi (2020).

The paper is organized as follows. Section 2 introduces the zero-inflated model and addresses the problem of its estimation. Section 3 focuses on panel data models and extends the previous results by considering DID estimation under zero inflation. Sections 4 and 5 provide a small simulation study and two illustrative examples, respectively. Finally, concluding remarks are given in Section 6 and proofs, additional details and information are gathered in an Appendix.

2 Model and estimation

Consider a real random variable Y and a random vector $\mathbf{X}$ taking values in $\mathbb{R}^p$. We suppose that the distribution of Y is a mixture of a continuous distribution and a Dirac at zero, so that

$$Y = ZY_c + (1-Z)0, \tag{1}$$

where Z is a Bernoulli random variable taking the value of 0 or 1 and Y_c is a continuous random variable. Given $\mathbf{X} = \mathbf{x}$, we suppose that

$$\mathbb{P}[Z = 1|\mathbf{X} = \mathbf{x}] = \pi(\mathbf{x}, \boldsymbol{\beta}) \tag{2}$$

for some known parametric function $\pi(.,.)$ with unknown parameter $\boldsymbol{\beta}$, and Y_c satisfies a standard linear regression model with expectation $\boldsymbol{\theta}^\top \mathbf{x}$ and variance σ^2,

$$Y_c = \boldsymbol{\theta}^\top \mathbf{x} + \epsilon, \tag{3}$$

where ϵ is a continuous noise component supposed to satisfy $\mathbb{E}(\epsilon|\mathbf{X} = \mathbf{x}) = 0$ and $\mathbb{V}ar(\epsilon|\mathbf{X} = \mathbf{x}) = \sigma^2$ almost surely. The parameters to be estimated are $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

We assume that the following conditional independence assumption holds:

$$(\mathrm{H}_1) \qquad Y_c \perp Z \mid \mathbf{X}.$$

Assumption (H_1) is closely related to the well-known missing-at-random (MAR) condition in missing data problems (see Little and Rubin, 2002) or the classical unconfoundedness assumption in program evaluation (see Imbens and Wooldridge, 2009). It ensures that we have a set of variables $\mathbf{X}$ such that Y_c and the fact that Y is not equal to zero are independent given $\mathbf{X}$. It can also be related to assumption (17.38) in Wooldridge (2010) for the hurdle model in which Y only takes positive values. With Model (3), assumption (H_1) can be expressed as

$$\epsilon \perp Z \mid \mathbf{X}.$$

We now mention the following basic properties that give the conditional expected outcome and the gradient, with respect to $\mathbf{x}$, of the expected outcome.

Proposition 2.1. *Assume that models (1) and (3) hold. Then, if H_1 is fulfilled,*

- $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}] = \pi(\mathbf{x},\boldsymbol{\beta})\mathbb{E}[Y_c|\mathbf{X}=\mathbf{x}] = \pi(\mathbf{x},\boldsymbol{\beta})\boldsymbol{\theta}^\top\mathbf{x}.$

If, furthermore, $\pi(\mathbf{x},\boldsymbol{\beta})$ is differentiable with respect to $\mathbf{x}$,

- $\frac{\partial\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]}{\partial\mathbf{x}} = \pi(\mathbf{x},\boldsymbol{\beta})\boldsymbol{\theta} + \boldsymbol{\theta}^\top\mathbf{x}\frac{\partial\pi(\mathbf{x},\boldsymbol{\beta})}{\partial\mathbf{x}}.$

The proof of Proposition 2.1 is direct and thus omitted. Note that in classical linear regression models, without zero inflation we have $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}] = \boldsymbol{\theta}^\top\mathbf{x}$ and $\frac{\partial\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]}{\partial\mathbf{x}} = \boldsymbol{\theta}$.

Suppose we observe a sample $(Y_i,\mathbf{x}_i)$, for $i=1,\ldots,n$, made of n independent and identically distributed copies of the random vector $(Y,\mathbf{X})$. We define $Z_i = \mathbf{1}_{\{Y_i\neq 0\}}$ for $i=1,\ldots,n$. The n random vectors $(Y_1,Z_1,\mathbf{x}_1),\ldots,(Y_n,Z_n,\mathbf{x}_n)$ are *i.i.d*, with the same distribution as $(Y,Z,\mathbf{X})$, where $Z=\mathbf{1}_{\{Y\neq 0\}}$. In the following, we denote by $\mathbf{Y}=(Y_1,\ldots,Y_n)$.

2.1 The particular case of constant probability of zeros

The ordinary least squares estimator for $\boldsymbol{\theta}$, which takes account of the zero-inflation phenomenon, is

$$\widetilde{\boldsymbol{\theta}} = \left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^\top\mathbf{Y}, \tag{4}$$

where $\widetilde{\mathbf{X}}$ is the $n\times p$ matrix with generic element $\mathbf{x}_{ij}$, for $i=1,\ldots,n$ and $j=1,\ldots,p$. This matrix is assumed to be a full column rank matrix. We have the following property:

Proposition 2.2. *Suppose that models (1) and (3) hold. If, Furthermore, $\pi(\mathbf{x},\boldsymbol{\beta})=\pi$ does not depend on $\mathbf{x}$, then*

$$\begin{aligned}\mathbb{E}\left[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1,\ldots,\mathbf{x}_n\right] &= \pi\boldsymbol{\theta},\\ \mathbb{V}ar\left[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1,\ldots,\mathbf{x}_n\right] &= \pi\left(\sigma^2\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1} + (1-\pi)\boldsymbol{\theta}\boldsymbol{\theta}^\top\right).\end{aligned}$$

If the conditions of Proposition 2.2 are fulfilled, the predicted value for Y given $\mathbf{X} = \mathbf{x}$ is unbiased,

$$\mathbb{E}\left[\widetilde{\boldsymbol{\theta}}^\top \mathbf{x}\right] = \pi \boldsymbol{\theta}^\top \mathbf{x} = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}],$$

with conditional variance

$$\mathbb{V}ar\left[\widetilde{\boldsymbol{\theta}}^\top \mathbf{x}|\mathbf{x}_1, \dots, \mathbf{x}_n\right] = \sigma^2 \pi \mathbf{x}^\top \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \mathbf{x}^\top + \pi(1-\pi)\left(\boldsymbol{\theta}^\top \mathbf{x}\right)^2.$$

On the other hand, if the probability of observing $Y = 0$ depends on $\mathbf{X}$ then $\widehat{\boldsymbol{\theta}}$ is a biased estimator of $\pi(\mathbf{x}, \boldsymbol{\beta})\boldsymbol{\theta}$ and classical OLS estimators will fail to provide consistent estimates of $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, except in very particular cases.

2.2 Subset estimator: definition and asymptotic properties

We now consider an estimator of parameter $\boldsymbol{\theta}$ based on the subsample corresponding to the non-null values of Y,

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \left(\sum_{i=1}^n Z_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i=1}^n Z_i Y_i \mathbf{x}_i\right) \\ &= \boldsymbol{\theta} + \left(\frac{1}{n}\sum_{i=1}^n Z_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n Z_i \epsilon_i \mathbf{x}_i\right), \qquad (5)\end{aligned}$$

using the fact that $Z_i Y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \epsilon_i$ when $Z_i = 1$, and $Z_i Y_i = 0$ otherwise.

We need to introduce the following additional assumption (H$_2$), which is an identification assumption ensuring that for a large enough n, $\sum_{i=1}^n Z_i \mathbf{x}_i \mathbf{x}_i^\top$ is a full rank matrix and $\widehat{\boldsymbol{\theta}}$ is uniquely defined.

$$(\mathrm{H}_2) \qquad \mathbb{E}\left(\pi(\mathbf{X}, \boldsymbol{\beta})\mathbf{X}\mathbf{X}^\top\right) = \mathbf{Q}_\pi \text{ where } \mathbf{Q}_\pi \text{ is non-singular.}$$

Our notations are borrowed from van der Vaart (1998), and we denote by $U_n = o_p(1)$ the fact that the sequence $(U_n)_{n \geq 1}$ of random variables (vectors or matrices) converges to zero in probability when n tends to infinity, whereas the convergence in distribution of the sequence towards a Gaussian random vector with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Gamma}$ is denoted by $U_n \rightsquigarrow \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$.

The subset estimator $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically Gaussian, as shown in the following proposition.

Proposition 2.3. *Suppose that models (1) and (3) hold. Assume also that hypotheses* (H_1) *and* (H_2) *are fulfilled. Then as n tends to infinity,*

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \rightsquigarrow \mathcal{N}\left(0,\sigma^2\mathbf{Q}_\pi^{-1}\right).$$

Note that the fact that parameter $\boldsymbol{\beta}$ is unknown is not an issue if the goal is only to make inference on $\boldsymbol{\theta}$, since, as shown in Proposition 2.4 below, the asymptotic variance $\sigma^2\mathbf{Q}_\pi^{-1}$ is consistently estimated by

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} \tag{6}$$

with

$$\widehat{\sigma}^2 = \frac{1}{\sum_{i=1}^{n} Z_i}\sum_{i=1}^{n} Z_i\left(Y_i-\mathbf{x}_i^\top\widehat{\boldsymbol{\theta}}\right)^2.$$

Proposition 2.4. *Suppose that models (1) and (3) hold. Assume also that hypotheses* (H_1) *and* (H_2) *are fulfilled. Then as n tends to infinity,*

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\theta}}) - \sigma^2\mathbf{Q}_\pi^{-1} = o_p(1).$$

An immediate consequence of Proposition 2.4 is that, thanks to Slutsky's Theorem, we are able to provide consistent confidence intervals for $\boldsymbol{\theta}$.

However, if the aim is to infer on the expected value of Y given $\mathbf{X}=\mathbf{x}$, then a consistent estimator of $\boldsymbol{\beta}$ is required to consistently estimate the conditional probability $\pi(\boldsymbol{\beta},\mathbf{x})$ of observing the continuous component Y_c.

2.3 Confidence intervals for prediction

We are now interested in computing a confidence interval for $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}] = \pi(\mathbf{x},\boldsymbol{\beta})\boldsymbol{\theta}^\top\mathbf{x}$.

We assume that $\pi(\boldsymbol{\beta},\mathbf{x})$ has a known parametric form. For example, $\log(\pi/(1-\pi)) = \boldsymbol{\beta}^\top\mathbf{x}$corresponds to logistic regression and $\pi=\Phi(\boldsymbol{\beta}^\top\mathbf{x})$ corresponds to probit regression when $\Phi(u)=\mathbb{P}(Z\le u)$, Z being a centered Gaussian random variable with unit variance.

Under previous hypotheses, parameter $\boldsymbol{\beta}$ can be estimated efficiently with classical maximum likelihood approaches (see Newey and McFadden (1994) for probit regression and Hjort and Pollard (2011) for logistic regression). The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ is such that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1)$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \rightsquigarrow \mathcal{N}(0,\mathbf{\Gamma}).$$

Even if the asymptotic covariance matrix $\mathbf{\Gamma}$ is unknown, it can be estimated consistently if the considered parametric models correspond to logit or probit link functions (see Newey and McFadden, 1994).

To get the joint distribution of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\beta}}$, note that these estimators of parameters are obtained by minimizing the functional

$$\Psi_n(\boldsymbol{\beta},\boldsymbol{\theta};(Y_1,\mathbf{x}_1),\ldots,(Y_n,\mathbf{x}_n)) = \Psi_{1n}(\boldsymbol{\beta};(Y_1,\mathbf{x}_1),\ldots,(Y_n,\mathbf{x}_n)) + \Psi_{2n}(\boldsymbol{\theta};(Y_1,\mathbf{x}_1),\ldots,(Y_n,\mathbf{x}_n)), \tag{7}$$

where

$$\Psi_{1n}(\boldsymbol{\beta};(Y_1,\mathbf{x}_1),\ldots,(Y_n,\mathbf{x}_n)) = -\frac{1}{n}\sum_{i=1}^{n}\left[Z_i \ln\left(\frac{\pi(\boldsymbol{\beta},\mathbf{x}_i)}{1-\pi(\boldsymbol{\beta},\mathbf{x}_i)}\right) + \ln\left(1-\pi(\boldsymbol{\beta},\mathbf{x}_i)\right)\right]$$

is the opposite of the averaged log likelihood and

$$\Psi_{2n}(\boldsymbol{\theta};(Y_1,\mathbf{x}_1),\ldots,(Y_n,\mathbf{x}_n)) = \frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i - \mathbf{x}_i^\top\boldsymbol{\theta}\right)^2$$

is a weighted least squares criterion. We clearly have $\dfrac{\partial^2\Psi_n}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}} = 0$, so that we can deduce, under previous assumptions,

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\beta}}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\beta}\end{pmatrix}\right) \rightsquigarrow \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2\mathbf{Q}_\pi^{-1} & 0\\ 0 & \mathbf{\Gamma}\end{pmatrix}\right). \tag{8}$$

A direct application of the delta method (see van der Vaart, 1998, Theorem 3.1) allows us to conclude that as n tends to infinity,

$$\sqrt{n}\left(\pi(\mathbf{x},\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\theta}}^\top\mathbf{x} - \pi(\mathbf{x},\boldsymbol{\beta})\boldsymbol{\theta}^\top\mathbf{x}\right) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\pi(\mathbf{x},\boldsymbol{\beta})^2\mathbf{x}^\top\mathbf{Q}_\pi^{-1}\mathbf{x} + \left(\boldsymbol{\theta}^\top\mathbf{x}\right)^2\frac{\partial\pi(\mathbf{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\mathbf{\Gamma}\frac{\partial\pi(\mathbf{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right). \tag{9}$$

The asymptotic distribution obtained in (9) can be used to get approximate confidence intervals for $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ by replacing the unknown coefficients in the variance terms by their estimates.

However, paired bootstrap approaches, which are reasonably time-consuming in this parametric framework, are generally preferred and are much simpler to use (see Wooldridge, 2010, Chapter 21, in the general context of policy evaluation and Cardot and Musolesi, 2020 for an illustration with zero-inflated data).

