Statistics and Probability Letters 133 (2018) 65–70

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A nonparametric test for covariate-adjusted models

Zhao Jingxin [b,c], Xie Chuanlong [a,*]

[a] Jinan University, Guangzhou, China
[b] Guangzhou University, Guangzhou, China
[c] Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 30 March 2017
Received in revised form 28 September 2017
Accepted 10 October 2017
Available online 24 October 2017

Keywords:
Covariate-adjusted model
Nonparametric test
Model checking

ABSTRACT

This paper provides a nonparametric test for covariate-adjusted models. The proposed test statistic, obtained by using the adjusted response and predictors, has the same limit distribution as when the response and predictors are observed directly.

1. Introduction

Consider a general covariate-adjusted model where the response variable Y, the p-dimensional predictor $X = (X_1, \dots, X_p)^{\mathrm{T}}$ and their observable surrogates $\tilde{Y}, \tilde{X}$ are related to each other by the relations

$$Y = \mu(X) + \varepsilon, \;\; \tilde{Y} = \psi(U)Y, \;\; \tilde{X}_r = \phi_r(U)X_r, \;\; r = 1, \dots, p. \tag{1.1}$$

Here U is an observable confounder and $\psi(U)$ and $\phi_r(U)$ are unknown distorting functions. The covariate-adjusted model was first introduced in Sentürk and Müller (2005), where they investigated the linear relationship of the fibrinogen level and the serum transferrin level in hemodialysis patients. However both fibrinogen level and serum transferrin level are measured with a confounding effect from body mass index. They pointed out that it is reasonable to assume that the effect is multiplicative, with a factor that is an unknown function of body mass index. Similarly, Cui et al. (2009) considered to estimate the glomerular filtration rate by the serum creatinine level. With the body surface area as the distortion effect, a study of nonlinear covariate-adjusted model was conducted.

Most of the literature is about the estimation of the conditional mean function $\mu(X)$. Sentürk and Müller (2005, 2006) proposed their estimators for the linear covariate-adjusted model. Sentürk (2006) extended the linear model to a varying coefficient one, and Sentürk and Müller (2009) discussed the generalized linear model. Cui et al. (2009) and Zhang et al. (2012) presented the parameter estimation in more general nonlinear regression models. Partial linear models were considered by Zhang et al. (2013b). Without the parametric structure assumption, Delaigle et al. (2016) put forward several nonparametric covariate-adjusted estimators of $\mu(X)$. Other literature about covariate-adjusted models include Li et al. (2010), Nguyen and Sentürk (2007, 2008), Nguyen et al. (2008), Sentürk and Müller (2006, 2009), Sentürk and Nguyen (2009) and Zhang et al. (2014, 2013a).

* Corresponding author.
E-mail address: clxie0929@jnu.edu.cn (C. Xie).

https://doi.org/10.1016/j.spl.2017.10.004
0167-7152/

With so many estimation methods for different models, it is natural to ask whether the covariate-adjusted model possesses certain structure. To avoid wrong inference, a formal model checking procedure should be processed before further analysis. However, there is few work that focused on model checking for covariate-adjusted model in literature. Zhang et al. (2015) proposed a residual based empirical process test statistic and used bootstrap to calculate the critical values. There is no valid local smoothing test in literature for testing a parametric assumption of covariate-adjusted model. In this paper, we make attempts to fill this void. To check whether the parametric function structure is plausible, the null and alternative hypotheses are formulated as follows:

$$\mathcal{H}_0 : \mu(X) = m(X;\theta_0) \ \text{ for } \theta_0 \in \Theta \subset \mathbb{R}^d, \quad \text{versus}$$
$$\mathcal{H}_1 : \mu(X) \neq m(X;\theta) \ \text{ for all } \theta \in \Theta \subset \mathbb{R}^d.$$

In this paper, we propose a test for checking $\mathcal{H}_0$ in the context of covariate-adjusted model. The observed variables $\tilde{Y}$ and $\tilde{X}$ are firstly adjusted to estimate the latent response Y and predictor X. By plugging the adjusted variables into a local smoothing test statistic, a test statistic is built for the covariate-adjusted model. Large sample properties are established to support the test. Compared to the test in Zhang et al. (2015), the proposed test is widely applicable and easy to implement.

The paper is organized as follows: we describe the covariate-adjusted model and propose a local smoothing test to check $\mathcal{H}_0$ in Section 2. In Section 3, we present the large sample properties under the local alternative hypothesis. Section 4 reports the simulation results and a real data application. The assumptions are postponed to Appendix A and the proofs are in the Supplement.

