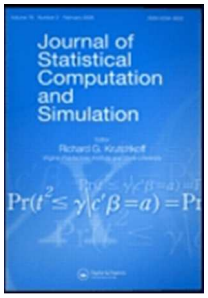

Model checks for nonparametric regression with missing data: A comparative study

Journal:	*Journal of Statistical Computation and Simulation*
Manuscript ID	GSCS-2015-0020.R2
Manuscript Type:	Original Paper
Areas of Interest:	MISSING DATA, BOOTSTRAP, COMPUTER INTENSIVE METHODS, CONTRAST TESTS, NONPARAMETRIC STATISTICS
2010 Mathematics Subject Classification:	62G08, 62G09, 62G10

To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–18

Model checks for nonparametric regression with missing data: A comparative study

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

This paper analyzes the behavior of the goodness of fit tests for regression models. To this end, it uses statistics based on an estimation of the integrated regression function with missing observations either in the response variable or in some of the covariates. It proposes several versions of one empirical process, constructed from a previous estimation, that uses only the complete observations or replaces the missing observations with imputed values. In the case of missing covariates a link model is used to fill the missing observations with other complete covariates. In all the situations, Bootstrap methodology is used to calibrate the distribution of the test statistics. A broad simulation study compares the different procedures based on empirical regression methodology, with smoothed tests previously studied in the literature. The comparison reflects the effect of the correlation between the covariates in the tests based on the imputed sample for missing covariates. In addition, the paper proposes a computational binning strategy to evaluate the tests based on an empirical process for large datasets. Finally, two applications to real data illustrate the performance of the tests.

Keywords: missing data in regression models, goodness of fit test, empirical process.

AMS Subject Classification: 62G08; 62G09; 62G10

1. Introduction

1.1. *Background*

A number of contributions using several methodologies have been published over the last years to test goodness of fit for regression models. Among them are smoothing-based tests or tests based on empirical regression processes (see a updated complete review in [1]). Most of these methods are applied to complete data. However, few studies have addressed behavior in tests when observations are missing in the response variable or in the covariates of the regression models.

The goodness of fit test is often based on the empirical estimator of the so-called integrated regression function. This method avoids the selection of a smoothing parameter when the regression model is estimated under the alternative with a parametric null hypothesis. Cramer-von-Mises or Kolmogorov-Smirnov type statistics are functionals over the estimated empirical processes marked by the residuals to check the model. Early references about this topic are [2, 3]. Numerous authors have used this methodology within different contexts.

During the last decades, several authors have used different approaches to study the missing data problem in different statistical areas, particularly in regression models. Among them are the complete case analysis, available case-analysis, imputation methods, multiple imputation, likelihood approach and so forth. For a broad review see [4]. Situations with missing observations in the response variable or incomplete covariates have generally been addressed separately. Papers scarcely consider both situations at a time (see [5]).

The goodness of fit test for regression models with missing data was addressed for the first time in [6]. This paper developed smoothing-based tests by applying the L^2 distance to check a parametric regression model with missing response. More recently, [7] proposes a type of test, based on minimum distances, to fit a parametric regression model with missing responses.

Following another methodological direction, several authors have used empirical processes. So, over an imputed sample, [8] tests the adequacy of partially linear models with missing response at random. It imputes the missing responses under the null hypothesis and uses inverse marginal probability weighted methods to fill the incomplete sample. The hypothesis concerning the nonparametric component, in partially linear models, is tested in [9] when the response has missing observations. Imputations are carried out as done in [8]. [10] uses score-type and empirical processes based test statistics over an imputed sample to check a general linear model. It uses the parametric model of the null hypothesis; thus it avoids the choice of the bandwith parameter, to make the imputation. A very important recent contribution to further methodological research with missing responses data is [11].

Studies within the context of goodness of fit test for regression models with missing covariates are even scarcer; the first publication we know of is [12]. Here the authors propose goodness-of-fit tests for generalized linear models based on conditional residuals. Later on, a paper checking general linear models with missing covariates [13] is published in which the authors use tests based on the residuals over the observed data weighting by parametric or nonparametric estimations of the missing data model. In general terms, when observations are missing in some of the covariates: Likelihood, Bayesian methods and inverse-probability weighting algorithms are the most frequently used approaches. However imputation methods are more difficult to address within this setting.

In this paper, our interest is to analyze, by means of a simulation study, the behaviour of several test statistics based on empirical processes in both settings: missing in response and missing in covariates. Test statistics are calculated for different situations: using the

complete and imputed data and weighting the observations by the missing data model. The study compares different empirical process tests with smoothed tests previously studied by [6] for the missing response case. Moreover, it compares the behaviour of the tests that assume known missing data model and those that estimate the missing data model. Bootstrap methodology is used to calibrate these tests. To our knowledge, no such similar comparative study with missing observations has been carried out as of yet.