More precisely, consider a bootstrap sample $(Y_1^*,\mathbf{x}_1^*),\ldots,(Y_n^*,\mathbf{x}_n^*)$ where $(Y_i^*,\mathbf{x}_i^*)$ are *i.i.d.* from the empirical distribution of $(Y_1,\mathbf{x}_1),\ldots,(Y_n,\mathbf{x}_n)$ and denote by $(\widehat{\boldsymbol{\beta}}^*,\widehat{\boldsymbol{\theta}}^*)$ the bootstrap estimate of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ defined as the minimizer of $\Psi_n(\boldsymbol{\beta}, \boldsymbol{\theta}; (Y_1^*, \mathbf{x}_1^*), \ldots, (Y_n^*, \mathbf{x}_n^*))$. For $\mathbf{u}_1 \in \mathbb{R}^p$ and $\mathbf{u}_2 \in \mathbb{R}^p$, we denote by $F_{n,B}(\mathbf{u}_1, \mathbf{u}_2)$ the conditional joint cumulative distribution function of the bootstrap estimator, given the data:

$$F_{n,B}(\mathbf{u}_1, \mathbf{u}_2) = \mathbb{P}\left[\sqrt{n}\left((\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\theta}}^*) - (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})\right) \leq (\mathbf{u}_1, \mathbf{u}_2) \mid (Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)\right]$$

where the inequality should be understood component-wise.

We denote by $F_n(\mathbf{u}_1, \mathbf{u}_2) = \mathbb{P}\left[\sqrt{n}\left((\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) - (\boldsymbol{\beta}, \boldsymbol{\theta})\right) \leq (\mathbf{u}_1, \mathbf{u}_2)\right]$ the joint cumulative distribution function corresponding to the multivariate Gaussian distribution given on the right-hand side of (8). We can state the following proposition, which ensures that the bootstrap procedures allow us to obtain a consistent approximation of the distribution of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ and to build confidence intervals for prediction with a given asymptotic confidence level.

Proposition 2.5. *Suppose that models (1), (2), and (3) hold and that $\pi(\mathbf{x}, \boldsymbol{\beta})$ is of a logit or probit shape. Assume also that hypotheses (H_1) and (H_2) are fulfilled. Then as n tends to infinity,*

$$\sup_{\mathbf{u}_1, \mathbf{u}_2} |F_{n,B}(\mathbf{u}_1, \mathbf{u}_2) - F_n(\mathbf{u}_1, \mathbf{u}_2)| = o_p(1).$$

Given that $\mathbf{X} = \mathbf{x}$, we also have

$$\sup_{u \in \mathbb{R}} \Bigg| \mathbb{P}\left[\sqrt{n}\left(\pi(\mathbf{x}, \widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\theta}}^\top \mathbf{x} - \pi(\mathbf{x}, \boldsymbol{\beta})\boldsymbol{\theta}^\top \mathbf{x}\right) \leq u\right] - \mathbb{P}\left[\sqrt{n}\left(\pi(\mathbf{x}, \widehat{\boldsymbol{\beta}}^*)\widehat{\boldsymbol{\theta}^*}^\top \mathbf{x} - \pi(\mathbf{x}, \widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\theta}}^\top \mathbf{x}\right) \leq u \mid (Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)\right] \Bigg| = o_p(1).$$

Note that we could consider other sufficiently smooth link functions to model the Bernoulli variable Z given $\mathbf{X} = \mathbf{x}$ and the previous result given in Proposition 2.5 would remain true provided that the criterion $\Psi_{1n}(.)$ is a twice differentiable convex function of $\boldsymbol{\beta}$. Indeed, in our context the estimators of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are obtained as minimizers of smooth convex functionals so that general consistency results for bootstrapping apply (see Theorem 2.4 in Bose and Chatterjee, 2003) and the "delta method" for bootstrapped estimates can be employed (see Theorem 23.9 in van der Vaart, 1998).

3 Unobserved effects panel data models and difference-in-differences estimation under zero inflation

In this section, we consider unobserved effects panel data models and DID estimation under zero inflation. We first deal with commonly used unobserved effects panel data models, focusing on the identification conditions of the unknown parameters, on the computation of partial effects, and on the ability of the paired bootstrap to provide consistent confidence intervals in this extended framework. Then we move to DID estimation, proposing an estimator of ATE that is proven to be consistent.

3.1 Unobserved effects panel data models

Consider now panel data, that is to say, the evolution of a size n sample observed over a time period $t = 1, \ldots, T$. It is often assumed that the data are generated according to a data-generating process governed by a linear relation of the following form, for $i = 1, \ldots, n$ and $t = 1, 2, \ldots, T$:

$$U_{it} = \eta_t + \boldsymbol{\theta}^\top \mathbf{x}_{it} + c_i + \epsilon_{it}, \tag{10}$$

$$\mathbb{E}\left(\epsilon_{it} \mid \mathbf{x}_i, c_i\right) = 0, \tag{11}$$

where η_t is a common time effect (or time period intercept), $\mathbf{x}_{it} \in \mathbb{R}^p$ is a vector of p explanatory variables observed at time t, and ϵ_{it} is the idiosyncratic error term supposed to be a continuous random variable. The term c_i denotes a time-constant unobserved effect and, under random sampling in the cross-section dimension, following Chamberlain (1984) and Wooldridge (2010) c_i can be viewed as random. Under this framework, i.e., an unobserved effects panel data Model (10) with strict exogeneity conditional on c_i (11), there is no restriction on the dependence between c_i and the $p \times T$ matrix $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, and the equalities $\mathbb{E}\left(c_i \mid \mathbf{x}_i\right) = \mathbb{E}\left(c_i\right) = 0$ do not generally hold since $\mathbb{E}\left(c_i \mid \mathbf{x}_i\right)$ is allowed to be any function of $\mathbf{x}_i$.

In order to estimate $\boldsymbol{\theta}$, the typical approach consists of adopting a transformation to eliminate the unobserved component c_i. Commonly used transformations are the *within* and the *first-difference* (FD, hereafter) transformations, and then OLS is applied.[1] For the FD estimator, we can consider the transformed variable

$$\begin{aligned}\Delta_{it} &= U_{it} - U_{it-1} \\ &= \left(\eta_t - \eta_{t-1}\right) + \boldsymbol{\theta}^\top \left(\mathbf{x}_{it} - \mathbf{x}_{it-1}\right) + \left(\epsilon_{it} - \epsilon_{it-1}\right),\end{aligned} \tag{12}$$

for $t = 2, \ldots, T$. Since most of the economic variables U, such as employment or wages, take non-negative values, a crucial consequence of modeling the individual variation of the outcome variable is that this can take either positive or negative values. Importantly, it may also be—especially at a micro data level—that for a non-negligible fraction of the statistical units the variable of interest U does not vary over time, so that we face a zero-inflated phenomenon that

[1] The same logic applies when the model contains individual time trends (Heckman and Hotz, 1989; Wooldridge, 2005).

cannot be dealt with properly by the Model (12). We therefore introduce a mixture model, and suppose that we observe Y_{it} instead of Δ_{it} with

$$Y_{it} = Z_{it}\Delta_{it} + (1 - Z_{it})0, \tag{13}$$

where Z_{it} is a binary random variable taking a value of 0 or 1. The conditional independence assumption H_1 assumed to hold for models (1) and (3) is now replaced by

$$(\mathrm{H}_{1\Delta}) \qquad (\Delta_2, \ldots, \Delta_T) \perp (Z_2, \ldots, Z_T) \mid \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T.$$

Note that there is no restrictive condition on the temporal dependence or heterogeneity of $\Delta_2, \ldots, \Delta_T$ and $Z_2, \ldots, Z_T$.

We suppose that we have a sample $(Y_{i1}, \ldots, Y_{iT}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, $i = 1, \ldots, n$ of n independent realizations of $(Y_1, \ldots, Y_T, \mathbf{X}_1, \ldots, \mathbf{X}_T)$. For each statistical unit, we can define the binary variables Z_{it} with $Z_{it} = 1$ if $Y_{it} \neq 0$ and $Z_{it} = 0$ otherwise, for $t = 2, \ldots, T$.

We can assume that for some known parametric model and unknown vectors of parameters $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_T$, the probability that Y_t equals zero only depends on the current values of the explanatory variables, so that the following relation holds:

$$\mathbb{P}\left[Z_t = 1 \mid \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_T = \mathbf{x}_T\right] = \pi(\mathbf{x}_t, \boldsymbol{\beta}_t), \quad t = 2, \ldots, T. \tag{14}$$

Then, with Model (13) and assumption $(\mathrm{H}_{1\Delta})$ we have that

$$\mathbb{E}\left[Y_t \mid \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_T = \mathbf{x}_T\right] = \pi(\mathbf{x}_t, \boldsymbol{\beta}_t)\ \boldsymbol{\theta}^\top (\mathbf{x}_t - \mathbf{x}_{t-1}), \quad t = 2, \ldots, T. \tag{15}$$

Consequently, the mean effect of variation in $\mathbf{x}_t$ on the conditional expectation of Y_t is equal to

$$\frac{\partial \mathbb{E}\left[Y_t \mid \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_T = \mathbf{x}_T\right]}{\partial \mathbf{x}_t} = \pi(\mathbf{x}_t, \boldsymbol{\beta}_t)\ \boldsymbol{\theta} + \boldsymbol{\theta}^\top (\mathbf{x}_t - \mathbf{x}_{t-1}) \frac{\partial \pi(\mathbf{x}_t, \boldsymbol{\beta}_t)}{\partial \mathbf{x}_t}, \quad t = 2, \ldots, T. \tag{16}$$

As in (7), the estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_T$ can be performed by minimizing the functional $\Psi_n^\Delta(\boldsymbol{\theta}, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_T)$, defined as follows:

$$\Psi_n^\Delta(\boldsymbol{\theta}, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_T) = \sum_{t=2}^{T} \Psi_{1n,t}^\Delta(\boldsymbol{\beta}_t) + \Psi_{2n}^\Delta(\boldsymbol{\theta}),$$

with

$$\Psi_{1n,t}^\Delta(\boldsymbol{\beta}_t) = -\frac{1}{n} \sum_{i=1}^{n} Z_{it} \ln\left(\frac{\pi(\boldsymbol{\beta}_t, \mathbf{x}_{it})}{1 - \pi(\boldsymbol{\beta}_t, \mathbf{x}_{it})}\right) + \ln\left(1 - \pi(\boldsymbol{\beta}_t, \mathbf{x}_{it})\right) \tag{17}$$

and

$$\Psi_{2n}^\Delta(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=2}^{T} Z_{it} \left(Y_{it} - \boldsymbol{\theta}^\top (\mathbf{x}_{it} - \mathbf{x}_{it-1})\right)^2. \tag{18}$$

Asymptotic identification of parameter $\boldsymbol{\theta}$ is ensured with the following assumption:

$$(\mathrm{H}_{2\Delta}) \qquad \mathbb{E}\left[\sum_{t=2}^{T} Z_t \left(\mathbf{X}_t - \mathbf{X}_{t-1}\right)\left(\mathbf{X}_t - \mathbf{X}_{t-1}\right)^\top\right] \quad \text{is a full rank matrix.}$$

Assumption ($\mathrm{H}_{2\Delta}$) is similar to assumption FD.2 in Wooldridge (2010) (Chapter 10) but also takes into account the zero-inflation phenomenon. It can be proven under hypotheses ($\mathrm{H}_{1\Delta}$) and ($\mathrm{H}_{2\Delta}$) that $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ that is asymptotically Gaussian as n tends to infinity.

If $\pi(\mathbf{x}, \boldsymbol{\beta})$ is of a logit or probit shape and if the set of assumptions

$$(\mathrm{H}_{3\Delta,t}) \qquad \mathbb{E}\left[\mathbf{X}_t \mathbf{X}_t^\top\right] \quad \text{is a full rank matrix}$$

hold for $t = 2, \ldots, T$, it can be proven that the estimators $\widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_T$ are consistent and asymptotically Gaussian as n tends to infinity.

As in Proposition 2.5, consistent confidence intervals for $\boldsymbol{\theta}$ and the conditional expected value in (15) can be obtained employing a paired bootstrap approach similar to that described in Section 2.3, since $\Psi_n^\Delta(\boldsymbol{\theta}, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_T)$ is a twice differentiable convex functional.

3.2 Evaluation of treatment effects with difference-in-differences

Unobserved effects panel data models have been shown to be useful in a variety of situations (Heckman and Hotz, 1989; Wooldridge, 2005). One relevant situation is the estimation of treatment effects when the model contains a discrete policy variable and selection on both observables and unobservables is an issue. More specifically, Model (10) relies on identification hypotheses similar to DID-type estimators. Indeed, the conventional DID estimator is often derived using a linear parametric model that only contains time and individual effects, besides the treatment variable (Abadie, 2005; Lechner, 2015), and a crucial underlying identification assumption is the existence of a common (non-linear) trend. This means that the average outcomes of the different groups of individuals, grouped according to the treatment status, would have followed parallel paths in the absence of treatment (Lee and Kang, 2006). Conditioning on some continuous variables makes the common-trend hypothesis more plausible (see, for example, Abadie, 2005; Heckman et al., 1997, 1998).

Recent works have provided further insights. Some of them have investigated the assumptions that are needed to yield estimated coefficients having a causal interpretation and, in particular, have considered various settings such as allowing for heterogeneous treatment effects, variation in treatment timing, and dynamic treatment effects (De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Han, 2021; Sun and Abraham, 2020). In this work we explore another direction, allowing the conditional distribution of the variation of the variable under study to be a mixture of a continuous distribution and a mass at zero.

More formally, suppose that we aim to evaluate a treatment effect on an outcome U_t at time t for a treatment that is set up at time t_τ with $1 < t_\tau < T$ among $R-1$ possible treatments. We denote by D^r, for $r \in \{0, 1, \ldots, R-1\}$, the binary treatment indicator variable that takes a value of 1 if treatment r has been applied and 0 otherwise, with the convention that $r = 0$ corresponds to no treatment. The $R-1$ possible treatments are supposed to be mutually exclusive, so that by definition $\sum_{r=0}^{R-1} D^r = 1$. The variable of interest is the potential outcome difference between time $t = 1$ and time $t \geq t_\tau$. It is denoted by Y_t^r and is defined as follows:

$$Y_t^r = Z_t^r \left(U_t^r - U_{t_1}\right) + \left(1 - Z_t^r\right) 0, \tag{19}$$

where Z_t^r is a binary variable indicating which regime governs the evolution of the outcome between t and t_1. If there has been a change, then Z_t^r is equal to one and Z_t^r is equal to zero in case of no variation. The quantity U_t^r is the potential outcome at time t under treatment r under the variation regime.

