Computational Statistics and Data Analysis 83 (2015) 52–64

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Checking the adequacy for a distortion errors-in-variables parametric regression model

Jun Zhang [a,*], Gaorong Li [b], Zhenghui Feng [c]

[a] Shen Zhen-Hong Kong Joint Research Center for Applied Statistical Sciences, College of Mathematics and Computational Science, Institute of Statistical Sciences, Shenzhen University, Shenzhen, China

[b] Beijing Center for Scientific and Engineering Computing, College of Applied Sciences, Beijing University of Technology, Beijing, China

[c] School of Economics and the Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China

ARTICLE INFO

Article history:
Received 17 October 2013
Received in revised form 26 May 2014
Accepted 20 September 2014
Available online 1 October 2014

Keywords:
Confounding variables
Errors-in-variables
Distorting functions
Empirical process

ABSTRACT

This paper studies tools for checking the validity of a parametric regression model, when both response and predictors are unobserved and distorted in a multiplicative fashion by an observed confounding variable. A residual based empirical process test statistic marked by proper functions of the regressors is proposed. We derive asymptotic distribution of the proposed empirical process test statistic: a centered Gaussian process under the null hypothesis and a non-centered one under local alternatives converging to the null hypothesis at parametric rates. We also suggest a bootstrap procedure to calculate critical values. Simulation studies are conducted to demonstrate the performance of the proposed test statistic and real examples are analyzed for illustrations.

1. Introduction

In many applications of regression analysis, the variables of interest may be observed with measurement errors. A distortion errors-in-variables model can be written as

$$\begin{cases} Y = g(\mathbf{X}) + \varepsilon, \\ \tilde{Y} = \phi(U)Y, \\ \tilde{\mathbf{X}} = \psi(U)\mathbf{X}, \end{cases} \tag{1.1}$$

where Y is an unobservable response, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\tau$ is an unobservable continuous predictor vector (the superscript τ denotes the transpose operator throughout this paper), $g(\cdot)$ is the regression function of Y on $\mathbf{X}$ and is usually unknown, $\tilde{Y}$ and $\tilde{\mathbf{X}}$ are the observed response and predictor vector and $\psi(\cdot)$ is a $p \times p$ diagonal matrix: $\mathrm{diag}\big(\psi_1(\cdot), \ldots, \psi_p(\cdot)\big)$, where $\phi(\cdot)$ and $\psi_r(\cdot)$ are unknown continuous distorting functions. The confounding variable $U \in \mathbb{R}^1$ is observable and independent of $(\mathbf{X}, Y)$. The diagonal form of $\psi(\cdot)$ indicates that the confounding variable U distorts unobserved X_r, $r = 1, \ldots, p$ in a multiplicative fashion. The model error ε has mean zero and finite variance, and is independent with $\mathbf{X}$ and U.

Both Y and $\mathbf{X}$ are unobserved but distorted by a confounding variable U. This type of measurement error data usually occurs in biomedical and health-related studies. For example, the body mass index (BMI), height or weight usually have some kinds of multiplicative effects on the primary variables of interest. Kaysen et al. (2002) analyzed that BMI plays the role of confounding variable that may simultaneously contaminate the fibrinogen level and serum transferrin level of hemodialysis

* Corresponding author.
E-mail address: zhangjunstat@gmail.com (J. Zhang).

http://dx.doi.org/10.1016/j.csda.2014.09.018
0167-9473/

patients, and Kaysen et al. (2002) further divided the observed fibrinogen level-$\tilde{Y}$ and observed serum transferrin level-$\tilde{X}$ by BMI to eliminate the possible biases for the target parameters. While such a simple way of dividing confounding variable BMI may lead to biased estimators. Şentürk and Müller (2006) proposed the distortion model $\tilde{Y} = \phi(U)Y$, $\tilde{\mathbf{X}} = \psi(U)\mathbf{X}$ as the flexible multiplicative adjustment by involving unknown smooth distorting functions $\phi(\cdot)$ and $\psi_r(\cdot)$'s to reduce possible biases and obtain proper estimators.

Recently, distortion errors-in-variables models have been received much attention in the literature. Readers can refer to Cui et al. (2009), Li et al. (2010), Nguyen and Şentürk (2007, 2008), Nguyen et al. (2008), Şentürk and Müller (2005a,b, 2006, 2009), Şentürk and Nguyen (2009), Zhang et al. (2014a, 2013a,b, 2012a). However, these existing literature focus on estimation rather than testing. To the best of our knowledge, there is little literature to consider model checking for distortion errors-in-variables data. Obviously, the correct model specification is fundamental and crucial in the data analysis.

