Computational Statistics and Data Analysis 59 (2013) 103–120

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

# Estimation in linear regression models with measurement errors subject to single-indexed distortion[☆]

Jun Zhang[a,*], Yujie Gai[b], Ping Wu[c]

[a] Shenzhen-Hong Kong Joint Research Centre for Applied Statistical Sciences, Shenzhen University, Shenzhen, Guangdong Province, China
[b] School of Economics, Central University of Finance and Economics, Beijing, China
[c] School of Finance and Statistics, East China Normal University, Shanghai, China

ARTICLE INFO

*Article history:*
Received 13 February 2012
Received in revised form 10 July 2012
Accepted 1 October 2012
Available online 9 October 2012

*Keywords:*
Confounding variables
Measurement errors
Profile least squares
Single index
Varying coefficient models

ABSTRACT

In this paper, we consider statistical inference for linear regression models when neither the response nor the predictors can be directly observed, but are measured with errors in a multiplicative fashion and distorted as single index models of observable confounding variables. We propose a semiparametric profile least squares estimation procedure to estimate the single index. Then we develop a global weighted least squares estimation procedure for parameters of linear regression models via the varying coefficient models. Asymptotic properties of the proposed estimators are established. The results combined with consistent estimators for the asymptotic variance can be employed to test whether the targeted parameters in the single index and linear regression models are significant. Finite-sample performance of the proposed estimators is assessed by simulation experiments. The proposed methods are also applied to a dataset from a Pima Indian diabetes data study.

## 1. Introduction

In many applications, variables may not be directly observed because of certain contamination. This type data are common in many disciplines, such as in health science and medicine research. As we know, the measurement error in covariates may cause large bias, sometimes seriously, in the estimated regression coefficient if we ignore the measurement error. The goal of measurement error modeling is to correct such bias, attainment of this goal requires considerable care. As such, the measurement error models have been widely studied and received great attention in the literature. Fuller (1987) is a comprehensive survey containing many linear measurement error models. Carroll et al. (2006) systematically summarized the recent research developments of nonlinear and semiparametric measurement error models.

In this paper, we consider the problem of estimating a $(p+1)$-vector of parameters $\boldsymbol{\beta}_0$ from the linear regression models

$$Y = \mathbf{X}^{\tau}\boldsymbol{\beta}_0 + \varepsilon, \tag{1}$$

where "$\tau$" denotes the transport operation throughout this paper, $Y$ is an univariate response, $\mathbf{X} = (X_0, X_1, \ldots, X_p)^{\tau}$ is a predictor vector with $X_0 \equiv 1$ for the intercept. $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p})^{\tau}$ is an unknown $p+1$ dimensional vector parameter

[☆] The first author and the third author are supported by the NSFC grant 11101157, China. The second author is supported by the NSF project ZR2011AQ007 of Shandong Province of China and NNSF grant 11201499, China. The authors thank an associate editor and three referees for their valuable comments and suggestions.

[*] Corresponding author.
*E-mail addresses:* zhangjunstat@gmail.com (J. Zhang), gaiyujie83@gmail.com (Y. Gai), wu_ping916@yahoo.com.cn (P. Wu).

0167-9473/$ – see front matter Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.10.001

in $\mathbb{R}^{p+1}$, and $\varepsilon$ is the model error satisfying $E(\varepsilon) = 0$ and $E(\varepsilon^2) < \infty$. Our interest in this paper is to estimate $\boldsymbol{\beta}_0$ when both the response and predictors are observed with measurement errors by certain multiplicative distorting functions. Specially:

$$\tilde{Y} = \phi(\theta_0^\tau \mathbf{U})Y, \qquad \tilde{X}_1 = \psi_1(\theta_0^\tau \mathbf{U})X_1, \ldots, \tilde{X}_p = \psi_p(\theta_0^\tau \mathbf{U})X_p, \tag{2}$$

$$E\left\{\phi(\theta_0^\tau \mathbf{U})\right\} = 1, \qquad E\left\{\psi_1(\theta_0^\tau \mathbf{U})\right\} = 1, \ldots, E\left\{\psi_p(\theta_0^\tau \mathbf{U})\right\} = 1, \tag{3}$$

where $(Y, X_1, \ldots, X_p) \perp\!\!\!\perp \mathbf{U}$, $\perp\!\!\!\perp$ indicates independence. $\phi(\cdot)$, $\psi_r(\cdot)$ are unknown continuous distorting functions. $\mathbf{U}$ is an observed continuous confounding variable. $\theta_0 = (\theta_{01}, \ldots, \theta_{0q})^\tau$ is a $q$-vector of parameters in $\mathbb{R}^q$ satisfying $\|\theta_0\| = 1$, where $\|\cdot\|$ stands for the Euclidean norm. The constraint $\|\theta_0\| = 1$ is used to identify single index $\theta_0$ because $\phi(\cdot)$, $\psi_r(\cdot)$ are all unknown and only the orientation of $\theta_0$ is identifiable (Zhu and Xue, 2006). Conditions (3) are the identifiability conditions on $\phi(\theta_0^\tau \mathbf{U})$, $\psi_r(\theta_0^\tau \mathbf{U})$ suggested by Nguyen and Şentürk (2008). The identifiability conditions ensure that the distorting effect vanishes with no average distortion, namely, $E(\tilde{Y}) = E(Y)$, $E(\tilde{X}_r) = E(X_r)$.