Note that it is only possible to observe one value of Y_t^r, which is equal to $Y_t = \sum_{r=0}^{R-1} D^r Y_t^r$, among the R potential outcome evolutions, $Y_t^0, \ldots, Y_t^{R-1}$. Our aim is to estimate the average treatment effect at time t under treatment r compared to no treatment, given that $\mathbf{X} = \mathbf{x}$:

$$\mathrm{ATE}^r(t, \mathbf{x}) = \mathbb{E}\left(Y_t^r - Y_t^0 | \mathbf{X} = \mathbf{x}\right). \tag{20}$$

We assume that for $r = 0, \ldots, R-1$, the continuous part of the response $Y_{ct}^r = U_t^r - U_{t_1}$ satisfies a standard linear regression model

$$Y_{ct}^r(t) = \boldsymbol{\theta}_{rt}^\top \mathbf{x} + \epsilon_{rt}, \tag{21}$$

where ϵ_{rt} is a noise component with a continuous distribution such that $\mathbb{E}(\epsilon_{rt}|\mathbf{X} = \mathbf{x}) = 0$ and $\mathbb{V}ar(\epsilon_{rt}|\mathbf{X} = \mathbf{x}) = \sigma_{rt}^2$ almost surely. Note that the probability that $Y_{ct}^r = 0$ is equal to $0\boldsymbol{\beta}_{rt}$, we have

$$\mathbb{P}\left[Z_t^r = 1 \mid \mathbf{X} = \mathbf{x}\right] = \pi(\mathbf{x}, \boldsymbol{\beta}_{rt}). \tag{22}$$

We assume that we have at hand a set of p confounding variables $\mathbf{X} = (X_1, \ldots, X_p)$ such that

$$(\mathrm{H}_{1t}) \qquad Y_{ct}^r \perp Z_t^r \mid \mathbf{X}.$$

Assumption (H_{1t}) is similar to assumption (H_1) discussed in Section 2 and allows obtaining the following decomposition for the conditional average treatment effect.

Proposition 3.1. *If assumption H_{1t} is in force and models (19), (21), and (22) hold, then*

$$ATE^r(t, \mathbf{x}) = (\pi(\mathbf{x}, \boldsymbol{\beta}_{rt})\boldsymbol{\theta}_{rt} - \pi(\mathbf{x}, \boldsymbol{\beta}_{0t})\boldsymbol{\theta}_{0t})^\top \mathbf{x}.$$

The proof of Proposition 3.1 is direct. With (19) and assumption (H_{1t}), we have

$$\begin{aligned}\mathrm{ATE}^r(t, \mathbf{x}) &= \mathbb{E}\left(Z_t^r Y_{ct}^r - Z_t^0 Y_{ct}^0 \mid \mathbf{X} = \mathbf{x}\right)\\ &= \mathbb{E}\left(Z_t^r Y_{ct}^r \mid \mathbf{X} = \mathbf{x}\right) - \mathbb{E}\left(Z_t^0 Y_{ct}^0 \mid \mathbf{X} = \mathbf{x}\right)\\ &= \mathbb{E}\left(Z_t^r \mid \mathbf{X} = \mathbf{x}\right)\mathbb{E}\left(Y_{ct}^r \mid \mathbf{X} = \mathbf{x}\right) - \mathbb{E}\left(Z_t^0 \mid \mathbf{X} = \mathbf{x}\right)\mathbb{E}\left(Y_{ct}^0 \mid \mathbf{X} = \mathbf{x}\right),\end{aligned}$$

and the announced result follows from (21) and (22).

Note that when there is no zero-inflation phenomenon, that is to say, $\pi(\mathbf{x}, \boldsymbol{\beta}_{rt}) = \pi(\mathbf{x}, \boldsymbol{\beta}_{0t}) = 1$, we get back to the classical result: $\mathrm{ATE}^r(t, \mathbf{x}) = (\boldsymbol{\theta}_{rt} - \boldsymbol{\theta}_{0t})^\top \mathbf{x}$.

3.2.1 Estimation of the conditional average treatment effect

Suppose we have a sample $(Y_{it}, D_i^0, \ldots, D_i^{R-1}, \mathbf{x}_i)$ for $i = 1, \ldots, n$, where the observed value of the outcome Y_{it} can be written as

$$Y_{it} = \sum_{r=0}^{R-1} D_i^r Y_{it}^r. \tag{23}$$

We define the binary variable Z_{it} as $Z_{it} = 1$ if $Y_{it} \neq 0$ and $Z_{it} = 0$ if $Y_{it} = 0$.

For treatment r and time t, the vectors of parameters $\boldsymbol{\beta}_{rt}$ and $\boldsymbol{\theta}_{rt}$ can be obtained by minimizing the function $\Psi_n^r(\boldsymbol{\beta}, \boldsymbol{\theta}) = \Psi_{1n}^r(\boldsymbol{\beta}) + \Psi_{2n}^r(\boldsymbol{\theta})$, where

$$\Psi_{1n}^r(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n D_i^r \left[Z_{it} \ln\left(\frac{\pi(\boldsymbol{\beta}, \mathbf{x}_i)}{1 - \pi(\boldsymbol{\beta}, \mathbf{x}_i)}\right) + \ln\left(1 - \pi(\boldsymbol{\beta}, \mathbf{x}_i)\right)\right] \tag{24}$$

is the opposite of the log likelihood and where, as in Section 2.3, the conditional probability $\pi(\boldsymbol{\beta}_{rt}, \mathbf{x}) = \mathbb{P}[Z_t^r = 1 | \mathbf{X} = \mathbf{x}]$ is supposed to be of a probit or logit shape. Function $\Psi_{2n}^r(\boldsymbol{\theta})$ corresponds to the least squares criterion

$$\Psi_{2n}^r(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n Z_{it} D_i^r \left(Y_{it} - \mathbf{x}_i^\top \boldsymbol{\theta}\right)^2. \tag{25}$$

Assuming that $\sum_{i=1}^n D_i^r Z_{it} \mathbf{x}_i \mathbf{x}_i^\top$ is a full rank matrix (which is true with high probability, as seen in Section 3.2.2 under hypothesis (H_{3t})), function Ψ_{2n}^r has a unique minimizer

$$\widehat{\boldsymbol{\theta}}_{rt} = \left(\sum_{i=1}^n D_i^r Z_{it} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i=1}^n D_i^r Z_{it} \mathbf{x}_i Y_{it}\right). \tag{26}$$

Replacing the unknown parameters $\boldsymbol{\beta}_{rt}$, $\boldsymbol{\beta}_{0t}$, $\boldsymbol{\theta}_{rt}$, and $\boldsymbol{\theta}_{0t}$ in the expression of $\mathrm{ATE}^r(t,\mathbf{x})$ given in Proposition 3.1 by their estimators $\widehat{\boldsymbol{\beta}}_{rt}$, $\widehat{\boldsymbol{\beta}}_{0t}$, $\widehat{\boldsymbol{\theta}}_{rt}$, and $\widehat{\boldsymbol{\theta}}_{0t}$, we get the estimate

$$\widehat{\mathrm{ATE}}^r(t,\mathbf{x}) = \left(\pi(\mathbf{x},\widehat{\boldsymbol{\beta}}_{rt})\widehat{\boldsymbol{\theta}}_{rt} - \pi(\mathbf{x},\widehat{\boldsymbol{\beta}}_{0t})\widehat{\boldsymbol{\theta}}_{0t}\right)^\top \mathbf{x} \tag{27}$$

for the conditional average treatment effect at time t for treatment r.

3.2.2 Consistency of the estimated conditional treatment effect

We denote by $\pi_t^r(\mathbf{x}) = \mathbb{P}\left[D^r Z_t^r = 1 | \mathbf{X} = \mathbf{x}\right]$ the probability of receiving treatment r and that a variation of $Y^r(t)$ occurs, given $\mathbf{X} = \mathbf{x}$.

We assume that

$$\begin{array}{ll} (\mathrm{H}_{2t}) & \left(Y_t^0,\dots,Y_t^{R-1}\right) \perp \left(D^0,\dots,D^{R-1}\right) \mid \mathbf{X}; \\ (\mathrm{H}_{3t}) & \mathbb{E}\left(\pi_t^r(\mathbf{X})\mathbf{X}\mathbf{X}^\top\right) = \mathbf{Q}_t^r \text{ where } \mathbf{Q}_t^r \text{ is non-singular.} \end{array}$$

Condition (H_{2t}) is classical in the econometric literature on policy evaluation and multiple treatment effects. With (21), it implies that

$$(\epsilon_{0t},\dots,\epsilon_{R-1,t}) \perp \left(D^0,\dots,D^{R-1}\right) \mid \mathbf{X} \tag{28}$$

and

$$(Z_t^0,\dots,Z_t^{R-1}) \perp \left(D^0,\dots,D^{R-1}\right) \mid \mathbf{X}. \tag{29}$$

Note that if $D_i^r = 1$, we only observe Z_{it}^r for unit i in the sample. Hypothesis (H_{2t}) implies that the distribution of Z_t^r given $\mathbf{X}$ and $D^{r'}$ does not depend on $D^{r'}$, for $r' \in \{0,\dots,R-1\}$. Note that when (H_{2t}) is true and Model (19) holds, we have $\pi_t^r(\mathbf{x}) = \pi(\mathbf{x},\boldsymbol{\beta}_{rt})\mathbb{P}\left[D^r|\mathbf{X}=\mathbf{x}\right]$. The joint distribution can be expanded as follows:

$$\begin{aligned} \mathbb{P}\left[D^r = r_i, Y^r = y_i, \mathbf{X} = \mathbf{x}_i\right] &= \mathbb{P}[D^r = r_i, Y^r = y_i|\mathbf{X} = \mathbf{x}_i]\mathbb{P}[\mathbf{X} = \mathbf{x}_i] \\ &= \mathbb{P}[Y^r = y_i|\mathbf{X} = \mathbf{x}_i]\mathbb{P}[D^r = r_i|\mathbf{X} = \mathbf{x}_i]\mathbb{P}[\mathbf{X} = \mathbf{x}_i]. \end{aligned}$$

Condition (H_{3t}) is fulfilled under the classical assumption that $\mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ is a full rank matrix and $\pi_t^r(\mathbf{X}) > 0$ almost surely. This identifiability condition ensures the existence of a unique estimator $\widehat{\boldsymbol{\theta}}_{rt}$ when the sample size n is large enough. It also implies that $\mathbb{P}[D^r = 1] > 0$, for $r = 0,1,\dots,R-1$.

We can now state the two following propositions, which establish the consistency and asymptotic normality of the estimators of the parameters defined in models (19) and (21). Note that there is no need to impose that $\boldsymbol{\beta}_{rt}$ belongs to some compact space, thanks to the concavity in the parameters of the log likelihood for probit (see Newey and McFadden, 1994) and logistic (see Hjort and Pollard, 2011) regression models.

Proposition 3.2. *Suppose that models (23), (19), and (21) hold. Assume also that hypotheses* (H_{1t}), (H_{2t}), *and* (H_{3t}) *are fulfilled. Then as* n *tends to infinity,*

$$\widehat{\boldsymbol{\theta}}_{rt} - \boldsymbol{\theta}_{rt} = o_p(1)$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{rt} - \boldsymbol{\theta}_{rt}\right) \rightsquigarrow \mathcal{N}\left(0, \sigma_{rt}^2 \left(\mathbf{Q}_t^r\right)^{-1}\right).$$

Proposition 3.3. *If hypotheses* (H_{1t}), (H_{2t}), *and* (H_{3t}) *are fulfilled, if* $\pi\left(\boldsymbol{\beta}_{rt}, \mathbf{x}\right) = \mathbb{P}[Z_t^r = 1|\mathbf{X} = \mathbf{x}]$ *is of a probit or logit shape, and if* n *tends to infinity, we have*

$$\widehat{\boldsymbol{\beta}}_{rt} - \boldsymbol{\beta}_{rt} = o_p(1)$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{rt} - \boldsymbol{\beta}_{rt}\right) \rightsquigarrow \mathcal{N}\left(0, \boldsymbol{\Sigma}_t^r\right),$$

where the expression of the asymptotic covariance matrix $\boldsymbol{\Sigma}_t^r$ *is given in the proof.*

We deduce the following corollary from the previous propositions.

Corollary 3.4. *Under the assumptions of Proposition 3.2 and Proposition 3.3, as* n *tends to infinity and for all* $\mathbf{x} \in \mathbb{R}^p$ *we have*

$$\widehat{ATE}^r(t, \mathbf{x}) - ATE^r(t, \mathbf{x}) = o_p(1).$$

Furthermore, if $(\boldsymbol{\beta}_{rt}, \boldsymbol{\theta}_{rt}) \neq (\boldsymbol{\beta}_{0t}, \boldsymbol{\theta}_{0t})$,

$$\sqrt{n}\left(\widehat{ATE}^r(t, \mathbf{x}) - ATE^r(t, \mathbf{x})\right) \rightsquigarrow \mathcal{N}\left(0, \boldsymbol{\Delta}_t^r(\mathbf{x})\right)$$

for some covariance matrix $\boldsymbol{\Delta}_t^r(\mathbf{x})$.

The expression for the asymptotic variance $\boldsymbol{\Delta}_t^r(\mathbf{x})$ of $\widehat{\mathrm{ATE}}^r(t, \mathbf{x})$ can be obtained with the delta method. It is complicated and not given here. As in Section 2.3, paired bootstrap approaches are not difficult to employ and give reliable (and consistent) confidence intervals since the estimators are obtained as minimizers of $\Psi_n^r(\boldsymbol{\beta}, \boldsymbol{\theta})$, which is a twice differentiable convex functional.

4 A simulated example

To illustrate with a very simple example the effect of zero inflation on the expected value of the response variable and to check the ability of paired bootstrap procedures to produce reliable confidence intervals, we consider the following univariate model:

$$
\begin{aligned}
Y &= ZY_c + (1-Z)0, \\
Y_c &= \theta_0 + \theta X + \epsilon, \\
Z^* &= \beta_0 + \beta X + v, \\
Z &= \begin{cases} 1 & \text{when } Z^* > 0, \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

where X is uniformly distributed in the interval $[-2, 2]$ and the two error terms ϵ and v are independent, normally distributed random variables with mean 0 and variance $\sigma_\epsilon^2 = 0.5$ and $\sigma_v^2 = 1$. The constant terms θ_0 and β_0 are both equal to 1, while the slope parameters θ and β take different values corresponding to different scenarios.

