Statistics and Probability Letters 123 (2017) 128–138

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Block empirical likelihood for partially linear panel data models with fixed effects

Bang-Qiang He [a], Xing-Jian Hong [b,*], Guo-Liang Fan [a]

[a] School of Mathematics and Physics, Anhui Polytechnic University, Wuhu 241000, China

[b] School of Data Sciences, Zhejiang University of Finance & Economics, Hangzhou 310018, China

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form 20 November 2016
Accepted 21 November 2016
Available online 19 December 2016

MSC:
62G15
62G20

Keywords:
Block empirical likelihood
Partially linear model
Panel data
Fixed effect

ABSTRACT

In this article, we consider a partially linear panel data models with fixed effects. In order to accommodate the within-group correlation, we apply the block empirical likelihood procedure to partially linear panel data models with fixed effects, and prove a nonparametric version of Wilks' theorem which can be used to construct the confidence region for the parametric. By the block empirical likelihood ratio function, the maximum empirical likelihood estimator of the parameter is defined and the asymptotic normality is shown. A simulation study and a real data application are undertaken to assess the finite sample performance of our proposed method.

1. Introduction

Panel data analysis has received a lot of attention during the last two decades due to applications in many disciplines, such as economics, finance, biology, engineering and social sciences. The double-index panel data models enable researchers to estimate complex models and extract information which may be difficult to obtain by applying purely cross section or time series models. There exists a rich literature on parametric linear and nonlinear panel data models. For an overview of statistical inference and econometric analysis of parametric panel data models, we refer to the books by Baltagi (2005) and Hsiao (2003). To avoid imposing the strong restrictions assumed in the parametric panel data models, some nonparametric methods have been used in both panel data model estimation and specification testing (e.g., Hjellvik et al., 2004; Henderson et al., 2008; Cai and Li, 2008). While the nonparametric approach is useful in exploring hidden structures and reducing modeling biases, they can be too flexible to draw concise conclusions, and face the *curse of dimensionality* due to a large number of covariates. To overcome these shortcomings, we use semiparametric approaches which are the compromises between the general nonparametric modeling and fully parametric specification.

Consider the following partially linear panel data models with fixed effects:

$$Y_{it} = X_{it}^{\tau}\beta + g(U_{it}) + \upsilon_i + \varepsilon_{it}, \quad i = 1, \dots, n, \ t = 1, \dots, T, \tag{1.1}$$

where Y_{it} is the response, $(X_{it}^{\tau}, U_{it}) \in R^p \times R$ are strictly exogenous variables, $\beta = (\beta_1, \dots, \beta_p)^{\tau}$ is a vector of p-dimensional unknown parameters, and the superscript τ denotes the transpose of a vector or matrix, $g(U_{it}) = (g_1(U_{it}), \dots, g_q(U_{it}))^{\tau}$ is

* Corresponding author.
E-mail address: myemailh@126.com (X.-J. Hong).

http://dx.doi.org/10.1016/j.spl.2016.11.021

a q-dimensional unknown functions and υ_i is the unobserved individual effects. Denote by $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})$ the random error vector of the ith subject and $\{\varepsilon_i, i = 1, \ldots, n\}$ are mutually independent with $E(\varepsilon_i|X_{it}, U_{it}) = 0$.

Model (1.1) is called a fixed effects model if υ_i is correlated with X_{it} and (or) U_{it} with an unknown correlation structure. Model (1.1) is called a random effects model if υ_i is uncorrelated with X_{it} and U_{it}. The fixed effects specification has the advantage of robustness compared to the random effects specification (e.g., Baltagi, 2005; Horowitz and Lee, 2004). However, the analysis of the fixed effects panel data is more challenging because of increasing number of parameters with the sample size, yielding the famous Neyman and Scott (1948) problem. In this paper, we are concerned with the fixed effects case.

Obviously, model (1.1) includes many usual parametric, nonparametric and semiparametric regression models. For example, when $\upsilon_i = 0$, model (1.1) reduces to the partially linear panel data model. Many researchers have explored the partially linear panel data model (e.g., Roy, 1997; Li and Ullah, 1998; You and Zhou, 2006a). Roy (1997) has used the partially linear panel data model to study the calorie and income relationship for two years panel data of rural south India. When $\upsilon_i = 0$ and $g = 0$, the model becomes the well-known parametric panel data regression model, which has been widely applied in economics (cf., Ahn and Schmidt, 2000; Hsiao, 2003). When $\upsilon_i = 0$ and $\beta = 0$, the model reduces to the panel data nonparametric model, which has been investigated by Ruckstuhl et al. (2000).

