Journal of Multivariate Analysis 141 (2015) 118–131

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Consistent test of error-in-variables partially linear model with auxiliary variables☆

Zhihua Sun [a,b,*], Xue Ye [c], Liuquan Sun [d]

[a] University of Chinese Academy of Sciences, Key Laboratory of Big Data Mining and Knowledge Management of CAS, Beijing, 100049, China

[b] Anhui Normal University, Wuhu, 241000, China

[c] University of Chinese Academy of Sciences, Beijing 100049, China

[d] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 11 October 2014
Available online 21 July 2015

AMS subject classification:
62F03
62G10

Keywords:
Auxiliary variable
Consistent test
Measurement error
Model check
Partial linear model

ABSTRACT

In this paper, we investigate the model checking problem of a partially linear model when some covariates are measured with error and some auxiliary variables are supplied. The often-used assumptions on the measurement error, such as a known error variance or a known distribution of the error variable, are not required. Also repeated measurements are not needed. Instead, a nonparametric calibration method is applied to deal with the measurement error. An estimating method for the null hypothetical model is proposed and the asymptotic properties of the proposed estimators are established. A testing method based on a residual-marked empirical process is then developed to check the null hypothetical partially linear model. The tests are shown to be consistent and can detect the alternative hypothesis close to the null hypothesis at the rate n^{-r} with $0 \leq r \leq 1/2$. Simulation studies and real data analysis are conducted to examine the finite sample behavior of the proposed methods.

1. Introduction

Measurement error data arise frequently in many fields, such as epidemiology, sociology and biomedical sciences. To deal with the measurement error, with some assumptions, for example, the error variables are normally distributed, the error variance has a known form or the repeated measurements are collected, the adverse influence of measurement error can be effectively eliminated. We can refer to Liang et al. [9], Cui et al. [5], Ma and Carroll [11], Liang et al. [10], Carroll et al. [3] and Pan et al. [13], where some classical results are presented. In this paper, we consider another method to deal with the measurement error: relaxing the above assumptions on the measurement error and assuming a supply of auxiliary variables. We take the widely-used partially linear regression model as an example to show the implementation of this method.

A partially linear regression model has the form of

$$Y = X^{\tau}\beta + g(T) + \varepsilon, \tag{1.1}$$

☆ This research was partly supported by the National Natural Science Foundation of China (Grant Nos. U1430103, 11231010 and 11171330), the President Fund of UCAS, Key Laboratory of RCSDS, CAS (No. 2008DP173182) and the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences.

* Correspondence to: School of Mathematics Science, University of Chinese Academy of Sciences, Beijing, China.
E-mail addresses: sunzh@amss.ac.cn (Z. Sun), leafasnow@sina.com (X. Ye), slq@amt.ac.cn (L. Sun).

http://dx.doi.org/10.1016/j.jmva.2015.07.007
0047-259X/

where Y is a scalar response variable, X and T are respectively p-vector and scalar covariates, β and $g(\cdot)$ are respectively p-dimensional and infinite dimensional regression parameters. The random error ε satisfies $E(\varepsilon|X,T)=0$ and $E(\varepsilon^2|X,T)<\infty$. Here we assume that T is scalar only for the simplicity of notation.

Assume that the covariate X is measured with error and a surrogate covariate $\tilde{X}$ is observed. The relationship between X and $\tilde{X}$ can be described as $X=E(\tilde{X}|V)$, where V is an observed d-vector of auxiliary variables. Actually, the above error model structure is a special case of the additive error model, since $X=E(\tilde{X}|V)$ implies that $\tilde{X}=X+e$ with $E(e|V)=0$. This error structure was first introduced by Cai et al. [2] and was further studied by Cui et al. [5], Zhou and Liang [21], Zhao and Xue [20] and Yang et al. [19]. Since $\tilde{X}$ and V are observed, the true covariate $X=E(\tilde{X}|V)$ can be estimated by the local smoothing method and then the estimation of the model parameters can be constructed for the null hypothetical partially linear model.

We mainly consider the model checking problem of the partial linear model since a correct model specification is fundamental and critical in statistical estimation and inference. Model checking problem with measurement error is very challenging and only sporadic research work can be found in the literature. See Cheng [4] and Koul and Song [8], among others. In this paper, the information of the auxiliary variable makes the model parameters estimable. An empirical-process-based test statistic can be constructed based on the estimated model error. Actually, for a partially linear model, several other popular model checking methods, such as the local smoothing method, can also be constructed. We can refer such methods to Fan and Li [6], Song [15], Gao and Gijbels [7], and so on. Comparatively, the proposed empirical-process-based method has the following merits: (i) The proposed test is consistent; (ii) The effect of the proposed test depends slightly on the choice of the smoothing parameters; (iii) It can detect the local alternative models close to the null hypothetical model at the rate $n^{-1/2}$. It is well known that the rate $n^{-1/2}$ is the possible fastest rate at which the lack-of-fit test can detect. More details about empirical-process-based test can refer to Zhu and Ng [22], Sun et al. [16], Xu and Zhu [18], Ma et al. [12].

