Ann Inst Stat Math (2015) 67:229–259
DOI 10.1007/s10463-014-0451-3

Model checking for parametric regressions with response missing at random

Xu Guo · Wangli Xu · Lixing Zhu

Received: 19 December 2012 / Revised: 24 November 2013 / Published online: 27 March 2014

Abstract This paper aims at investigating model checking for parametric models with response missing at random which is a more general missing mechanism than missing completely at random. Different from existing approaches, two tests have normal distributions as the limiting null distributions no matter whether the inverse probability weight is estimated parametrically or nonparametrically. Thus, p values can be easily determined. This observation shows that slow convergence rate of nonparametric estimation does not have significant effect on the asymptotic behaviors of the tests although it may have impact in finite sample scenarios. The tests can detect the alternatives distinct from the null hypothesis at a nonparametric rate which is an optimal rate for locally smoothing-based methods in this area. Simulation study is carried out to examine the performance of the tests. The tests are also applied to analyze a data set on monozygotic twins for illustration.

X. Guo
Department of Mathematics, Hong Kong Baptist University, Fong Shu-Chuen Library 1110, Kowloon Tong, Hong Kong
e-mail: liushengjunyi@163.com

X. Guo
College of Economics and Management, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, 210016, Nanjing, China

W. Xu
School of Statistics, Center for Applied Statistics, Renmin University of China, Zhongguancun Street 59, 100872 Beijing, China
e-mail: xwlbnu@163.com

L. Zhu (✉)
Department of Mathematics, Hong Kong Baptist University, Fong Shu Chuen Library 1208 Waterloo Road 224, Kowloon Tong, Hong Kong
e-mail: lzhu@hkbu.edu.hk

Keywords Inverse probability weight · Response missing at random · Model checking

1 Introduction

The parametric regression model has received considerable attention, and relationship between the scalar response Y and the covariates X of dimension m is described as

$$Y = f(X, \theta_0) + \epsilon, \tag{1}$$

where $f(\cdot, \theta_0)$ is a known parametric function, θ_0 is an unknown parameter vector of p-dimension. It is assumed that the conditional expectation of ϵ given X is zero.

To prevent wrong conclusion and improve estimation efficiency, it is important to develop testing methods to ascertain whether the hypothetical parametric model is satisfied. When the response measurements are all available, there are a number of proposals available in the literature. For example, Härdle and Mammen (1993) constructed a test statistic that is based on the L^2 distance between parametric and nonparametric estimators with the assistance from the wild bootstrap for critical value determination. Zheng (1996) suggested a consistent test of functional form of nonlinear regression models. Stute et al. (1998b) and Stute and Zhu (2002) considered to check the parametric regression models by replacing the residual cusum processes by their innovation martingale, and the resulting tests are asymptotically distribution free. Aerts et al. (1999) constructed tests that are based on orthogonal series that involved selecting a nested model sequence in the bivariate regression. Fan and Huang (2001) introduced a method called Neyman threshold test using the fact that the Fourier transform of the residuals compresses useful signals into low frequencies so that the power of the adaptive Neyman test can be enhanced. Stute et al. (2008) constructed a test that is based on the residual empirical process marked by proper functions of the regressors to deal with large dimension of the regressor vector. Eubank et al. (2005) proposed data-driven lack-of-fit tests using fit comparison statistics that are based on nonparametric linear smoothers. All these tests can usually be classified into two categories: using local smoothing methods (nonparametric function fitting) to construct test statistics and using global smoothing methods (empirical process) to define tests. It is well known that the two types of methodologies have their own pros and cons. The former can be more sensitive to high-frequency alternative models than the tests based on the latter methodology, whereas the latter is more sensitive to smooth alternatives, and can detect the alternatives distinct from the null at faster convergence rate. Thus, both methodologies have been the main methodologies popularly used in practice. In this paper, we construct tests that can be classified into the first category in our setting. We will see that the limiting null distributions are normal and thus determining p values is easily implemented. We will also make a limited comparison with a test in the latter category to see their advantages and disadvantages in the simulation study.

In practice, it is often the case that not all response measurements are observable. For example, due to limited budget, only the responses for a part of subjects among the fully cohort are measured. Individuals may refuse to answer certain questions, or investigators forget to write down the related information. Thus, it is of interest for us to investigate model checking with missing response. Among others, González-Manteiga and Pérez-González (2006) constructed a test that is based on the L_2 distance between the nonparametric and parametric fits. Xu et al. (2012) defined a residual-marked empirical process to construct a test. Based on two completed samples, which are constructed by imputation and inverse probability weighting methods, Sun and Wang (2009) introduced two score-type tests and two empirical process-based tests for the general linear models with missing response. Recently, Li (2012) proposed a test that is based on minimum integrated square distances between the nonparametric and parametric fits, which can be viewed as an extension of the minimum distance test proposed by Koul and Ni (2004) to handle missing responses.

In this paper, we propose two tests for model (1). For a known parameter function $f(\cdot, \cdot)$, almost surely, the null hypothesis is

$$H_0\colon E(Y|X) = f(X, \theta_0), \tag{2}$$

for some θ_0 against alternative hypothesis

$$H_1\colon E(Y|X) \neq f(X, \theta), \tag{3}$$

for any θ. The interesting feature of the newly proposed tests is that although the tests are also dependent on nonparametric smoothing, belonging to the category of local smoothing methodologies, the limiting null distributions are tractable for p value determination. This advantage makes the tests easy to implement compared with existing ones. The tests can be regarded as an extension of Zheng (1996)'s test. As discussed above, there are many other possible approaches which can be used to handle the problem. We focus on the Zheng (1996)'s test in this article due to its technical tractability and easy computation. Dette and von Lieres und Wilkau (2001) compared several tests for additivity by kernel-based methods. They pointed out that, for realistic sample sizes, the bias has to be taken into account. For Zheng (1996)'s method, its standardized version has no bias converging to infinity and thus no bias-correction is needed. Gao et al. (2011) argued that a major advantage of Zheng (1996)'s method over its competitors is that an indirect estimator of the unknown nonparametric $\sigma^2(X) = E(\epsilon^2|X)$ is used to replace $\sigma^2(X)$. They believed such a feature is attractive when the conditional variance function $\sigma^2(X)$ is a generally smooth function. Further, no matter whether the inverse probability is estimated parametrically or nonparametrically, the tests interestingly have the same asymptotic properties although in finite sample scenarios, they should have different performances. Compared with the test developed by Li (2012), both higher-order kernel functions and trimming on the boundary of a density function that are used in many applications of nonparametric regressions are not needed. Also compared with Sun and Wang (2009), we do not need to construct the completed data set first for test statistic construction.

The rest of this paper is organized as follows. In Sect. 2, we construct the test statistics and derive their asymptotic properties under the null hypothesis and local alternatives. In Sect. 3, simulation results are reported to examine the performance of the tests and a real data analysis is carried out for illustration. The proofs are presented in the Appendix.

2 Test procedures

2.1 Construction of test statistics

For model (1), it is assumed that the response Y is missing at random (MAR), while the observations for the covariate X are available. Let δ be the missing indicator for the individual whether Y is observed ($\delta = 1$) or not ($\delta = 0$). Then, MAR implies

$$P(\delta = 1|Y, X) = P(\delta = 1|X) = \pi(X).$$

MAR is an usual missing mechanism in practice, which is more general than missing completely at random (MCAR), see Little and Rubin (1987).

Denote $\epsilon = Y - f(X, \theta_0)$; under the MAR assumption, we have

$$E\left(\frac{\delta}{\pi(X)}\epsilon|X\right) = E\left[\epsilon E\left(\frac{\delta}{\pi(X)}|X, Y\right)|X\right] = E[\epsilon|X].$$

Or equivalently

$$E(\delta\epsilon|X) = E[\epsilon E(\delta|X, Y)|X] = \pi(X)E(\epsilon|X).$$

Consequently, under H_0 of (2), we have

$$E\left(\frac{\delta}{\pi(X)}\epsilon E\left(\frac{\delta}{\pi(X)}\epsilon|X\right)W(X)\right) = E\left(E^2\left(\frac{\delta}{\pi(X)}\epsilon|X\right)W(X)\right) = 0,$$
$$E(\delta\epsilon E(\delta\epsilon|X)W(X)) = E(E^2(\delta\epsilon|X)W(X)) = 0, \tag{4}$$

where $W(X)$ is some positive weight function which will be discussed below. Under the alternative hypothesis H_1, $E(\epsilon|X) \neq 0$, we have

$$E\left(\frac{\delta}{\pi(X)}\epsilon E\left(\frac{\delta}{\pi(X)}\epsilon|X\right)W(X)\right) = E\left(E^2\left(\frac{\delta}{\pi(X)}\epsilon|X\right)W(X)\right) > 0,$$
$$E(\delta\epsilon E(\delta\epsilon|X)W(X)) = E(E^2(\delta\epsilon|X)W(X)) > 0. \tag{5}$$

Thus, the null hypothesis H_0 holds if and only if the Eq. (4) are zero. In other words, both can be used to be the bases for constructing test statistics. The empirical version of the left-hand side in (4) can then be used to define test statistics. We will discuss their pros and cons later.

Let $(x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)$ be an i.i.d. sample from (X, Y, δ). Estimate the terms $E(\delta\epsilon/\pi(X)|X = x)$ and $E(\delta\epsilon|X = x)$ by, respectively,

$$\hat{E}\left(\frac{\delta}{\pi(X)}\epsilon|x_i\right) = \frac{1}{n-1}\sum_{j\neq i}^{n}\frac{\delta_j}{\hat{\pi}(x_j)}K_h(x_i - x_j)\hat{\epsilon}_j/\hat{p}(x_i),$$

$$\hat{E}(\delta\epsilon|x_i) = \frac{1}{n-1}\sum_{j\neq i}^{n}\delta_j K_h(x_i - x_j)\hat{\epsilon}_j/\hat{p}(x_i).$$

where $\hat{\epsilon}_j = y_j - f(x_j, \hat{\theta}_N)$ with $\hat{\theta}_N$ being an estimator of θ_0, which will be specified later, $\hat{\pi}(x)$ is a nonparametric estimator of $\pi(x)$, $K_h(\cdot) = K(\cdot/h)/h^m$ with $K(\cdot)$ being a kernel function and h being the bandwidth, and $\hat{p}(x)$ is the estimator of the density of X $p(x)$ defined as

$$\hat{p}(x_i) = \frac{1}{n-1}\sum_{j\neq i}^{n}\frac{\delta_j}{\hat{\pi}(x_j)}K_h(x_i - x_j).$$

Since our aim is to construct some efficient and simple tests, a natural selection of the weight function will be the density function $p(x)$ because it can eliminate the boundary effect of the kernel estimation. Two test statistics are defined as follows,

$$\begin{aligned} T_n^N &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(x_i)}\hat{\epsilon}_i\sum_{j\neq i}^{n}\frac{\delta_j}{\hat{\pi}(x_j)}K_h(x_i - x_j)\hat{\epsilon}_j \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i}{\hat{\pi}(x_i)}\frac{\delta_j}{\hat{\pi}(x_j)}K_h(x_i - x_j)\hat{\epsilon}_i\hat{\epsilon}_j; \\ R_n^N &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\delta_i\hat{\epsilon}_i\sum_{j\neq i}^{n}\delta_j K_h(x_i - x_j)\hat{\epsilon}_j \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\delta_i\delta_j K_h(x_i - x_j)\hat{\epsilon}_i\hat{\epsilon}_j. \end{aligned} \tag{6}$$