Figure 2 considers alternative DGPs, which are obtained by allowing θ to take the following values: $2, 1, 0.6, 0.2, -0.2, -0.6, -1, -2$, while $\beta = 2$ and depicts the expected value arising from the DGPs described above, i.e.,

$$\mathbb{E}[Y|X = x] = \Phi\left(\beta_0 + \beta x\right) \times (\theta_0 + \theta x),$$

$\Phi(.)$ being the normal c.d.f function.[2] Figure 2 also depicts the predicted values obtained using either the zero-inflated model, which uses probit estimation for the conditional probability of not observing zero and OLS for the continuous part of the zero-inflated model, and the standard linear model, which is estimated using OLS and does not account for the zero-inflated feature of the data.

As far as the zero-inflated model is concerned, we also employ a non-parametric bootstrap procedure to build confidence intervals for the predicted value of the dependent variable Y. The algorithm is the following, considering $B = 1000$ bootstrap replications:

- Repeat for $b = 1$ to $b = B$
  - Draw from the initial sample a paired bootstrap sample $(Y_1^*, X_1^*), \ldots, (Y_n^*, X_n^*)$ with equal probability sampling with replacement.
  - Estimate, with probit and OLS, the conditional probability of not observing zero and the continuous part of the zero-inflated model, respectively, and then compute predicted value $\widehat{Y}^b = \pi(\mathbf{x}, \widehat{\boldsymbol{\beta}}^b)(\widehat{\boldsymbol{\theta}}^b)^\top \mathbf{x}$.

Then, non-parametric bootstrap confidence intervals with confidence $\alpha$are built by considering the quantiles of order $\alpha/2$ and $1 - \alpha/2$ for the estimated expected value.

[2] Additional results obtained by considering other values for β, are available upon request.

===== Figure 2 =====

It clearly appears from the plots in Figure 2 that the zero-inflated phenomenon can produce very different functional relations depending on the parameters β and θ. When β and θ have the same sign the relation is monotonic, otherwise, when they have opposite signs, the resulting relation can also be non-monotonic. Clearly, as β (resp. θ) gets closer to 0 the resulting relation approaches linearity (resp. a probit shape).

As far as the estimation is concerned, the proposed estimator, which handles the zero inflation, provides a very faithful description of the underlying DGP. Additionally, the true underlying relation is always within the bootstrapped bands, which closely follow the DGP. In contrast, the linear model always provides misleading results, and in particular, when the underlying relation is non-monotonic it clearly provides a senseless fit.

By considering samples with moderate sizes, $n = 200$, we also evaluate the ability of the bootstrap procedure to build reliable confidence intervals. Results are plotted in Figure 3, for a nominal level of $1 - \alpha = 0.95$. We note that irrespective of the values of the parameters β and θ, the empirical coverages are most often very close to the nominal ones. The only exception is when X is in the range between -2 and -1, where the empirical coverage is about 0.9. Overall, these results offer evidence of the validity of the non-parametric bootstrap in the proposed zero-inflated framework even in the case of a very small sample size.

===== Figure 3 =====

5 Empirical illustrations

The purpose of this section is to provide two empirical illustrations of the proposed approach. We first revisit the problem of estimating a classical Mincer wage equation with zero-inflated data and exploit the panel data of Baltagi and Khanti-Akom (1990). As a second illustrative example, we consider DID estimation of the effects of two distinct public policies that were devoted to boosting rural development in France over a similar period of time and were recently investigated in Cardot and Musolesi (2020). In both cases, differencing over time to eliminate the unobserved component produces a zero-inflated phenomenon that can be described by equations (1) and (3).

5.1 Revisiting the Mincer wage equation with zero-inflated data

In this subsection, we revisit the classical problem of estimating a wage equation with panel data and then provide evidence that the standard FD estimator suffers from a bias due to the zero-inflated phenomenon that was previously discussed and that this bias is sizeable despite the fraction of observations equal to zero being relatively small.

We consider the dataset described in Baltagi and Khanti-Akom (1990), which corresponds to a panel of 595 individuals observed over the 1976–1982 period and is drawn from the Panel Study of Income Dynamics. A Mincer wage equation (Mincer, 1974) is adjusted, with the logarithm of earnings, $\log(WAGE)$, modeled as the sum of a linear function of years of education (EDU) and a quadratic function of full-time work experience (EXP), and the model is extended by considering additional explanatory variables. This set of variables includes the number of weeks worked (WKS) and some dummy variables: occupation ($OCC = 1$ if the individual is a blue-collar worker), industry ($IND = 1$ if the individual works in manufacturing), geographical location ($SOUTH = 1$ and $SMSA = 1$ if the individual resides in the South and in a metropolitan area, respectively), marital status ($MS = 1$ if the individual is married), union coverage ($UNION = 1$), sex ($FEM = 1$ if the individual is female), and race ($BLK = 1$ if the individual is Black).

The Mincer wage equation is the cornerstone of a huge literature in empirical economics, probably because it is derived from a theoretical model of schooling choice and post-schooling training decisions, because it is simple enough, and because it captures reality quite well (Card, 1999). Previous studies have extensively discussed the empirical validity of this specification and its implications (Heckman et al., 2006). A relevant debate has emerged regarding functional form and the adoption of a quadratic form for experience. In particular, according to Murphy and Welch (1990) the quadratic specification provides a poor approximation of the underlying concave function as it overstates initial earnings, overstates earnings at mid-career, and understates earnings at retirement; using higher-order polynomial functions was subsequently proposed (Lemieux, 2006). More recently, studies adopting non-parametric regression models have provided further interesting insights. Henderson and Souto (2018) provide evidence of a concave but monotonic relation using both splines and kernels, which is consistent with the main findings of Murphy and Welch (1990). Su et al. (2009) further criticize the Mincer specification because of the inadequacy of a simple linear separable model to capture more complex and more realistic non-linear interactive effects.

To estimate the model, we adopt the framework given in equations (10) and (11) and then apply the FD estimator. This allows for arbitrary correlation between the unobserved effects and the observed explanatory variables but does not allow for identifying the effect of education and other time-invariant variables. Therefore, as an illustrative example we focus our attention on the effect of experience.

As far as functional form is concerned, we adopt a log-log specification. This choice provides a number of relative advantages. First, it allows for the identification of the parameter of work experience when time dummies are introduced into the model, while the log-level specification does not. Indeed, while $EXP_{it} = a_i + t$ is perfectly collinear with respect to the time dummies, $\log(EXP_{it})$ is not as $\log(a_i + t) \neq \log(a_i) + \log(t)$. Second, the log-log specification also encompasses a variety of non-linear relations between $WAGE$ and EXP, and in particular, it may allow for a decreasing marginal return of experience. We are not claiming that the log-log model provides the best approximation to the underlying function, but we adopt it because it is consistent with a concave and monotonic relation, as suggested by the literature discussed above, and it is simple enough for our illustration purposes.

As previously discussed, we apply the FD estimator, such that the estimating equation can be written as

$$\begin{aligned}\log\left(WAGE_{it}\right) - \log\left(WAGE_{it-1}\right) = {} & \left(\eta_t - \eta_{t-1}\right) + \theta_1\left(\log(EXP_{it}) - \log(EXP_{it-1})\right) + \\ & + \theta_2\left(\log(WKS_{it}) - \log(WKS_{it-1})\right) + \theta_3\left(OCC_{it} - OCC_{it-1}\right) + \\ & + \theta_4\left(IND_{it} - IND_{it-1}\right) + \theta_5\left(SOUTH_{it} - SOUTH_{it-1}\right) + \\ & + \theta_6\left(SMSA_{it} - SMSA_{it-1}\right) + \theta_7\left(MS_{it} - MS_{it-1}\right) + \\ & + \theta_8\left(UNION_{it} - UNION_{it-1}\right) + \left(\epsilon_{it} - \epsilon_{it-1}\right). \qquad (30)\end{aligned}$$

As far as the data are concerned, it is interesting to note that while the response variable in levels, $\log\left(WAGE_{it}\right)$, takes only positive values and can be supposed to be continuous, the first-difference variable, $\log\left(WAGE_{it}\right) - \log\left(WAGE_{it-1}\right)$, can no longer be considered to be continuous since we observe that around 6.7% of the observations have a null value. As displayed in Figure 4, a mixture model combining a mass at zero and a continuous distribution is more appropriate to describe that distribution.

===== Figure 4 =====

The estimation results are presented in Table 1. In column (i), we give the estimated values of the parameters considering the standard FD estimator. This estimator assumes a continuous density function, however, and under the zero-inflated phenomenon described by (1) and (3), it is generally a biased estimator of $\pi\boldsymbol{\theta}$ unless $\pi(\mathbf{x}_t, \boldsymbol{\beta}) = \pi$ does not depend on $\mathbf{x}_t$. As we will see below, in this empirical application the conditional probability of observing zero is significantly affected by some of the explanatory variables that are in the continuous part of the model.

Our main goal is to recover partial effects (PEs) and average partial effects (APEs) (see (16)) of the considered zero-inflated model, which is intrinsically non-linear. The vector of unknown parameters $\boldsymbol{\theta}$ is estimated by applying the subset estimator (5) for the FD equation (column (ii)), while the conditional probability $\pi(\mathbf{x}_t, \boldsymbol{\beta})$ and the partial effects from the binary model $\frac{\partial \pi(\mathbf{x}_t, \boldsymbol{\beta})}{\partial \mathbf{x}_t}$ can be obtained by adopting either a probit or a logit regression model.

When estimating a binary response regression model with panel data, one would ideally estimate the quantities of interest without putting restrictions on the conditional distribution of the unobserved effects given the explanatory variables, $D\left(c_i \mid \mathbf{X} = \mathbf{x}_i\right)$. However, the standard fixed effects approach that consists in viewing the components c_i as parameters to be estimated provides inconsistent estimates of the parameters for a fixed T and a sample size n growing to infinity, because of the incidental parameter problem (Neyman and Scott, 1948).[3] Interestingly, only in the logit case only is it possible to allow c_i and $\mathbf{x}_i$ to be arbitrarily related, by adopting a similar strategy that is used in the linear framework to eliminate c_i from the estimating equation. This approach leads to considering a conditional maximum likelihood estimator (CMLE). Unfortunately, PEs are not identified. Therefore, we instead consider a correlated random effects (CRE) framework (see, for example, the seminal work by Mundlak, 1978), which places some restrictions on $D\left(c_i \mid \mathbf{x}_i\right)$, and adopt the Chamberlain CRE probit model (Chamberlain, 1980). Wooldridge (2010) proposes both a joint and a pooled MLE. We specifically adopt the pooled MLE, which is a simple probit model supplemented with time averages of the continuous explanatory variables. Moreover, while the MLE is not robust to the violation of the conditional independence assumption, meaning that serial independence of the idiosyncratic shocks is needed for consistency, the pooled MLE is robust to such a violation, serial dependence can be handled by standard robust inference, and obtaining PEs is straightforward.

===== Table 1 =====

The results are as follows. When considering the standard FD estimator and assuming a continuous density function (i), the estimated coefficient of $\log(EXP_{it})$ is close to 0.2, suggesting a concave monotonic wage–experience relation (i.e., diminishing returns to experience). This result is broadly consistent with the above-cited literature, which mainly exploits cross-sectional

[3] Fernández-Val and Weidner (2016) propose bias corrections for panels where both n and T are moderately large.

data. Comparing this result with that obtained without including the time effects may provide some interesting insight into the possible bias that arises because of the omission of time-related factors. In that case, the estimated coefficient of $\log(EXP_{it})$ increases up to 0.82, indicating an almost linear wage–experience relation and suggesting a sizeable omitted common factors bias. When the model does not contain time effects, we can also apply the FD estimator to a typical Mincer log-level equation that contains experience and its square as regressors instead of the logarithm of experience. In this case, the FD estimator provides estimates of the coefficients of experience and of its square equal to 0.116 and -0.0005, respectively, which suggests an unsatisfactorily increasing exponential relation between wage and experience, thus reinforcing the idea that including time effects in the econometric specification is of crucial empirical relevance.

However, the standard FD estimator assuming an underlying continuous response may suffer from a bias because of the zero-inflation phenomenon. From the probit regression Model (iii), it emerges that $\log(EXP_{it})$ also significantly affects the conditional probability $\pi(\mathbf{x}, \boldsymbol{\beta})$, i.e., the conditional probability of observing zero (i.e., a null variation in wages), with an estimated APE equal to -0.08, thus implying that the naive FD estimator is a biased estimator of $\pi\boldsymbol{\theta}$. From the probit model, we can also observe that other factors have a significant effect. These factors are IND, UNION, and FEM, all positively affecting the conditional probability with estimated ATEs equal to 0.024, 0.044, and 0.032, respectively. Estimating the probit Model (iii) not only provides the basis for the computation of the PEs of the zero-inflated model but also gives interesting insight from an economic viewpoint.

We finally compute the PE of $\log(EXP)$ according to (16). It is found that the proposed mixture model provides a PE of $\log(EXP)$ ranging from 0.104 to 0.181, with an APE equal to 0.168. The kernel density estimate of such a PE is depicted in Figure 5.

===== Figure 5 =====

These results suggest i) a sizeable overestimation of the APE when erroneously adopting a standard FD approach and that, ii) in any case, assuming an underlying continuous density function does not allow capturing the heterogeneity of the PE that is due to the zero inflation.

5.2 Estimating ATEs with zero-inflated data: rural development policies in France

5.2.1 Description of the programs, data, and ATEs of interest

In France, as in other countries, enterprise-zone programs have been implemented to boost job creation. Such policies are based on fiscal incentives to firms located in deprived areas. Specifically designed to boost employment in rural areas, the ZRR (*Zones de Revitalisation Rurale)* program started the 1st of September, 1996, and covered the 1996–2004 period. At a supranational level, territorial cohesion, convergence, and a harmonious development across regions are among the objectives the European Union tries to pursue through these structural funds. Specifically devoted to boosting rural development, the objective 5B programs (1991–1993 and 1994–1999) allocated financial subsidies to firms and public actors located in eligible "*rural areas in decline*". A notable feature of both programs is that the selection process of the treated units was clearly not random, and sources of selection on both observables and unobservables are expected to be relevant.

