COMSTA: 6439 Model 3Gsc pp. 1–15 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation and hypothesis test on partial linear models with additive distortion measurement errors

Jun Zhang [a,b], Yan Zhou [a,*], Bingqing Lin [a,*], Yao Yu [c]

[a] College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen University, Shenzhen, China
[b] Shenzhen-Hong Kong Joint Research Center for Applied Statistical Sciences, Shenzhen University, Shenzhen, China
[c] Department of Biostatistics and Computational Biology, University of Rochester, NY, USA

ARTICLE INFO

Article history:
Received 30 August 2016
Received in revised form 15 January 2017
Accepted 4 March 2017
Available online xxxx

Keywords:
Additive distortion measurement errors
Partial linear models
Local linear smoothing
Profile least squares estimators
Lack-of-fit test

ABSTRACT

We consider estimation and hypothesis test for partial linear measurement errors models when the response variable and covariates in the linear part are measured with additive distortion measurement errors, which are unknown functions of a commonly observable confounding variable. We propose a transformation based profile least squares estimator to estimate unknown parameter under unrestricted and restricted conditions. Asymptotic properties for the estimators are established. To test a hypothesis on the parametric components, a test statistic based on the normalized difference between the residual sums of squares under the null and alternative hypotheses is proposed, and we further show that its limiting distribution is a standard chi-squared distribution. Lastly, we suggest a lack-of-fit test of score type for checking the validity of partial linear models. The quadratic form of the scaled test statistic is asymptotically chi-squared under the null hypothesis and a non-centered one under local alternatives converging to the null hypothesis at parametric rates. Simulation studies are conducted to demonstrate the performance of the proposed procedure and a real example is analyzed for an illustration.

1. Introduction

Various semiparametric regression models have been proposed to balance model interpretability and flexibility when a full nonparametric model does not work, and have been widely studied and used to explore complicated relations between the response variable and covariates of interest in data analysis. Some useful semi-parametric regression models include partial linear models (Härdle et al., 2000; Heckman, 1986; Liu et al., 2011; Xu and Guo, 2013), partial linear single index models (Carroll et al., 1997; Liang et al., 2010; Yu and Ruppert, 2002), and partial linear varying-coefficient models (Fan and Huang, 2005; Kai et al., 2011; Zhou and Liang, 2009). In this article, we focus on the partial linear models (PLMs), which are expressed as

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + g(Z) + \epsilon. \quad (1.1)$$

In model (1.1), Y is a scalar response variable, covariates $(\boldsymbol{X}^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p \times \mathbb{R}$, $\boldsymbol{\beta}_0$ is an unknown vector in $\mathbb{R}^p$, ϵ is the error term with mean zero and finite variance, and $g(\cdot)$ is an unknown smooth function.

Measurement errors are common in many disciplines, such as medical research, health science and economics, due to improper instrument calibration or many other reasons. When some variables have been measured with errors, estimation

* Corresponding authors.
E-mail addresses: zhouy1016@szu.edu.cn (Y. Zhou), bqlin@szu.edu.cn (B. Lin).

http://dx.doi.org/10.1016/j.csda.2017.03.009

based on the standard assumption may cause large bias and leads to inconsistent estimates, meaning that the parameter estimates do not tend to the true values even in a large sample. For example, the effect of the measurement errors in a simple linear regression is an underestimate of the coefficient, known as the attenuation bias (Fuller, 1987). In non-linear models, the structure of the bias is more complicated (Carroll et al., 2006). The classical statistical estimation and inference for measurement errors models is challenging, and it requires particular care to eliminate such bias when estimating target parameters. Research on measurement errors models has been widely studied. In this paper, we consider the following additive distortion measurement errors models

$$\widetilde{Y} = Y + \phi(U), \quad \widetilde{\boldsymbol{X}} = \boldsymbol{X} + \psi(U) \quad U \perp\!\!\!\perp (Y, \boldsymbol{X}, Z), \tag{1.2}$$