In general, the function $\pi(X)$ is unknown and we can estimate it by a kernel estimator:

$$\hat{\pi}(x_i) = \frac{\sum_{j=1}^{n}\delta_j K_h(x_i - x_j)}{\sum_{j=1}^{n}K_h(x_i - x_j)}. \tag{7}$$

When $\pi(X)$ follows a parametric structure, that is, $\pi(X) = \pi(X, \alpha)$, we then only need to estimate the parameter α. As an example, for the logistic regression expressed as $\pi(x_i, \alpha) = (1 + \exp(-\alpha_0 - \alpha_1^{\tau}x_i))^{-1}$, where $\alpha = (\alpha_0, \alpha_1)^{\tau}$ is an unknown vector parameter, we can obtain consistent estimators of the regression coefficients $\hat{\alpha}$ by the maximum likelihood estimation. Then, the corresponding estimator of $\pi(x, \alpha)$ follows as

$$\pi(x_i, \hat{\alpha}) = (1 + \exp(-\hat{\alpha}_0 - \hat{\alpha}_1 x_i))^{-1}. \tag{8}$$

Below we analyze the estimation of the regression parameter θ_0 using the inverse probability weight least-squares method:

$$\hat{\theta}_N = \arg\min \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i)} \{y_i - f(x_i, \theta)\}^2,$$

when $\pi(X)$ is estimated nonparametrically; and

$$\hat{\theta}_P = \arg\min \sum_{i=1}^{n} \frac{\delta_i}{\pi(x_i, \hat{\alpha})} \{y_i - f(x_i, \theta)\}^2,$$

when $\pi(X, \alpha)$ is estimated parametrically. Correspondingly, when the parameter function $\pi(X, \alpha)$ is estimated by $\pi(X, \hat{\alpha})$, we denote the statistics in (6) as, respectively,

$$\begin{aligned} T_n^P &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i}{\pi(x_i, \hat{\alpha})} \frac{\delta_j}{\pi(x_j, \hat{\alpha})} K_h(x_i - x_j) \hat{\epsilon}_i \hat{\epsilon}_j, \\ R_n^P &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \delta_i \delta_j K_h(x_i - x_j) \hat{\epsilon}_i \hat{\epsilon}_j, \end{aligned} \tag{9}$$

where $\hat{\epsilon}_j = y_j - f(x_j, \hat{\theta}_P)$.

Remark 1 The proposed tests R_n^N and R_n^P are some modifications of the complete case-based tests. Though we only use the completely observed units in the test constructions, an inverse probability weight method is adopted as seen on page 6 to estimate the parameter θ_0 and get $\hat{\epsilon}_i$. Use of the inverse probability weight method requires estimation of $\pi(\cdot)$ or $\pi(\cdot, \alpha)$ by all the available data. Thus, the asymptotic properties may not be directly derived from the transfer principle which was recently developed by Koul et al. (2012). In that paper, they proved the efficiency of complete case statistics in the situation of missing response at random situation, see also Müller and Van Keilegom (2012) and Chown and Müller (2013) for more discussions about the transfer principle. On the other hand, we note that the asymptotic properties of R_n^N and R_n^P are easier to develop compared with those for T_n^N and T_n^P. Thus in the appendix, we focus on the asymptotic properties of T_n^N and T_n^P.

2.2 Asymptotic behavior of the test statistics

Interestingly, we find that both the test statistics T_n^N in (6) and T_n^P in (9) have the same asymptotic properties. Also, the asymptotic properties of R_n^N in (6) and R_n^P in (9) are equivalent. To state the theorems, we introduce some notations that are related to the limiting variances of the test statistics. Let

$$
\begin{aligned}
\hat{\Sigma}^{\mathrm{TN}} &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^m} \frac{\delta_i \delta_j}{\hat{\pi}^2(x_i)\hat{\pi}^2(x_j)} K^2\left(\frac{x_i - x_j}{h}\right) \hat{\epsilon}_i^2 \hat{\epsilon}_j^2, \\
\hat{\Sigma}^{\mathrm{RN}} &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^m} \delta_i \delta_j K^2\left(\frac{x_i - x_j}{h}\right) \hat{\epsilon}_i^2 \hat{\epsilon}_j^2,
\end{aligned} \tag{10}
$$

$\hat{\Sigma}^{\mathrm{TP}}$ and $\hat{\Sigma}^{\mathrm{RP}}$ are similarly defined as $\hat{\Sigma}^{\mathrm{TN}}$ and $\hat{\Sigma}^{\mathrm{RN}}$ respectively except for using $\pi(x_i, \hat{\alpha})$ and $\hat{\epsilon}_i = y_i - f(x_i, \hat{\theta}_P)$ instead of $\hat{\pi}(x_i)$ and $\hat{\epsilon}_i = y_i - f(x_i, \hat{\theta}_N)$ respectively. The asymptotic normalities for T_n^N, T_n^P and R_n^N, R_n^P under H_0 are stated below.

Theorem 1 *Under H_0 and the conditions in Appendix, we have*

$$
nh^{m/2} T_n^N \to N(0, \Sigma^T), \quad \textit{and} \quad nh^{m/2} T_n^P \to N(0, \Sigma^T),
$$

$$
nh^{m/2} R_n^N \to N(0, \Sigma^R), \quad \textit{and} \quad nh^{m/2} R_n^P \to N(0, \Sigma^R),
$$

where

$$
\begin{aligned}
\Sigma^T &= 2 \int K^2(u) du \cdot \int \frac{(\sigma^2(x))^2 p^2(x)}{\pi^2(x)} \mathrm{d}x. \\
\Sigma^R &= 2 \int K^2(u) du \cdot \int (\sigma^2(x))^2 p^2(x) \pi^2(x) \mathrm{d}x.
\end{aligned}
$$

Moreover, Σ^T can be consistently estimated by $\hat{\Sigma}^{\mathrm{TN}}$ or $\hat{\Sigma}^{\mathrm{TP}}$, which depends on whether $\pi(x_i)$ is estimated parametrically or nonparametrically, and Σ^R can be consistently estimated by $\hat{\Sigma}^{\mathrm{RN}}$ or $\hat{\Sigma}^{\mathrm{RP}}$, that is, in probability

$$
\begin{aligned}
\hat{\Sigma}^{\mathrm{TN}} &\to \Sigma^T, \quad \textit{and} \quad \hat{\Sigma}^{\mathrm{TP}} \to \Sigma^T; \\
\hat{\Sigma}^{\mathrm{RN}} &\to \Sigma^R, \quad \textit{and} \quad \hat{\Sigma}^{\mathrm{RP}} \to \Sigma^R.
\end{aligned}
$$

When there are no missing data, that is, $\pi(x) \equiv 1$, Σ^T and Σ^R are identical to Σ in Zheng (1996). Theorem 1 then is the same as Lemma 3.3 in Zheng (1996). Further compared with the results in the models without missing data, we can see clearly that the test statistics T_n^N and T_n^P induce larger asymptotic variances whereas R_n^N and R_n^P can obtain smaller asymptotic variances. However, this does not mean that R_n^N and R_n^P generally are more powerful compared with T_n^N and T_n^P, since the powers of the tests also depend on the non-random drifts. We will discuss this point later. According to Theorem 1, the standardized versions of the test statistics V_n^{TN}, V_n^{TP}, V_n^{RN} and V_n^{RP} can be defined as follows

$$
\begin{aligned}
V_n^{\mathrm{TN}} &= nh^{m/2} T_n^N / \sqrt{\hat{\Sigma}^{\mathrm{TN}}}, \quad \text{and} \quad V_n^{\mathrm{TP}} = nh^{m/2} T_n^P / \sqrt{\hat{\Sigma}^{\mathrm{TP}}} \\
V_n^{\mathrm{RN}} &= nh^{m/2} R_n^N / \sqrt{\hat{\Sigma}^{\mathrm{RN}}}, \quad \text{and} \quad V_n^{\mathrm{RP}} = nh^{m/2} R_n^P / \sqrt{\hat{\Sigma}^{\mathrm{RP}}}.
\end{aligned}
$$

By Slusky Theorem, we have the following corollary.

Corollary 1 *Under H_0 and the conditions in Appendix, we have*

$$V_n^{\mathrm{TN}} \to N(0,1), \quad \text{and} \quad V_n^{\mathrm{TP}} \to N(0,1);$$
$$V_n^{\mathrm{RN}} \to N(0,1), \quad \text{and} \quad V_n^{\mathrm{RP}} \to N(0,1).$$

Thus, different from existing ones, it is easy to determine p values when our tests are applied. We now investigate the power behaviors of the tests under alternatives. We consider the following local alternatives:

$$H_{1n}: Y = f(X,\theta_0) + C_n G(X) + \eta, \tag{11}$$

where $E(\eta|X) = 0$, the function $G(\cdot)$ satisfies $E(G^2(X)) < \infty$ and $\{C_n\}$ is a constant sequence. We have the following theorem.

Theorem 2 *Assume the same conditions as Theorem 1. Under the local alternatives H_{1n}, we have when $C_n = n^{-1/2}h^{-m/4}$,*

$$nh^{m/2}T_n^N \to N(\mu^T, \Sigma^T), \quad \textit{and} \quad nh^{m/2}T_n^P \to N(\mu^T, \Sigma^T)$$
$$nh^{m/2}R_n^N \to N(\mu^R, \Sigma^R), \quad \textit{and} \quad nh^{m/2}R_n^P \to N(\mu^R, \Sigma^R),$$

where

$$\mu^T = E\left[l^2(X)p(X)\right], \quad \textit{and} \quad \mu^R = E\left[l^2(X)\pi^2(X)p(X)\right],$$

with $\Sigma_1 = E(f'(X,\theta_0)f'^{\tau}(X,\theta_0))$ and $l(X) = G(X) - f'^{\tau}(X,\theta_0)\Sigma_1^{-1}E[G(X)f'(X,\theta_0)]$. The definitions of Σ^T and Σ^R are the same as those in Theorem 1.

When $n^{-1/2}h^{-m/4} = o(C_n)$, the test statistics converge in probability to infinity.

Theorem 2 indicates that the proposed tests have asymptotic power 1 for the local alternatives which are distinct from the null hypothesis at the rate slower than $n^{-1/2}h^{-m/4}$. Also, the tests can still detect the alternatives converging to the null hypothesis at the rate $n^{-1/2}h^{-m/4}$, which is the same rate as that in Li (2012). However, since the asymptotic variances in Li (2012) and the present paper are very different, it is not easy to tell which one can outperform the other in theory. Thus, a comparison will be made through simulation studies. Denote $D_1 = \mu^T/\sqrt{\Sigma^T}$ and $D_2 = \mu^R/\sqrt{\Sigma^R}$, from the above theorems, it can be also shown that the asymptotic powers of T_n^N (or T_n^P) and R_n^N (or R_n^P) are $2 - \Phi(z_{\alpha/2} - D_1) - \Phi(z_{\alpha/2} + D_1)$ and $2 - \Phi(z_{\alpha/2} - D_2) - \Phi(z_{\alpha/2} + D_2)$ respectively for the alternatives, which are distinct from the null ones at rate $n^{-1/2}h^{-m/4}$. Here, $\Phi(\cdot)$ is the standard normal distribution function, and $z_{\alpha/2}$ is the $\alpha/2$-th quantile. When the response is missing completely at random, that is, $0 < \pi(X) = c \le 1$, after some simple calculations, we have $D_1 = D_2$. As a result, the test statistics T_n^N (or T_n^P) and R_n^N (or R_n^P) have the same asymptotic power in this special case.