Municipalities, which correspond to the finest-available spatial level, are the statistical units of analysis, and the dependent variable U_{it} is the number of employees at time t. This variable has been observed over a period of ten years, from 1993 to 2002. As policy variables, we use ZRR zoning during the period and 5B zoning over the 1994–1999 period. The set of confounding variables comes from the French census of 1990. We specifically gather information on demographics, education, and work qualifications aggregated at the municipality level. We also have at hand information on land use, obtained thanks to satellite images that were also taken in 1990. These variables are indicated as relevant by the related literature on local employment growth. We consider pre-treatment covariates to ensure that D causes $\mathbf{X}$ and U causes $\mathbf{X}$ does not occur (Lechner, 2011; Lee, 2005). Another relevant variable that is worth mentioning is the initial level of employment. Including the initial outcome as a regressor implies assuming unconfoundedeness given a lagged outcome. This inclusion avoids an omitted variable bias, which would be particularly relevant if the average outcome of the treated and control groups differ substantially in the first period (Imbens and Wooldridge, 2009), as in this case. The sample is made up of $n = 25,593$ municipalities.

We focus on the assessment of ZRR and 5B as well as their joint effect and thus adopt a framework with $R = 4$ multiple potential outcomes. These potential outcomes are associated with the potential treatments $\{0, ZRR, 5B, ZRR\&5B\}$ indicating the program in which each municipality actually participated. The modality 0 indicates that the municipality was not endowed with either policy measure, whereas ZRR (respectively, $5B$) indicates that the municipality received incentives only from the ZRR initiative (respectively, only from the 5B initiative) and $ZRR\&5B$ indicates that the municipality received incentives from both ZRR and 5B. Specifically, we focus on the estimation of the following ATEs:

$$\mathrm{ATE}^{5B}(t,\mathbf{x}) = \mathbb{E}\left(Y_t^{5B} - Y_t^0|\mathbf{X}=\mathbf{x}\right),$$
$$\mathrm{ATE}^{ZRR\&5B}(t,\mathbf{x}) = \mathbb{E}\left(Y_t^{ZRR\&5B} - Y_t^0|\mathbf{X}=\mathbf{x}\right).$$

As far as the effect of ZRR is concerned, it can be noted that only a few municipalities (precisely 722) are treated. Consequently, we prefer to focus our attention on the 7014 municipalities that received incentives both from 5B and ZRR, and we calculate the following differential effect:

$$\mathrm{ATE}^{ZRR}(t,\mathbf{x}) = \mathbb{E}\left(Y_t^{ZRR\&5B} - Y_t^{5B}|\mathbf{X}=\mathbf{x}\right).$$

This differential effect simply represents the expected difference between the outcome when a municipality receives incentives both from ZRR and 5B and when it receives incentives only from 5B.

As for the pre-treatment period t_0, we set $t_0 = 1993$, which is before the introduction of both policies. When setting t, in principle we could use all of the available information in the data. In particular, by setting $t = 1994, 1995$ we could conduct placebo tests on ZRR, which was introduced in 1996, and use the remaining time periods, $t = 1996, ..., 2002$, to estimate the temporal treatment effects for ZRR and 5B as well as their interaction, as in Cardot and Musolesi (2020). With the aim of providing an illustration of the proposed approach, we set $t = 1999$, which is the last time period under the 5B program.

5.2.2 Estimation results and comparison with the continuous response model

The statistical units under study are generally demographically small, and for a non-negligible fraction of the municipalities we observe no variation whatsoever in the dependent variable, i.e., local employment along time: $U_{it} = U_{it_1}$. Descriptive statistics in the Appendix of Cardot and Musolesi (2020) show that the modal value of $U_{it} - U_{it_1}$ is indeed 0 for all values of t, with t varying between 1994 and 2002. We can also remark that the fraction of zeros decreases with t and varies with the treatment status. The distribution of the dependent variable, $U_{it} - U_{it_1}$, can be considered to be a mixture of a mass at 0 and a continuous density function, as depicted in Figure 1. In particular, for $t = 1999$ the fraction of municipalities for which $U_{it} = U_{it_1}$ varies between 9% when there is no treatment and 15% under treatment $ZRR\&5B$.

We compare the estimated values of the ATE defined in (27) obtained with the proposed mixture approach, which combines information from both the continuous and the discrete parts of the model, with those obtained with a naive method that does not account for the mass at zero and only assumes a continuous response model (Imbens and Wooldridge, 2009). This may provide relevant insight into the size of the bias when neglecting the zero-inflation feature of the data. We consider alternative specifications for the regression function (21), which are presented in more detail below. The estimation results are presented in Table 2.

We first follow a common practice in the econometric literature that consists in adopting a linear specification for the confounding variables and assuming that only the intercept varies between treated groups, while the slope parameters do not (Model (i)). This is a simple extension of the DID estimator that allows for temporal policy effects and takes account of linear effects of the initial conditions (Abadie, 2005). We consider the same set of variables as in Cardot and Musolesi (2020), who employ a backward variable selection procedure to select the variables to be introduced in the regression functions. We then consider more flexible models. In the second model (Model (ii)), because the linearity assumption is strong and a misspecification of the relation between $Y^r(t)$ for $r \in \{0, ZRR, 5B, ZRR\&5B\}$ and the regressors may lead to incorrect results and a misinterpretation of the policy effect, we allow for non-linear effects of the confounding variables. This is achieved by adopting natural cubic regression splines, i.e., piecewise-cubic splines with the constraint that they are linear in their tails beyond the boundary knots, which are generally preferred to cubic splines because of less problematic edge effects (Harrell Jr, 2015). This also makes the underlying identification conditions less restrictive (Lechner, 2011). Finally, in the third model (Model (iii)) we rely on a linear regression model, but it is assumed that both the intercepts and the slope parameters of some confounding variables vary between treated groups (see, for example, Heckman and Hotz, 1989, eq. 3.9). Following Cardot and Musolesi (2020), we retain only two significant interactions of the policy variable: the first one with the initial level of employment (variable `SIZE`) in the municipality and the second one with its population density (variable `DENSITY`).

In order to build confidence intervals, we consider the flexible non-parametric bootstrap approach (see, for example, Efron and Tibshirani, 1993) to approximate the distribution of the conditional counterfactual outcome of each municipality i having the characteristic $\mathbf{X} = \mathbf{x}$. We draw $B = 1000$ bootstrap samples, and for each bootstrap sample b, with $b = 1, \dots B$, we make the following estimation of the ATE (see (27)):

$$\widehat{\text{ATE}}^{r,b}(t, \mathbf{x}) = \left(\pi(\mathbf{x}, \widehat{\boldsymbol{\beta}}^b_{rt}) \widehat{\boldsymbol{\theta}}^b_{rt} - \pi(\mathbf{x}, \widehat{\boldsymbol{\beta}}^b_{0t}) \widehat{\boldsymbol{\theta}}^b_{0t} \right)^\top \mathbf{x}.$$

Bootstrap confidence intervals are then deduced using the percentile method.

Average treatment effects When comparing the proposed conditional mixture model with the naive DID model, it can be noted in Table 2 that accounting for a mass of observations at zero increases the estimated ATEs by about 5%–10%. This happens for the three specifications considered (Models (i), (ii), and (iii)), providing robust evidence that accounting for the mass of observations at zero is important to avoid a significant underestimation of the average effect of the policies. It is interesting to note that here we face a moderate zero-inflated phenomenon, with less than 15% of the observations being equal to zero. In many other empirical frameworks, the mass of observations at zero can be more important, so that even more severe biases are likely to appear when assuming a continuous response model. Although this is outside of the scope this paper, these results also indicate that the estimated ATEs are very sensitive to the functional form and that this issue should be carefully addressed.

===== Table 2 =====

Distributional treatment effects The results discussed above hide another important feature of the proposed zero-inflated model. Indeed, a relevant consequence of the model described in equations (1) and (3) is that even though it is assumed that only the intercept varies between treated groups, while the slope parameters do not, as in Models (i) and (ii) the resulting treatment effects are heterogeneous across individuals. Distributional treatment effects are reported in Table 3.

===== Table 3 =====

First, by focusing on Models (i) and (ii) it can be noted that when handling the zero-inflated phenomenon the estimated treatment effects vary greatly across units, with the estimated treatment effects for the 99th percentile often being more than twice those of the 1st percentile, while the estimated treatment effects based on a continuous response model obviously do not vary across units. When focusing on Model (iii), we can note that the estimated treatment effects vary even more than those obtained from Models $(i) - -(ii)$ and that the distribution of the estimated treatment effects is similar when comparing the two estimators. This, however, does not ensure that at an individual level the two approaches provide similar estimates. With the aim of highlighting possible individual differences between the estimates obtained with the two methods, we build a new variable defined as the relative change between the treatment effect obtained from the zero-inflated approach ($\widehat{tez_i^r}$) and that obtained from the naive estimator ($\widehat{ten_i^r}$). The variable is defined as $\widehat{rc_i^r} = \left(\widehat{tez_i^r} - \widehat{ten_i^r}\right)/\widehat{ten_i^r}$, the estimated density functions of which—with bandwidths selected using biased cross-validation—are depicted in Figure 6. For Models (i) and (ii), all the estimated densities are left-skewed, with the mode around 0.15–0.2. For Model (iii), the estimated densities are rather symmetric, with bimodal shapes in two cases out of three. Overall, these results highlight that when focusing on distributional treatment effects (rather than only focusing on the mean effect), the naive estimator faces a sizeable bias and the sign of this bias can be either positive or negative.

===== Figure 6 =====

6 Conclusion

In this paper, we introduce a statistical model combining a continuous response regression model, which can take either positive or negative values, and a mass at zero.

This type of zero-inflated model can be appropriate in situations when the interest lies in modeling outcome variation over time. This includes unobserved effects panel data models, difference-in-differences treatment effect estimation, and cross-sectional models when studying growth rates or temporal evolution as a function of explanatory variables observed at a given point in time.

We first provide a mathematical formalization by means of conditional mixtures. In particular, it is shown that when this kind of zero-inflated phenomenon occurs, the classical OLS estimator is generally biased. We then propose a subset estimator based on the subsample of units for which the dependent variable has non-null values, and we derive its asymptotic properties under a conditional independence assumption. Such an estimator can be used, along with a binary response model for the conditional probability of facing a mass at zero, to compute the partial effects arising from zero-inflated regression models. We prove the asymptotic normality of the estimator as well as consistency of the empirical bootstrap. Then, we move specifically to unobserved effects panel data models, focusing on the identification conditions of the unknown parameters, on the computation of partial effects, and on the ability of the paired bootstrap to provide consistent confidence intervals in this extended framework. We also address the problem of DID estimation under zero inflation and propose an estimator of the ATE that is proven to be consistent.

We also bring new evidence based both on simulated and real data. The simulated example illustrates the effect of zero inflation on the expected value of the response variable, and it clearly shows that the zero-inflated phenomenon can produce very different functional relations that depend on the underlying parameters and that the linear model fails to provide a faithful description of the underlying DGP. The simulation study also provides evidence of the validity of non-parametric paired bootstrapping with small samples. Finally, by using real data we first revisit a classical Mincer wage equation and then address the problem of estimating the ATE of two distinct public policies that were devoted to boosting rural development in France. In both cases, the estimation results provide additional insight into the usefulness of the proposed estimator and also indicate that commonly used regression models, which are based on the assumption that the response variable is continuous, may face a sizeable bias with respect to average effects and that, in any case, assuming an underlying continuous density function does not allow for capturing the heterogeneity of PEs that arises because of the non-linear shape of the zero-inflation model.

The present work could be extended in many directions. For instance, further studies may consider instrumental variables estimation under zero inflation or may focus on more flexible non-parametric regression models. These interesting extensions are outside the scope of this paper and certainly deserve further investigation.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

Baltagi, B. H. and S. Khanti-Akom (1990). On efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied econometrics 5*(4), 401–406.

Bose, A. and S. Chatterjee (2003). Generalized bootstrap for estimators of minimizers of convex functions. *J. Statist. Plann. Inference 117*(2), 225–239.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics 3*, 1801–1863.

Cardot, H. and A. Musolesi (2020). Modeling temporal treatment effects with zero inflated semi-parametric regression models: the case of local development policies in france. *Econometric Reviews 39*, 135–157.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The review of economic studies 47*(1), 225–238.

Chamberlain, G. (1984). Panel data. *Handbook of econometrics 2*, 1247–1318.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap*, Volume 57 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, New York.

Fernández-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics 192*(1), 291–312.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Han, S. (2021). Identification in nonparametric models for dynamic treatment effects. *Journal of Econometrics 225*(2), 132–147.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer.

Heckman, J. and V. Hotz (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *J. Amer. Statist. Assoc. 84*, 862–874.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998, September). Characterizing Selection Bias Using Experimental Data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies 64*(4), 605–654.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. *Handbook of the Economics of Education 1*, 307–458.

Henderson, D. J. and A.-C. Souto (2018). An introduction to nonparametric regression for labor economists. *Journal of Labor Research 39*(4), 355–382.

Hjort, N. L. and D. Pollard (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.

Imbens, G. W. and J. M. Wooldridge (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature 47*(1), 5–86.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics 4*(3), 165–224.

Lechner, M. (2015). Treatment effects and panel data. In B. Baltagi (Ed.), *The Oxford Handbook of Panel Data*. Oxford University Press.

Lee, M.-J. (2005). *Micro-econometrics for policy, program, and treatment effects*. Oxford University Press on Demand.

Lee, M.-j. and C. Kang (2006). Identification for difference in differences with cross-section and panel data. *Economics letters 92*(2), 270–276.

Lemieux, T. (2006). The "mincer equation" thirty years after schooling, experience, and earnings. In *Jacob Mincer a pioneer of modern labor economics*, pp. 127–145. Springer.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data* (Second ed.). Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Mincer, J. (1974). *Schooling, experience and earnings*. Columbia University Press for National Bureau of Economic Research, New York.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Murphy, K. M. and F. Welch (1990). Empirical age-earnings profiles. *Journal of Labor economics 8*(2), 202–229.

Newey, K. and D. McFadden (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1–32.

Sala-i Martin, X. (1997). I just ran two million regressions. *The American Economic Review 87*(2), 178–183.

Sheather, S. J. (2004). Density estimation. *Statistical Science 19*(4), 588–597.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Su, L., Y. Chen, and A. Ullah (2009). Functional coefficient estimation with both categorical and continuous data. In *Nonparametric Econometric Methods.* Emerald Group Publishing Limited.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random coefficient and treatment-effect panel data models. *The Review of Economics and Statistics 87*, 395–390.

