Journal of Statistical Planning and Inference 150 (2014) 49–65

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Nonlinear measurement error models subject to additive distortion

Jun Zhang [a], Yao Yu [b], Bu Zhou [c], Hua Liang [d,*]

[a] Shen Zhen-Hong Kong Joint Research Center for Applied Statistical Sciences, College of Mathematics and Computational Science, Institute of Statistical Sciences at Shenzhen University, Shenzhen University, Shenzhen, China
[b] Department of Biostatistics and Computational Biology, University of Rochester, NY 14642, USA
[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore
[d] Department of Statistics, George Washington University, Washington, DC 20052, USA

ARTICLE INFO

Article history:
Received 17 September 2013
Received in revised form
12 March 2014
Accepted 15 March 2014
Available online 28 March 2014

Keywords:
Bootstrap approximation
Confounding variables
Covariate-adjusted regression
Distorting function
Empirical likelihood
Empirical process
Measurement error models
Model checking
Multiplicative effect
Regression spline

ABSTRACT

We study nonlinear regression models when the response and predictors are unobservable and distorted in a multiplicative fashion by additive models of some observed confounding variables. After approximating the additive nonparametric components via polynomial splines and calibrating the error-prone response and predictors, we develop a semi-parametric profile nonlinear least squares procedure to estimate the parameters of interest. We show that the resulting estimators are asymptotically normal. We further suggest an empirical likelihood-based statistic for statistical inference to improve the accuracy of the associated normal approximation with the aim to avoid estimating the asymptotic covariance matrix that involves infinite-dimensional nuisance of additive distorting functions. We also show that the empirical likelihood statistic is asymptotically chi-squared. Moreover, a test procedure is proposed to check whether the parametric model is adequate or not under this distorted measurement error setting. A wild bootstrap procedure is suggested to compute p-values. Simulation studies are conducted to examine the performance of the proposed procedures. The methods are applied to analyze real data from a low birth infants weight for an illustration.

1. Introduction

Using nonlinear regression models for study of the relationships between the response and the predictors of interest in biomedical research, we may encounter situations that the response and the predictors cannot be directly observed and may be measured with errors. Simply ignoring the errors can cause estimation bias and lead to loss of power for accurately detecting the relationship among variables. See Carroll et al. (2006) for a variety of such real-world examples. In this paper, we consider a situation that both the response and predictors are unobservable and distorted by general multiplicative effects of some observable confounding variables in the following covariate-adjusted setting:

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

* Corresponding author.
E-mail addresses: zhangjunstat@gmail.com (J. Zhang), yuyaour@gmail.com (Y. Yu), zhoubu1988@yahoo.com.cn (B. Zhou), hliang@gwu.edu (H. Liang).

http://dx.doi.org/10.1016/j.jspi.2014.03.005
0378-3758/

where $f(\cdot,\cdot)$ is a known continuous nonlinear function, $\boldsymbol{\beta}$ is an unknown $q\times 1$ parameter vector on a compact parameter space $\Theta_\beta \subset \mathbb{R}^q$. Y is the unobservable response, $\boldsymbol{X}=(X_1,X_2,\ldots,X_p)^\tau$ is the unobservable continuous predictor vector (the superscript τ denotes the transpose operator throughout this paper), while $\tilde{Y}$ and $\tilde{\boldsymbol{X}}$ are available distorted response and predictors, $\psi(\cdot)$ is a $p\times p$ diagonal matrix $\mathrm{diag}(\psi_1(\cdot),\ldots,\psi_p(\cdot))$, where $\phi(\cdot)$ and $\psi_r(\cdot)$ are unknown contaminating functions of an observed confounding vector $\boldsymbol{U}=(U_1,\ldots,U_d)^\tau$. The diagonal form of $\psi(\cdot)$ indicates that the confounding vector $\boldsymbol{U}$ distorts each component of the unobserved predictors $\boldsymbol{X}$ in a multiplicative fashion. The confounding vector $\boldsymbol{U}$ and model error ε are independent of $(\boldsymbol{X},Y)$.

The distortion framework within model (1.1) was motivated from biomedical and health-related studies, in which confounding variables such as the body mass index (BMI), height or weight usually have multiplicative effects on the primary response and predictor variables. Recently, substantial literature has evolved to study distortion measurement errors. Two main estimation procedures include the transformation techniques proposed by Şentürk and Müller (2005, 2006, 2009), and the direct plug-in method proposed by Cui et al. (2009). The transformation technique was designed for models with linear structure such as linear models (Nguyen and Şentürk, 2008; Nguyen et al., 2008; Şentürk and Müller, 2005, 2006), generalized linear models (Şentürk and Müller, 2009) and partial linear single index models (Zhang et al., 2013). The direct plug-in method can be easily adopted in more general model settings such as semi-parametric models. See for example Cui et al. (2009), Li et al. (2010), Zhang et al. (2012a,b).