On the other hand, the computational evaluation of an empirical process is known to be rather heavy for large datasets. So we propose a strategy based on binning techniques to reduce computational cost. This implementation can be seen in one application to real data. This context poses a challenge for future developments. Finally, let us also point out here that, the behaviour of the specification tests is known to be affected by a high covariate dimension. Several authors have proposed modified tests to address this problem in the complete data case: Escanciano [14], Lavergne and Patilea [15] and Stute *et al.* [16]. In the context of missing data Zhu *et al.* [17] and Bravo [18] use the same approach proposed by Escanciano [14] to address this fact for missing covariates and responses, respectively.

1.2. *Introduction of the tests*

Let (X, Y), a random vector in $\mathbb{R}^{d+1}$ such that Y has a finite expectation. Denote with $m(x) = \mathbb{E}[Y|X = x]$, $x \in \mathbb{R}^d$, be the associated regression function.

In the context of parametric regression, the function m is assumed to belong to a certain family $M_\Theta = \{m_\theta(\cdot), \theta \in \Theta \subset \mathbb{R}^p\}$, depending on some p-dimensional parameter θ. Important examples are the linear models.

The objective here is to check the regression model within the context of missing observations, by testing

$$H_0 : m \in M_\Theta \quad \text{versus} \quad H_1 : m \notin M_\Theta.$$

The pilot integrated regression function is defined as:

$$I(x) = \int_{-\infty}^{x} m(t)\, dF(t), \qquad x \in \mathbb{R}^d,$$

where F is the unknown distribution function of the covariate X.

An empirical consistent estimator of $I(\cdot)$ is given by

$$I_n(x) = n^{-1} \sum_{i=1}^{n} I_{\{X_i \leq x\}} Y_i.$$

The resulting empirical process is

$$R_n(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left(Y_i - \widehat{m}_\theta(X_i)\right),$$

where, $\widehat{m}_\theta(\cdot)$ is one parametric estimator of the regression function under the null hypothesis.

For the construction of different tests for this hypothesis we must choose a functional of R_n. In the case of complete data, [2] considers the weak convergence of R_n to a Gaussian process. Later, [3] uses bootstrap calibration to approximate the distribution of

R_n by considering different statistics as functionals of R_n. In particular, the Kolmogorov-Smirnov and the Cramér-von Mises statistics based on R_n are used to do this as follows. That is

$$D_n = \sup_x |R_n(x)|,$$

and

$$W_n^2 = \int_{\mathbb{R}^d} [R_n(x)]^2 F_n(dx),$$

where F_n is the empirical distribution function of $X_1, \ldots, X_n$.

Our objective is to design test statistics based on R_n adapted to the case of missing observations in the response variable or in any of the covariates.

Section 2 addresses the goodness of fit test with missing data in the response variable. Section 3 studies the case of missing data in some of the covariates. Section 4 discusses some of the computational aspects. Section 5 uses a simulation study to compare the performance of the proposed tests. Section 6 shows two applications to real data with missing observations. Finally, Section 7 gives some conclusions.

2. Missing data in the response variable

2.1. *Introduction to the test*

Let us consider the regression model:

$$Y = m(X) + \eta,$$

where η is the error, with $E[\eta|X] = 0$. A particular model is a heteroscedastic regression where $\eta = \sigma(X)\varepsilon$, with ε of mean 0 and variance 1, and $\sigma^2(x) = Var[Y|X = x]$. In the case of no missing observations, we have a sample $\{(X_i, Y_i)\}_{i=1}^n$ representing independent vectors with a distribution identical to that of the random vector $(X, Y) \in \mathbb{R}^{d+1}$. In missing response data, it may be that Y_i is not observed for any index i. This implies that we are faced with: (X_i, Y_i) if Y_i is observed, and $(X_i, ?)$ if Y_i is missing.

To control for the presence of a complete observation, we introduce a new variable δ into the model, as an indicator of the missing observations. Thus for each index i, $\delta_i = 1$ if Y_i is observed, and zero if Y_i is missing.

Following the guidelines laid down in previous publications (see [4] among others), it is necessary to establish whether or not the loss of a datum is independent of the value of the observed data and/or missing data. This paper models the aforementioned loss assuming they are Missing At Random (MAR), i.e.,

$$P(\delta = 1|Y, X) = P(\delta = 1|X) = p(X), \quad X \in \mathbb{R}^d. \tag{1}$$

Guo *et al.* [19] also use this model to check parametric regression. Niu *et al.* [20] consider a more general model, nonignorable missing response, to study confidence intervals for the parameters in a linear regression model. This model requires a correct specification of the mechanism of missing data; a misspecification could lead to biased estimations.

Two modifications of the process R_n are considered. One is based on the complete observations, i.e $(X_i, Y_i, \delta_i = 1)$ ("Simplified" version), and the other one is based on a

weighted expression using the model of missing data (1) ("Weighted Simplified" version). The processes are, respectively:

$$R_{n,S}(x) = n_1^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \delta_i \left(Y_i - \widehat{m}_{\theta,S}(X_i)\right),$$

and

$$R_{n,S,w}(x) = n_1^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \frac{\delta_i}{\hat{p}(X_i)} \left(Y_i - \widehat{m}_{\theta,S}(X_i)\right),$$

where $\widehat{m}_{\theta,S}(\cdot)$ is a parametric estimator of the regression function under the null hypothesis with the complete data, $n_1 = \sum_{i=1}^{n} \delta_i$, and $\hat{p}$ is an estimator of the missing data model (1).

