Journal of Multivariate Analysis 124 (2014) 116–129

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A revisit to correlation analysis for distortion measurement error data

Jun Zhang [a], Zhenghui Feng [b,*], Bu Zhou [c]

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

[b] The Wang Yanan Institute for Studies in Economics, The School of Economics, Xiamen University, Xiamen, China

[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 10 January 2013
Available online 31 October 2013

AMS subject classifications:
62G08
62G20
62F12

Keywords:
Correlation coefficient
Distorting function
Measurement error models
Kernel smoothing

ABSTRACT

In this paper, we consider the estimation problem of a correlation coefficient between unobserved variables of interest. These unobservable variables are distorted in a multiplicative fashion by an observed confounding variable. Two estimators, the moment-based estimator and the direct plug-in estimator, are proposed, and we show their asymptotic normality. Moreover, the direct plug-in estimator is shown asymptotically efficient. Furthermore, we suggest a bootstrap procedure and an empirical likelihood-based statistic to construct the confidence interval. The empirical likelihood statistic is shown to be asymptotically chi-squared. Simulation studies are conducted to examine the performance of the proposed estimators. These methods are applied to analyze the Boston housing price data as an illustration.

1. Introduction

Measurement error is common in many disciplines, such as economics, health science and medical research, due to improper instrument calibration or many other reasons. Generally, an estimation procedure which ignores measurement error may cause large bias, sometimes seriously large bias. The classical statistical estimation and inference become very challenging. Therefore, it requires particular care to eliminate such bias when estimating target parameters. Research on classical errors-in-variables have been widely studied in the last two decades, for example by Li and Hsiao [12] and Schennach [26] using replicate data, and Carroll et al. [37], Schennach [27] and Wang and Hsiao [3], using instrumental variable methods. Others considered nonparametric or semi-parametric approaches (Carroll et al. [25]; Delaigle et al. [36]; Liang et al. [46]; Liang and Li [17]; Liang and Ren [19]; Liang and Wang [20]; Schafer [16], Taupin [1]; Zhou and Liang [6]). In addition, Fuller [9] and Carroll et al. [4] give comprehensive reviews containing many parametric and semi-parametric measurement error models.

In this paper, we consider multiplicative effect type errors, namely, distorting measurement errors. Both the response and predictors are unobservable and distorted by general multiplicative effects of some observable confounding variable as

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

* Corresponding author.
E-mail addresses: zhangjunstat@gmail.com (J. Zhang), zhfengwise@gmail.com (Z. Feng), zhoubu1988@163.com (B. Zhou).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.10.004

where $(Y, X)^\tau$ are the unobservable continuous variables of interest, (the superscript τ denotes the transpose operator throughout this paper), while $(\tilde{Y}, \tilde{X})^\tau$ are available distorted variables. $\phi(\cdot)$ and $\psi(\cdot)$ are unknown contaminating functions of an observed confounding variable U, and U is independent of $(X, Y)^\tau$.

The distortion measurement errors model (1.1) usually occurs in biomedical and health-related studies. The confounding variable (for example, it can be the body mass index (BMI), height or weight) usually has some kind of multiplicative effect on the primary variables of interest. Kaysen et al. [11] analyzed the relationship between fibrinogen level and serum transferrin level among hemodialysis patients, and realized that BMI plays the role of confounding variable that may contaminate the fibrinogen level and the serum transferrin level simultaneously. To eliminate the potential bias, Kaysen et al. [11] simply divided the observed fibrinogen level-$\tilde{Y}$ and observed serum transferrin level-$\tilde{X}$ by BMI-U. Şentürk and Müller [30] noticed that the exact relationship between the confounding variable (BMI) and primary variables is hardly known in practice. Such a simple way of dividing confounding variable BMI may not be appropriate and may lead to an inconsistent estimator of the target parameters. So Şentürk and Müller [30] proposed model (1.1) as a flexible multiplicative adjustment by involving unknown smooth distorting functions $\phi(\cdot)$, $\psi(\cdot)$ for the confounding variable U.

To estimate the correlation coefficient between Y and X, denoted as $\rho_{(Y,X)}$, Şentürk and Müller [28] observed that the direct calculation of correlation coefficient between $\tilde{Y}$ and $\tilde{X}$ will result in an arbitrarily large biased estimator of $\rho_{(Y,X)}$. Therefore, a proper adjustment method for estimating $\rho_{(Y,X)}$ needs to be addressed. To solve this problem, Şentürk and Müller [28] established the relationship between $\tilde{Y}$ and $\tilde{X}$ through a varying coefficient model, and then employed the binning technique to estimate $\rho_{(Y,X)}$. Such a transformation procedure can be generalized to regression models with linear structure—for example, linear models [28,30,22,21], generalized linear models [31] and partial linear single index models [43].