where the notation $\perp\!\!\!\perp$ indicates independence. Here, $(\boldsymbol{X}, Y)$ are unobserved, $\widetilde{Y}$ and $\widetilde{\boldsymbol{X}}$ are observed distorted variables, Z is also observed without distortion, U is an observed continuous scalar confounding variable, $\psi(U)$ is a p-dimensional vector $\left(\psi_1(U), \ldots, \psi_p(U)\right)^{\mathrm{T}}$, where $\phi(\cdot)$ and $\psi_r(\cdot)$ denote the unknown continuous distortion functions. The confounding variable U, for example, it can be the body mass index (BMI), height or weight in health or medical studies, usually has some kind of additive or multiplicative effects on the primary variables of interest. The simultaneous dependence of the original variables on the same confounding variable may result in any artificial relations which do not exist between the unobserved true variables (Şentürk and Müller, 2005). The additive distortion measurement errors model (1.2) is introduced by Şentürk and Müller (2005), but they did not make an intensive study of the estimation and statistical inference for model (1.2). Şentürk and Müller (2005) focused on another multiplicative distortion measurement errors model $\widetilde{Y} = \phi(U)Y$, $\widetilde{X}_r = \psi_r(U)X_r$, $r = 1, \ldots, p$ under the identifiability condition $E[\phi(U)] = E[\psi_r(U)] = 1$. Recently, a number of authors have studied the multiplicative distortion measurement errors models in various parametric or semi-parametric setting. See for example, Şentürk and Müller (2006, 2009) considered the linear regression models and generalized linear models. Cui et al. (2009) studied the nonlinear regression models. Li et al. (2010) considered the PLMs when $\boldsymbol{X}$ is distorted with multiplicative distortion measurement errors. Delaigle et al. (2016) obtained a fundamental work of nonparametric estimation of a regression curve when the data are observed with multiplicative distortion. Toward this end, there is little systematic studies on the additive distortion measurement errors models, Nguyen and Şentürk (2007) proposed some graphical techniques for assessing departures from or violations of assumptions regarding the type and form of the additive or multiplicative distortion. For the estimation and hypothesis test problems, there are no studies for the PLMs with the additive distortion measurement errors.

In this article, we investigate the estimation and hypothesis test for PLMs when distorted variables are available. Our first goal is to estimate $\boldsymbol{\beta}_0$ and $g(z)$ consistently. We transform the PLMs into partial linear additive models by using the observed data $\left\{\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i, Z_i, U_i\right\}_{i=1}^n$, and then we propose a transformation based profile least squares estimator of $\boldsymbol{\beta}_0$. An interesting result is that the transformation based profile least squares estimator of $\boldsymbol{\beta}_0$ is *efficient*, i.e., the asymptotic variance of the estimator is the same as the classical asymptotic variance obtained in Härdle et al. (2000) when data have no additive distortion effects (i.e., $\phi(\cdot) \equiv 0$, $\psi(\cdot) \equiv \boldsymbol{0}$). In other words, the transformation estimation procedure eliminates the effect caused by the additive distorting measurement errors $\phi(U)$ and $\psi(U)$. Moreover, we also consider the problem of testing whether $\boldsymbol{\beta}_0$ satisfies some linear restriction conditions or not. A restricted transformation based profile least squares estimator is proposed by using Lagrange multipliers under the null hypothesis. Finally, a test statistic based on the normalized difference between the residual sums of squares under the null and alternative hypotheses is proposed. Under the null hypothesis, the limiting distribution of the test statistic is shown to be a standard chi-squared distribution. Lastly, we aim to develop a lack-of-fit test for checking the adequacy of PLMs in the context of additive distortion measurement errors. We suggest lack-of-fit tests of score type statistic, which is shown to be asymptotically a centered normal distribution under null hypothesis. Another advantage for this test statistic is that it can detect local alternatives converging to null hypothesis at the rate $n^{-1/2}$. After estimating the asymptotic variance of the test statistic, the quadratic form of the scaled test statistic is asymptotically chi-squared under the null hypothesis and a non-centered one under local alternatives converging to the null hypothesis at parametric rates. We conduct Monte Carlo simulation experiments to examine the performance of the proposed estimation and test procedures. Our simulation results show that the proposed estimators and test statistics perform well both in estimation and hypothesis testing.

The paper is organized as follows. In Section 2, we propose the estimation procedure and hypothesis test for the parameter $\boldsymbol{\beta}_0$, introduce the estimator of $g(z)$ and present their asymptotic results. In Section 3, we develop a lack-of-fit test for checking the adequacy of PLMs, associated with the theoretical results of test statistic. In Section 4, we report the results of simulation studies and present the results of our statistical analysis of a real data. Technical proofs of theorems are provided in the supplementary materials (see Appendix A).

2. Methodology

2.1. Transformation based estimators for $\boldsymbol{\beta}_0$ and $g(z)$

As the variables $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ are unobservable and measured with errors, we transform the model (1.1) with the observed sample $\{\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i, U_i, Z_i\}_{i=1}^n$ into the following model:

$$\begin{aligned}\widetilde{Y} &= \widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\beta}_0 + g(Z) + \phi(U) - \psi^{\mathrm{T}}(U)\boldsymbol{\beta}_0 + \epsilon, \\ &\stackrel{\text{def}}{=} \alpha_0 + \widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\beta}_0 + g_1(Z) + g_2(U) + \epsilon,\end{aligned} \tag{2.1}$$