We also propose three different processes based on the imputed samples. The first one, called "Imputed", is defined as follows:

$$R_{n,I}(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left[\left(\delta_i Y_i + (1-\delta_i)\widehat{m}_{h,S}(X_i)\right) - \widehat{m}_{\theta,S}(X_i)\right],$$

where $\widehat{m}_{h,S}(\cdot)$ is a nonparametric estimator of the regression function with the complete data and a bandwidth parameter h. The second considers the nonparametric estimations of all the responses. We denote it as the "Averaged" process defined by

$$R_{n,A}(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left(\widehat{m}_{h,S}(X_i) - \widehat{m}_{\theta,S}(X_i)\right).$$

Finally, we obtain the following mixture of the two previous processes ("Weighted Imputed" version)

$$R_{n,I,w}(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left[\frac{\delta_i}{\hat{p}(X_i)} \left(Y_i - \widehat{m}_{\theta,S}(X_i)\right) + \left(1 - \frac{\delta_i}{\hat{p}(X_i)}\right) \left(\widehat{m}_{h,S}(X_i) - \widehat{m}_{\theta,S}(X_i)\right)\right].$$

As test statistics, the Kolmogorov-Smirnov and the Cramér-von Mises functionals are applied to the previously defined processes.

2.2. *Bootstrap calibration of the tests*

We used a method based on Wild Bootstrap to approximate the distributions of the statistics in order to obtain the critical values. In [3] we may see that Wild bootstrap yields consistent approximations of distribution of R_n. [6] adapts this method to the situation with missing observations in the response variable. Following is the bootstrap procedure:

Step 1. From a random sample $\{(X_i, Y_i, \delta_i)\}_{i=1}^{n}$, where Y_i may not be observed for some i index, the residuals are defined by

$$\widehat{\eta}_i = Y_i - \widehat{m}_{\theta,S}(X_i), \text{ if } \delta_i = 1,$$

where $\widehat{m}_{\theta,S}$ is the least squares estimator of m under the null hypothesis using only the complete observations.

Step 2. Wild Bootstrap methodology [3] is used to resample the available residuals. That is to say, the bootstrap errors $\{\eta_i^*\}_{i\in I}$ obtained are such that:

$$\mathbb{E}_*\left[(\eta_i^*)\right] = 0 \text{ and } \mathbb{E}_*\left[(\eta_i^*)^2\right] = \widehat{\eta}_i^2, i \in I,$$

with I being the set of indices such that $\delta_i = 1$ and $\mathbb{E}_*$ denotes the expectation taken under bootstrap resampling.

Step 3. Repeat *Step 2*, B times, to generate B bootstrap samples, defined as: if $\delta_i = 1$ then $Y_i^* = \widehat{m}_{\theta,S}(X_i) + \eta_i^*$ and if $\delta_i = 0$, Y_i^* is missing. For each b, with $b = 1, \ldots, B$, the resultant bootstrap resample is

$$\left\{\left(X_i^{*,b}, Y_i^{*,b}, \delta_i^{*,b}\right) = \left(X_i, Y_i^{*,b}, \delta_i\right)\right\}_{i=1}^{n}.$$

The five versions of the empirical process are evaluated over the bootstrap sample. Then the Kolmogorov-Smirnov or the Cramér-von Mises functionals are applied to them to obtain the bootstrap test statistics, say T_b^* with $b = 1, \ldots, B$.

Step 4. The p-value is estimated by $\frac{k}{B+1}$ where k is the number of T_b^*, $b = 1, \ldots, B$, larger than or equal to T, the corresponding Kolmogorov-Smirnov or the Cramér-von Mises functional applied to the empirical process under study with the original sample. H_0 is rejected when p-value $\leq \alpha$ for a designed level α.

3. Missing data in the covariate

3.1. *Introduction to the test*

Now the missing data are in the covariables. We split the covariate X as $X = (X^c, X^m)$, where X^c contains these variables that are completely observed in all individuals whereas X^m represents the variables with missing observations. Let us reconsider the regression model:

$$Y = m(X^c, X^m) + \eta, \tag{2}$$

where η is the error term, with $E[\eta | X^c, X^m] = 0$.

To control whether an observation is complete, we introduce a new variable ξ into the model, as an indicator of the missing observations. Just like [21], we assume that the whole vector X^m is not observed. Thus, $\xi_i = 1$ for each index i if X_i^m is observed, and it is zero if X_i^m is missing. Without loss of generality, we suppose that X^m is one-dimensional.

In order to obtain information about the missing covariate, we consider the following link regression model:

$$X^m = g(X^c) + \zeta, \tag{3}$$

where X^m is now considered as the response variable, X^c represents the covariates totally observed and ζ has a conditional mean of zero, $E[\zeta | X^c] = 0$.

