Statistics and Probability Letters 126 (2017) 219–229

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Hypothesis tests in partial linear errors-in-variables models with missing response

Hong-Xia Xu[a,b], Guo-Liang Fan[c,b], Zhen-Long Chen[a,*]

[a] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China
[b] School of Mathematics & Physics, Anhui Polytechnic University, Wuhu, China
[c] Research Center of Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, China

ARTICLE INFO

Article history:
Received 2 November 2016
Received in revised form 28 February 2017
Accepted 2 March 2017
Available online 22 March 2017

MSC:
62G10
62G20

Keywords:
Partial linear model
Missing at random
Errors-in-variables
U-statistics
Quadratic conditional moment test

ABSTRACT

In this paper, we investigate the problem of testing nonparametric function in partial linear errors-in-variables models with response missing at random. In order to overcome the bias produced by measurement errors, two bias-corrected test statistics based on the quadratic conditional moment method are proposed. The limiting null distributions of the test statistics are established respectively and p values can be easily determined which show that the proposed test statistics have similar theoretical properties. Moreover, our tests can detect the alternatives distinct from the null hypothesis at the optimal nonparametric rate for local smoothing-based methods in this area. Simulation studies are conducted to demonstrate the performance of the proposed test methods and the proposed two tests give similar performances. A real data set from the ACTG 175 study is used for illustrating the proposed test methods.

1. Introduction

Partial linear regression model (PLM) first introduced by Engle et al. (1986) can reduce high risk of misspecification related to a fully parametric model and avoid some serious drawbacks of fully nonparametric methods. In general, a partially linear or semiparametric regression model can be written as

$$Y = X^{\tau}\beta + g(T) + \varepsilon, \tag{1.1}$$

where Y is the response, $X \in R^p$ and $T \in R$ are explanatory variables, $\beta = (\beta_1, \ldots, \beta_p)^{\tau}$ is an unknown p-dimensional parameter vector, $g(\cdot)$ is an unknown smooth function defined on R and ε is the model error with $E(\varepsilon|X, T) = 0$.

It is well known that missing data are often encountered for various reasons such as unwillingness of some sampled units to supply the desired information, loss of information caused by uncontrollable factors, failure on the part of the investigator to gather correct information when the explanatory variables can be controlled. Many statisticians have investigated statistical inference with missing data. For example, Wang and Rao (2002), Zou et al. (2015) and Chown (2016) for estimation and Sun et al. (2009), Xu and Zhu (2013), Xu et al. (2017) and Cotos-Yáñez et al. (2016) for hypothesis testing. Hence, we assume the response Y is missing at random (MAR), and let $\delta = 0$ if Y is missing and $\delta = 1$ otherwise. The MAR assumption implies that δ and Y are conditionally independent given X and T, i.e., $P(\delta = 1|X, T, Y) = P(\delta = 1|X, T) = \Delta(X, T)$, where

[*] Corresponding author.
E-mail address: zlchenv@163.com (Z.-L. Chen).

http://dx.doi.org/10.1016/j.spl.2017.03.001
0167-7152/© 2017 Elsevier B.V. All rights reserved.

$\Delta(X, T)$ is called the propensity score or selection probability function representing the heterogeneity in the missingness mechanism. Further we define $\Delta_t(T) = P(\delta = 1|T)$.

In many practical fields, we may encounter the situation that the covariates are measured with errors. For example, it has been well documented that covariates such as blood pressure, urinary sodium chloride level, and exposure to pollutants are subject to measurement errors, and these cause difficulties in conducting a statistical analysis that involves them. Simply ignoring measurement errors will result in biased estimators. The measurement errors (or errors-in-variables, EV) models have been surveyed by lots of researchers, such as Fuller (1987), Carroll et al. (1995), Liang et al. (1999), Wang (1999), You et al. (2006), Fan et al. (2013), Feng and Xue (2014), Sun et al. (2015), Fan et al. (2016) and De Nadai and Lewbel (2016) among others. However, the above mentioned EV models mainly discussed estimation problems. There are few EV models related to hypothesis test issues especially for missing response.

