Journal of Multivariate Analysis 154 (2017) 147–161

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Improved model checking methods for parametric models with responses missing at random

Zhihua Sun [a,b], Feifei Chen [c], Xiaohua Zhou [d], Qingzhao Zhang [e,*]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Key Laboratory of Big Data Mining and Knowledge Management of CAS, Beijing, China
[c] School of Statistics, Renmin University of China, Beijing, China
[d] Department of Biostatistics, University of Washington and HSR&D Center of Excellence, VA Puget Sound Health Care System, WA, USA
[e] School of Economics and the Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China

ARTICLE INFO

Article history:
Received 22 April 2016
Available online 14 November 2016

AMS subject classifications:
62F03
62G10

Keywords:
Empirical process
High dimensional covariates
Missing at random
Model checking
Projection

ABSTRACT

In this paper, we consider the lack-of-fit test of a parametric model when the response variable is missing at random. The popular imputation and inverse probability weighting methods are first employed to tackle the missing data. Then by employing the projection technique, we propose empirical-process-based testing methods to check the appropriateness of the parametric model. The asymptotic properties of the test statistics are obtained under the null and local alternative hypothetical models. It is shown that the proposed testing methods are consistent, and can detect local alternative hypothetical models converging to the null model at the parametric rate. To determine the critical values, a consistent bootstrap method is proposed, and its asymptotic properties are established. The simulation results show that the tests outperform the existing methods in terms of empirical sizes and powers, especially under the situation with high dimensional covariates. Analysis of a diabetes data set of Pima Indians is carried out to demonstrate the application of the testing procedures.

1. Introduction

Parametric models are valuable tools for exploring the relationship between a response and a set of covariates (explanatory variables). Due to its appealing advantages such as the high precision, the good interpretation and the accurate prediction, a correctly specified parametric model is usually a "first-best" solution [10]. However, a misspecified parametric model would cause misleading analysis results. Therefore, it is important to perform a model checking procedure to validate whether the specified parametric model is acceptable.

The lack-of-fit test of parametric models has attracted a lot of attention. The existing consistent tests can be mainly divided into two categories: "local approach" and "integrated approach" [2]. The local approach uses nonparametric smoothing techniques. See [7,9,26] for some examples. Recent works by [5,8] aim to avoid the curse of dimensionality for the local approach. Specially, [5] creatively employed the sufficient dimension reduction technique to lack-of-fit test and proposed a dimension reduction model-adaptive test procedure. The integrated approach is based on estimated empirical processes. We can refer to [1,17,18] and the references within for more details of the integrated approach. Recently, [2]

* Corresponding author.
E-mail address: zhangqingzhao@amss.ac.cn (Q. Zhang).

http://dx.doi.org/10.1016/j.jmva.2016.11.003

proposed a projection-based integrated approach to avoid the curse of dimensionality. More discussions and related literature can be found in [3,15,24].

When some subjects suffer from missing responses, many researchers have investigated model checking problem of parametric regression models. For the local approach, [4,6] respectively extended the methods in [7,26] to the issues of model checking with response missing at random. For the integrated approach, [19] constructed an empirical process based test with a simple indicator weighting function, and the numerical results showed their proposed method outperforms that of [4] in terms of the empirical powers and sizes. However, the methods of [4,6] involve the kernel smoothingness and suffer from the typical curse of dimensionality. The empirical process based tests with the simple indicator weighting function [19] tend to degenerate zero and would lose effect when the dimension of the explanatory variable is high or even moderate.

In this paper, we aim to study the lack-of-fit tests of parametric models with response missing at random. We construct tests in the framework of the integrated approach by resorting to the idea of projection. The tests can be regarded as an extension of the test in [2]. The proposed methods are valid even if the covariates are medium or high dimensional. Besides this merit, it is shown that the proposed tests inherit the advantages of the empirical process-based tests with the simple indicator weighting function: they are consistent, and they can detect the local alternative hypothetical models converging to the null model with the rate n^{-r}, $0 \le r \le 1/2$. It is well known that the parametric rate $n^{-1/2}$ is the fastest rate that the lack-of-fit test can detect.

The rest of this paper is organized as follows. In Section 2, two new empirical-process testing methods using projection are constructed. The asymptotic properties of the test statistics under the null and alternative models are rigorously investigated. We further propose the wild bootstrap method for calculating the critical value. Simulation studies and a real data analysis are conducted in Sections 3 and 4, respectively. The article concludes with discussions in Section 5. The estimation procedures of the response probability and the null hypothetical model, the conditions and proofs of the main results are presented in Appendix.

2. Testing methods using projection

2.1. Construction of test methods

In this study, we consider the null hypothesis,

$$\mathcal{H}_0 : \Pr\{\mathrm{E}(Y|X) = g(X, \beta_0)\} = 1 \quad \text{for some } \beta_0, \tag{2.1}$$

against the alternatives,

$$\mathcal{H}_1 : \ \Pr\{\mathrm{E}(Y|X) = g(X, \beta)\} < 1 \quad \text{for all } \beta \in \mathbb{R}^P$$

where Y is a scalar response variable with some potential missingness, X is a p-dimensional covariate, $g(x, \beta_0)$ is a known function and β_0 is an unknown parameter. We have a sample $\{(Y_i, X_i, \delta_i), i = 1, \dots, n\}$ as a copy of (Y, X, δ), where δ is a missing indicator variable with $\delta = 1$ if the response variable is observed and 0, otherwise. We assume that the response variable is missing at random (MAR); that is, $\Pr(\delta = 1|X, Y) = \Pr(\delta = 1|X)$, denoted by $p(X)$. This missing mechanism means that the missing process is unrelated to the missing variable itself. [12] provides more details about the MAR mechanism.

Under the MAR assumption and the null hypothesis $\mathcal{H}_0$, by employing the techniques of imputation and inverse probability weight, we have

$$\mathrm{E}\Big\{\delta Y + (1-\delta)g(X, \beta_0)\Big\} = \mathrm{E}(Y), \quad \text{and} \quad \mathrm{E}\left\{\frac{\delta}{p(X)}Y + \left(1 - \frac{\delta}{p(X)}\right)g(X, \beta_0)\right\} = \mathrm{E}(Y).$$

Thus we can construct the following two completed data sets:

$$\left\{\Big(\delta_i Y_i + (1-\delta_i)g(X_i, \hat{\beta}_n), X_i, \delta_i\Big),\ i = 1, \dots, n\right\},$$

and

$$\left\{\left(\frac{\delta_i}{\hat{p}_n(X_i)}Y_i + \left(1 - \frac{\delta_i}{\hat{p}_n(X_i)}\right)g(X_i, \hat{\beta}_n), X_i, \delta_i\right),\ i = 1, \dots, n\right\}.$$

The definitions of $\hat{p}_n(x)$ and $\hat{\beta}_n$ are presented in Appendix A.1.

Let the model error $e(Z, \beta) = Y - g(X, \beta)$ with $Z = (X, Y)$. Note that the null hypothesis (2.1) is equivalent to that $\mathrm{E}\{e(Z, \beta_0)\mathbf{1}(X < x)\} = 0$ for any x. We consider the following estimated empirical processes of $\mathrm{E}\{e(Z, \beta)\mathbf{1}(X < x)\}$:

$$R^s_{n,j}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{e}_j(Z_i, \hat{\beta}_n)\mathbf{1}(X_i < x), \quad j = 1, 2,$$

where

$$\hat{e}_1(Z_i, \hat{\beta}_n) = \delta_i Y_i + (1-\delta_i)g(X_i, \hat{\beta}_n) - g(X_i, \hat{\beta}_n)$$