The previous model (3) is used to impute the missing observations of the covariate X^m. In order to obtain a consistent estimation of g in (3), we must assume a missing

model MAR:

$$P\left(\xi = 1|Y, X^c, X^m\right) = P\left(\xi = 1|X^c\right) = \pi\left(X^c\right).$$

First, we consider one "Simplified" process using only the complete observations

$$R_{n,S}\left(x^c, x^m\right) = n_1^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{(X_i^c, X_i^m) \leq (x^c, x^m)\}} \xi_i \left(Y_i - \widehat{m}_{\theta,S}\left(X_i^c, X_i^m\right)\right),$$

where $n_1 = \sum_{i=1}^{n} \xi_i$ and $\widehat{m}_{\theta,S}(\cdot)$ is a parametric estimator of the regression function under the null hypothesis using only the complete observations.

The other proposal imputes the missing observations. We use model (3) to impute the missing observations in the covariate X^m. The new "Imputed" process is defined as follows:

$$R_{n,I}\left(x^c, x^m\right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\left\{\left(X_i^c, \hat{X}_i^m\right) \leq (x^c, x^m)\right\}} \left[\left(Y_i - \widehat{m}_{\theta,I}\left(X_i^c, \hat{X}_i^m\right)\right)\right],$$

where $\hat{X}_i^m = \xi_i X_i^m + (1 - \xi_i)\, \hat{g}_S\left(X_i^c\right)$, $\hat{g}_S$ an estimation of the (3) with the complete data (simplified version). Finally, $\widehat{m}_{\theta,I}(\cdot)$ is an estimator of the regression function, under the null hypothesis, using the imputed covariate. We once again apply the statistics of Kolmogorov-Smirnov and the Cramér-von Mises to the previous processes.

3.2. *Bootstrap calibration of the tests*

Starting from a random sample $\{(X_i^m, X_i^c, Y_i, \xi_i)\}_{i=1}^n$ where X_i^m may not be observed for some i index, we follow the following steps in the bootstrap algorithm.

Step 1. We use the regression model (3) to impute the missing observations of the covariate X^m, $\hat{X}_i^m = \xi_i X_i^m + (1 - \xi_i)\, \hat{g}_S\left(X_i^c\right)$. Then, the residuals of (2) are obtained by

$$\widehat{\eta}_i = Y_i - \widehat{m}_\theta\left(X_i^c, \hat{X}_i^m\right),\ 1 \leq i \leq n,$$

where $\widehat{m}_\theta$ is the least squares estimator of m under the null hypothesis with the imputed data.

Step 2. We resample the available residuals following the Wild Bootstrap methodology [3], analogously to the missing response case (see *Step 2* in subsection 2.2).

Step 3. Repeat *Step 2*, B times, to generate B bootstrap samples defined as: $X_i^{c*} = X_i^c$, $\xi_i^* = \xi_i$, if $\xi_i = 1$ then $X_i^{m*} = X_i^m$, in other case X_i^{m*} is missing and $Y_i^* = \widehat{m}_\theta\left(X_i^c, \hat{X}_i^m\right) + \eta_i^*$. For each b, with $b = 1, \ldots, B$, the resultant bootstrap resample is

$$\left\{\left(X_i^{m*,b}, X_i^{c*,b}, Y_i^{*,b}, \xi_i^*\right) = \left(X_i^m, X_i^c, Y_i^{*,b}, \xi_i\right)\right\}.$$

We evaluate the empirical process $R_{n,S}$ or $R_{n,I}$ over the bootstrap sample and then apply the Kolmogorov-Smirnov or the Cramér-von Mises functionals to obtain the bootstrap test statistic, say T_b^* with $b = 1, \ldots, B$.

Step 4. We follow the same procedure used in the missing response case (see *Step 4* in subsection 2.2) to estimate the p-value.

4. Computational aspects

This section outlines a computational method that is suitable for evaluating the empirical process, R_n, for large datasets. Subsequently, subsection 6.2 applies this approach to a real case of missing data in the covariate. However, we only describe the application in the case of complete data. The application for the case with missing observations would be similar to that of the complete sample.

The evaluation of the empirical process in the multidimensional case can be heavy if the sample size is rather large. Evaluating R_n at a point x, requires $\binom{n}{d}$ comparatives of the indicator function for irregularly spaced data. We propose binning approximations to reduce computational cost. See the following references for more details [22–24].

Let the parametric residuals under the null hypothesis be: $\widehat{\eta}_i = Y_i - \widehat{m}_\theta(X_i)$.

We have applied the binning techniques to the sample $\{X_i, \widehat{\eta}_i\}_{i=1}^n$, where $X = \left(X^1, \ldots, X^d\right)^t$.

For $j = 1, \ldots, d$, let $g_{j,1} < \cdots < g_{j,M_j}$ be middle points of the equally spaced grid in the j-th covariate X^j,where M_j is the corresponding grid size. Also let

$$g_{l_1,\ldots,l_d} = (g_{1,l_1}, \ldots, g_{d,l_d}),$$

the multivariate grid point where $1 \leq l_j \leq M_j$.

