Statistics and Probability Letters 83 (2013) 1595–1602

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Semiparametric estimation of fixed effects panel data single-index model

Peng Lai [a], Gaorong Li [b], Heng Lian [c,*]

[a] School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

[b] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

[c] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

ARTICLE INFO

Article history:
Received 14 October 2012
Received in revised form 11 March 2013
Accepted 11 March 2013
Available online 19 March 2013

Keywords:
Fixed effects
Local linear smoothing
Panel data
Single-index model
Test statistic

ABSTRACT

We consider the fixed effects panel data single-index model. For estimation of the link function and the index parameter, the local linear smoothing and the least squares method are used. We also propose a test for the presence of the fixed effects. Finite sample performances are illustrated using simulations.

1. Introduction

A panel data set is one that follows a given sample of individuals over time, and thus provides multiple observations on each individual. Panel data involve two dimensions: a cross-sectional dimension and a time-series dimension. Such two-dimensional information enables researchers to extract information and make inferences which may not be possible using pure time-series data or cross-sectional data. This type of data is often produced in fields such as biology, engineering, medicine and social sciences. With the increasing availability of panel data, both theoretical and applied works in panel data analysis with parametric, semiparametric or nonparametric models have become more popular in recent years. In addition, for these panel data models, one needs to consider whether they should contain fixed effects or random effects, which may be 'ancillary' parameters. Arellano (2003), Baltagi (2005) and Hsiao (2003) provided an excellent overview of parametric panel data model analysis, Henderson and Ullah (2005), Lin and Ying (2001) and Wu and Zhang (2002) considered the random effects nonparametric panel data model, Sun et al. (2009) considered the problem of estimating a varying coefficient panel data model with fixed effects, Henderson et al. (2008) considered the nonparametric fixed effects panel data model, and Su and Ullah (2006) considered the fixed effects partially linear panel data model, among others.

It is well known that nonparametric models are more flexible than parametric models in practice. However, nonparametric regression suffers from "curse of dimensionality" when the dimension of the regressors is high. The single-index models trade off flexibility of modeling with interpretability of linear models and thus circumvent the curse of dimensionality. Treatments of the single-index models include Härdle et al. (1993), Ichimura (1993), Lai and Wang (2011), Lin (2011),

* Corresponding author. Tel.: +65 65137175.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0167-7152/$ – see front matter
http://dx.doi.org/10.1016/j.spl.2013.03.005

Lin and Kulasekera (2007), Lu (2010) and Wang et al. (2010), among others. In this paper, we consider the following fixed effects single-index model

$$Y_{ik} = g(X_{ik}^\top \beta) + \mu_i + \varepsilon_{ik}, \quad i = 1, \dots, n, \ k = 1, \dots, m, \tag{1.1}$$

where the covariate $X_{ik} = (X_{ik1}, \dots, X_{ikp})^\top$ is of dimension p, $g(\cdot)$ is an unknown link function, $\mu_i, i = 1, \dots, n$ are the fixed effects with variance σ_μ^2, the random errors ε_{ik} are assumed to be i.i.d with mean zero, $\mathrm{Var}(\varepsilon_{ik}) = \sigma_\varepsilon^2$, and independent of the fixed effects μ_j and X_{js} for all (i, j, s, k), β is an unknown index vector in $\mathbb{R}^p$ with $\|\beta\| = 1$ (where $\|\cdot\|$ denotes the Euclidean norm), and the true value of β is denoted by β_0. We allow for the individual effects $\mu_i, i = 1, \dots, n$, to be correlated with the covariates $X_{i1}, \dots, X_{im}$. The random effect models can be regarded as a special case where $\mu_i, i = 1, \dots, n$ are not correlated with $X_{i1}, \dots, X_{im}$ (Sun et al., 2009).