When there are no missing data, we get similar results as Theorem 3 in Zheng (1996). We should note that we model the local alternatives slightly different from Zheng (1996). Zheng (1996) sets the alternative as $m(X) = f(X, \tilde{\theta}_0) + C_n G(X)$. Here, $\tilde{\theta}_0$ is the value of θ that minimizes $\tilde{S}_{0n}(\theta) = E[(m(X) - f(X, \theta))^2]$ with $m(X) = E(Y|X)$. Under H_0, $\tilde{\theta}_0 = \theta_0$. If the alternative hypothesis holds, $\tilde{\theta}_0$ will typically depend on $p(X)$. Note that under the local alternatives we design, $m(X) = f(X, \tilde{\theta}_0) + f(X, \theta_0) - f(X, \tilde{\theta}_0) + C_n G(X)$ and $\tilde{\theta}_0 - \theta_0 = C_n \Sigma_1^{-1} E[G(X) f'(X, \theta_0)]$. Thus $m(X) = f(X, \tilde{\theta}_0) + C_n l(X)$. Here $l(X) = G(X) - f'^{\tau}(X, \theta_0) \Sigma_1^{-1} E[G(X) f'(X, \theta_0)]$.

Recall that $\tilde{\theta}_0$ is the value of θ that minimizes $\tilde{S}_{0n}(\theta) = E[(m(X) - f(X, \theta))^2]$. Thus, in Zheng's setting, we can have $E[G(X) f'(X, \tilde{\theta}_0)] = 0$. While, in our setting, there is also an orthogonality condition, that is, $E[l(X) f'(X, \theta_0)] = 0$. If we adopt the setting used in Zheng (1996), μ^T and μ^R will be equal to $E[G^2(X) p(X)]$ and $E[G^2(X) \pi^2(X) p(X)]$ respectively. Compared with the situations with no missing data, though the asymptotic variances of T_n^N and T_n^P are larger, the drift of them is the same. That is, though the asymptotic variances of R_n^N and R_n^P are smaller, the drift of them is also smaller.

Consider the fixed alternative, $H_1 : m(X) = f(X, \theta_0) + G(X) = f(X, \tilde{\theta}_0) + \Delta(X)$, here $\Delta(X) = f(X, \theta_0) - f(X, \tilde{\theta}_0) + G(X)$. Note that even under the fixed alternative, according to White (1981), $\hat{\theta}_N$ is still a root-n consistent estimator of $\tilde{\theta}_0$. It is easy to see that $T_n^N = E[\Delta^2(X) p(X)] + o_p(1)$ and $R_n^N = E[\pi^2(X) \Delta^2(X) p(X)] + o_p(1)$. Similar results can be obtained for T_n^P and R_n^P. Thus, the consistencies of the proposed tests are proved. Dette (1999) found an interesting phenomenon, that is, generally for the local smoothing-based test procedures, the rate of convergence is different under the null hypothesis and the fixed alternatives. To be precise, while the rate is $(n^2 h^m)^{-1}$ under H_0, it is of order n^{-1} under the fixed alternative. Dette and his coauthors showed that this is generally true in many different testing problems, see also Dette (2002), Dette and Spreckelsen (2003), Dette and Spreckelsen (2004) and Dette and Hildebrandt (2012). We can prove that this is still true even when there are some missing data. Specially, we can have $\sqrt{n}(T_n^N - E[\Delta^2(X) p(X)]) \to N(0, \sigma_T^2)$ and $\sqrt{n}(R_n^N - E[\pi^2(X) \Delta^2(X) p(X)] \to N(0, \sigma_R^2)$. This can be proven using Lemma 1 in appendix and the fact that $\sqrt{n}(\hat{\theta}_N - \tilde{\theta}_0) = O_p(1)$. We omit the details for saving space.

3 Numerical analysis

3.1 Simulation study

We now carry out simulations to examine the performance of the proposed test statistics and to compare the proposed statistics with the tests proposed by Li (2012) and Sun and Wang (2009) respectively. There are two statistics with the notations T_{n1}^S and T_{n2}^S in Sun and Wang (2009) for checking the adequacy of general linear models with missing response. Note that the behaviors of the T_{n1}^S and T_{n2}^S are very similar according to the simulation results in Sun and Wang (2009), we then only consider T_{n2}^S. As suggested by a referee, it is also interesting to compare with the tests based on empirical process. We note that Sun and Wang (2009) also developed two empirical process-based tests with notations T_{n1}^E and T_{n2}^E respectively. For the tests based on empirical process, see also Sun et al. (2009) where one procedure is given for testing the general partial linear model. Following their simulation studies, we know that T_{n1}^E and T_{n2}^E perform similarly, and thus we only consider T_{n2}^E in the following.

Study 1 To make the simulations comparable, we consider the same setting as that in Li (2012). The hypothetical model is linear as

$$Y = \theta^\tau l(X) + \epsilon, \tag{12}$$

where $\theta = (0.5, 0.8)^\tau$ and $l(X) = X = (X_1, X_2)$. The covariates $X_i = (X_{1i}, X_{2i})$, $i = 1, 2, \ldots, n$, are i.i.d. from bivariate normal distribution $N(0, \Sigma_j)$, $j = 1, 2$ with

$$\Sigma_1 = \begin{pmatrix} 0.36 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1.00 & 0.64 \\ 0.64 & 1.00 \end{pmatrix},$$

respectively. As for the random error term, we consider two distributions as Li (2012) did: $N(0, 0.3^2)$ and the double exponential distribution $DE(0, 3/10\sqrt{2})$ with density $f(x) = 5\sqrt{2}/3 \exp(-10\sqrt{2}/3|x|)$. Two missing probability mechanisms are considered for model (12), that is,

Case 1. $\pi_1(x) = P(\delta = 1|X = x) = 1/(1 + \exp(-(0.8 + 0.5x_1 + 0.5x_2))$.
Case 2. $\pi_2(x) = P(\delta = 1|X = x) = 1/(1 + \exp(-(0.2 + 0.3x_1 + 0.3x_2))$.

For the above different cases, the missing rates are 0.320 and 0.450 respectively. Here, we consider the power performance of the tests under certain alternatives as follows:

$$\begin{aligned}
H_{11}&: Y = \theta^\tau l(X) + 0.5(X_1 - 0.2)(X_2 - 0.4) + \epsilon;\\
H_{12}&: Y = \theta^\tau l(X) + 0.5(X_1 X_2 - 1) + \epsilon;\\
H_{13}&: Y = \theta^\tau l(X) + 2(\exp -0.4X_1^2 - \exp 0.6X_2^2) + \epsilon;\\
H_{14}&: Y = \theta^\tau l(X) + X_1 I_{X_2>0.2} + \epsilon,
\end{aligned}$$

where $I_{X_2>0.2}$ is an indicator, which equals to one if $X_2 > 0.2$ and otherwise zero. The kernel function takes the form $K(u, v) = K^1(u)K^1(v)$ with $K^1(u) = 0.75(1 - u^2)I_{|u|\leq 1}$. The sample sizes are $n = 50$, 100 and 200. As for the bandwidth, we set it to be $n^{-1/4.5}$ which was used in Li (2012). Obviously, this bandwidth satisfies the conditions in Appendix. All the simulations are based on 1,000 replications. The nominal level is set to be $\alpha = 0.05$. We denote the test proposed by Li (2012) as LI, the tests T_{n2}^S and T_{n2}^E by Sun and Wang (2009) as SW^S and SW^E, and for our tests T_n^N as GXZ_{TN}, T_n^P as GXZ_{TP}, R_n^N as GXZ_{RN} and R_n^P as GXZ_{RP} respectively.

Table 1 gives the empirical sizes and powers for testing H_0 against H_{1i}, $i = 1, \ldots, 4$ with design $X \sim N(0, \Sigma_1)$, $\epsilon \sim N(0, 0.3^2)$ when the data are randomly missing under either of the two missing mechanisms. The empirical sizes of these tests are all very close to the nominal level $\alpha = 0.05$. It is reasonable that the larger the sample size is, the closer the empirical sizes is to the nominal level. Among these tests, SW^S and SW^E have the best control on empirical size. About the power performance, it can

Table 1 Study 1: empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \ldots, 4$ with $X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 0.3^2)$

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.020	0.027	0.029	0.029	0.033	0.034
SW^S	0.049	0.052	0.058	0.052	0.053	0.062
SW^E	0.045	0.046	0.054	0.047	0.052	0.047
GXZ_{TN}	0.029	0.036	0.043	0.041	0.042	0.042
GXZ_{TP}	0.031	0.028	0.036	0.034	0.046	0.041
GXZ_{RN}	0.035	0.041	0.045	0.039	0.050	0.042
GXZ_{RP}	0.031	0.036	0.041	0.044	0.042	0.041
H_{11} LI	0.102	0.079	0.278	0.176	0.633	0.513
SW^S	0.069	0.080	0.098	0.102	0.137	0.142
SW^E	0.183	0.136	0.454	0.333	0.857	0.725
GXZ_{TN}	0.151	0.096	0.351	0.228	0.745	0.508
GXZ_{TP}	0.159	0.114	0.311	0.313	0.811	0.632
GXZ_{RN}	0.139	0.115	0.396	0.291	0.786	0.673
GXZ_{RP}	0.185	0.119	0.459	0.323	0.842	0.714
H_{12} LI	0.993	0.941	1.000	1.000	1.000	1.000
SW^S	1.000	1.000	1.000	1.000	1.000	1.000
SW^E	1.000	1.000	1.000	1.000	1.000	1.000
GXZ_{TN}	0.989	0.937	1.000	1.000	1.000	1.000
GXZ_{TP}	0.988	0.939	1.000	1.000	1.000	1.000
GXZ_{RN}	0.992	0.943	1.000	1.000	1.000	1.000
GXZ_{TP}	0.997	0.950	1.000	1.000	1.000	1.000
H_{13} LI	0.315	0.203	0.351	0.271	0.375	0.338
SW^S	0.555	0.498	0.686	0.681	0.747	0.695
SW^E	0.465	0.453	0.635	0.639	0.701	0.724
GXZ_{TN}	0.811	0.717	0.911	0.878	0.929	0.931
GXZ_{TP}	0.786	0.673	0.870	0.844	0.891	0.903
GXZ_{RN}	0.799	0.661	0.861	0.817	0.884	0.853
GXZ_{TP}	0.774	0.704	0.869	0.848	0.905	0.891
H_{14} LI	0.241	0.159	0.671	0.497	0.983	0.921
SW^S	0.055	0.043	0.045	0.040	0.046	0.054
SW^E	0.091	0.068	0.166	0.129	0.467	0.357
GXZ_{TN}	0.208	0.132	0.588	0.386	0.950	0.828
GXZ_{TP}	0.252	0.154	0.600	0.439	0.957	0.874
GXZ_{RN}	0.241	0.153	0.634	0.501	0.968	0.894
GXZ_{TP}	0.261	0.189	0.653	0.472	0.975	0.893

be seen from Table 1 that all of our proposed tests are more powerful than LI and SW^S under all the designed alternative hypotheses except H_{14}. Under H_{14}, the most powerful one is the test LI. However, our tests are still competitive, that is, the gains of LI over our methods are limited. The power performance of SW^S is the worst under H_{11} and H_{14}. However, we also note that SW^E improves SW^S greatly under H_{11} and H_{14}. Specifically, under H_{11}, SW^E has similar power performance as those of our proposed tests. Under H_{14}, though there is some improvement of SW^E over SW^S, SW^E is still inferior to LI and our proposed tests. Further, under H_{13}, LI is the worst and our gain over LI, SW^S and SW^E is obvious. The impact of missing mechanisms on the empirical powers is evident. Generally, with the first missing mechanism, all these tests have greater powers. This is reasonable since there are less missing data with the first mechanism. Comparing the test GXZ_{TN} with GXZ_{TP}, GXZ_{TP} works better under H_{11} and H_{14} with larger power, whereas GXZ_{TN} is better under H_{13}. It seems using a parametric estimation of π gains not much compared with that using a nonparametric estimation. When we compare T_n with R_n, we can have the following results: under H_{11} and H_{14}, generally, R_n has larger power; while under H_{13}, T_n performs slightly better. In other words, overall, R_n works better. This seems to suggest that T_n may be affected by boundary effect more seriously to deteriorate its performance although in theory, it does not have such an issue. However, these comparisons cannot firmly say that among our proposed test statistics, which one is the best. But it seems that the tests with parametric estimation of $\pi(\cdot)$ are recommendable although it may have a misspecification issue.