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

A Appendix : proofs

Proof. of Proposition 2.2.
Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *we have*

$$\begin{aligned}
\mathbb{E}[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1, \ldots, \mathbf{x}_n] &= \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \mathbb{E}[\mathbf{Y}|\mathbf{x}_1, \ldots, \mathbf{x}_n] \\
&= \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \overline{\boldsymbol{\pi}} \mathbb{E}[\mathbf{Y}_c|\mathbf{x}_1, \ldots, \mathbf{x}_n] \\
&= \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \overline{\boldsymbol{\pi}} \widetilde{\mathbf{X}} \boldsymbol{\theta} \qquad (31)
\end{aligned}$$

where $\overline{\boldsymbol{\pi}}$ *is the* $n \times n$ *diagonal matrix with diagonal elements* $(\pi(\mathbf{x}_1, \boldsymbol{\beta}), \ldots, \pi(\mathbf{x}_n, \boldsymbol{\beta}))$*.*

If $\pi(\mathbf{x}_i, \boldsymbol{\beta})$ *does not depend on* $\mathbf{x}_i$ *then previous expression simplifies to*

$$\mathbb{E}[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1, \ldots, \mathbf{x}_n] = \pi \boldsymbol{\theta}.$$

For the conditional variance, we have

$$\begin{aligned}
\mathbb{V}ar\left[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1, \ldots, \mathbf{x}_n\right] &= \mathbb{E}\left[\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^\top|\mathbf{x}_1, \ldots, \mathbf{x}_n\right] - \mathbb{E}[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1, \ldots, \mathbf{x}_n]\left(\mathbb{E}[\widetilde{\boldsymbol{\theta}}|\mathbf{x}_1, \ldots, \mathbf{x}_n]\right)^\top \\
&= \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\top|\mathbf{x}_1, \ldots, \mathbf{x}_n\right] \widetilde{\mathbf{X}} \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} - \pi^2 \boldsymbol{\theta}\boldsymbol{\theta}^\top \\
&= \pi \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top \left(\sigma^2 \mathbf{I}_n + \widetilde{\mathbf{X}}\boldsymbol{\theta}\boldsymbol{\theta}^\top \widetilde{\mathbf{X}}^\top\right) \widetilde{\mathbf{X}} \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} - \pi^2 \boldsymbol{\theta}\boldsymbol{\theta}^\top \\
&= \pi \left(\sigma^2 \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\right)^{-1} + (1-\pi)\boldsymbol{\theta}\boldsymbol{\theta}^\top\right)
\end{aligned}$$

since $\mathbf{Y} = (Z_1(\mathbf{x}_1^\top \boldsymbol{\theta} + \epsilon_1), \ldots, Z_n(\mathbf{x}_n^\top \boldsymbol{\theta} + \epsilon_n))$ *and* $\mathbb{E}[Y_i Y_j|\mathbf{x}_1, \ldots, \mathbf{x}_n] = \pi\left(\mathbf{x}_i^\top \boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{x}_j + \sigma^2 \mathbf{1}_{\{i=j\}}\right)$*.* □

Proof. of Proposition 2.3
Consider the random matrix $Z\mathbf{X}\mathbf{X}^\top$*. We have,*

$$\begin{aligned}
\mathbb{E}[Z\mathbf{X}\mathbf{X}^\top] &= \mathbb{E}\left(\mathbb{E}[Z|\mathbf{X}]\ \mathbf{X}\mathbf{X}^\top\right) \\
&= \mathbb{E}\left(\pi(\mathbf{X}, \boldsymbol{\beta})\mathbf{X}\mathbf{X}^\top\right) \\
&= \mathbf{Q}_\pi,
\end{aligned}$$

and the weak law of large numbers gives us

$$\left(\frac{1}{n}\sum_{i=1}^n Z_i \mathbf{x}_i \mathbf{x}_i^\top\right) - \mathbf{Q}_\pi = o_p(1).$$

The application of the continuous mapping theorem (see van der Vaart (1998), Theorem 2.3) together with assumptions (H_2) *which implies that inversion is continuous in a neighborhood of* $\mathbf{Q}_\pi$ *gives us*

$$\left(\frac{1}{n}\sum_{i=1}^n Z_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} - \mathbf{Q}_\pi^{-1} = o_p(1). \qquad (32)$$

Consider now the random vector $ZY\mathbf{X}$*. We have, with assumption* (H_1) *under model (3)*

$$\begin{aligned}
\mathbb{E}\left[ZY\mathbf{X}\right] &= \mathbb{E}\left([Z|\mathbf{X}]\,\mathbb{E}\left[Y_c|\mathbf{X}\right]\mathbf{X}\right) \\
&= \mathbb{E}\left[\pi(\boldsymbol{\beta}, \mathbf{X})\mathbf{X}\mathbf{X}^\top\right]\boldsymbol{\theta}, \\
&= \mathbf{Q}_\pi \boldsymbol{\theta},
\end{aligned}$$

and the weak law of large numbers gives us

$$\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i \mathbf{x}_i - \mathbf{Q}_\pi \boldsymbol{\theta} = o_p(1). \tag{33}$$

The classical decomposition, together with (32) and (33), permits to prove the first result,

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_c - \boldsymbol{\theta} &= \left(\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} - \mathbf{Q}_\pi^{-1}\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i \mathbf{x}_i\right) - \mathbf{Q}_\pi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i \mathbf{x}_i - \mathbf{Q}_\pi \boldsymbol{\theta}\right)\\ &= o_p(1). \end{aligned} \tag{34}$$

To get the asymptotic normality consider now the decomposition (5) of $\widehat{\boldsymbol{\theta}}_c$. The random vectors $(Z_i\epsilon_i\mathbf{x}_i)$, $i = 1,\dots,n$ are i.i.d, with expectation 0 when (H_1) is true, and variance-covariance matrix

$$\begin{aligned}\mathbb{V}ar\,(Z\epsilon\mathbf{X}) &= \mathbb{E}\left(Z\epsilon^2\mathbf{X}\mathbf{X}^\top\right)\\ &= \mathbb{E}\left(\mathbb{E}[Z|\mathbf{X}]\ \mathbb{E}[\epsilon^2|\mathbf{X}]\ \mathbf{X}\mathbf{X}^\top\right)\\ &= \sigma^2\mathbf{Q}_\pi.\end{aligned}$$

We deduce from the central limit theorem that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\epsilon_i\mathbf{x}_i\right) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\mathbf{Q}_\pi\right) \tag{35}$$

and with (5), (32) and Slutsky's Lemma (see van der Vaart (1998), Proposition 2.8),

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}\right) &\rightsquigarrow \mathbf{Q}_\pi^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\epsilon_i\mathbf{x}_i\right)\mathbf{Q}_\pi^{-1}\\ &\rightsquigarrow \mathcal{N}\left(0, \sigma^2\mathbf{Q}_\pi^{-1}\right).\end{aligned}$$

□

Proof. of Proposition 2.4.
The proof is based on successive applications of the weak law of large numbers, the continuous mapping theorem and Slustky's Lemma. First note that $n^{-1}\sum_{i=1}^n Z_i - \mathbb{E}[\pi(\boldsymbol{\beta},\mathbf{X})] = o_p(1)$ and since $\mathbb{E}[\pi(\boldsymbol{\beta},\mathbf{X})] > 0$, we deduce that

$$\frac{n}{\sum_{i=1}^n Z_i} - \frac{1}{\mathbb{E}[\pi(\boldsymbol{\beta},\mathbf{X})]} = o_p(1). \tag{36}$$

We also have $\left(Y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\theta}}_c\right)^2 = \left(\epsilon_i + \mathbf{x}_i^\top(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_c)\right)^2$ and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\theta}}_c\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} Z_i\epsilon_i^2\\ &\quad + \left(\frac{2}{n}\sum_{i=1}^{n} Z_i\epsilon_i\mathbf{x}_i^\top\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_c\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_c\right)^\top\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\mathbf{x}_i\mathbf{x}_i^\top\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_c\right) \qquad (37)\\ &= \sigma^2\mathbb{E}[\pi(\boldsymbol{\beta},\mathbf{X})] + o_p(1), \qquad (38)\end{aligned}$$

because the weak law of large numbers gives us $\frac{1}{n}\sum_{i=1}^n Z_i\epsilon_i^2 - \sigma^2\mathbb{E}[\pi(\boldsymbol{\beta},\mathbf{X})] = o_p(1)$ and it can be checked easily that the two terms at the righthand side of (37) tend to zero in probability. We finish the proof by using the decomposition (32), the convergence in probability given in (36) and (38) and an application of Slutsky's Lemma. □

Proof. of Proposition 2.5.
The first part of the Proposition is a direct consequence of Theorem 2.1 and Theorem 2.4 in Bose and Chatterjee (2003), remarking that if we assume that the link function for $\pi(\boldsymbol{\beta}, \mathbf{x})$ has a logit or probit shape, the objective function $\Psi_n(\boldsymbol{\beta}, \boldsymbol{\theta}; (Y_1, \mathbf{x}_1), \dots, (Y_n, \mathbf{x}_n))$ is a convex function, of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ that is also twice differentiable. The Hessian matrix is positive definite at the true value of the parameter $(\boldsymbol{\beta}, \boldsymbol{\theta})$ thanks to hypothesis (H_2).

The second part of the proof is a direct consequence of the delta method for bootstrapped estimates (see Theorem 23.9 in van der Vaart (1998)) considering the function $\pi(\boldsymbol{\beta}, \mathbf{x})\boldsymbol{\theta}^\top \mathbf{x}$, which is differentiable with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta})$. □

Proof. of Proposition 3.2.
We follow the same lines as in the proof of Proposition 2.3 and thus omit some details. With (H_{2t}), the law of large numbers, the continuity of the inverse application for non-singular matrices, and hypothesis (H_{3t}), we have that

$$\left(\frac{1}{n}\sum_{i=1}^{n} D_i^r Z_{it}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} - (\mathbf{Q}_t^r)^{-1} = o_p(1). \tag{39}$$

With H_{1t} and H_{2t}, we also have that $\mathbb{E}[D^r Z_t Y_t|\mathbf{X}] = \mathbb{E}[D^r Z_t Y_{ct}|\mathbf{X}] = \mathbb{E}[D^r Z_t|\mathbf{X}]\mathbb{E}[Y_{ct}|\mathbf{X}]$. We deduce, under model (21), that

$$\begin{aligned}\mathbb{E}[D^r Z_t Y_t \mathbf{X}] &= \mathbb{E}\left[\mathbb{E}[D^r Z_t Y_t|\mathbf{X}]\mathbf{X}\right] \\ &= \mathbf{Q}_t^r \boldsymbol{\theta}_{rt}. \end{aligned}\tag{40}$$

With the weak law of large numbers, as n tends to infinity we then get

$$\left(\frac{1}{n}\sum_{i=1}^{n} D_i^r Z_{it}\mathbf{x}_i Y_{it}\right) - \mathbf{Q}_t^r\boldsymbol{\theta}_{rt} = o_p(1) \tag{41}$$

and, finally, $\widehat{\boldsymbol{\theta}}_{rt} - \boldsymbol{\theta}_{rt} = o_p(1)$.

The asymptotic normality is based on the following expression of $\widehat{\boldsymbol{\theta}}_{rt}$:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{rt} &= \left(\sum_{i=1}^{n} D_i^r Z_{it}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\left(\sum_{i=1}^{n} D_i^r Z_{it}\mathbf{x}_i\left(\mathbf{x}_i^\top\boldsymbol{\theta}_{rt} + \epsilon_{i,rt}\right)\right) \\ &= \boldsymbol{\theta}_{rt} + \left(\frac{1}{n}\sum_{i=1}^{n} D_i^r Z_{it}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} D_i^r Z_{it}\epsilon_{i,rt}\mathbf{x}_i\right).\end{aligned}$$

Under hypotheses (H_{1t}) to (H_{3t}) and model (21), the random vectors $D_i^r Z_{it}\epsilon_{i,rt}\mathbf{x}_i$, for $i = 1, \dots, n$ are i.i.d copies of a centered distribution with covariance matrix $\sigma_{rt}^2\mathbf{Q}_t^r$. The central limit theorem gives us

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} D_i^r Z_{it}\epsilon_{i,rt}\mathbf{x}_i\right) \rightsquigarrow \mathcal{N}(0, \sigma_{rt}^2\mathbf{Q}_t^r),$$

and result (39) combined with Slutsky's Lemma allow us to conclude the proof. □

Proof. of Proposition 3.3.
This is not a standard maximum likelihood framework because the number of observations, $n_r(n) = \sum_{i=1}^n D_i^r$, used for estimating $\boldsymbol{\beta}$ is not deterministic. If n_r was not random, we would directly get under previous assumptions that the maximum likelihood estimator of $\boldsymbol{\beta}_{rt}$ is consistent and asymptotically Gaussian.

First note that, for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathbb{E}\left[\Psi_{1n}^r(\boldsymbol{\beta})\right] = -\mathbb{E}\left(Z^r \ln\left(\frac{\pi(\boldsymbol{\beta},\mathbf{X})}{1-\pi(\boldsymbol{\beta},\mathbf{X})}\right) + \ln\left(1-\pi(\boldsymbol{\beta},\mathbf{X})\right)|D^r=1\right)\mathbb{P}[D^r=1]. \tag{42}$$

By assumption H_{2t} we have, given $D^r = 1$,

$$\mathbb{P}\left[Z_t^r = 1|\mathbf{X}=\mathbf{x}\right] = \pi(\boldsymbol{\beta}_{rt},\mathbf{x})$$

so that, with assumption (H_{3t}),

$$\begin{aligned}\mathbb{E}\left[\Psi_{1n}^r(\boldsymbol{\beta})\right] &= -\mathbb{E}\left(\pi(\boldsymbol{\beta}_{rt},\mathbf{X})\ln\left(\frac{\pi(\boldsymbol{\beta},\mathbf{X})}{1-\pi(\boldsymbol{\beta},\mathbf{X})}\right) + \ln\left(1-\pi(\boldsymbol{\beta},\mathbf{X})\right)|D^r=1\right)\mathbb{P}[D^r=1]\\ &> \mathbb{E}\left[\Psi_{1n}^r(\boldsymbol{\beta}_{rt})\right],\end{aligned}$$

for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_{rt}$ (see e.g Lemma 2.2 in Newey and McFadden (1994)). We also get, with the strong law of large numbers that for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\Psi_{1n}^r(\boldsymbol{\beta}) - \mathbb{E}\left[\Psi_{1n}^r(\boldsymbol{\beta})\right] \to 0, \quad \textit{almost surely}$$

and we can deduce, by Theorem 2.7 in Newey and McFadden (1994) that the sequence $\widehat{\boldsymbol{\beta}}_{rt}$ of minimizers of Ψ_{1n}^r tends to $\boldsymbol{\beta}_{rt}$ almost surely.