For each point of the grid, $g_{l_1,\ldots,l_d}$, we define the grid count $(c_{l_1,\ldots,l_d}, \tilde{\eta}_{l_1,\ldots,l_d})$ which represents the amount of $(X, \widehat{\eta})$ data near each grid point. This can be defined as

$$c_{l_1,\ldots,l_d} = \sum_{i=1}^{n} w\left(X_i - g_{l_1,\ldots,l_d}\right), \quad \tilde{\eta}_{l_1,\ldots,l_d} = \frac{\sum_{i=1}^n w\left(X_i - g_{l_1,\ldots,l_d}\right)\widehat{\eta}_i}{c_{l_1,\ldots,l_d}},$$

where w are the weights according to the binning strategy. Several strategies are used to obtain grid counts in a multivariate setting (simple, linear), see reference [22] once again.

So, the empirical process can be approximated as

$$R_n(x) \approx \left(\frac{1}{\prod_{j=1}^d M_j}\right)^{-\frac{1}{2}} \sum_{l_1=1}^{M_1} \cdots \sum_{l_d=1}^{M_d} I_{\{g_{l_1,\ldots,l_d} \leq x\}} c_{l_1,\ldots,l_d} \tilde{\eta}_{l_1,\ldots,l_d}.$$

In the previous equation, $\prod_{j=1}^d M_j$ is the number of comparisons of the indicator function.

5. Simulation Study

Two different simulations evaluate the performance of the test statistics proposed in the previous sections: a model with missing data in the response and another one with missing data in the covariate.

5.1. *Missing in the response*

This simulation study :

- Compares the behaviour of the test statistics based on the $R_{n,S}$, $R_{n,S,w}$, $R_{n,I}$,$R_{n,A}$ and $R_{n,I,w}$.
- Compares them all with the tests that would have been computed with the complete data. These are denoted by $D_{n,C}$ and $W^2_{n,C}$. Table 1 shows a scheme of the test statistics

Table 1. Kolmogorov-Smirnov and the Cramér-von Mises statistics over empirical processes when the response variable has missing observations.

	Kolmogorov-Smirnov	Cramér-von Mises
Complete sample	$D_{n,C} = \sup_x \lvert R_n(x)\rvert$	$W^2_{n,C} = \int_{\mathbb{R}^d} [R_n(x)]^2 F_n(dx)$
Simplified	$D_{n,S} = \sup_x \lvert R_{n,s}(x)\rvert$	$W^2_{n,S} = \int_{\mathbb{R}^d} [R_{n,S}(x)]^2 F_n(dx)$
Simplified Weighted	$D_{n,S,w} = \sup_x \lvert R_{n,s,w}(x)\rvert$	$W^2_{n,S,w} = \int_{\mathbb{R}^d} [R_{n,S,w}(x)]^2 F_n(dx)$
Imputed	$D_{n,I} = \sup_x \lvert R_{n,I}(x)\rvert$	$W^2_{n,I} = \int_{\mathbb{R}^d} [R_{n,I}(x)]^2 F_n(dx)$
Averaged	$D_{n,A} = \sup_x \lvert R_{n,A}(x)\rvert$	$W^2_{n,A} = \int_{\mathbb{R}^d} [R_{n,A}(x)]^2 F_n(dx)$
Imputed Weighted	$D_{n,I,w} = \sup_x \lvert R_{n,I,w}(x)\rvert$	$W^2_{n,I,w} = \int_{\mathbb{R}^d} [R_{n,I,w}(x)]^2 F_n(dx)$

under study for this context.

- Finally, it also compares the behaviour of these statistics tests with the procedures based on smoothing proposed in [6]. The simplified version

$$T_{n,S} = n\,|H|^{\frac{1}{4}} \int \left(\widehat{m}_{S,H}(x) - \widehat{m}_{\theta,S}(x)\right)^2 w(x)\,dx,$$

and the imputed version

$$T_{n,I} = n\,|H|^{\frac{1}{4}} \int \left(\widehat{m}_{I,H,G}(x) - \widehat{m}_{\theta,S}(x)\right)^2 w(x)\,dx.$$

where H and G are multidimensional bandwidth parameters.

We consider the following regression model:

$$Y_i = 5X_i + aX_i^2 + \sigma(X_i)\,\varepsilon_i,\ 1 \le i \le n,$$

where X_i were generated from the uniform distribution in the interval $[0,1]$, $\varepsilon_i \sim Normal(0,1)$ and $a = 0,1,3,5,7,9$, the parameter $a \neq 0$ represents the deviation of a linear model. The sample sizes were $n = 50$ and $n = 100$. The missing data model was MAR given by $p(x) = 0.4 + 0.5\,(\cos(2x + 0.4))^2$. In a first step, we assumed that $p(x)$ was known. In a second step, we used a local linear kernel smoothing with bandwidth selected by cross-validation to estimate the missing probabilities.

Our objective was to test the null hypothesis:

$$H_0 : m \in \{m_\theta(x) = \theta_0 + \theta_1 x\}$$

Wild Bootstrap resampling was performed $B = 1000$ times for each sample in order to approximate the quantile of order $1-\alpha$, with $\alpha = 0.01, 0.05$ and 0.1. The experiment was iterated 1000 times and the percentage of rejections was calculated.

