Journal of Multivariate Analysis 154 (2017) 96–111

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Improved estimation of fixed effects panel data partially linear models with heteroscedastic errors

Jianhua Hu [a,b,*], Jinhong You [a,b], Xian Zhou [c]

[a] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

[b] Key Laboratory of Mathematical Economics (SUFE), Ministry of Education, Shanghai 200433, China

[c] Department of Applied Finance and Actuarial Studies, Macquarie University, Sydney, NSW 2109, Australia

ARTICLE INFO

Article history:
Received 6 April 2016
Available online 4 November 2016

AMS 2000 classifications:
primary 62H12
secondary 62A01

Keywords:
Consistent estimator
Fixed effects
Heteroscedastic errors
Incidental parameter
Partially linear

ABSTRACT

Fixed effects panel data regression models are useful tools in econometric and microarray analysis. In this paper, we consider statistical inferences under the setting of fixed effects panel data partially linear regression models with heteroscedastic errors. We find that the usual local polynomial estimator of the error variance function based on residuals is inconsistent, and develop a consistent estimator. Applying this consistent estimator of error variance and spline series approximation of the nonparametric component, we further construct a weighted semiparametric least squares dummy variables estimator for the parametric and nonparametric components. Asymptotic normality of the proposed estimator is derived and its asymptotic covariance matrix estimator is provided. The proposed estimator is shown to be asymptotically more efficient than those ignoring heteroscedasticity. Simulation studies are conducted to demonstrate the finite sample performances of the proposed procedure. As an application, a set of economic data is analyzed by the proposed method.

1. Introduction

Panel data refer to the pooling of observations on a cross-section of subjects, such as households, countries, firms, etc., over a time period, which can be achieved by surveying a sample of subjects and following them over time; see, e.g., Baltagi [4]. Such a two-dimensional information set enables researchers to estimate complex models and draw efficient statistical inferences that may not be possible using pure time-series data or cross-section data. Both theoretical developments and applied works in panel data analysis have become more popular in recent years.

Panel data parametric (mainly linear) regression models have been the dominant framework for analyzing panel data; see [1,4] for summaries of early work and [18] for a recent comprehensive survey. If correctly specified, the parametric model has the advantages of easy interpretation and efficient estimation. In practice, however, correct parameterization is often difficult or unavailable, and a misspecification of the model could lead to biased and misleading estimates of the underlying parameters. To address this issue, various more flexible models have been introduced in literature of statistics and econometrics. Among the most important is the panel data partially linear regression model, which allows unspecified relationship between the response variable and some covariate(s).

* Corresponding author at: School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.
E-mail address: frank.jianhuahu@gmail.com (J. Hu).

http://dx.doi.org/10.1016/j.jmva.2016.10.010
0047-259X/

Let $(\mathbf{X}_{it}, \mathbf{U}_{it}, Y_{it})$ denote the observations collected from the ith subject, $i = 1, \ldots, n$, at time t, $t = 1, \ldots, T$, where $Y_{it} \in \mathbb{R}$ is the response of interest, $\mathbf{X}_{it} = (X_{it1}, \ldots, X_{itp})^\top \in \mathbb{R}^p$ is a p-vector of linear predictors and $\mathbf{U}_{it} = (U_{it1}, \ldots, U_{itq})^\top$ is a q-vector of nonlinear predictors. A typical panel data partially linear regression model has the form

$$Y_{it} = \mathbf{X}_{it}^\top \boldsymbol{\beta} + g(\mathbf{U}_{it}) + \varepsilon_{it}, \quad \varepsilon_{it} = \mu_i + \nu_{it}, \ i = 1, \ldots, n, \ t = 1, \ldots, T, \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is an unknown parameter vector, $g(\cdot)$ is an unknown function, μ_i is the individual effect of subject i and ν_{it} is an *idiosyncratic* error. Depending on whether μ_i is correlated with the observed explanatory variables $(\mathbf{X}_{it}^\top, \mathbf{U}_{it}^\top)^\top$ or not, model (1.1) can be divided into two classes. One is the *random effects* model in which μ_i is uncorrelated with $(\mathbf{X}_{it}^\top, \mathbf{U}_{it}^\top)^\top$, so that $\mathrm{E}\{\mu_i(\mathbf{X}_{it}^\top, \mathbf{U}_{it}^\top)^\top\} = 0$, where E is the symbol of expectation; and another is the *fixed effects* model in which μ_i is correlated with $(\mathbf{X}_{it}^\top, \mathbf{U}_{it}^\top)^\top$, i.e., $\mathrm{E}\{\mu_i(\mathbf{X}_{it}^\top, \mathbf{U}_{it}^\top)^\top\} \neq 0$. Considering a large number of random draws from the cross section, it makes sense for us to treat the individual effects, μ_i, as random draws from the population. Fixed effects panel data model, however, is appropriate if the interest is in a specific set of subjects, such as specific firms, a set of OECD countries, all American states, and so on.