2. Methodology

To identify $\psi(U)$ and $\phi_r(U)$, we assume that ε, X and U are mutually independent, $U \in [0, 1]$, and $\psi(U), \phi_r(U)$ are positive functions that satisfy

$$E[\psi(U)] = 1, \quad E[\phi_r(U)] = 1, \ \text{ for } r = 1, \dots, p. \tag{2.1}$$

The condition (2.1) means that there is no distortion effect on average, which is similar to $E[U] = 0$ for the classical additive measurement error $W = X + U$. Then according to above assumptions, we know

$$E[|\tilde{Y}|\big|U = u] = \psi(u)E[|Y|], \ \text{ and } E[|\tilde{X}_r|\big|U = u] = \phi_r(u)E[|X_r|].$$

Since (2.1) holds, $E[|Y|]$ and $E[|X_r|]$ are equal to $E[|\tilde{Y}|]$ and $E[|\tilde{X}_r|]$ respectively. In addition, the two conditional means at left hand side can be estimated by nonparametric estimators. Assume the observed data $\{(\tilde{y}_i, \tilde{x}_i, u_i), i = 1, \dots, n\}$ are generated by model (1.1) where $\tilde{x}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{ip})^{\mathrm{T}}$. Let $M(.)$ be a 4-th order kernel function (Cui et al., 2009; Zhu and Fang, 1996) and $M_g(.) = M(./g)/g$ where g is a bandwidth. To proceed further, denote

$$\hat{\psi}_0(u) = \frac{1}{n}\sum_{i=1}^{n} M_g(u-u_i)|\tilde{y}_i|, \quad \hat{\phi}_{r0}(u) = \frac{1}{n}\sum_{i=1}^{n} M_g(u-u_i)|\tilde{x}_{ir}|, \tag{2.2}$$
$$\hat{p}(u) = \frac{1}{n}\sum_{i=1}^{n} M_g(u-u_i), \quad \hat{E}[|X|] = \frac{1}{n}\sum_{i=1}^{n}|\tilde{x}_i|, \quad \hat{E}[|Y|] = \frac{1}{n}\sum_{i=1}^{n}|\tilde{y}_i|.$$

Then $\hat{\psi}(u) = \hat{\psi}_0(u)/(\hat{p}(u)\hat{E}[|Y|])$ and $\hat{\phi}_r(u) = \hat{\phi}_{r0}(u)/(\hat{p}(u)\hat{E}[|X_r|])$. Therefore $(y_i, x_i), i = 1, \dots, n$ can be estimated by

$$\hat{y}_i = \tilde{y}_i/\hat{\psi}(u_i), \quad \hat{x}_{ir} = \tilde{x}_{ir}/\hat{\phi}_r(u_i), \quad r = 1, \dots, p. \tag{2.3}$$

This covariate-adjusted procedure is slightly different to that in Zhang et al. (2015), because we use the method suggested in Delaigle et al. (2016) to drop the conditions that $E[X_r] > 0$ and $E[Y] > 0$. Cui et al. (2009) indicated that under $\mathcal{H}_0$, a $\sqrt{n}$-consistent estimate $\hat{\theta}_n$ of θ_0 can be obtained by nonlinear least square method with $\{(\hat{y}_i, \hat{x}_i)\}$ where $\hat{x}_i = (\hat{x}_{i1}, \dots, \hat{x}_{ip})^{\mathrm{T}}$. In other words, $\hat{\theta}_n$ has the same first-order asymptotic properties as the classical least squares estimator when Y and X are observable. This motivates us to replace $\{(y_i, x_i)\}$ in a test statistic and to check whether the convergence rate can be kept. Denote $\hat{\varepsilon}_i = \hat{y}_i - m(\hat{x}_i, \hat{\theta}_n)$. We define a test statistic as

$$V_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{x}_i - \hat{x}_j)\hat{\varepsilon}_i\hat{\varepsilon}_j, \tag{2.4}$$

where $K_h(.) = K(./h)/h$ and $K(.)$ is a kernel function which may be different to $M(.)$. Note V_n is an empirical version of $E\{\varepsilon E[\varepsilon|X]f_X(X)\} = E\{E^2[\varepsilon|X]f_X(X)\}$ which is zero if $\mathcal{H}_0$ is true and positive under $\mathcal{H}_1$. So if V_n is a large positive scale, we should reject the null hypothesis $\mathcal{H}_0$. In next section, we will investigate the asymptotic properties of the test statistic V_n and specify its reject criterion.