The bandwidth parameter of $\widehat{m}_{h,S}$ was selected by Cross-validation method over the complete subsample, i.e. $\{(X_i, Y_i, \delta_i = 1)\}$.

First, the results were expected to be better in the case of the complete data: $D_{n,C}$ and $W^2_{n,C}$. However, sometimes when the calibration of the tests was not optimal, their behaviour was not as good as expected. In the homoscedastic case (Table 2), we can observe that the Kolmogorov-Smirnov statistics based on the the weighted processes ($D_{n,S,w}$ and $D_{n,I,w}$) behave better than the others, especially as we move away from the null hypothesis ($a > 0$).

The Cramér-von Mises tests, generally, perform better than the Kolmogorov-Smirnovs tests. The statistics ($W_{n,S}$, $W_{n,S,w}$ and $W_{n,I,w}$) obtain the best results within the context

of missing data. The power of smoothing statistics used in the paper [6], $T_{n,S}$ and $T_{n,I}$, is similar to that of the tests based on empirical processes.

Table 2. Percentage of times that the null hypothesis is rejected with $\sigma(x) = 1$ and $n = 100$.

$\sigma(x) = 1$	a=0	a=1	a=3	a=5	a=7	a=9
$D_{n,C}$	5.3	8.1	42.8	83.2	98.8	100.0
$D_{n,S}$	5.4	8.3	26.5	57.2	85.9	97.7
$D_{n,I}$	6.2	8.3	27.4	58.1	84.1	95.7
$D_{n,S,w}$	6.4	7.0	31.8	64.7	92.0	98.1
$D_{n,A}$	6.6	6.6	23.4	53.4	79.7	93.0
$D_{n,I,w}$	6.6	8.9	32.7	63.3	88.6	96.7
$W^2_{n,C}$	4.6	9.7	50.4	90.8	99.9	100.0
$W^2_{n,S}$	6.0	8.6	34.1	68.9	94.1	99.7
$W^2_{n,I}$	6.1	8.0	32.3	65.0	89.4	98.0
$W^2_{n,S,w}$	6.0	7.3	34.4	69.4	94.2	99.4
$W^2_{n,A}$	6.1	7.3	27.4	59.6	84.9	95.7
$W^2_{n,I,w}$	6.2	8.6	35.2	68.1	90.5	98.0
$T_{n,S}$	6.1	8.9	31.5	66.7	92.6	98.4
$T_{n,I}$	6.1	9.0	33.0	66.8	92.5	98.5

Next we consider the heteroscedastic model $\sigma(x) = b(x+1)$ with $b = 0.5, 1$ and $n = 100$.

When $b = 0.5$ (see Table 3), the behaviour of the test statistics is similar to the homoscedastic case, because the dispersion is very small. Yet, we observe the effect of heteroscedasticity in the power of the tests when we consider the case $b = 1$ (Table 4). As in the homocedastic case, the Cramér-von Mises statistics generally perform better than the statistics based on Kolmogorov-Smirnov. It should be noted that the "Averaged" test, $D_{n,A}$, is generally more powerful than the tests based on Kolmogorov-Smirnov.

Table 3. Percentage of times that the null hypothesis is rejected with $\sigma(x) = 0.5(x+1)$ and $n = 100$.

$b = 0,5$	a=0	a=1	a=3	a=5	a=7	a=9
$D_{n,C}$	5.8	11.1	57.9	95.2	100.0	100.0
$D_{n,S}$	4.4	8.7	38.3	79.9	97.1	99.7
$D_{n,I}$	6.3	9.7	39.9	82.3	98.3	99.7
$D_{n,S,w}$	5.2	10.7	41.4	83.4	98.1	99.7
$D_{n,A}$	6.0	11.5	49.9	89.0	96.9	99.0
$D_{n,I,w}$	6.4	10.8	44.7	84.8	98.0	99.7
$W^2_{n,C}$	6.3	13.6	72.8	99.0	100.0	100.0
$W^2_{n,S}$	4.4	10.0	50.9	91.6	99.5	100.0
$W^2_{n,I}$	6.1	9.7	45.1	87.4	99.1	99.9
$W^2_{n,S,w}$	5.1	10.5	48.4	89.0	99.2	100.0
$W^2_{n,A}$	5.8	11.6	50.9	90.4	98.9	99.7
$W^2_{n,I,w}$	6.0	10.8	48.2	88.0	98.7	99.9
$T_{n,S}$	6.6	10.1	44.7	86.3	99.0	99.6
$T_{n,I}$	6.6	10.1	45.0	86.1	98.9	99.7

Table 5 shows the behaviour of the test statistics for $n = 50$ and $n = 100$ with levels $\alpha = 0.01, 0.05$ and 0.1. As expected, the results are better with $n = 100$ and the test gives a better approximation of the values of α.

