Computational Statistics and Data Analysis 118 (2018) 112–125

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Consistent test for parametric models with right-censored data using projections☆

Zhihua Sun [a,b,*], Xue Ye [a], Liuquan Sun [c]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Key Laboratory of Big Data Mining and Knowledge Management of CAS, Beijing, China
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 8 February 2017
Received in revised form 7 September 2017
Accepted 13 September 2017
Available online 22 September 2017

Keywords:
Consistent test
Curse-of-dimensionality-free
Empirical process
Projection
Right-censored data

ABSTRACT

In the literature, there are several methods to test the adequacy of parametric models with right-censored data. However, these methods will lose effect when the predictors are medium-high dimensional. In this study, a projection-based test method is built, which acts as if the predictors were scalar even if they are multidimensional. The proposed test is shown to be consistent and can detect the alternative hypothesis converging to the null hypothesis at the rate n^{-r} with $0 \leq r \leq 1/2$. Also, it is free from the choices of the subjective parameters such as bandwidth, kernel and weighting function. A wild bootstrap method is developed to determine the critical value of the test, which is shown to be robust to the model conditional heteroskedasticity. Simulation studies and real data analyses are conducted to validate the finite sample behavior of the proposed method.

1. Introduction

In recent years, there has been an upsurge of study on statistic models, especially on semiparametric models. However, parametric models are still valuable tools because they have prominent advantages such as the high precision, the good interpretability, the desiring prediction capability and the easy accessibility via existing procedures in statistic software. Li and Racine (2007) concluded that the correctly specified parametric models are usually a "first-best" solution for statistic inference. But if the parametric models are misspecified, the statistical analysis results will be erroneous. Therefore, it is significant to develop a formal testing procedure for parametric models.

The adequacy check of parametric models has attracted a lot of attention since (Bierens, 1982) first proposed a consistent test. Escanciano (2006) divided the existing methods into two categories: local approaches and integrated ones. For the local approaches, see Dette (1999), Härdle et al. (1998), Horowitz and Härdle (1994), Eubank and Hart (1992), Härdle and Mammen (1993), and Zheng (1996), among others. For the integrated method, we can refer to Bierens (1982), Bierens and Ploberger (1997), Stute et al. (1998), Stute and Zhu (2002), Escanciano (2006), Sun and Wang (2009) and the references within.

In medical, biologic and economic research fields, one may find that the response variable is often rightly censored because of the end of the study or the loss of follow-up. Several existing approaches aim at checking the adequacy of parametric models with right-censored data. For example, Stute et al. (2000) extended the method of Stute et al. (1998)

☆ The research was supported by the National Natural Science Foundation of China (Grant Nos. 11231010, 11690015, 11571340, U1430103), the National Center for Mathematics and Interdisciplinary Sciences and the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, CAS.

* Corresponding author.
E-mail address: sunzh@amss.ac.cn (Z. Sun).

https://doi.org/10.1016/j.csda.2017.09.005
0167-9473/

to deal with the right-censored data. Pardo-Fernández et al. (2007) constructed a testing method based on the distribution difference of the estimated residuals of the parametric and nonparametric models. Lopez and Patilea (2009) employed a U-process to build a test statistic. All these methods perform well in obtaining reasonable empirical sizes and high empirical powers for a low dimensional parametric model. However, they will lose effect when the predictors are medium-high dimensional. The methods of Pardo-Fernández et al. (2007) and Lopez and Patilea (2009) applied the local smoothing method, which causes the testing methods suffer from "curse of dimensionality". For the integrated method of Stute et al. (2000), the test statistic based on the indicator function tends to degenerate to zero when the covariates are medium-high dimensional.