In this paper, we consider a model checking problem for a distortion errors-in-variables parametric regression model. Let $\mathcal{M} = \{g(\cdot, \beta) : \beta \in \Theta \subset \mathbb{R}^q\}$ be a given parametric family of functions, and β be an unknown $q \times 1$ parameter vector on a compact parameter space $\Theta \subset \mathbb{R}^q$. It is interesting to test whether $g(\cdot)$ belongs to $\mathcal{M}$ or not. The null hypothesis is given as

$$\mathcal{H}_0 : g(\mathbf{X}) = g(\mathbf{X}, \beta_0) \quad \text{for some } \beta_0 \text{ and } g(\cdot, \cdot), \tag{1.2}$$

against the alternative hypothesis

$$\mathcal{H}_1 : g(\mathbf{X}) \neq g(\mathbf{X}, \beta_0) \quad \text{for any } \beta_0 \text{ and } g(\cdot, \cdot). \tag{1.3}$$

If $(Y, \mathbf{X})$ can be observed, there exist many tests for testing $\mathcal{H}_0$ against $\mathcal{H}_1$. For example, when model $g(\mathbf{X}, \beta_0)$ is to a linear regression model $\mathbf{X}^\tau \beta_0$, Azzalini and Bowman (1993) used the smoothed residuals and proposed a pseudo-likelihood ratio test statistic, which measures the distance between the nonparametric and the parametric models. Lin et al. (2002) proposed objective and informative model-checking techniques based on the cumulative sum of residuals over certain coordinates. For more general nonlinear parametric models, Härdle and Mammen (1993) and Stute et al. (1998a) proposed model checking methods based on the squared deviation between the nonparametric and parametric fits and the empirical process of the regressors marked by the residuals. Dette (1999) proposed a variance difference based test, Stute et al. (1998b) proposed an innovation approach to conveniently calculate p-values, and Fan and Huang (2001) suggested an adaptive Neyman threshold test when model errors are normally distributed. For the use of nonparametric techniques for the model checking problems, the book by Hart (1997) gave an extensive overview and useful references.

In the present article, we aim to develop a lack-of-fit test for checking $\mathcal{H}_0$ in the context of distortion measurement errors. A residual empirical process based test statistic is proposed. Firstly, we calibrate unobserved $(Y, \mathbf{X})$ by using the direct plug-in method proposed by Cui et al. (2009). The direct plug-in method is to estimate distortion functions $\phi(\cdot)$, $\psi(\cdot)$ by the traditional nonparametric kernel smoothing, say $\hat{\phi}(\cdot)$, $\hat{\psi}(\cdot)$. Then we calibrated $\mathbf{X}$ and Y by $\hat{\psi}^{-1}\tilde{\mathbf{X}}$, $\tilde{Y}/\hat{\phi}$, denoted as $\hat{\mathbf{X}}$, $\hat{Y}$ respectively. Secondly, using these calibrated quantities, a nonlinear least squares estimator of β_0 is obtained under the null hypothesis $\mathcal{H}_0$. Large-sample properties of this nonlinear least squares estimator are investigated under the null hypothesis and the local alternative hypothesis. It is noted that if $\mathcal{H}_0$ holds, the asymptotic results are the same as those obtained in Zhang et al. (2012a) when the confounding variable is univariate. Thirdly, we propose a residual based empirical process test statistic. This process is shown to be a centered Gaussian process under $\mathcal{H}_0$. Another advantage for this test process is that it can detect local alternatives converging to $\mathcal{H}_0$ at the rate $n^{-1/2}$. Lastly, we propose a test statistic based on the empirical process test statistic, and a bootstrap procedure is also proposed to define p-values.

The rest of this paper is organized as follows. In Section 2, a test procedure is proposed and its asymptotic properties are investigated. A bootstrap method is further proposed to define the critical values. In Section 3, simulation studies and two real data analyses are respectively conducted to evaluate the proposed test statistic proposed in Section 2. The detailed proofs of the theorems are given in the Appendix.

2. The testing procedure and main results

2.1. Covariate calibration

As the distorted $\tilde{Y}$ and $\tilde{\mathbf{X}}$ are available, we first calibrate unobserved Y and $\mathbf{X}$ by using the observed i.i.d. sample $\{\tilde{Y}_i, \tilde{\mathbf{X}}_i, U_i\}_{i=1}^n$. In this subsection, we introduce the calibration estimation procedure. To ensure identifiability for the model (1.1), it is assumed that

$$E[\phi(U)] = 1, \qquad E[\psi_r(U)] = 1, \tag{2.1}$$

$r = 1, \ldots, p$. This identifiability condition is introduced by Şentürk and Müller (2005a), and it is analogous to the classical additive measurement error setting: $E(\eta) = 0$ for $W = Z + \eta$, where W is error-prone and Z is error-free. Under the independence condition between U and $(Y, \mathbf{X})$, the identifiability condition (2.1) entails that

$$\begin{cases} E[\tilde{Y}|U] = \phi(U)E[Y|U] = \phi(U)E[Y] = \phi(U)E[\tilde{Y}], \\ E[\tilde{X}_r|U] = \psi_r(U)E[X_r|U] = \psi_r(U)E[X_r] = \psi_r(U)E[\tilde{X}_r], \end{cases} \tag{2.2}$$

for $r = 1, \ldots, p$. From (2.2), the calibration procedure for (Y, X_r) can be implemented by $Y = \tilde{Y}E[\tilde{Y}]/E[\tilde{Y}|U]$ and $X_r = \tilde{X}_r E[\tilde{X}_r]/E[\tilde{X}_r|U]$ in the population level. Cui et al. (2009) and Zhang et al. (2012a) proposed a moment-based procedure for those unobserved variables $\{Y_i, X_{ri}\}_{i=1}^n$. The moment-based procedure is summarized as follows.