The above scenario is common in practice due to the distortion from the effects of the confounding variable. For example, Kaysen et al. (2002) collected data on hemodialysis patients involved in medical studies and realized that the fibrinogen level and serum transferrin level should be divided by the body mass index (BMI). This adjustment by division implies a multiplicative fashion of the relationship between the unobserved primary variables and the confounding variable. Nevertheless, the precise knowledge of the confounding variable and the primary variables is hardly known in practice. The naive division of the confounding variable to estimate original response $Y$ and predictors $X_r$'s may cause large bias or lead to model misspecification. Thus, Şentürk and Müller (2005) introduced a more flexible model, namely, covariate adjusted regression (CAR), in which the unknown continuous distortion functions $\phi(\cdot)$, $\psi_r(\cdot)$ are allowed from a practical point of view. In many applications, more than one confounding variable may simultaneously affect the primary variables of interest. One example is the Pima Indian diabetes data. In this dataset, Nguyen and Şentürk (2008) found that body mass index (BMI) and triceps skin fold thickness (SFT) are two potential distorting covariates to affect plasma glucose concentration (GLU) and diastolic blood pressure (DBP). To examine the underlying relationship between GLU and DBP, they considered the single index distortions (2) and modeled GLU and DBP as a linear regression model.

To eliminate the effect caused by distortions, Şentürk and Müller (2005, 2006, 2009) transformed the distorted response and distorted predictors via a connection to a varying coefficient regression and apply a binning method similar to that proposed by Fan and Zhang (2000) for longitudinal data. This binning method is designed for the models with linear structure, such as linear regression models (Şentürk and Müller, 2005, 2006), generalized linear models Şentürk and Müller (2009) and partial linear single index models (Zhang et al., 2012a). As for the nonlinear models, the transformation technique to the varying coefficient models may not work well and may lead to the non-identifiability of some parameters (Cui et al., 2009). As a remedy, Cui et al. (2009) proposed a direct plug-in method by using the calibrated arguments $\hat{Y} = \tilde{Y}/\hat{\phi}$, $\hat{X}_r = \tilde{X}_r/\hat{\psi}_r$, here $\hat{\phi}$ and $\hat{\psi}_r$ are the traditional nonparametric kernel smoothing estimators. Any further estimation is then based on the calibrated quantities.

The primary goal is to estimate $\boldsymbol{\beta}_0$ in the linear regression models (1) and uncover the true relationship between $Y$ and $\mathbf{X}$. To achieve this goal, Nguyen and Şentürk (2008) extends the transformation method to an adaptive varying single index coefficient model (Xia and Li, 1999). Nguyen and Şentürk (2008) used a hybrid backfitting algorithm to simultaneously estimate the unknown single index and varying coefficient functions. The final estimator of $\boldsymbol{\beta}_0$ is a weighted-average of the estimated coefficient functions. However, Nguyen and Şentürk (2008) did not provide theoretical justification for their approach. Another issue is about the hybrid backfitting algorithm they adopted. This algorithm needs to take derivatives with respect to the single index $\theta_0$ when updated it (see p. 818 of Nguyen and Şentürk, 2008). Nevertheless, as noted in Zhu and Xue (2006) and Zhu et al. (2010), the restriction of $\|\theta_0\| = 1$ leads to a non-differential problem at the point $\theta_0$ lying on the boundary of a unit ball. What should be right hybrid backfitting algorithm for estimating the single index $\theta_0$ under this situation is much less well understood.

In this paper, we propose a different estimation proposal in the multivariate covariate adjusted setting. We use the popular "delete-one-component" method to overcome the non-differential difficulty and propose a semiparametric profile least squares method to estimate the single index $\theta_0$. Next, we establish a connection to varying coefficient models. Unlike the binning method used in Şentürk and Müller (2005, 2006, 2009), a global weighted least squares method is adopted to estimate these varying coefficient functions, and an estimator of $\boldsymbol{\beta}_0$ can be constructed by using these estimated varying coefficient functions. The asymptotic normality of the parameter estimators which we are interested in is also obtained. Furthermore, we propose consistent estimators of asymptotic variance to construct a test statistic for testing whether the targeted parameters $\theta_0$, $\boldsymbol{\beta}_0$ are significant. A simulation study is conducted to examine the performance of the proposed procedures with moderate sample sizes. In this simulation, we also compare our method with some existing methods, such as the binning method (Şentürk and Müller, 2005, 2006), direct plug-in method (Cui et al., 2009) and dimension reduction based method (Zhang et al., 2012b). A re-visit to the Pima Indian diabetes data shows a more reasonable explanation as compared to Nguyen and Şentürk (2008).

The remainder of the paper is organized as follows. In Section 2, we propose the semiparametric profile least squares estimation procedure for the single index $\theta_0$, and further introduce a global weighted least squares estimation procedure for the parameters $\boldsymbol{\beta}_0$. The asymptotic properties of the proposed estimators are investigated in this section. The estimators of asymptotic variance are also given in Section 2. In Section 3, we report the results of a simulation study. In Section 4, we