In this paper, we develop a testing method free from the curse of dimensionality for parametric models with right-censored data. A projection-based testing statistic is constructed by applying a linear indicator weighting function. We show that even if the covariates are multivariant, the proposed testing method acts as if they were scalar. Furthermore, we carry through extensive simulation studies to validate that the proposed method performs over the existing tests in terms of the empirical sizes and powers when the predictors are medium or high dimensional. Besides the powerful advantage to deal with medium or high dimensional data, the proposed method has the following merits: it is consistent; it can detect the alternative hypothesis converging to the null hypothesis at the rate n^{-r} with $0 \leq r \leq 1/2$; it is free from the subjective parameters such as bandwidth, kernel and weighting function. To determine the critical value, a wild bootstrap method is proposed, which is shown to be robust to the conditional heteroskedasticity.

The rest of this paper is organized as follows. We describe the testing problem and an estimating procedure in Section 2. In Section 3, we propose the testing method and study its asymptotic properties. The analysis of the power is conducted in Section 4. In Section 5, we consider the calculation of the testing statistic and develop a bootstrap method to calculate the critical value. Simulation studies and two real data analyses are conducted in Sections 6 and 7, respectively. The proofs of the main results are collected in Appendix.

2. Existing estimation of the null hypothetical model

Let X be a p-dimensional predictor and Y be a scalar response. Consider the regression model: $Y = m(X) + e$, where $m(x) = E(Y|X = x)$ and e is the error variable with $E(e|X) = 0$. In this study, we consider the null hypothesis:

$$H_0 : P\{m(X) = g(X, \beta)\} = 1 \quad \text{for some} \quad \beta, \tag{2.1}$$

where g is a known function and β is an unknown parameter. The alternative hypothesis can be written as:

$$H_1 : \quad P\{E(Y|X) = g(X, \beta)\} < 1 \ \text{ for all } \ \beta \in R^p.$$

In presence of random right censoring, the response Y is not always available. Let C denote the censoring time variable with the distribution $G(\cdot)$ and $\delta = I(Y \leq C)$. Then, instead of Y, the variable $Z = Y \wedge C$ is observed.

We first describe an estimating procedure of the null hypothetical model. To deal with randomly-censored data, the inverse probability weighting method is an effective way. There are two means to employ the inverse probability: One aims for adjusting the response variable; and the other aims at adjusting the objective least square function or equivalently for adjusting the estimating equation. The former is called the synthetic data (SD) method and the latter is called the weighted least squares (WLS) method. It is well known that the estimating procedure is very critical to the effect of the model checking method. Lopez and Patilea (2009) showed that the testing method based on the WLS estimating procedure outperforms the test based on the SD method. The WLS method calibrates the model error directly, which is appropriate for the model checking problem. Therefore, we apply the WLS method to estimate the null hypothetical model.

Assume that we have an i.i.d. sample $\{(Z_i, \delta_i, X_i), i = 1, 2, \ldots, n\}$ from (Z, δ, X). The WLS method defines an estimator of β, denoted by $\hat{\beta}_n$, which minimizes the following weighted least squared objective function:

$$M_n(\beta) = \sum_{i=1}^{n} \frac{\delta_i}{1 - \hat{G}(Z_i-)}(Z_i - g(X_i, \beta))^2, \tag{2.2}$$

where $\hat{G}(z) = 1 - \prod_{j:Z_j \leq z}(1 - 1/\sum_{k=1}^{n} I(Z_j \leq Z_k))^{1-\delta_j}$ is the Kaplan–Meier estimator of $G(z)$. More details on $\hat{G}(z)$ can refer to Koul et al. (1981).

3. Testing methods

3.1. Existing testing methods

Based on the estimated null hypothetical model, an estimated model error can be constructed: $\hat{e}_i = nW_{in}(Z_i - g(X_i, \hat{\beta}_n))$ with $W_{in} = \delta_i/(n(1 - \hat{G}(Z_i-)))$. Observe that the null hypothesis (2.1) is equivalent to $E(eI(X < x)) = 0$ for any x with $e = Y - g(X, \beta)$. An empirical-process based test statistic can be defined as:

$$\mathcal{T}_{n,e} = \int [R_{n,e}(x)]^2 F_n(dx),$$