Table 2 summaries the empirical sizes and powers for testing H_0 against $H_{1i}, i = 1, \ldots, 4$ when $X \sim N(0, \Sigma_2)$, $\epsilon \sim N(0, 0.3^2)$. In this case, X_1 is dependent of X_2. The only difference from the previous example is that X is from $N(0, \Sigma_2)$: the components are correlated. However, we can see, by a comparison with Table 1, that the performance on maintaining the nominal level is similar to that in the independent case. We can also see that the powers of all our proposed tests increase under H_{11} and H_{14}, while decrease under the other alternatives. Under H_{11}, SW^E is the most powerful and SW^S is also more powerful than our proposed tests and LI which is much different from that in the previous example. However, the opposite phenomenon happens under H_{12} for SW^S and SW^E. That is, the independence between X_1 and X_2 has a great influence on the behavior of SW^S and SW^E. LI is slightly more powerful than our tests under H_{12} whereas the proposed statistics are the most powerful under H_{13} and H_{14} in the most cases. Overall speaking, the proposed test is powerful under all the scenarios, SW^S and SW^E is not robust to the distribution of the covariates, and further SW^E generally performs better than SW^S.

We now report the results about the testing for H_0 against $H_{1i}, i = 1, \ldots, 4$ with $\epsilon \sim DE(0, 3/10\sqrt{2})$ and $X \sim N(0, \Sigma_1)$ in Table 3, and $X \sim N(0, \Sigma_2)$ in Table 4 respectively. The comparisons between those in Tables 1 and 3, or those in Tables 2 and 4, are made to see the impact from the distributions. The powers of the proposed tests increase greatly under H_{11} and H_{14} in Table 3 when comparing with those in Table 1, while the performance of SW^S and SW^E is slightly affected by the error distribution. Compared with LI, SW^S and SW^E, our tests have highest power under any alternatives in Table 3, and under H_{12}, and H_{13} in Table 4. From all these four tables, we can know that the proposed tests perform well and are the most powerful

Table 2 Study 1: empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \ldots, 4$ with $X \sim N(0, \Sigma_2)$ and $\epsilon \sim N(0, 0.3^2)$

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.031	0.023	0.041	0.036	0.045	0.043
SW^S	0.047	0.048	0.038	0.047	0.055	0.051
SW^E	0.043	0.045	0.046	0.043	0.044	0.044
GXZ_{TN}	0.045	0.033	0.042	0.037	0.041	0.056
GXZ_{TP}	0.045	0.043	0.043	0.041	0.045	0.041
GXZ_{RN}	0.032	0.031	0.046	0.035	0.049	0.047
GXZ_{RP}	0.041	0.036	0.031	0.042	0.039	0.045
H_{11} LI	0.115	0.103	0.199	0.164	0.479	0.373
SW^S	0.627	0.559	0.919	0.917	0.999	0.999
SW^E	0.733	0.646	0.985	0.975	1.000	1.000
GXZ_{TN}	0.334	0.248	0.762	0.676	0.993	0.968
GXZ_{TP}	0.422	0.277	0.812	0.679	0.960	0.969
GXZ_{RN}	0.381	0.269	0.783	0.703	0.961	0.985
GXZ_{RP}	0.441	0.361	0.883	0.773	1.000	0.995
H_{12} LI	0.965	0.831	0.999	0.991	1.000	1.000
SW^S	0.693	0.581	0.902	0.846	0.989	0.981
SW^E	0.714	0.610	0.927	0.882	0.999	0.996
GXZ_{TN}	0.923	0.783	1.000	1.000	1.000	1.000
GXZ_{TP}	0.911	0.769	0.998	0.992	1.000	0.999
GXZ_{TN}	0.956	0.809	1.000	0.995	1.000	1.000
GXZ_{TP}	0.934	0.803	1.000	0.993	1.000	1.000
H_{13} LI	0.237	0.187	0.272	0.209	0.274	0.227
SW^S	0.495	0.502	0.628	0.634	0.683	0.682
SW^E	0.538	0.526	0.668	0.680	0.751	0.767
GXZ_{TN}	0.770	0.654	0.897	0.862	0.936	0.910
GXZ_{TP}	0.728	0.622	0.848	0.811	0.893	0.862
GXZ_{RN}	0.765	0.624	0.868	0.804	0.887	0.870
GXZ_{RP}	0.767	0.656	0.833	0.822	0.888	0.878
H_{14} LI	0.203	0.144	0.596	0.471	0.957	0.892
SW^S	0.491	0.448	0.806	0.776	0.987	0.981
SW^E	0.636	0.545	0.955	0.923	1.000	0.999
GXZ_{TN}	0.448	0.315	0.890	0.785	1.000	0.993
GXZ_{TP}	0.499	0.345	0.907	0.837	1.000	0.996
GXZ_{RN}	0.465	0.337	0.908	0.839	1.000	0.997
GXZ_{RP}	0.552	0.380	0.954	0.874	1.000	0.996

Table 3 Study 1: empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \ldots, 4$ with $X \sim N(0, \Sigma_1)$ and $\epsilon \sim DE(0, 3/10\sqrt{2})$

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.026	0.030	0.041	0.035	0.049	0.047
SW^S	0.045	0.041	0.044	0.058	0.045	0.050
SW^E	0.038	0.032	0.043	0.047	0.054	0.046
GXZ_{TN}	0.032	0.029	0.039	0.025	0.043	0.041
GXZ_{TP}	0.031	0.036	0.029	0.034	0.038	0.034
GXZ_{RN}	0.032	0.034	0.038	0.039	0.051	0.042
GXZ_{RP}	0.031	0.034	0.038	0.044	0.036	0.036
H_{11} LI	0.083	0.076	0.222	0.227	0.682	0.549
SW^S	0.122	0.101	0.148	0.136	0.262	0.271
SW^E	0.432	0.329	0.861	0.775	0.998	0.993
GXZ_{TN}	0.745	0.544	0.991	0.951	1.000	1.000
GXZ_{TP}	0.746	0.575	0.972	0.952	1.000	1.000
GXZ_{RN}	0.752	0.562	0.996	0.954	1.000	1.000
GXZ_{RP}	0.770	0.598	0.991	0.972	1.000	1.000
H_{12} LI	0.986	0.931	1.000	1.000	1.000	1.000
SW^S	1.000	1.000	1.000	1.000	1.000	1.000
SW^E	1.000	1.000	1.000	1.000	1.000	1.000
GXZ_{TN}	1.000	0.986	1.000	1.000	1.000	1.000
GXZ_{TP}	1.000	0.988	1.000	1.000	1.000	1.000
GXZ_{RN}	1.000	0.994	1.000	1.000	1.000	1.000
GXZ_{RP}	1.000	0.993	1.000	1.000	1.000	1.000
H_{13} LI	0.281	0.208	0.335	0.274	0.417	0.329
SW^S	0.548	0.529	0.681	0.683	0.738	0.757
SW^E	0.484	0.459	0.641	0.636	0.708	0.734
GXZ_{TN}	0.814	0.683	0.896	0.875	0.942	0.922
GXZ_{TP}	0.752	0.728	0.864	0.854	0.907	0.892
GXZ_{RN}	0.777	0.712	0.852	0.814	0.984	0.880
GXZ_{RP}	0.780	0.703	0.863	0.847	0.897	0.901
H_{14} LI	0.192	0.176	0.688	0.524	0.979	0.905
SW^S	0.052	0.053	0.041	0.035	0.045	0.033
SW^E	0.131	0.115	0.418	0.292	0.927	0.802
GXZ_{TN}	0.741	0.581	0.994	0.956	1.000	1.000
GXZ_{TP}	0.748	0.624	0.992	0.959	1.000	1.000
GXZ_{RN}	0.776	0.602	0.994	0.969	1.000	1.000
GXZ_{RP}	0.754	0.577	0.986	0.968	1.000	1.000

Table 4 Study 1: empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \ldots, 4$ with $X \sim N(0, \Sigma_2)$ and $\epsilon \sim DE(0, 3/10\sqrt{2})$

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.028	0.032	0.033	0.042	0.035	0.046
SW^S	0.043	0.039	0.041	0.045	0.042	0.047
SW^E	0.036	0.044	0.043	0.045	0.044	0.047
GXZ_{TN}	0.026	0.027	0.053	0.042	0.067	0.048
GXZ_{TP}	0.031	0.032	0.035	0.039	0.037	0.038
GXZ_{RN}	0.035	0.027	0.046	0.043	0.049	0.040
GXZ_{RP}	0.033	0.032	0.040	0.049	0.032	0.054
H_{11} LI	0.117	0.095	0.237	0.194	0.523	0.433
SW^S	0.762	0.704	0.981	0.968	1.000	1.000
SW^E	0.861	0.800	0.997	0.989	1.000	1.000
GXZ_{TN}	0.714	0.561	0.957	0.901	1.000	0.998
GXZ_{TP}	0.692	0.528	0.914	0.871	0.966	0.979
GXZ_{RN}	0.717	0.565	0.969	0.916	1.000	0.992
GXZ_{RP}	0.756	0.602	0.970	0.921	1.000	0.998
H_{12} LI	0.949	0.826	1.000	0.99	1.000	1.000
SW^S	0.776	0.686	0.941	0.905	1.000	0.991
SW^E	0.801	0.726	0.970	0.944	0.999	1.000
GXZ_{TN}	0.993	0.958	1.000	1.000	1.000	1.000
GXZ_{TP}	0.988	0.949	0.998	1.000	1.000	1.000
GXZ_{RN}	0.993	0.955	1.000	1.000	1.000	1.000
GXZ_{RP}	0.993	0.950	1.000	1.000	1.000	1.000
H_{13} LI	0.244	0.184	0.265	0.225	0.272	0.242
SW^S	0.488	0.497	0.581	0.621	0.672	0.718
SW^E	0.538	0.525	0.687	0.696	0.752	0.773
GXZ_{TN}	0.799	0.661	0.891	0.861	0.914	0.919
GXZ_{TP}	0.712	0.627	0.838	0.809	0.887	0.873
GXZ_{RN}	0.768	0.615	0.852	0.828	0.888	0.872
GXZ_{RP}	0.759	0.614	0.869	0.809	0.879	0.868
H_{14} LI	0.259	0.207	0.626	0.483	0.944	0.881
SW^S	0.613	0.589	0.932	0.908	1.000	1.000
SW^E	0.801	0.730	0.990	0.985	1.000	1.000
GXZ_{TN}	0.782	0.599	0.998	0.988	1.000	1.000
GXZ_{TP}	0.806	0.625	0.982	0.979	1.000	1.000
GXZ_{RN}	0.793	0.654	0.994	0.975	1.000	1.000
GXZ_{RP}	0.807	0.659	0.994	0.984	1.000	1.000

under many scenarios. For T_n and R_n, it seems that the latter performs better overall.