A commonly used procedure to eliminate the effect of the contamination caused by confounding variables is to divide the response and predictors by these confounding variables. For instance, in a study of the relationship between fibrinogen level and serum transferrin level among haemodialysis patients, Kaysen et al. (2002) realized that BMI generally has influence on the fibrinogen level and serum transferrin level and may contaminate these primary variables, and suggested a simple calibration procedure, dividing the observed fibrinogen level and observed serum transferrin level by BMI. From a practical point of view, since the exact relationship between the confounding variable and primary variables is unknown, naively dividing confounding variable BMI may be too rough and may lead to an inconsistent estimator of the parameter $\boldsymbol{\beta}$. As a remedy, Şentürk and Müller (2005) introduced a flexible multiplicative adjustment by adopting unknown smooth distorting functions $\phi(\cdot)$, $\psi_r(\cdot)$ of the confounding variables.

When more than one confounding variables affect the response and predictors, i.e., a d-dimensional confounding vector $\boldsymbol{U}$, we face the 'curse of dimensionality' problem if the distorting functions are directly modeled as $\phi(\boldsymbol{U})=\phi(U_1,\ldots,U_d)$, $\psi_r(\boldsymbol{U})=\psi_r(U_1,\ldots,U_d)$ with $\phi(\cdot)$, $\psi_r(\cdot)$ being unknown d-dimensional smooth functions. Nguyen and Şentürk (2008) modeled the distortions by single-index models and considered to model the primary variables as a linear regression model. The authors proposed a hybrid backfitting algorithm to simultaneously estimate unknown single-index parameters and varying coefficient functions, and derived final estimators of the parameters with a weighted-average of the estimated coefficient functions. Recently, Zhang et al. (2012a) extended the single-index distortions to the nonlinear regression models, and proposed a two-step estimation procedure. The estimating function method (EFM) proposed by Cui et al. (2011) was applied to estimate the single-index parameters, and the direct plug-in procedure was used to give a nonlinear least squares estimator of $\boldsymbol{\beta}$.

In this paper, we use additive models to model distortion functions $\phi(\boldsymbol{U})$ and $\psi(\boldsymbol{U})$ for their unique feature of interpretability and flexibility. See Hastie and Tibshirani (1990), Huang (2003), Li and Ruppert (2008), Opsomer and Ruppert (1997), and Stone (1985) for studies and applications of this kind of models in theory and in exploring complicated relationship between response and covariates of interest in data analysis. The spline approximation will be used to approximate the unknown distorting functions.

For the parameters estimation, we first calibrate distorted response and predictors by using spline approximations, then we estimate the parameter $\boldsymbol{\beta}$ by profile nonlinear least squares procedure and establish asymptotic normality for the proposed estimator. The linear regression models, a special case of nonlinear models, are discussed as well. As empirical likelihood method is more convenient and efficient for statistical inference than the normal approximation in practice, and can avoid estimating the complex asymptotic covariance matrix in the asymptotic distribution of the estimator of $\boldsymbol{\beta}$, we propose an empirical likelihood based statistic to construct confidence regions, and show that it is asymptotically chi-squared distributed.

For model checking in the distorted measurement error problems, the adequacy of nonlinear models has not been investigated to the best of our knowledge. The existing literature, as mentioned above, is devoted to estimation rather than testing. Clearly, testing the adequacy of the parametric models in the distorted errors-in-variables context is important and relevant. In the present paper, we apply the Cramér–von Mises (CvM) test proposed by Stute et al. (1998a) for model checking, and further investigate the bootstrap approximation for calculating critical values.

The paper is organized as follows. In Section 2, we propose an estimation procedure for the additive distortion setting and derive related asymptotic results. In Section 3, we develop an empirical log-likelihood ratio statistic for the parameter $\boldsymbol{\beta}$, and show that the empirical likelihood statistic has an asymptotic chi-squared distribution. In Section 4, we introduce a CvM test for model checking. In Section 5, we conduct simulation studies to examine the performance of the proposed methods. In Section 6, an analysis of a dataset of low birth weight infants is presented. All the technical proofs of the asymptotic results are given in the Appendix.