The asymptotic normality of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{rt} - \boldsymbol{\beta}_{rt}\right)$ is based on an application of Theorem 3.3 in Newey and McFadden (1994) for probit regression, with

$$\boldsymbol{\Sigma}_t^r = \frac{1}{\mathbb{P}[D^r=1]}\left[\mathbb{E}\left(\lambda(\boldsymbol{\beta}_{rt}^\top\mathbf{X})\lambda(-\boldsymbol{\beta}_{rt}^\top\mathbf{X})\mathbf{X}\mathbf{X}^\top|D^r=1\right)\right]^{-1}$$

where $\lambda(u) = \Phi'(u)/\Phi(u)$, $u \in \mathbb{R}$. In case of logistic regression, it can be deduced from Theorem 5.1 in Hjort and Pollard (2011) for logistic regression, with

$$\boldsymbol{\Sigma}_t^r = \frac{1}{\mathbb{P}[D^r=1]}\left[\mathbb{E}\left(\pi(\boldsymbol{\beta}_{rt},\mathbf{X})(1-\pi(\boldsymbol{\beta}_{rt},\mathbf{X}))\mathbf{X}\mathbf{X}^\top|D^r=1\right)\right]^{-1}.$$

□

Proof. of Corollary 3.4.
The proof of the first point is a direct application of the continuous mapping theorem in (27). For the second point, one needs to consider the joint asymptotic normality of $(\widehat{\boldsymbol{\beta}}_{rt}, \widehat{\boldsymbol{\theta}}_{rt}, \widehat{\boldsymbol{\beta}}_{0t}, \widehat{\boldsymbol{\theta}}_{0t})$ and a direct application of the delta method with the function

$$g\left(\boldsymbol{\beta}_1, \boldsymbol{\theta}_1, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0\right) = \left(\pi(\mathbf{x},\boldsymbol{\beta}_1)\boldsymbol{\theta}_1 - \pi(\mathbf{x},\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\right)^\top \mathbf{x}$$

which is differentiable with a non vanishing gradient if $(\boldsymbol{\beta}_{rt}, \boldsymbol{\theta}_{rt}) \neq (\boldsymbol{\beta}_{0t}, \boldsymbol{\theta}_{0t})$. □

B Appendix: description of the variables

We present here the variables that were considered in Section 5.2. A detailed description of the definition of these variables as well as some descriptive statistics can be found in the Appendix of Cardot and Musolesi (2020).

The dependent variable Y_{it} corresponds to the number of employees at time t for municipality i. The socio-economic and demographic variables come from standard INSEE sources while the variables measuring land use have been obtained from the "Corine Land Cover" base. By starting from a set of sixteen possible explanatory variables, the final set of variables, which were selected by employing a backward variable selection procedure, contains the following eleven variables:

- SIZE $\equiv Y_{t_0}$ is the initial outcome, i.e the level of employment at t_0, with t_0 equals to 1993.
- DENSITY $\equiv (\textit{total population}) \,/\, \left(\textit{total surface in terms of } km^2\right);$
- INCOME $\equiv (\textit{net taxable income}) \,/\, (\textit{total population});$
- OLD $\equiv (\textit{population over 65}) \,/\, (\textit{total population});$
- FACT $\equiv (\textit{number of factory workers}) \,/\, (\textit{total population});$
- BTS $\equiv \frac{(\textit{number of people with a technical degree called ``Brevet de Technicien Supérieur''})}{(\textit{total population})};$
- AGRI $\equiv (\textit{farmland surface}) \,/\, (\textit{total surface});$
- CULT $\equiv (\textit{cultivated land surface}) \,/\, (\textit{total surface});$
- URB $\equiv (\textit{urban surface}) \,/\, (\textit{total surface});$
- IND $\equiv (\textit{industrial surface}) \,/\, (\textit{total surface});$
- ARA $\equiv (\textit{arable surface}) \,/\, (\textit{total surface});$

where the total surface and the total population should be understood within the considered municipality.

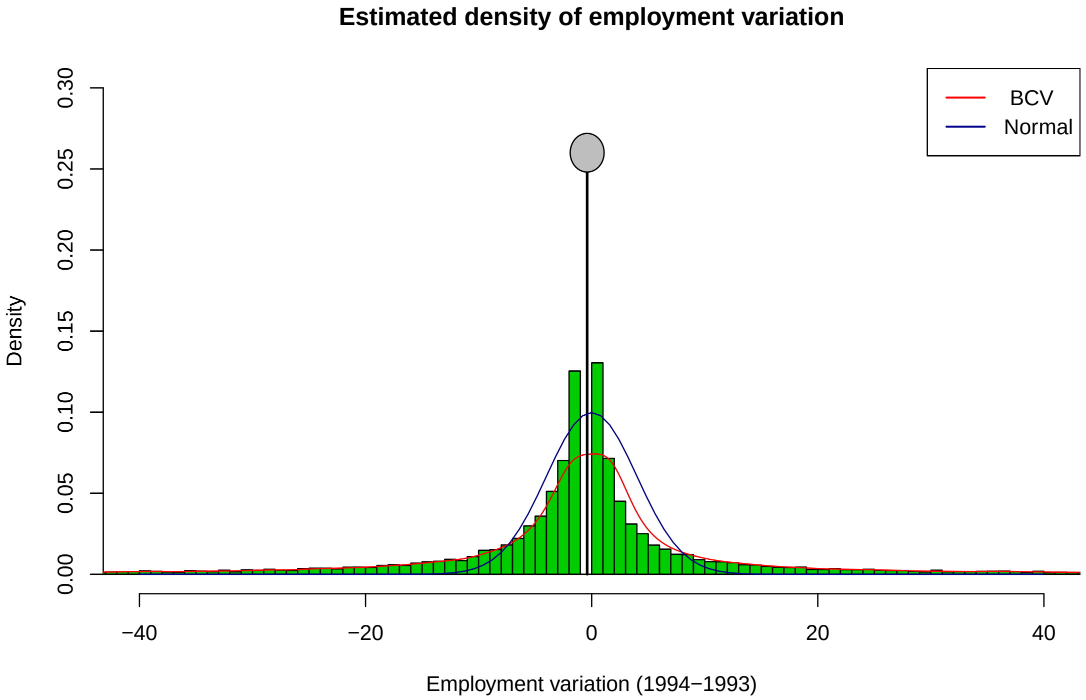

Figure 1: The estimated distribution of $EMP_{it} - EMP_{it-1}$ for $t = 1994$. The probability of observing no variation is estimated by the proportion of observations such that $EMP_{it} - EMP_{it-1} = 0$. The vertical bars represent the probability of observing a given value when $EMP_{it} - EMP_{it-1} \neq 0$. We also consider a continuous density estimation of $EMP_{it} - EMP_{it-1} \neq 0$ thanks to a kernel estimator; BCV: biased cross-validation (see Sheather, 2004; Silverman, 1986).

	CONTINUOUS RESPONSE MODEL	CONDITIONAL MIXTURE MODEL		
	(i)	(ii)	(v)	
		Continuous part	Discrete part	
	FD	FD - Subset	CRE Probit - Pooled MLE	
	Coefficient	Coefficient	Coefficient	APE
log(EXP)	0.199***	0.182***	-0.697*	-0.081*
	(0.037)	(0.038)	(0.374)	(0.043)
log(WKS)	0.001	0.002	-0.028	-0.003
	(0.037)	(0.039)	(0.214)	(0.025)
OCC	-0.022	-0.025	0.168*	0.020*
	(0.019)	(0.020)	(0.093)	(0.011)
IND	0.022	0.027	0.205***	0.024***
	(0.021)	(0.022)	(0.080)	(0.009)
SOUTH	-0.003	-0.010	0.155*	0.0181*
	(0.079)	(0.085)	(0.081)	(0.009)
SMSA	-0.054**	-0.059**	-0.018	-0.002
	(0.027)	(0.028)	(0.080)	(0.009)
MS	-0.053**	-0.054**	0.203	0.024
	(0.025)	(0.027)	(0.128)	(0.014)
UNION	0.012	0.015	0.379***	0.044***
	(0.020)	(0.021)	(0.012)	(0.012)
FEM			0.278*	0.032*
			(0.170)	(0.020)
BLK			-0.207	-0.024
			(0.148)	(0.0172)
ED			-0.003	-0.001
			(0.018)	(0.002)

All specifications include a full set of time dummies.

The standard errors of the estimated coefficients (in brackets) are robust to arbitrary serial correlation.

The standard errors of the APEs in the CRE probit model are obtained using the delta method.

***, **, *: significant at 1%, 5%, and 10% level, respectively.

Table 1: Wage equation

	CONTINUOUS RESPONSE MODEL			CONDITIONAL MIXTURE MODEL		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)
ATE$^{ZRR\&5B}$	2.021 [0.664-3.303]	3.001 [1.688-4.358]	2.955 [1.105-4.828]	2.110 [0.710-3.401]	3.134 [1.860-4.557]	3.016 [1.160-4.941]
ATE^{5B}	0.896 [-0.452-2.331]	1.571 [0.213-2.981]	0.781 [-0.530-2.144]	0.943 [-0.400-2.356]	1.604 [0.245-2.997]	0.821 [-0.485-2.180]
ATE^{ZRR}	1.125 [-0.194-2.318]	1.430 [0.173-2.785]	2.174 [0.187- 4.140]	1.167 [-0.154-2.387]	1.530 [0.271-2.935]	2.195 [0.205-4.201]

Model (i): DID with linear regression function.
Model (ii): DID with natural cubic regression splines.
Model (iii): DID with linear regression function and policy interaction with DENSITY and SIZE.
Between brackets: 95% confidence bands computed by Non-parametric bootstrap (percentile method)

Table 2: Average treatment effects

	CONTINUOUS RESPONSE MODEL														
Model	(i)					(ii)					(iii)				
Percentile	1	25	50	75	99	1	25	50	75	99	1	25	50	75	99
$ZRR\&5B$	2.021	2.021	2.021	2.021	2.021	3.001	3.001	3.001	3.001	3.001	-4.721	2.289	2.786	3.558	12.384
$5B$	0.896	0.896	0.896	0.896	0.896	1.571	1.571	1.571	1.571	1.571	-7.951	0.604	1.350	1.724	2.99
ZRR	1.125	1.125	1.125	1.125	1.125	1.430	1.430	1.430	1.430	1.430	-7.391	0.566	1.380	2.866	20.336
	CONDITIONAL MIXTURE MODEL														
$ZRR\&5B$	1.576	1.916	2.148	2.349	2.368	1.382	3.142	3.371	3.470	3.524	-5.167	2.269	2.269	3.594	12.963
$5B$	0.628	0.889	0.987	1.011	1.101	0.683	1.605	1.724	1.774	1.800	-8.072	0.746	1.352	1.655	3.056
ZRR	0.782	1.036	1.180	1.338	1.357	0.683	1.536	1.646	1.696	1.724	-7.830	0.651	1.356	2.774	20.894
	CONDITIONAL MIXTURE MODEL vs. CONTINUOUS RESPONSE MODEL (Relative change in the treatment effect)														
$ZRR\&5B$	-0.220	-0.052	0.063	0.162	0.172	-0.539	0.046	0.123	0.156	0.174	-0.356	-0.088	0.056	0.155	0.678
$5B$	-0.299	-0.007	0.101	0.128	0.229	-0.564	0.022	0.097	0.130	0.146	-1.937	-0.077	0.022	0.104	2.135
ZRR	-0.303	-0.078	0.049	0.190	0.206	-0.522	0.074	0.151	0.186	0.205	-2.279	-0.136	0.037	0.1431	2.831

Model (i): DID with linear regression function.

Model (ii): DID with cubic regression splines.

Model (iii): DID with linear regression function and policy interaction with DENSITY and SIZE.