In study 1, normal distribution is used to determine critical values such that the empirical sizes and powers of the proposed tests can be conveniently computed. However, it is also well known that the rate of convergence to the normal limit is slow, see also Härdle and Mammen (1993), Stute et al. (1998a) and González-Manteiga and Crujeiras (2013). Thus, the use of the asymptotic normality may be inappropriate for small sample sizes. This can also be seen from Tables 1, 2, 3 and 4 in study 1. To be precise, though the simulated sizes are very close to the nominal level, they generally underestimate it. As an alternative for calibrating critical values, we consider the residual-based bootstrap in the following. Let the bootstrap errors $\epsilon_1^*, \ldots, \epsilon_n^*$ be an independent sample from the empirical distribution function of the centered residuals $\tilde{\epsilon}_i = \hat{\epsilon}_i - n^{-1}\sum_{l=1}^n \hat{\epsilon}_l$. Then, we generate the bootstrap observations:

$$y_i^* = f(x_i, \hat{\theta}_N) + \epsilon_i^*.$$

Let T_n^{N*} be defined similarly as T_n^N, basing on the bootstrap sample $(x_1, y_1^*), \ldots, (x_n, y_n^*)$. The null hypothesis is rejected if T_n^N is bigger than the corresponding quantile of the bootstrap distribution of T_n^{N*}. We denote the bootstrap version of T_n^N as GXZ_{TN}^*. The bootstrap version of other proposed tests can be similarly developed and is denoted as GXZ_{TP}^*, GXZ_{RN}^* and GXZ_{RP}^* respectively. The study of the asymptotic validity of this procedure in the presence of missing response will be undertaken elsewhere. Here, we investigate the empirical properties of this bootstrap procedure when the same settings with $X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 0.3^2)$ as in study 1 are considered. For comparison, we use the same bandwidth $n^{-1/4.5}$ first. The number of replications was 1,000 and for each replication 500 bootstrap samples were generated. The results are presented in Table 5. From this table, we can see clearly that even with the sample size $n = 25$, the resampling method can control the type I error very well. As for the power performance, comparing Table 1 with Table 5, we can conclude that the bootstrap performs better than the normal approximation under the alternatives H_{11}, H_{12} and H_{14}. While under H_{13}, the normal approximation is the winner. Overall, the resampling method is recommendable for calibration especially when the sample size is small. Due to the computational intensiveness, when we have sufficient observations, the normal approximation is still recommendable.

Notice in the above study, we do not investigate the impact of the bandwidth on the performance of our tests and the considered alternative hypothesizes are limited to the fixed alternatives. In the following study, we aim to study the bandwidth selection problem and the performances of our proposed tests under some local alternative.

Study 2 The local alternative is taken to be

$$H_{1n}: Y = \theta^\tau l(X) + C_n(X_1 - 0.2)(X_2 - 0.4) + \epsilon, \tag{13}$$

here we adopt the same setting as that in study 1 except we consider the local alternative indexed by C_n instead of the fixed alternative H_{11} in study 1. It is evident that the null hypothesis $H_0: Y = \theta^\tau l(X) + \epsilon$ is valid if and only if $C_n = 0$. In this study, we set

Table 5 Simulated size and power under different sample sizes n =25, 50 and $n = 100$, missing mechanisms $\pi_1(x)$ and $\pi_2(x)$ for bootstrap calibration

	$n = 25$		$n = 50$		$n = 100$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 GXZ^*_{TN}	0.057	0.049	0.047	0.048	0.047	0.049
GXZ^*_{TP}	0.046	0.059	0.056	0.051	0.056	0.051
GXZ^*_{RN}	0.059	0.050	0.044	0.052	0.048	0.050
GXZ^*_{RP}	0.044	0.058	0.053	0.055	0.055	0.048
H_{11} GXZ^*_{TN}	0.129	0.124	0.269	0.175	0.481	0.366
GXZ^*_{TP}	0.144	0.115	0.251	0.216	0.526	0.455
GXZ^*_{RN}	0.137	0.131	0.293	0.192	0.531	0.427
GXZ^*_{RP}	0.142	0.118	0.271	0.228	0.559	0.455
H_{12} GXZ^*_{TN}	0.844	0.706	0.999	0.978	1.000	1.000
GXZ^*_{TP}	0.844	0.681	0.999	0.983	1.000	1.000
GXZ^*_{RN}	0.844	0.709	0.999	0.983	1.000	1.000
GXZ^*_{RP}	0.854	0.708	0.999	0.984	1.000	1.000
H_{13} GXZ^*_{TN}	0.509	0.382	0.598	0.585	0.682	0.598
GXZ^*_{TP}	0.445	0.393	0.580	0.542	0.632	0.574
GXZ^*_{RN}	0.491	0.358	0.576	0.557	0.648	0.565
GXZ^*_{RP}	0.454	0.382	0.588	0.538	0.654	0.565
H_{14} GXZ^*_{TN}	0.195	0.118	0.357	0.262	0.673	0.513
GXZ^*_{TP}	0.168	0.154	0.372	0.261	0.739	0.604
GXZ^*_{RN}	0.188	0.119	0.381	0.294	0.741	0.604
GXZ^*_{RP}	0.163	0.156	0.375	0.267	0.764	0.628

$X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 0.3^2)$ for space considerations. We only consider the first missing probability mechanism in study 1 to save space.

Zhu and Ng (2003) pointed out that it is still an open problem about how to select optimal bandwidth in the testing problems. Though in nonparametric estimation literature the bandwidth selection has been discussed extensively, the selected optimal bandwidth for estimation may not yard the optimal power and size performance under different alternatives. To investigate the impact of bandwidth selection on our proposed tests, we take the bandwidth h to be $j/100$ for $j = 11, 15, 19, \ldots, 99$. Based on the 1,000 simulations, we plot the estimated size and power curve against the above bandwidth sequences with the sample size 50, missing mechanism $\pi_1(x)$ and $C_n = Cn^{-1/2}$ with $C = 0, 2, 4$, which is shown in Fig. 1. This strategy is also conducted by many authors, such as Sun and Wang (2009) and Lopez and Patilea (2009). From Fig. 1, we have the following observations: (1) with the sample size $n = 50$, the proposed tests with different bandwidths can control the size reasonably. To be precise, the empirical sizes are all contained in the range of (0.04,0.06). In one word, the bandwidth selection for the size control for our proposed tests is not too critical. (2) Generally, we can get larger powers if we use a relatively larger bandwidth. However, when we take the bandwidth not too small, the gain by employing a larger bandwidth is marginal.

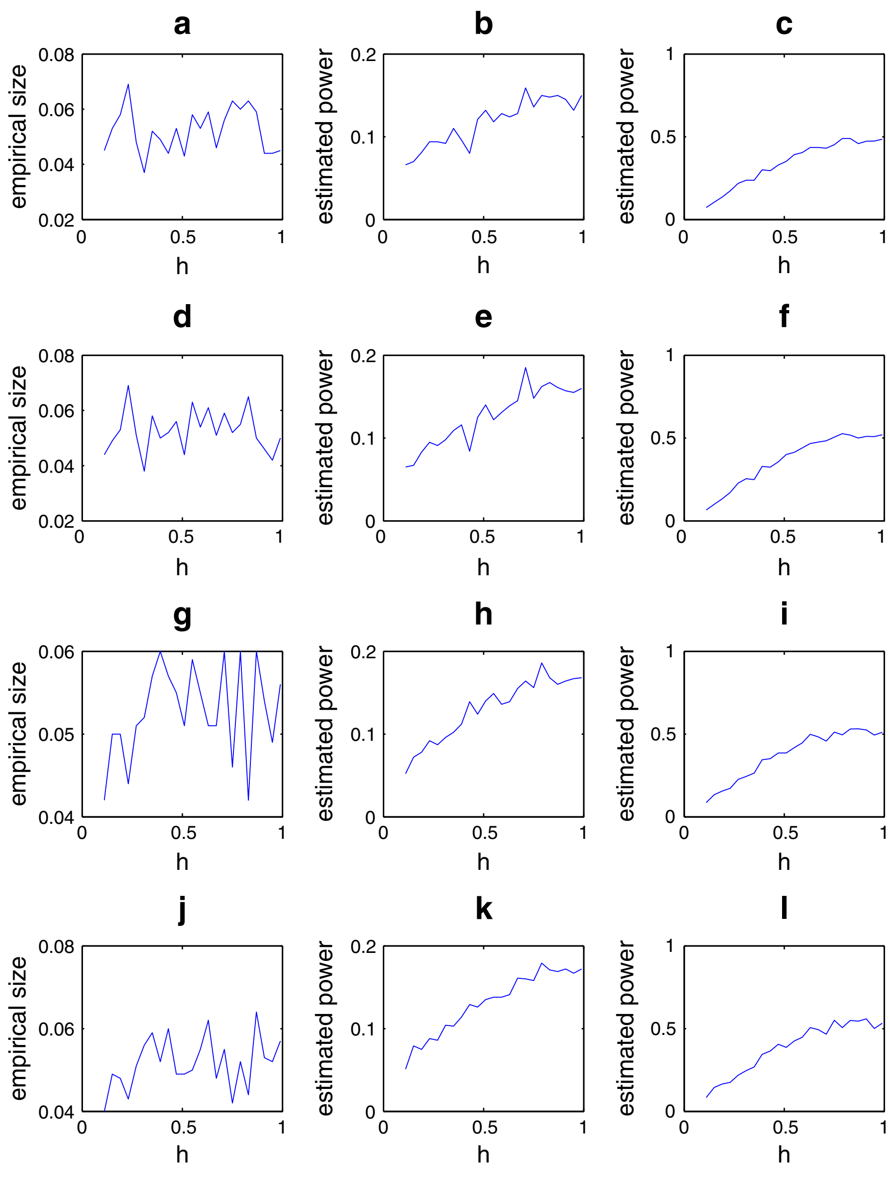

Fig. 1 The estimated size and power curves of the tests GXZ^*_{TN}, GXZ^*_{RN}, GXZ^*_{TP}, and GXZ^*_{RP} against the bandwidth h with missing mechanisms $\pi_1(x)$ and sample size 50 under different choices of C_n for testing problem (13). **a** GXZ^*_{TN}, $C_n = 0$; **b** GXZ^*_{TN}, $C_n = 2n^{-1/2}$; **c** GXZ^*_{TN}, $C_n = 4n^{-1/2}$. **d** GXZ^*_{RN}, $C_n = 0$; **e** GXZ^*_{RN}, $C_n = 2n^{-1/2}$; **f** GXZ^*_{RN}, $C_n = 4n^{-1/2}$. **g** GXZ^*_{TP}, $C_n = 0$; **h** GXZ^*_{TP}, $C_n = 2n^{-1/2}$; **i** GXZ^*_{TP}, $C_n = 4n^{-1/2}$. **j** GXZ^*_{RP}, $C_n = 0$; **k** GXZ^*_{RP}, $C_n = 2n^{-1/2}$; **l** GXZ^*_{RP}, $C_n = 4n^{-1/2}$

As discussed by Sperlich (2014), most, if not all, of the known methods are computationally expensive and somewhat very complex. Based on the above observations and suggestions by Lavergne and Vuong (2000), in the following simulations, we choose $h = 1.25 \times n^{-1/6}$ which is at the same rate as the optimal bandwidth derived in non-

Table 6 Simulated size and power under different sample sizes $n = 25, 50$ and $n = 100$, missing mechanisms $\pi_1(x)$, and different C_n for Study 2

C_n	GXZ^*_{TP}	GXZ^*_{RP}	GXZ^*_{TN}	GXZ^*_{RN}
$n = 25$				
0.0	0.045	0.049	0.050	0.051
0.2	0.076	0.071	0.074	0.075
0.4	0.136	0.140	0.126	0.132
0.6	0.232	0.231	0.205	0.215
0.8	0.314	0.321	0.305	0.322
1.0	0.386	0.394	0.396	0.410
$n = 50$				
0.0	0.055	0.056	0.048	0.047
0.2	0.089	0.093	0.086	0.096
0.4	0.267	0.274	0.233	0.248
0.6	0.520	0.536	0.470	0.515
0.8	0.683	0.693	0.659	0.701
1.0	0.844	0.855	0.814	0.833
$n = 100$				
0.0	0.056	0.055	0.052	0.057
0.2	0.149	0.157	0.126	0.138
0.4	0.537	0.543	0.481	0.541
0.6	0.893	0.903	0.852	0.879
0.8	0.983	0.987	0.972	0.978
1.0	0.994	0.997	0.997	0.997

parametric estimation. Also note in study 1, we set $h = n^{-1/4.5}$ which is smaller than the selected bandwidth $h = 1.25 \times n^{-1/6}$. Further from Fig. 1, we know the powers of our proposed tests with $h = n^{-1/4.5}$ are smaller than that with $h = 1.25 \times n^{-1/6}$. In other words, the power performance of the proposed tests in study 1 can be improved using the selected bandwidth $h = 1.25 \times n^{-1/6}$.