For partially linear panel data models, Li and Ullah (1998) constructed a feasible semiparametric generalized least squares estimator for the coefficient of the linear component and derived the asymptotic normality of the proposed estimator. For the model (1.1), Su and Ullah (2006) adapted a local linear dummy variable approach to remove the unknown fixed effects. In this paper, we make statistical inference for the parametric β in partially linear panel data models with fixed effects. Following the estimation procedure proposed by Li and Ullah (1998) and Su and Ullah (2006), the least-squared estimations of β can be obtained. Based on this, a normal-based confidence region for the parametric is constructed. But such a construction is inconvenient because it involves estimating complex asymptotic covariance of the estimators. To end this, we recommend using the empirical likelihood (EL) method to construct the confidence regions for β. Based on the EL method, we can construct immediately an approximate confidence region for the regression parameter. One motivation is that empirical likelihood inference does not involve the asymptotic covariance of the estimators, which is rather complex structure for the partially linear panel data models with fixed effects. Another motivation is that the confidence region based on EL approach does not impose prior constraints on the region shape, and the shape and orientation of confidence regions are determined completely by the data. Therefore, The EL method has been used by many authors, such as Shi and Lau (2000), Wang and Jing (1999), You and Zhou (2006b), Fan et al. (2012) and so on.

The usual empirical likelihood method cannot be applied, however, to partially linear panel data models with fixed effects due to correlation within groups. To accommodate the within-group correlation, we apply the block empirical likelihood procedure proposed by You et al. (2006) to model (1.1), establish a block empirical log-likelihood ratio for the parametric component, and derive a nonparametric version of Wilks' theorem which can be used to construct the block empirical likelihood confidence region with asymptotically correct coverage probability for the parametric component.

The rest of this paper is organized as follows. Section 2 introduces the methodology and empirical log-likelihood ratio function for β. Assumptions and the main result are given in Section 3. Some simulation studies and a real-data example are conducted in Section 4. The proofs of the main results are relegated to Section 5.

2. The model and methodology

To introduce our estimation, we assume that model holds with the restriction $\sum_{i=1}^n \upsilon_i = 0$. Let $\upsilon = (\upsilon_2, \ldots, \upsilon_n)^\tau$ and $\upsilon_0 = (-\sum_{i=2}^n \upsilon_i, \upsilon_2, \ldots, \upsilon_n)^\tau$. We rewrite model (1.1) in a matrix format yields

$$Y = X\beta + g(U) + M\upsilon + \varepsilon, \tag{2.1}$$

where $M = [-i_{n-1}\, I_{n-1}]^\tau \otimes i_T$ is an $nT \times (n-1)$ matrix, I_n denotes the $n \times n$ identity matrix, and i_n denotes the $n \times 1$ vector of ones. There are many approaches to estimating the parameters $\{\beta_j, j = 1, \ldots, p\}$ and the functions $\{g_i(\cdot), i = 1, \ldots, q\}$. The main idea is from the profile least squares approach proposed by Fan and Huang (2005): suppose that we have a random sample $\{(U_{it}, X_{it1}, \ldots, X_{itp}, Y_{it}), i = 1, \ldots, n, t = 1, \ldots, T\}$ from model (2.1). Let $\theta = (\upsilon^\tau, \beta^\tau)^\tau$. Given θ, one can apply a local linear regression technique to estimate the nonparametric component $\{g_j(\cdot), j = 1, \ldots, q\}$ in (2.1). For U_{it} in a small neighborhood of u, one can approximate $g_j(U_{it})$ locally by a linear function as below

$$g_j(U_{it}) \approx g_j(u) + g_j'(u)(U_{it} - u) \equiv a_j + b_j(U_{it} - u), \quad j = 1, \ldots, q,$$

where $g_j'(u) = \partial g_j(u)/\partial u$. This leads to the following weighted local least-squares problem: find $\{(a_j, b_j), j = 1, \ldots, q\}$ to minimize

$$\sum_{i=1}^n \sum_{t=1}^T \left\{ \left(Y_{it} - X_{it}^\tau \beta - \upsilon_i \right) - \sum_{j=1}^q \Big[a_j + b_j(U_{it} - u) \Big] \right\}^2 K_h(U_{it} - u), \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a sequence of positive numbers tending to zero, called bandwidth. Simple calculation yields that

$$(\hat{g}_1(u), \ldots, \hat{g}_q(u), h\hat{g'}_1(u), \ldots, h\hat{g'}_q(u))^\tau = (D_u^\tau W_u D_u)^{-1} D_u^\tau W_u (Y - X\beta - M\upsilon),$$