For model (1.1) with fixed effects, the individual effect is often viewed as a parameter to be estimated. It is typically assumed that T is finite and n is large. Consequently, the number of parameters grows with the sample size, and the ordinary least squares (OLS) or maximum likelihood estimator (MLE) would lead to inconsistent estimates of the common parameter of interest. This is well-known as the *incidental parameter problem*; see [19] for a general discussion on this problem. Due to the incidental parameter problem, it is a great challenge to construct consistent estimators for the parametric and nonparametric components in the fixed effects panel data partially linear regression model, and few results were available until Baltagi and Li [5], who proposed a difference-based series estimation (DSE) for the parametric component and nonparametric component. They established the asymptotic normality of the former and derived the convergence rate of the latter. This DSE, however; is not efficient when $T > 2$, see [2].

Fan et al. [9] found that the model (1.1) with fixed effects is also useful to conduct the microarray analysis of the neuroblastoma cell in response to macrophage migration inhibitory factor (MIF). Fan et al. [8] proposed a novel profile least squares estimation (PLSE) for the parametric component and a local linear estimation for the nonparametric component by the back fitting method. They established the asymptotic normality of the former and the MSE upper bound of the latter. In addition, the estimation problem of the model (1.1) with fixed effects was considered in [13,24], while the problem of estimating a varying-coefficient panel data model with fixed effects was studied in [21].

All above-mentioned results assumed that the idiosyncratic errors ν_{it} are independent and identically distributed (i.i.d.). In practice, however, the random errors are often *heteroscedastic* (with unequal variances). For example, heteroscedasticity has been found in gasoline demands across Organization for Economic Co-operation and Development (OECD) countries, in steam electricity generations across utilities of different sizes, in cost functions for US airline firms, and in medical expenditures [4]. It is well known that when heteroscedasticity is present, ignoring its impact will result in inefficient estimators of the regression coefficients and biased estimators of covariance. Under the setting of fixed effects panel data linear regression model, Kézdi [14] and Stock and Watson [20] investigated the consistent estimations of the regression parameters and asymptotic properties. They found that, due to the incidental parameter problem, the conventional heteroscedasticity-robust (HR) asymptotic covariance matrix estimator is inconsistent, and further provided a $\sqrt{n}$ consistently bias-adjusted HR estimator. However, [14,20] did not make any assumption about the heteroscedasticity. Therefore, the error variance in their case could not be estimated and the information of the heteroscedasticity could not be taken into account to improve the estimation of the mean parameter. There is another important situation of heteroscedasticity where the error variance is a function of some of the predictors, see [3,10,23]. In this situation, the error variance could usually be estimated and the information of heteroscedasticity could be taken into account to improve the estimation of the mean parameter, see Amemiya [3] for the cross-sectional data, Fan and Yao [10] for the time series data and You et al. [23] for the random effects panel data. To date, however, whether the error variance could be estimated consistently and whether the information of heteroscedasticity could be utilized to improve the estimation of the mean parameter remains unsolved even in the setting of fixed effects panel data linear regression model. In this paper, we address these issues under the more general partially linear model (1.1).

As in [8], we assume that the errors are heteroscedastic and the error variance is a smoothing function of $\mathbf{V}_{it}$ in the form:

$$\varepsilon_{it} = \mu_i + \sigma(\mathbf{V}_{it})\nu_{it}, \quad i = 1, \ldots, n \text{ and } t = 1, \ldots, T, \tag{1.2}$$

where $\sigma(\cdot)$ is an unknown function, $\mathrm{var}(\nu_{it}) = 1$, and $\mathbf{V}_{it} = (V_{it1}, \ldots, V_{itm})^\top$ is a known vector. $\mathbf{V}_{it}$ may be a function of $\mathbf{X}_{it}$ and $\mathbf{U}_{it}$, such as $\mathbf{V}_{it} = \mathbf{U}_{it}$ in [8]. We find that the usual residuals-based local polynomial estimator of the error variance function $\sigma(\cdot)$ is not consistent, and will propose an alternative consistent estimator. Applying the proposed estimator together with spline series approximation of the nonparametric component, we further construct a weighted semiparametric least squares dummy variables estimator for the parametric components and a weighted spline series estimator for nonparametric components. Asymptotic normal distributions for the proposed estimators are derived and asymptotic covariance matrix estimators are provided. The proposed estimator is shown to be asymptotically more efficient than those ignoring the heteroscedasticity. The results can be extended to more general situation, such as the case where the noise level may be a smoothing function of the mean, and so on.

Throughout this paper, we choose $\mathrm{E}\{g(\mathbf{U}_{it})\} = 0$ as our identification condition and assume that n is large and T is small and fixed with $T \geq 2$. The remainder of the paper is as follows. The pilot estimators of the parametric and nonparametric