Table 3: Distributional treatment effects

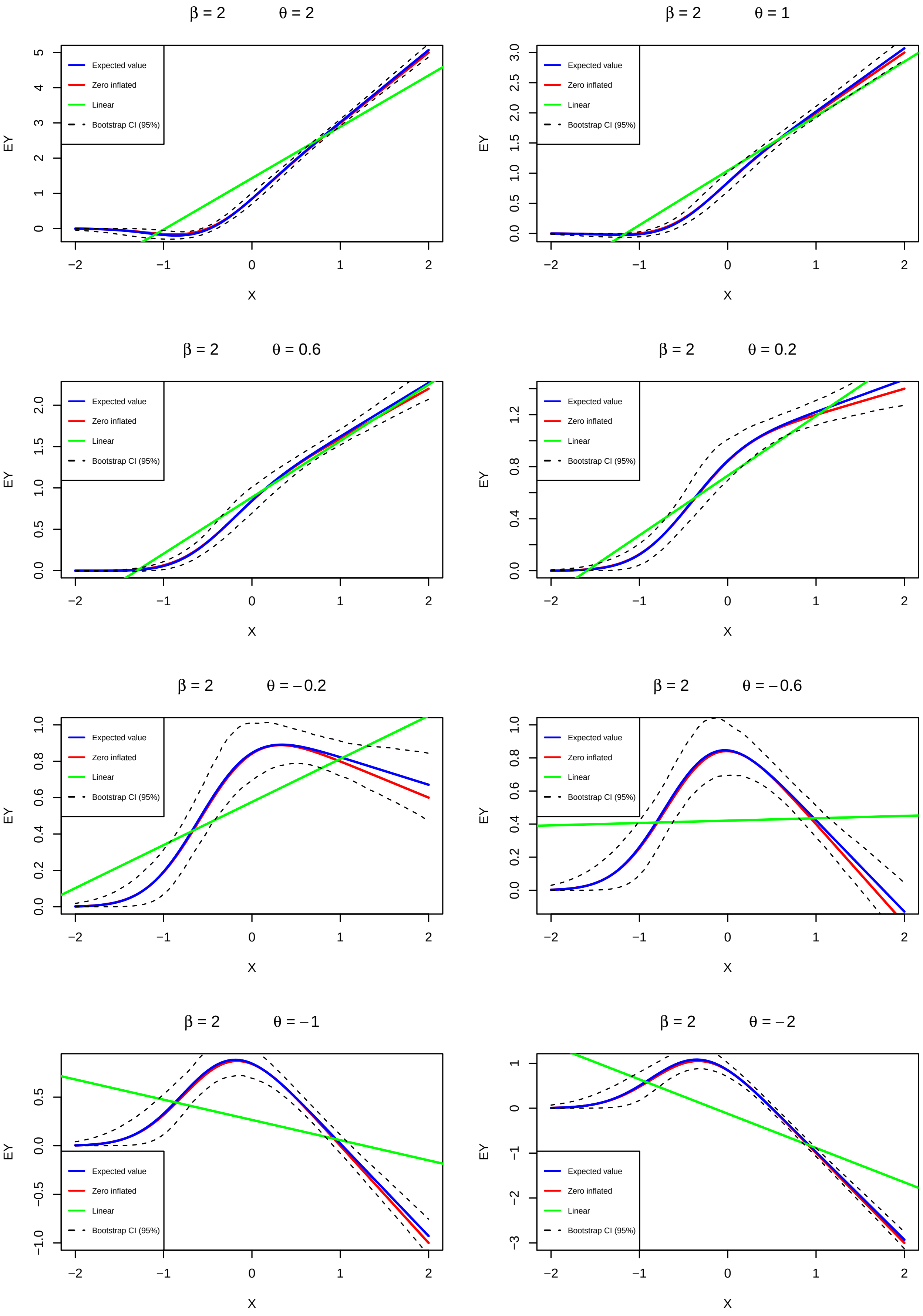

Figure 2: Simulation

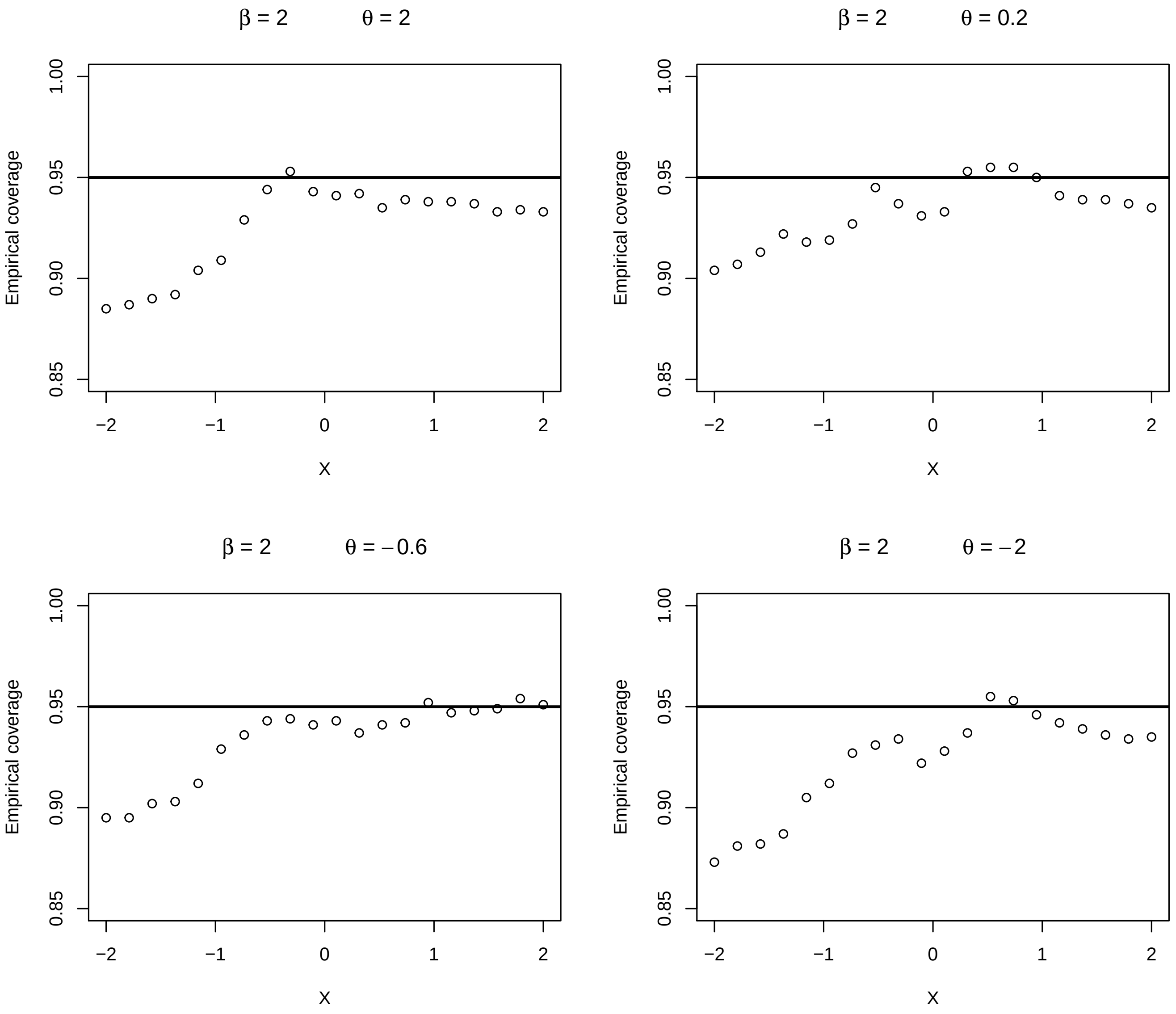

Figure 3: Empirical coverage

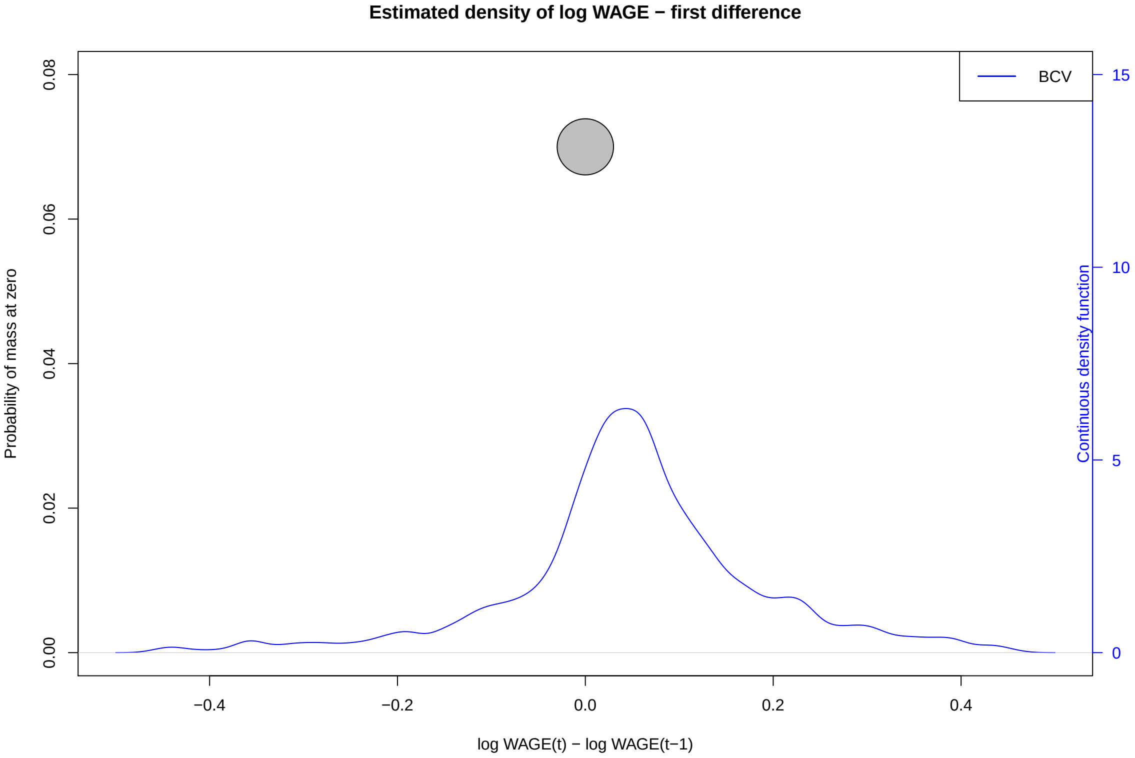

Figure 4: The estimated distribution of $\log(WAGE)$ in first differences. The probability of a mass at zero is estimated by the proportion of observations taking this value (indicated by the circle). We also consider a continuous density estimation for the continuous part of the response variable thanks to a kernel estimator; BCV: biased cross-validation (see Sheather, 2004; Silverman, 1986).

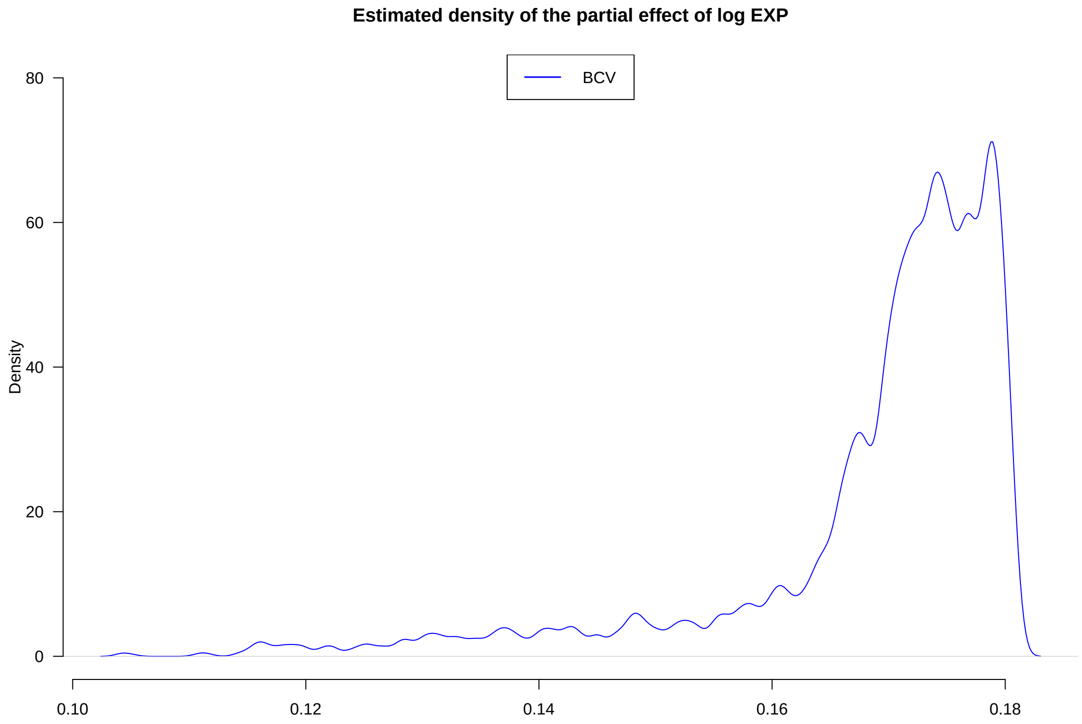

Figure 5: The estimated distribution of the partial effect of $\log(EXP)$ in the wage equation from the zero-inflated model. Bandwidth selected using biased cross-validation.

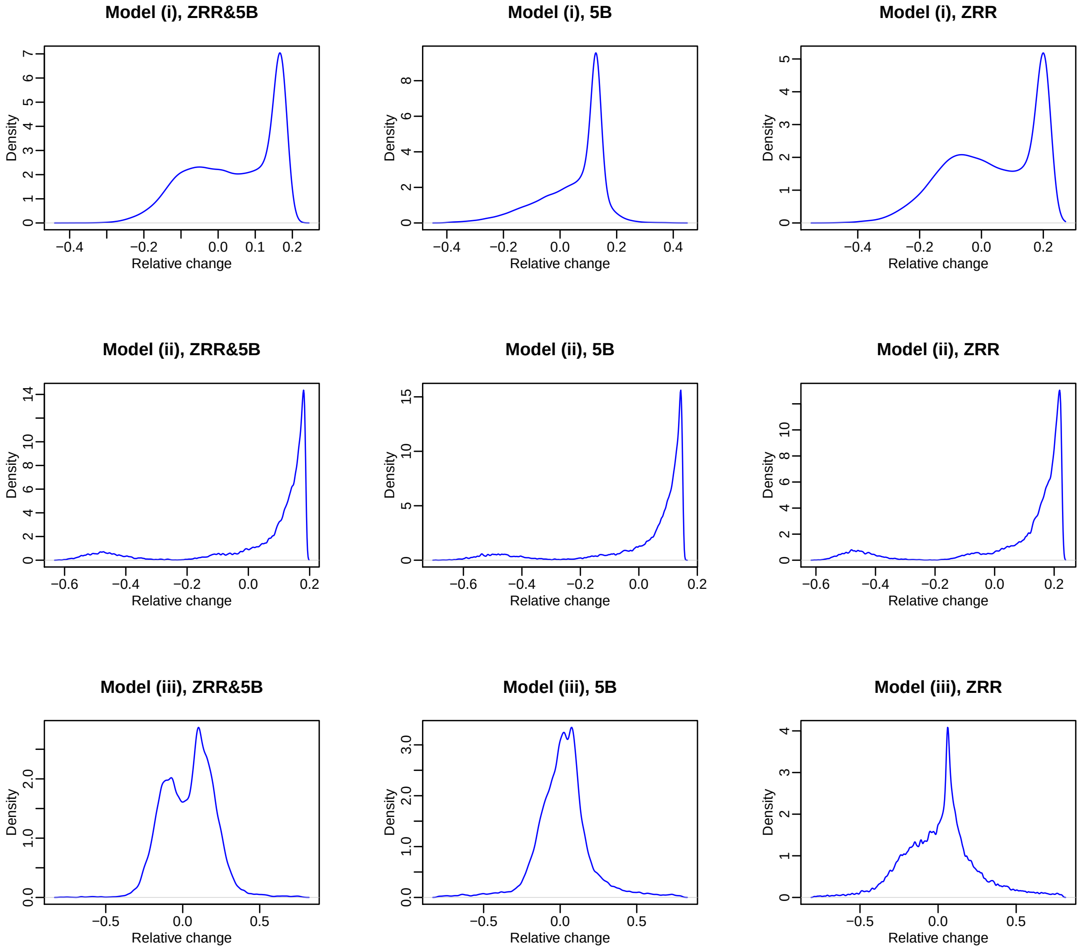

Figure 6: The estimated distribution of the relative change between the treatment effect obtained using the zero-inflated approach and the one obtained adopting the naive estimator. Bandwidth selected using biased cross-validation.