The rest of this paper is organized as follows. Section 2 proposes the estimation method for the model parameters. In Section 3, we construct an empirical-process-based model checking method and study its asymptotic properties. Section 4 reports some results from simulation studies conducted for evaluating the proposed methods. An application to a data set of Duchenne Muscular Dystrophy is provided in Section 5. All proofs are given in the Appendix.

2. Estimating procedure

Suppose that we have a sample $(Y_i,\tilde{X}_i,T_i,V_i)_{i=1}^n$ from $(Y,\tilde{X},T,V)$. For model (1.1), Robinson [14] showed that

$$\beta=\Big[E\{(X-E(X|T))(X-E(X|T))^\tau\}\Big]^{-1}E\Big[\{X-E(X|T)\}\{Y-E(Y|T)\}\Big].$$

Since X is not observed, we replace X by a nonparametric estimation, say $\hat{\gamma}_n(V)$, which can be defined as $\sum_{j=1}^n\tilde{X}_j\lambda_v(V-V_j)/\sum_{j=1}^n\lambda_v(V-V_j)$ with $\lambda(\cdot)$ a kernel function, h_v a bandwidth and $\lambda_v(\cdot)=1/h_v^d\lambda(\cdot/h_v)$. Then by the sampling moment method, we have an estimator of β as follows:

$$\hat{\beta}_n=\Big[\sum_{i=1}^n\{\hat{\gamma}_n(V_i)-\hat{g}_{1,n}(T_i)\}\{\hat{\gamma}_n(V_i)-\hat{g}_{1,n}(T_i)\}^\tau\Big]^{-1}\sum_{i=1}^n\{\hat{\gamma}_n(V_i)-\hat{g}_{1,n}(T_i)\}\{Y_i-\hat{g}_{2,n}(T_i)\},$$

where $\hat{g}_{1,n}(t)$ and $\hat{g}_{2,n}(t)$ are respectively the local linear estimators of $g_1(t)=:E[X|T=t]$ and $g_2(t)=:E[Y|T=t]$, which are defined in the following. Let $\omega_t=\mathrm{diag}(K_t(T_1-t),\ldots,K_t(T_n-t))$,

$$\hat{\gamma}_n=\begin{pmatrix}\hat{\gamma}_n(V_1)^\tau\\ \hat{\gamma}_n(V_2)^\tau\\ \vdots\\ \hat{\gamma}_n(V_n)^\tau\end{pmatrix},\qquad T_t=\begin{pmatrix}1 & T_1-t\\ 1 & T_2-t\\ \vdots & \vdots\\ 1 & T_n-t\end{pmatrix},\qquad \hat{Y}=\begin{pmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix},$$

where $K(\cdot)$ is a kernel function, h_t is a bandwidth and $K_t(\cdot)=1/h_tK(\cdot/h_t)$. By local linear regression technique, we can give the estimator of $g_1(t)$, denoted by $\hat{g}_{1,n}(t)$, and that of its derivative $g_1'(t)$, denoted by $\hat{g}'_{1,n}(t)$, as follows:

$$\begin{pmatrix}\hat{g}_{1,n}(t)^\tau\\ \hat{g}'_{1,n}(t)^\tau\end{pmatrix}=(T_t'\omega_tT_t)^{-1}T_t'\omega_t\hat{\gamma}_n.$$

Let $I_{j,p}$ be the $1\times p$ vector with the first j components ones and the remaining zeros. Then we have $\hat{g}_{1,n}(t)^\tau=I_{1,2}^\tau(T_t'\omega_tT_t)^{-1}T_t'\omega_t\hat{\gamma}_n$. Similarly, an estimator of $g_2(t)=E(Y|T=t)$ is given by $\hat{g}_{2,n}(t)=I_{1,2}^\tau(T_t'\omega_tT_t)^{-1}T_t'\omega_t\hat{Y}$. The following theorem summarizes the asymptotic expansion of $\hat{\beta}_n$ with the proof given in the Appendix.

Theorem 1. *Under the conditions* (C.1)–(C.6) *in the Appendix, we have that under model* (1.1),

$$n^{1/2}(\hat{\beta}_n-\beta)=n^{-1/2}\Sigma_0^{-1}\sum_{j=1}^n\Big[(\tilde{X}_j-X_j)E(\varepsilon_j|V_j)+\{X_j-E(X_j|T_j)\}\varepsilon_j+E\{X_j-E(X_j|T_j)|V_j\}(X_j-\tilde{X}_j)^\tau\beta\Big]+o_p(1),\quad(2.1)$$