The goal of this paper is to construct consistent estimation and do inference for correlation coefficient $\rho_{(Y,X)}$. Two estimators for $\rho_{(Y,X)}$ are proposed. One is the moment-based estimator, and the other is the direct plug-in estimator. The basic ideas and motivations of these two methods are summarized as follows.

- Our first estimator is based on $\rho_{(Y,X)} = \rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}/\Delta$ for some unknown constant Δ, where $\rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}$ is the correlation coefficient between $e_{\tilde{Y}U} = \tilde{Y} - E[\tilde{Y}|U]$ and $e_{\tilde{X}U} = \tilde{X} - E[\tilde{X}|U]$, i.e., $\rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})} = \frac{\mathrm{Cov}(e_{\tilde{Y}U}, e_{\tilde{X}U})}{\sqrt{\mathrm{Var}(e_{\tilde{Y}U})\mathrm{Var}(e_{\tilde{X}U})}}$ in the population level. This relationship $\rho_{(Y,X)} = \rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}/\Delta$ was first revealed in Appendix D of Şentürk and Müller [28], but they did not study the statistical properties and simulations further. However, $\rho_{(Y,X)} = \rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}/\Delta$ implies that $\hat{\rho}_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}/\hat{\Delta}$ is also an estimator of $\rho_{(Y,X)}$, where $\hat{\rho}_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}$ and $\hat{\Delta}$ are some estimators of $\rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}$ and Δ, respectively. So it is worth studying the estimation procedure for $\hat{\rho}_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}/\hat{\Delta}$ and its associated sample properties. We propose moment-based estimators for $\rho_{(e_{\tilde{Y}U}, e_{\tilde{X}U})}$, Δ and further establish their asymptotic normality properties. Moreover, some simulations are also evaluated. To construct a confidence interval for $\rho_{(Y,X)}$, a wild bootstrap procedure is proposed.
- The second estimator is based on the recent methodology *direct plug-in*, proposed by Cui et al. [5]. The direct plug-in method is a convenient tool for distorting measurement error data. The basic idea is to obtain estimators of the distortion functions, say $\hat{\phi}$, $\hat{\psi}$ and then calibrate Y and X by $\tilde{Y}/\hat{\phi}$, $\tilde{X}/\hat{\psi}$, and finally construct estimation by using these calibrated quantities. The direct plug-in method can be easily adopted in parametric and semi-parametric models, see for instance [5,45,44,13,42]. In this paper, an estimator of $\rho_{(Y,X)}$ based on the direct plug-in method is also investigated. An interesting result is that the direct plug-in estimator for $\rho_{(Y,X)}$ is *efficient*, i.e., the asymptotic variance of the direct plug-in estimator is the same as the classical asymptotic variance of the sample correlation coefficient (see for example [8], Section 8) when data has no distortion effect ($\phi(\cdot) \equiv 1$, $\psi(\cdot) \equiv 1$). In other words, this direct plug-in estimation procedure for $\rho_{(Y,X)}$ eliminates the effect caused by the multiplying distorting measurement error $\phi(U)$, and $\psi(U)$. Moreover, an empirical likelihood-based statistic is proposed to construct a confidence interval.

Further, we use our proposed estimators to re-analyze Boston housing price data. In [28], the authors used the education level 'Lstat' as the confounding variable to investigate the correlation between 'Crime' and 'price'. Zhang et al. [45] indicated that another choice of confounding variable is 'Ptratio', pupil–teacher ratio by town. We will make a comparison for those estimators of $\rho_{(Y,X)}$ under these two different choices of confounding variables to see which estimator is more informative and reasonable.

The paper is organized as follows. In Section 2, we propose the moment-based estimator and derive related asymptotic results. A wild bootstrap procedure to construct a confidence interval is also investigated. In Section 3, we give the direct plug-in estimator and present some asymptotic results. We develop a calibrated empirical log-likelihood ratio statistic and show that the statistic has an asymptotic chi-squared distribution. In Section 4, simulation studies are conducted to examine the performance of the proposed methods. In Section 5, the analysis of Boston housing price data is presented. All technical proofs of the asymptotic results are given in the Supplementary material.

2. Estimation procedure and asymptotic results

To ensure identifiability for model (1.1), Şentürk and Müller [28] introduced that

$$E[\phi(U)] = 1, \qquad E[\psi(U)] = 1. \tag{2.2}$$