Owing to potential correlations between the unobservable effects μ_i and the regressors, if we ignore the fixed effects when estimating the unknown link function $g(\cdot)$ and the index vector β, the resulting estimators will be inconsistent for model (1.1). To obtain consistent estimators of $g(\cdot)$ and β, an important step is to remove the fixed individual effects from the models. The usual first-difference estimation method can remove the time-invariant fixed effects by subtracting one equation from another. For example, subtracting equation for Y_{i1} from that for Y_{ik}, we get

$$\tilde{Y}_{ik} = Y_{ik} - Y_{i1} = g(X_{ik}^\top \beta) - g(X_{i1}^\top \beta) + \tilde{\varepsilon}_{ik}, \quad k = 2, \dots, m,$$

where $\tilde{\varepsilon}_{ik} = \varepsilon_{ik} - \varepsilon_{i1}$. Alternatively, we can also remove the fixed effects by subtracting the average of $Y_{ik}, k = 1, \dots, m$, giving

$$\tilde{Y}_{ik} = Y_{ik} - \frac{1}{m}\sum_{s=1}^{m} Y_{is} = g(X_{ik}^\top \beta) - \frac{1}{m}\sum_{s=1}^{m} g(X_{is}^\top \beta) + \tilde{\varepsilon}_{ik}, \quad t = 2, \dots, m,$$

where $\tilde{\varepsilon}_{ik} = \varepsilon_{ik} - \frac{1}{m}\sum_{s=1}^{m} \varepsilon_{is}$. Instead, we impose the restriction that $\sum_{i=1}^{n} \mu_i = 0$ for identifiability, as adopted in Sun et al. (2009) for varying coefficient models.

The rest of this paper is organized as follows. In Section 2, we will propose a method for estimating the unknown link function $g(\cdot)$ and the index vector β. In addition, we also present the asymptotic properties of the estimators. In Section 3, we construct a test statistic to test the presence of fixed effects. We report in Section 4 some simulation results to demonstrate the finite sample performance of the estimators. The proofs of the main results are given in the supplementary Appendix.

2. Methodology and main results

We will use vector and matrix notations as in the following. Let $\mathbf{Y} = (Y_1^\top, \dots, Y_n^\top)^\top$, $\mathbf{g}(\mathbf{X}^\top \beta) = (\mathbf{g}^\top (X_1^\top \beta), \dots, \mathbf{g}^\top (X_n^\top \beta))^\top$ and $\mathbf{V} = (\varepsilon_1^\top, \dots, \varepsilon_n^\top)^\top$ be $(nm) \times 1$ vectors, where $Y_i = (Y_{i1}, \dots, Y_{im})^\top$, $\mathbf{g}(X_i^\top \beta) = (g(X_{i1}^\top \beta), \dots, g(X_{im}^\top \beta))^\top$ and $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^\top$. $\mu_0 = (\mu_1, \dots, \mu_n)^\top$ is an $n \times 1$ vector of the fixed effects, and $D_0 = I_n \otimes e_m$ is an $(nm) \times n$ matrix, where e_m is an $m \times 1$ vector with all elements being ones, I_n is the $n \times n$ identity matrix and $\otimes$ denotes the Kronecker product. Then, model (1.1) can be rewritten as

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}^\top \beta) + D_0 \mu_0 + \mathbf{V}. \tag{2.1}$$

Noting that $\sum_{i=1}^{n} \mu_i = 0$, (2.1) can be reduced to

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}^\top \beta) + D\mu + \mathbf{V},$$

where $D = [-e_{n-1}\ I_{n-1}]^\top \otimes e_m$ is an $(nm) \times (n-1)$ matrix and $\mu = (\mu_2, \dots, \mu_n)^\top$.

Our approach for estimating g (together with its derivative) and β is iterative in nature, with g or β fixed in turn. For any fixed $\beta \in \mathbf{B}$, $\mathbf{B} = \{\beta : \|\beta\| = 1\}$, to estimate the unknown link function $g(t)$ (here $t \in \mathbf{T}$, $\mathbf{T} = \{t : t = X^\top \beta, X \in A_X\}$ and A_X is the compact support set of X), we solve the following optimization problem

$$\min_{g,\mu} [\mathbf{Y} - \mathbf{g}(\mathbf{X}^\top \beta) - D\mu]^\top \mathbf{K}(t, h) [\mathbf{Y} - \mathbf{g}(\mathbf{X}^\top \beta) - D\mu], \tag{2.2}$$

where $\mathbf{K}(t, h)$ is an $(nm) \times (nm)$ diagonal weight matrix

$$\mathbf{K}(t, h) = \begin{pmatrix} K(X_1, t, h) & & \\ & \ddots & \\ & & K(X_n, t, h) \end{pmatrix},$$

$$K(X_i, t, h) = \begin{pmatrix} K(X_{i1}, t, h) & & \\ & \ddots & \\ & & K(X_{im}, t, h) \end{pmatrix},$$

$K(X_{is}, t, h) = K((X_{is}^\top \beta - t)/h)$ for all i and s, and h is the bandwidth.