We evaluate the performance of our proposed tests under the above-defined local alternative (13) through varying the values of C_n, different sample sizes $n = 25, 50$ and 100, and missing mechanism $\pi_1(x)$. The simulation results are shown in Table 6. From the table, we can have similar findings except the following observations. When the local alternative hypothesis holds, that is, $C_n \neq 0$, the powers of our tests increase quickly as C_n in (13) increases. To be precise, the power of R_n^P is 0.903 under the sample size $n = 100$, missing mechanism $\pi_1(x)$ and $C_n = 6n^{-1/2} = 0.6$. In other words, the tests are very sensitive to the alternatives. Moreover from these two tables, we can also conclude that R_n^P performs best among these four proposed tests.

3.2 Real data analysis

Consider the data set about monozygotic twins with the sample size 50. In this data set, the response Y stands for birth-weight of a baby and two corresponding covariates

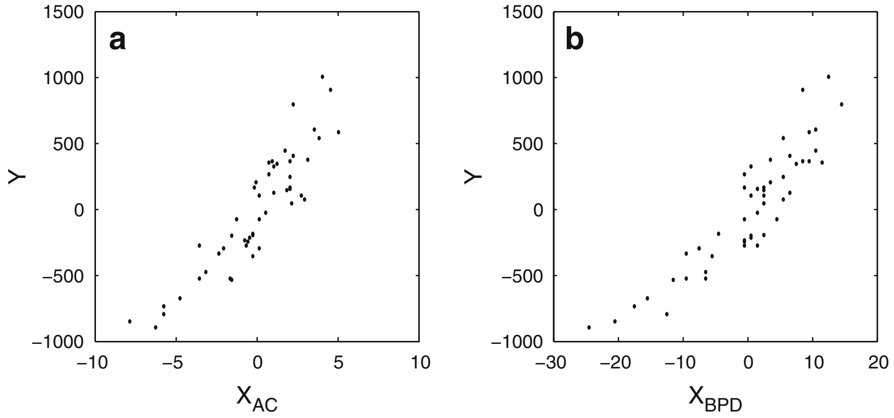

Fig. 2 The plot for real data set: **a** for X_{AC} and Y; **b** for X_{BPD} and Y

X_{AC} (=AC) and X_{BDP} (=BDP) respectively for abdominal circumference and biparietal diameter. The data set has been used to test whether the nonparametric part in a partial linear model with missing response is of parametric form by Xu et al. (2012). They found that a parametric model is plausible. However, we got the residual plots and found that a linear model would be further plausible. Thus, we make a further check to see whether a linear model is adequate.

Consider the null hypothesis

$$H_0\colon E(Y|X) = X_{\text{BDP}}\beta_{\text{BDP}} + X_{\text{AC}}\beta_{\text{AC}} \tag{14}$$

for some β_{BDP} and β_{AC}. First all the variables are centered and the same notations are given without confusion. We illustrate our methods by missing 20 % of the response randomly. Then, we try to obtain the results by 2,000 simulation runs in which each time we use the kernel function in Sect. 3.1 for computation. We present the scatter plot for the covariates and the outcome shown in Fig. 2. From our simulations in the study 2 in the Sect. 3.1, we set the bandwidth to $\text{std}(x_{\text{AC}}) \times n^{-1/6}$. Under these settings, the p values for T_n^N, R_n^N, T_n^P and R_n^P are 0.900, 0.865, 0.802 and 0.875 respectively. As a result, the null hypothesis in (14) cannot be rejected.

4 Discussion

In this paper, we extend Zheng's1996 method to adopt to missing response at random due to its technical tractability and easy computation. The asymptotic properties are developed for our proposed tests under null and local alternatives. The intensive simulation studies suggest that our proposed tests can perform well. To better control the empirical size, we also propose to use the residual-based bootstrap. Through our simulations, we also find that the use of different bandwidth has almost no effect on the empirical sizes of our proposed tests. On the other hand, the powers of the tests can be improved if we use a slightly larger bandwidth. Based on these observations, we suggest to use a rule of thumb which may be not optimal in all directions. The problem of choosing the bandwidth to optimize the power remains an open problem faced by all smoothing-based tests due to the obstacle that there are infinitely many alternatives. In this paper, we prefer to use the suggested simple method instead of using other computation intensive and complex methodologies to find the optimal bandwidth.

Another direction which needs more attention is the model checking with missing covariates at random. We discuss this issue here briefly. Denote $X = (U, V)$, here U and V are $m_1$1- and m_2-dimensional random vectors with $m_1 + m_2 = m$. Consider the situation that U is missing at random, whereas other variables Y and V are observed completely. Let δ be the missing indicator for the individual whether U is observed ($\delta = 1$) or not ($\delta = 0$). Assume that U is missing at random which implies

$$P(\delta = 1|Y, X) = P(\delta = 1|Y, V) = \pi(Z),$$

here $Z = (Y, V)$. Note that in this situation, $E(\delta\epsilon|X) = E[E(\delta\epsilon|X, Y)|X] = E(\epsilon\pi(Z)|X)$ may not be equal to zero. This further implies that

$$E(\delta\epsilon E(\delta\epsilon|X)W(X)) = E(E^2(\delta\epsilon|X)W(X))$$

may not be zero even under the null hypothesis. Thus, we cannot construct test statistics similarly as R_n^N and R_n^P. However, we note that $E(\delta/\pi(Z)\epsilon|X) = E[E(\delta/\pi(Z)\epsilon|X, Y)|X] = E(\epsilon|X) = 0$ under the null hypothesis. This suggests that test statistics could be constructed similarly as T_n^N and T_n^P. For the formal development of the related results, we leave them to further studies.

Appendix: Proofs of Theorems

The following conditions are required for the theorems in Sect. 2.

(1) $f(x, \theta)$ is a Borel measurable function on R^m for each θ and a twice continuously differentiable real function on a compact subset of R^p Θ for each $x \in R^m$; $\tilde{\theta}_0$, the value of θ that minimizes $\tilde{S}_{0n}(\theta) = E[(E(Y|X) - f(X, \theta))^2]$, is an interior point of Θ and is the unique minimizer of the function $\tilde{S}_{0n}$; $\Sigma_1 = E(f'(X, \theta_0) f'^{\tau}(X, \theta_0))$ is nonsingular.
(2) $\pi(x)$ has bounded partial derivatives up to order 2 almost surely and $\inf_x \pi(x) > 0$;
(3) $\sup E(\varepsilon^4|X = x) < \infty$, $E|X|^4 < \infty$ and $E|Y|^4 < \infty$;
(4) $nh^{3m/2} \to \infty$ and $h \to 0$;
(5) The density of X, say $p(x)$ on support $\mathcal{C}$, exists and has bounded derivatives up to order 2 and satisfies

$$0 < \inf_{x \in \mathcal{C}} p(x) \leq \sup_{x \in \mathcal{C}} p(x) < \infty;$$

6) The continuous kernel function $K(\cdot)$ satisfies the following: (i) the support of $K(\cdot)$ is the interval $[-1, 1]$; (ii) $K(\cdot)$ is symmetric about 0; (iii) $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} |u|K(u)du \neq 0$.

Remark 2 Conditions (4) and (6) are typical for obtaining convergence rates when nonparametric estimation is applied. Condition (2) is a common assumption in missing data study, for example, Sun and Wang (2009). The conditions (1) and (3) are necessary for the asymptotic normality of the least-squares estimator. Condition (5) is aimed for avoiding tedious proofs of the theorems, see, e.g. Xue (2009). Without this condition, we have to resort to some truncation techniques to control small values in the denominators.

Lemma 1 *Under the null hypothesis and conditions above, we have*

$$W_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i M(x_j) = O_p(1/\sqrt{n}), \quad (15)$$

where $M(\cdot)$ is continuously differentiable and $|M(x)| \leq b(x)$ for $x \in R^m$ and some $b(x)$ satisfying $E[b^2(X)] < \infty$.

This can be obtained following the same argument as Zheng (1996), so we omit the details.

Lemma 2 *Under the conditions in Appendix and the alternative H_{1n}, the asymptotic properties of $\sqrt{n}(\hat{\theta}_N - \theta_0)$ is as follows*

$$\begin{aligned} \sqrt{n}(\hat{\theta}_N - \theta_0) &= C_n \sqrt{n} \Sigma_1^{-1} E(f'(X, \theta_0)G(X)) \\ &\quad + \frac{\Sigma_1^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i f'(x_i, \tilde{\theta}_0)(y_i - f(x_i, \tilde{\theta}_0))}{\pi(x_i)} + o_p(1). \end{aligned} \quad (16)$$

where $\Sigma_1 = E(f'(X, \theta_0) f'^{\tau}(X, \theta_0))$.

Lemma 3 *Under conditions in Appendix and the alternative H_{1n}, the asymptotic properties of $\sqrt{n}(\hat{\theta}_P - \theta_0)$ is*

$$\begin{aligned} \sqrt{n}(\hat{\theta}_P - \theta_0) &= C_n \sqrt{n} \Sigma_1^{-1} E(f'(X, \theta_0)G(X)) \\ &\quad + \frac{\Sigma_1^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i f'(x_i, \tilde{\theta}_0)(y_i - f(x_i, \tilde{\theta}_0))}{\pi(x_i, \alpha)} + o_p(1). \end{aligned} \quad (17)$$

Lemmas 2 and 3 can be similarly obtained from the Lemma 4.2 in Van Keilegom et al. (2008) and the Lemmas in Guo and Xu (2012) respectively, and we omit the detailed proof here.

The proof for R_n^N is similar to that for T_n^N, so we omit the detail for R_n^N in Theorems 1 and 2. Below we give the proof for T_n^N in Theorems 1 and 2.