In this paper, we assume the covariate X in model (1.1) is measured with additive error and one can only observe the surrogate variable W. The response Y is assumed to be MAR, i.e., δ and Y are conditionally independent given W and T and $P(\delta = 1|W, T, Y) = P(\delta = 1|W, T) = \Delta(W, T)$. Specifically, we consider the following partial linear errors-in-variables model (PLEVM),

$$\begin{cases} Y = X^\tau \beta + g(T) + \varepsilon, \\ W = X + \eta, \end{cases} \tag{1.2}$$

where η is the measurement error, which is independent and identically distributed (i.i.d.) with mean zero and known covariance matrix Σ_η, and is independent of (X, T, ε). Our aim is to test whether the nonparametric part in (1.2) is a parametric function:

$$\mathcal{H}_0 : g(T) = G(T, \theta_0) \quad \text{for some } \theta = \theta_0 \quad \text{vs.} \quad \mathcal{H}_1 : g(T) \neq G(T, \theta) \quad \text{for any } \theta, \tag{1.3}$$

where $G(\cdot, \theta)$ is a known function form. Since the measurement errors often make the estimator have bias, a bias-corrected estimation method is proposed. Then inspired by Zheng (1996), we introduce two quadratic conditional moment test statistics to test the testing problem (1.3). Our results show that the proposed two tests behave similarly in theoretical and numerical analysis.

The rest of the paper is organized as follows. Section 2 constructs two test statistics and establishes asymptotic properties of test statistics under the null hypothesis and local alternatives. Simulation studies and a real data analysis are carried out to reveal the performances of the tests in Section 3. The technical proofs are relegated to Appendix.

2. Test procedures

2.1. Bias-corrected estimation

Suppose that $\{(y_i, \delta_i, x_i, w_i, t_i), 1 \le i \le n\}$ is an i.i.d. random sample which comes from (Y, δ, X, W, T). For model (1.1), under the MAR assumption, Niu et al. (2016) adopted the following estimator $\tilde{\beta}_N$ to estimate β, i.e.,

$$\tilde{\beta}_N = \left[\sum_{i=1}^n \delta_i (x_i - \tilde{g}_1(t_i))(x_i - \tilde{g}_1(t_i))^\tau\right]^{-1} \sum_{i=1}^n \delta_i (x_i - \tilde{g}_1(t_i))(y_i - \hat{g}_2(t_i)), \tag{2.1}$$

where

$$\tilde{g}_1(t_i) = \frac{\sum_{j=1}^n \delta_j x_j K_h(t_i - t_j)}{\sum_{j=1}^n \delta_j K_h(t_i - t_j)} \quad \text{and} \quad \hat{g}_2(t_i) = \frac{\sum_{j=1}^n \delta_j y_j K_h(t_i - t_j)}{\sum_{j=1}^n \delta_j K_h(t_i - t_j)}$$

are the nonparametric estimators of $g_1(t_i) = \frac{E(\delta_i x_i|t_i)}{E(\delta_i|t_i)}$ and $g_2(t_i) = \frac{E(\delta_i y_i|t_i)}{E(\delta_i|t_i)}$, respectively for $i = 1, \ldots, n$. Here $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a sequence of positive numbers tending to zero, called bandwidth.

However, in our case, instead of observing x_i, we observe its surrogate w_i with $w_i = x_i + \eta_i$, for $i = 1, \ldots, n$. If one ignores the measurement error and replaces x_i by w_i directly, the resulting estimator is inconsistent and biased. Considering for this, we introduce a bias-corrected estimator $\hat{\beta}$ to estimate β as follows.

$$\hat{\beta} = \left[\sum_{i=1}^n \delta_i (w_i - \hat{g}_1(t_i))(w_i - \hat{g}_1(t_i))^\tau - \Sigma_\eta \sum_{i=1}^n \delta_i\right]^{-1} \sum_{i=1}^n \delta_i (w_i - \hat{g}_1(t_i))(y_i - \hat{g}_2(t_i)), \tag{2.2}$$

where $\hat{g}_1(t_i)$ has the same form as $\tilde{g}_1(t_i)$ except that x_j are replaced by w_j.

In order to estimate θ, we note that under the null hypothesis,

$$G(T, \theta) = E\left\{\frac{\delta(Y - X^\tau \beta)}{\Delta_t(T)}\bigg|T\right\} = E\left\{\frac{\delta(Y - W^\tau \beta)}{\Delta_t(T)}\bigg|T\right\}. \tag{2.3}$$