As previously mentioned, prior simulations were carried out using the true missing data model (1). Our interest now lies in analyzing the effect of the estimation of the probability of missing data. This estimation affects the statistics $D_{n,S}, D_{n,I}, W^2_{n,S}$ and $W^2_{n,I}$. Tables 6 to 8 show the results obtained when the missing probability is true (rows

Table 4. Percentage of times that the null hypothesis is rejected with $\sigma(x) = (x+1)$ and $n = 100$

b = 1	**a=0**	**a=1**	**a=3**	**a=5**	**a=7**	**a=9**
$D_{n,C}$	6.0	6.9	18.4	42.9	72.6	92.2
$D_{n,S}$	4.9	7.1	11.9	29.6	50.7	72.3
$D_{n,I}$	4.5	8.4	14.0	32.1	54.6	76.9
$D_{n,S,w}$	5.1	7.8	15.0	32.9	55.2	78.3
$D_{n,A}$	5.3	8.2	18.4	39.8	62.6	83.7
$D_{n,I,w}$	5.8	8.1	15.7	34.7	56.5	80.7
$W^2_{n,C}$	5.1	7.7	25.4	55.5	84.3	97.5
$W^2_{n,S}$	5.3	6.8	17.8	39.4	65.7	85.3
$W^2_{n,I}$	4.4	7.9	16.0	35.4	60.0	82.9
$W^2_{n,S,w}$	5.2	7.8	16.8	36.1	62.3	84.5
$W^2_{n,A}$	5.6	7.7	18.3	40.2	64.4	86.0
$W^2_{n,I,w}$	5.4	7.6	16.3	36.2	60.6	83.8
$T_{n,S}$	5.8	8.7	16.9	36.1	58.9	80.7
$T_{n,I}$	6.4	8.9	17.6	35.8	59.2	81.3

Table 5. Percentage of times that the null hypothesis is rejected with $\sigma(x) = x + 1$, with $n = 50$, $n = 100$ and $\alpha = 1\%, 5\%, 10\%$

	n=50			n=100		
a = 0	1%	5%	10%	1%	5%	10%
$D_{n,C}$	0.3	4.2	11.7	1.3	6.0	11.0
$D_{n,S}$	0.3	4.8	9.8	0.3	4.9	9.8
$D_{n,I}$	1.2	6.4	11.4	0.8	4.5	10.1
$D_{n,S,w}$	0.8	6.5	12.5	1.2	5.1	10.4
$D_{n,A}$	0.9	6.4	12.1	0.9	5.3	10.7
$D_{n,I,w}$	1.1	6.7	13.4	1.3	5.8	11.6
$W^2_{n,C}$	0.9	4.7	10.5	0.7	5.1	10.4
$W^2_{n,S}$	0.5	5.2	10.9	0.9	5.3	9.2
$W^2_{n,I}$	1.5	6.8	11.6	1.5	4.4	9.3
$W^2_{n,S,w}$	0.9	6.3	13.3	1.5	5.2	10.3
$W^2_{n,A}$	0.8	6.6	13.2	1.2	5.6	11.1
$W^2_{n,I,w}$	1.0	6.2	12.5	1.1	5.4	10.6

with p) and nonparametically estimated (rows with $\widehat{p}$). The percentage of times that the null hypothesis is rejected is similar both for the true p and the estimated p. The behaviour, in terms of power, is slightly better for true p . This shows that the effect of estimating the missing data model (1) generally does not introduce major changes in the behaviour of the tests.

Table 6. Percentage of times that the null hypothesis is rejected with $\sigma(x) = 1$ when $p(x)$ is a real p or a nonparametrically estimated $\widehat{p}$ with $n = 100$.

$\sigma(x) = 1$		**a=0**	**a=1**	**a=3**	**a=5**	**a=7**	**a=9**
$D_{n,S,w}$	p	6.4	7.0	31.8	64.7	92.0	98.1
	$\widehat{p}$	6.3	7.7	29.8	60.2	87.7	96.6
$D_{n,I,w}$	p	7.2	8.1	34.2	67.1	93.1	98.2
	$\widehat{p}$	6.9	7.5	31.9	63.3	89.9	97.0
$W^2_{n,S,w}$	p	6.3	7.3	34.4	69.4	94.2	99.4
	$\widehat{p}$	6.5	7.1	32.3	67.0	92.3	99.0
$W^2_{n,I,w}$	p	6.9	7.2	35.4	69.4	93.4	98.8
	$\widehat{p}$	6.8	7.4	33.7	67.1	91.4	98.2

Table 7. Percentage of times that the null hypothesis is rejected with $\sigma(x) = 0.5(x+1)$ when $p(x)$ is a real p or a nonparametrically estimated $\widehat{p}$ with $n = 100$.

b = 0.5		**a=0**	**a=1**	**a=3**	**a=5**	**a=7**	**a=9**
$D_{n,S,w}$	p	5.2	10.7	41.4	83.4	98.1	99.7
	$\widehat{p}$	5.5	9.4	38.2	78.3	95.4	99.3
$D_{n,I,w}$	p	6.4	10.8	44.7	84.8	98.0	99.7
	$\widehat{p}$	6.6	9.8	40.3	81.7	95.9	97.9
$W^2_{n,S,w}$	p	5.1	10.5	48.4	89.0	99.2	100.0
	$\widehat{p}$	5.1	9.7	46.9	87.3	98.8	99.9
$W^2_{n,I,w}$	p	6.0	10.8	48.2	88.0	98.7	99.9
	$\widehat{p}$	5.5	9.6	46.2	86.2	98.0	99.1