Proof of Theorem 1 For T_n^N in (6), it can be decomposed as follows

$$
\begin{aligned}
T_n^N &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\hat{\epsilon}_i \hat{\epsilon}_j \\
&\quad - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))}{\hat{\pi}(x_i)\hat{\pi}(x_j)\pi(x_j)} K_h(x_i - x_j)\hat{\epsilon}_i \hat{\epsilon}_j \\
&\quad - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j (\hat{\pi}(x_i) - \pi(x_i))}{\hat{\pi}(x_j)\hat{\pi}(x_i)\pi(x_i)} K_h(x_i - x_j)\hat{\epsilon}_i \hat{\epsilon}_j \\
&\quad - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j (\hat{\pi}(x_i) - \pi(x_i))(\hat{\pi}(x_j) - \pi(x_j))}{\hat{\pi}(x_j)\hat{\pi}(x_i)\pi(x_j)\pi(x_i)} K_h(x_i - x_j)\hat{\epsilon}_i \hat{\epsilon}_j \\
&=: T_{n1} - T_{n2} - T_{n3} - T_{n4}.
\end{aligned} \tag{18}
$$

Below we analyze the term T_{n1} in (18) first. It can be further divided as

$$
\begin{aligned}
T_{n1} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i \epsilon_j \\
&\quad - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i (f(x_j, \hat{\theta}_N) - f(x_j, \theta_0)) \\
&\quad - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_j (f(x_i, \hat{\theta}_N) - f(x_i, \theta_0)) \\
&\quad + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \Bigg(\frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)(f(x_i, \hat{\theta}_N) - f(x_i, \theta_0)) \\
&\quad \times (f(x_j, \hat{\theta}_N) - f(x_j, \theta_0)) \Bigg) \\
&=: T_{n1,1} - T_{n1,2} - T_{n1,3} + T_{n1,4}.
\end{aligned} \tag{19}
$$

Notice that $T_{n1,1}$ is a second-order degenerate U-statistic. By some tedious calculations and according to Theorem 1 of Hall (1984), we can have

$$
nh^{m/2} T_{n1,1} \rightarrow N(0, \Sigma^T), \tag{20}
$$

where $\Sigma^T = \int K^2(u)du \cdot \int (\sigma^2(x))^2 p^2(x)\pi^{-2}(x)\mathrm{d}x$.

Below we prove that $nh^{m/2}T_{n1,2} = o_p(1)$. It can be divided as

$$
\begin{aligned}
T_{n1,2} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i \frac{\partial f(x_j, \theta_0)}{\partial \theta^\tau} (\hat{\theta}_N - \theta_0) \\
&\quad + (\hat{\theta}_N - \theta_0)^\tau \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i \frac{\partial^2 f(x_j, \tilde{\theta})}{\partial \theta \partial \theta^\tau} (\hat{\theta}_N - \theta_0) \\
&= R_{n1,1}(\hat{\theta}_N - \theta_0) + (\hat{\theta}_N - \theta)^\tau R_{n1,2}(\hat{\theta}_N - \theta_0),
\end{aligned}
$$

where $\tilde{\theta}$ lies between $\hat{\theta}_N$ and θ_0.

Recalling Lemma 1, we have $R_{n1,1} = O_p(1/\sqrt{n})$. Let $\tilde{A}_{j,st}$ and $A_{j,st}$ denote the (s, t) element of $\partial^2 f(x_j, \tilde{\theta})/\partial \theta \partial \theta^\tau$ and $\partial^2 f(x_j, \theta_0)/\partial \theta \partial \theta^\tau$ respectively. Due to the fact that $\hat{\theta}_N - \theta_0 = o_p(1)$ and the continuity of $\partial^2 f(x_j, \theta)/\partial \theta \partial \theta^\tau$ as a function of θ, we can assert that

$$
\begin{aligned}
& E \left| \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i \tilde{A}_{j,st} \right| \\
&\quad = E \left(\frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)|\epsilon_i||A_{j,st}| \right) \\
&\quad\quad + o_p(1) = E\left(K_h(x_i - x_j)|A_{j,st}|E(|\epsilon_i||x_i)\right) = O(1).
\end{aligned}
$$

Thus, we can have $R_{n1,2} = O_p(1)$. According to Lemma 2, we can have $\sqrt{n}(\hat{\theta}_N - \theta_0) = O_p(1)$. Then, we can conclude that

$$
T_{n1,2} = O_p(n^{-1/2}) \cdot O_p(n^{-1/2}) + O_p(n^{-1/2}) \cdot O_p(n^{-1/2}) = O_p(n^{-1}).
$$

Thus

$$
nh^{m/2}T_{n1,2} = O_p(h^{m/2}) = o_p(1). \tag{21}
$$

Similarly as the derivation for $T_{n1,2}$, we have

$$
nh^{m/2}T_{n1,3} = O_p(h^{m/2}) = o_p(1). \tag{22}
$$

For $T_{n1,4}$ in (24), we have

$$
\begin{aligned}
T_{n1,4} &= (\hat{\theta}_N - \theta_0)^\tau \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(\frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j) \frac{\partial f(x_i, \tilde{\theta}_1)}{\partial \theta} \right. \\
&\quad \left. \times \frac{\partial f(x_j, \tilde{\theta}_2)}{\partial \theta^\tau} \right) (\hat{\theta}_N - \theta_0) \\
&=: (\hat{\theta}_N - \theta_0)^\tau R_{n1,3}(\hat{\theta}_N - \theta_0).
\end{aligned}
$$

Similar to the argument for $R_{n1,2}$, we can conclude that $R_{n1,3} = O(1)$. We have

$$T_{n1,4} = O_p(n^{-1/2}) \cdot O_p(1) \cdot O_p(n^{-1/2}) = O_p(n^{-1}).$$

Thus,

$$nh^{m/2} T_{n1,4} = O_p(h^{m/2}) = o_p(1). \tag{23}$$

Based on (24), (20), (21), (22) and (23), we have

$$nh^{m/2} T_{n1} = nh^{m/2} T_{n1,1} + o_p(1) \to N(0, \Sigma^T). \tag{24}$$

Following the argument for T_{n1}, $nh^{m/2} T_{n2} = o_p(1)$ can be proved by proving $nh^{m/2} T_{n2,1} = o_p(1)$, here

$$T_{n2,1} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))}{\hat{\pi}(x_i)\hat{\pi}(x_j)\pi(x_j)} K_h(x_i - x_j)\epsilon_i \epsilon_j.$$

Note that by computing the variance of $T_{n2,1}$, we can have

$$\begin{aligned}
\mathrm{Var}(T_{n2,1}) = \frac{1}{n^2(n-1)^2 h^{2m}} & \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k=1}^{n} \sum_{l \neq k}^{n} \\
& \times E\left[\frac{\delta_i \delta_j \delta_k \delta_l (\hat{\pi}(x_j) - \pi(x_j))(\hat{\pi}(x_l) - \pi(x_l))}{\hat{\pi}(x_i)\hat{\pi}(x_j)\pi(x_j)\hat{\pi}(x_k)\hat{\pi}(x_l)\pi(x_l)} \right. \\
& \left. \times K(\frac{x_i - x_j}{h}) K(\frac{x_k - x_l}{h}) \epsilon_i \epsilon_j \epsilon_k \epsilon_l \right].
\end{aligned}$$

For the above summands, only the terms with $i = k$, $j = l$ and $i = l$, $j = k$ are non-zero. When $i = k$, $j = l$, we can have

$$\mathrm{Var}(T_{n2,1}) = \frac{1}{n^2(n-1)^2 h^{2m}} \sum_{i=1}^{n} \sum_{j \neq i}^{n} E\left[\frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))^2}{\hat{\pi}^2(x_i)\hat{\pi}^2(x_j)\pi^2(x_j)} K^2(\frac{x_i - x_j}{h}) \epsilon_i^2 \epsilon_j^2 \right].$$

Further note that

$$\begin{aligned}
& E\left[\frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))^2}{\hat{\pi}^2(x_i)\hat{\pi}^2(x_j)\pi^2(x_j)} K^2\left(\frac{x_i - x_j}{h}\right) \epsilon_i^2 \epsilon_j^2 \right] \\
& = E\left[\frac{(\hat{\pi}(x_j) - \pi(x_j))^2}{\pi(x_i)\pi^3(x_j)} K^2\left(\frac{x_i - x_j}{h}\right) \sigma^2(x_i)\sigma^2(x_j) \right] + o(1) \\
& \leq E\left[\frac{\sigma^2(x_i)\sigma^2(x_j)}{\pi(x_i)\pi^3(x_j)} K^2\left(\frac{x_i - x_j}{h}\right) \times \sup_x (\hat{\pi}(x) - \pi(x))^2 \right]
\end{aligned}$$

$$\le E^{1/2}\left[\frac{\sigma^4(x_i)\sigma^4(x_j)}{\pi^2(x_i)\pi^6(x_j)}K^4\left(\frac{x_i-x_j}{h}\right)\right]\times E^{1/2}[\sup_x(\hat{\pi}(x)-\pi(x))^4]$$
$$=\left[h^m\int K^4(u)du\cdot\int(\sigma^4(x))^2p^2(x)\pi^{-8}(x)\mathrm{d}x\right]^{1/2}\times O\left(\sqrt{\frac{\ln(n)}{nh^m}}^2\right)=o(h^m).$$

For the last equation to hold, we need the condition that $nh^{3m/2}\to\infty$. Thus, we can have $\mathrm{Var}(T_{n2,1})=o(n^{-2}h^{-m})$. Similarly, for the term with $i=l$, $j=k$, we can also obtain that $\mathrm{Var}(T_{n2,1})=o(n^{-2}h^{-m})$. Thus, $T_{n2,1}=o_p(n^{-1}h^{-m/2})$ which is also true for T_{n2}. Consequently,

$$nh^{m/2}T_{n2}=O_p(1)\cdot o_p(1)=o_p(1). \tag{25}$$

Similarly, we can obtain that

$$nh^{m/2}T_{n3}=o_p(1)\quad\text{and}\quad nh^{m/2}T_{n4}=o_p(1). \tag{26}$$

Combining the equations (18), (24), (25) and (26) together, we have

$$nh^{m/2}T_n=nh^{m/2}T_{n1,1}+o_p(1)\to N(0,\Sigma^T). \tag{27}$$

Note that

$$\sup_{a\le x\le b}|\pi(x,\hat{\alpha})-\pi(x,\alpha)|=O_p(\frac{1}{\sqrt{n}})=o_p(1),$$

by a similar derivation as T_n^N, we have

$$nh^{m/2}T_n^P\to N(0,\Sigma^T). \tag{28}$$

Below we prove the consistency of $\hat{\Sigma}^{\mathrm{TN}}$ based on U-statistics theory. Similarly as the derivation for T_n^N, we can verify

$$\hat{\Sigma}^{\mathrm{TN}}=\frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i}^n\frac{1}{h^m}\frac{\delta_i\delta_j}{\pi^2(x_i)\pi^2(x_j)}K^2\left(\frac{x_i-x_j}{h}\right)\epsilon_i^2\epsilon_j^2+o_p(1).$$

Invoking the U-statistics theory, if

$$E\left[\frac{1}{h^m}\frac{\delta_i\delta_j}{\pi^2(x_i)\pi^2(x_j)}K^2\left(\frac{x_i-x_j}{h}\right)\epsilon_i^2\epsilon_j^2\right]^2=o(n), \tag{29}$$

we can have

$$\hat{\Sigma}^{\mathrm{TN}} = 2E\left(\frac{1}{h^m}\frac{\delta_i\delta_j}{\pi^2(x_i)\pi^2(x_j)}K^2\left(\frac{x_i-x_j}{h}\right)\epsilon_i^2\epsilon_j^2\right) + o_p(1)$$
$$= 2\int K^2(u)du\cdot\int\frac{(\sigma^2(x))^2p^2(x)}{\pi^2(x)}\mathrm{d}x + o_p(1).$$

In fact, the equation (29) can be proved based on the following fact

$$E\left[\frac{1}{h^{2m}}\frac{\delta_i\delta_j}{\pi^4(x_i)\pi^4(x_j)}K^4(\frac{x_i-x_j}{h})\epsilon_i^4\epsilon_j^4\right]$$
$$= \frac{1}{h^{2m}}\int\int\frac{K^4((x_1-x_2)/h)\tilde{\sigma}^4(x_1)\tilde{\sigma}^4(x_2)}{\pi^3(x_1)\pi^3(x_2)}p(x_1)p(x_2)\mathrm{d}x_1\mathrm{d}x_2$$
$$= \frac{1}{h^m}\int\int\frac{K^4(u)\tilde{\sigma}^4(x)\tilde{\sigma}^4(x-hu)}{\pi^3(x)\pi^3(x-hu)}p(x)p(x-hu)\mathrm{d}x du$$
$$= \frac{1}{h^m}\int K^4(u)du\cdot\int\frac{(\tilde{\sigma}^4(x))^2p^2(x)}{\pi^6(x)}\mathrm{d}x + o(1/h^m)$$
$$= O(1/h^m) = o(n),$$

here $\tilde{\sigma}^4(x) = E(\epsilon_1^4|x_1)$. The consistence of $\hat{\Sigma}^{\mathrm{TP}}$ can be similarly derived when $\pi(x,\alpha)$ is a parametric function. We finish the proof for Theorem 1. □

Proof of Theorem 2 Denote

$$\bar{T}_n^N = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j.$$

Under the local alternatives (11) and recalling that $\epsilon_i = y_i - f(x_i,\theta_0)$, the term T_n^N can be decomposed as $T_n^N = \bar{T}_n^N + o_p(\bar{T}_n^N)$. As for $\bar{T}_n^N$, we have the expansion

$$\bar{T}_n^N = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i\epsilon_j$$
$$-\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i(f(x_j,\hat{\theta}_N)-f(x_j,\theta_0))$$
$$-\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_j(f(x_i,\hat{\theta}_N)-f(x_i,\theta_0))$$

$$+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Big(\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)(f(x_i,\hat\theta_N)-f(x_i,\theta_0))$$
$$\times(f(x_j,\hat\theta_N)-f(x_j,\theta_0))\Big)$$
$$=\bar T_{n1}-\bar T_{n2}-\bar T_{n3}+\bar T_{n4}+o_p(1). \tag{30}$$

For the term $\bar T_{n2}$ in (30), it follows that

$$\begin{aligned}\bar T_{n2}&=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_i f'(x_j,\tilde\theta)(\hat\theta_N-\theta_0)\\&\quad+C_n\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)G(x_i)f'(x_j,\tilde\theta)(\hat\theta_N-\theta_0)\\&=\bar T_{n2,1}(\hat\theta_N-\theta_0)+C_n\bar T_{n2,2}(\hat\theta_N-\theta_0),\end{aligned}$$

where $\tilde\theta$ lies between $\hat\theta_N$ and θ_0.

Based on the conclusion from Lemma 1, we have $\bar T_{n2,1}=O_p(n^{-1/2})$. It can also be proved that

$$\begin{aligned}\bar T_{n2,2}&=E(G(X_1)f'(X_2,\theta_0)K_h(X_1-X_2))+o_p(1)\\&=E(G(X)f'(X,\theta_0)p(X))+o_p(1).\end{aligned}$$

When $C_n=n^{-1/2}h^{-m/4}$, Lemma 2 implies that

$$\begin{aligned}nh^{m/2}\bar T_{n2}&=nh^{m/2}\Big[O_p(n^{-1/2})O_p(C_n)\\&\qquad+C_n^2E(G(X)f'(X,\theta_0)p(X))\Sigma_1^{-1}E(G(X)f'(X,\theta_0))\Big]\\&=E(G(X)f'(X,\theta_0)p(X))\Sigma_1^{-1}E(G(X)f'(X,\theta_0))+o_p(1).\end{aligned} \tag{31}$$

For $\bar T_{n3}$ in (30), we can similarly derive that

$$nh^{m/2}\bar T_{n3}=E(G(X)f'(X,\theta_0)p(X))\Sigma_1^{-1}E(G(X)f'(X,\theta_0))+o_p(1). \tag{32}$$

Using a similar argument as that for proving Theorem 1 and Lemma 2, we have the expansion for $\bar T_{n4}$:

$$\begin{aligned}\bar T_{n4}&=(\hat\theta_N-\theta_0)^\tau E[f'(X,\theta_0)f'^\tau(X,\theta_0)p(X)](\hat\theta_N-\theta_0)+o_p(C_n^2)\\&=C_n^2E^\tau[G(X)f'(X,\theta_0)]\Sigma_1^{-1}E[f'(X,\theta_0)f'^\tau(X,\theta_0)p(X)]\\&\qquad\times\Sigma_1^{-1}E[G(X)f'(X,\theta_0)]+o_p(C_n^2).\end{aligned}$$

As a result, when $C_n = n^{-1/2}h^{-m/4}$, we can obtain

$$nh^{m/2}\bar{T}_{n4} = E^{\tau}[G(X)f'(X,\theta_0)]\Sigma_1^{-1}E[f'(X,\theta_0)f'^{\tau}(X,\theta_0)p(X)] \\ \times\Sigma_1^{-1}E[G(X)f'(X,\theta_0)] + o_p(1). \tag{33}$$

Now we turn to investigate the term $\bar{T}_{n1}$ in (30); it can be decomposed as follows

$$\begin{aligned} \bar{T}_{n1} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_i\eta_j \\ &\quad + C_n\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_i G(x_j) \\ &\quad + C_n\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_j G(x_i) \\ &\quad + C_n^2\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)G(x_i)G(x_j) \\ &= \bar{T}_{n1,1} + C_n\bar{T}_{n1,2} + C_n\bar{T}_{n1,3} + C_n^2\bar{T}_{n1,4}. \end{aligned}$$

From the proof for Theorem 1 and the conclusion of Lemma 1, we know that

$$\begin{aligned} &nh^{m/2}\bar{T}_{n1,1} \to N(0,\Sigma^T); \\ &\bar{T}_{n1,2} = O_p(n^{-1/2}); \\ &\bar{T}_{n1,3} = O_p(n^{-1/2}); \\ &\bar{T}_{n1,4} = E(G^2(X)p(X)) + o_p(1). \end{aligned}$$

Consequently, when $C_n = n^{-1/2}h^{-m/4}$, we can obtain that

$$nh^{m/2}T_n^N \to N(\mu^T,\Sigma^T) \tag{34}$$

where

$$\mu^T = E\left[\left(G(X) - f'^{\tau}(X,\theta_0)\Sigma_1^{-1}E[G(X)f'(X,\theta_0)]\right)^2 p(X)\right].$$

Combining equations (31),(33) and (34), we can have

$$nh^{m/2}T_n^N \to N(\mu^T,\Sigma^T).$$

It can be similarly verified for $\pi(X,\alpha)$ to be estimated as $\pi(X,\hat{\alpha})$. When C_n has a slower convergence rate than $n^{-1/2}h^{-m/4}$, the above proof can show that the test statistic goes to infinity in probability. We omit the details. Theorem 2 is proved. □

Acknowledgments The authors thank the editor, the associate editor and two referees for their constructive comments and suggestions which led a substantial improvement of an early manuscript. The research described here was supported by a grant from the University Grants Council of Hong Kong, Hong Kong and National Natural Science Foundation of China (No. 11071253).

References

Aerts, M., Claeskens, G., Hart, J. D. (1999). Testing lack of fit in multiple regression. *Journal of the American Statistical Association*, *94*, 869–879.
Chown, J., Müller, U. U. (2013). Efficiently estimating the error distribution in nonparametric regression with responses missing at random. *Journal of Nonparametric Statistics*, *25*, 665–677.
Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Annals of Statistics*, *27*, 1012–1050.
Dette, H. (2002). A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *Journal of Statistical Planning and Inference*, *103*, 311–329.
Dette, H., Hildebrandt, T. (2012). A note on testing hypotheses for stationary processes in the frequency domain. *Journal of Multivariate Analysis*, *104*, 101–114.
Dette, H., Spreckelsen, I. (2003). A note on a specification test for time series models based on spectral density estimation. *Scandinavian Journal of Statistics*, *30*, 481–491.
Dette, H., Spreckelsen, I. (2004). Some comments on specification tests in nonparametric absolutely regular processes. *Journal of Time Series Analysis*, *25*, 159–172.
Dette, H., von Lieres und Wilkau, C. (2001). Testing additivity by kernel-based methods-what is a reasonable test? *Bernoulli*, *7*, 669–697.
Eubank, R. L., Li, C. S., Wang, S. J. (2005). Testing lack-of-fit of parametric regression models using nonparametric regression techniques. *Statistica Sinica*, *15*, 135–152.
Fan, J., Huang, L. (2001). Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association*, *96*, 640–652.
Gao, J., Wang, Q., Yin, J. (2011). Specification testing in nonlinear time series with long-range dependence. *Econometric Theory*, *27*, 260–284.
González-Manteiga, W., Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models. *Test*, *22*, 361–411.
González-Manteiga, W., Pérez-González, A. (2006). Goodness-of-fit tests for linear regression models with missing response data. *Canadian Journal of Statistics*, *34*, 149–170.
Guo, X., Xu, W. L. (2012). Goodness-of-fit tests for general linear models with covariates missed at random. *Journal of Statistical Planning and Inference*, *142*, 2047–2058.
Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, *14*, 1–16.
Härdle, W., Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, *21*, 1926–1947.
Koul, H. L., Ni, P. P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference*, *119*, 109–141.
Koul, H. L., Müller, U. U., Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Annals of Statistics*, *40*, 3031–3049.
Lavergne, P., Vuong, Q. H. (2000). Nonparametric significance testing. *Econometric Theory*, *16*, 576–601.
Li, X. Y. (2012). Lack-of-fit testing of regression model with response missing at random. *Journal of Statistical Planning and Inference*, *142*, 155–170.
Little, R. J. A., Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
Lopez, O., Patilea, V. (2009). Nonparametric lack-of-fit tests for parametric mean-regression models with censored data. *Journal of Multivariate Analysis*, *100*, 210–230.
Müller, U. U., Van Keilegom, I. (2012). Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics*, *6*, 1200–1219.
Sperlich, S. (2014). On the choice of regularization parameters in specification testing: a critical discussion. *Empirical Economics*. doi:10.1007/s00181-013-0752-z.
Stute, W., Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, *29*, 535–546.

Stute, W., Gonzáles-Manteiga, W., Presedo-Quindimil, M. (1998a). Bootstrap approximation in model checks for regression. *Journal of American Statistical Association*, *93*, 141–149.
Stute, W., Thies, S., Zhu, L. X. (1998b). Model checks for regression: An innovation process approach. *Annals of Statistics*, *26*, 1916–1934.
Stute, W., Xu, W. L., Zhu, L. X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, *95*, 451–467.
Sun, Z. H., Wang, Q. H. (2009). Checking the adequacy of a general linear model with responses missing at random. *Journal of Statistical Planning and Inference*, *139*, 3588–3604.
Sun, Z., Wang, Q., Dai, P. (2009). Model checking for partially linear models with missing responses at random. *Journal of Multivariate Analysis*, *100*, 636–651.
Van Keilegom, I., Gonzáles-Manteiga, W., Sánchez Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test*, *17*, 401–415.
White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, *76*, 419–433.
Xu, W. L., Guo, X., Zhu, L. X. (2012). Goodness-of-fitting for partial linear model with missing response at random. *Journal of Nonparametric Statistics*, *24*, 103–118.
Xue, L. G. (2009). Empirical likelihood for linear models with missing responses. *Journal of Multivariate Analysis*, *100*, 1353–1366.
Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*, 263–289.
Zhu, L. X., Ng, K. W. (2003). Checking the adequacy of a partial linear model. *Statistica Sinica*, *13*, 763–781.