Table 8. Percentage of times that the null hypothesis is rejected with $\sigma(x) = x+1$ when $p(x)$ is a real p or a nonparametrically estimated $\widehat{p}$ with $n = 100$.

b = 1		**a=0**	**a=1**	**a=3**	**a=5**	**a=7**	**a=9**
$D_{n,S,w}$	p	5.3	7.8	15.0	32.9	55.2	78.3
	$\widehat{p}$	4.8	8.0	12.5	30.4	49.8	74.0
$D_{n,I,w}$	p	6.9	8.1	15.7	34.7	56.5	80.7
	$\widehat{p}$	5.5	8.6	14.7	32.0	52.4	75.7
$W^2_{n,S,w}$	p	5.8	7.8	16.8	36.1	62.3	84.5
	$\widehat{p}$	5.7	7.1	14.9	35.2	61.4	83.0
$W^2_{n,I,w}$	p	6.0	7.6	16.3	36.2	60.6	83.8
	$\widehat{p}$	5.9	7.0	15.8	35.3	60.4	81.9

5.2. *Missing covariates*

We now consider the following bidimensional regression model:

$$Y_i = X_i^c + X_i^m + a\,(X_i^c)^2 + \eta_i,$$

where η_i and X_i^c are independent variables $Normal\,(0, 0.5)$. The link model between X_i^m and X_i^c

$$X_i^m = \gamma X_i^c + \zeta, \tag{4}$$

with $\zeta \sim N(0, 0.5)$ is used. We chose the values a=0, 0.05 and 0.1. Again the sample size was $n = 50, 100$, and the missing data model was $p(x) = (1 + \exp(-0.5 - 2x))^{-1}$. The number of bootstrap resamples was $B = 1000$ in order to approximate $\alpha-$level ($\alpha = 0.01, 0.05, 0.10$), and the empirical power of the goodness of fit test was

$$H_0 : m \in \{m_\theta(x) = \theta_0 + \theta_1 x^c + \theta_2 x^m\}$$

If we are to draw conclusions concerning the behaviour of different procedures, it is very important to understand the relationship between the variables (X^c, X^m). To this end we used different scenarios with different correlations between (X^c, X^m). We chose γ, in our simulation model (4), to obtain a correlation coefficient of $\rho_{(X^c, X^m)} = 0.64, 0.80, 0.98$.

We compared the complete (subscript C) and simplified case (subscript S) with various chosen imputation techniques.

- Subscript *true*: corresponds to imputation under the true model, i.e., if $\delta_i = 0$, $\hat{X}_i^m = \gamma X_i^c$.

Table 9. Kolmogorov-Smirnov and the Cramér-von Mises statistics over empirical processes when the a covariate has missing observations.

	Kolmogorov-Smirnov	Cramér-von Mises
Complete sample	$D_{n,C} = \sup_x \lvert R_n(x)\rvert$	$W^2_{n,C} = \int_{\mathbb{R}^d} [R_n(x)]^2 F_n(dx)$
Simplified	$D_{n,S} = \sup_x \lvert R_{n,s}(x)\rvert$	$W^2_{n,S} = \int_{\mathbb{R}^d} [R_{n,S}(x)]^2 F_n(dx)$
True Model Imputed	$D_{n,I_{true}} = \sup_x \lvert R_{n,I}(x)\rvert$	$W^2_{n,I_{true}} = \int_{\mathbb{R}^d} [R_{n,I}(x)]^2 F_n(dx)$
Least Squares Imputed	$D_{n,I_{ls}} = \sup_x \lvert R_{n,I}(x)\rvert$	$W^2_{n,I_{ls}} = \int_{\mathbb{R}^d} [R_{n,I}(x)]^2 F_n(dx)$
Nonparametric Imputed	$D_{n,I_{np}} = \sup_x \lvert R_{n,I}(x)\rvert$	$W^2_{n,I_{np}} = \int_{\mathbb{R}^d} [R_{n,I}(x)]^2 F_n(dx)$

- Subscript ls: corresponds to imputation under the null hypothesis, i.e., in this case $\delta_i = 0$, $\hat{X}_i^m = \hat{\gamma} X_i^c$ with $\hat{\gamma}$ estimated by least squares.
- Subscript np: corresponds to a nonparametric imputation, i.e, if $\delta_i = 0$, $\hat{X}_i^m = \hat{g}(X_i^c)$ with $\hat{g}(\cdot)$ being a nonparametric regression estimator.

Table 9 shows a scheme of the test statistics for this scenario.

In all cases, we applied the bootstrap given in Section 3.2, to approximate the percentage of rejections under the null hypothesis.

Table 10 shows the percentage of rejections under the null hypothesis and the alternative hypothesis, i.e. $a = 0, 0.05, 0.1$. We can observe that the imputed tests approximate the level (5%) better and are more powerful, than the "Simplified" test, as long as the correlation coefficient grows. The alternative gives rise to an interesting situation. When the correlation coefficient is not very large, the percentage of rejection in the "Simplified" test is higher than that of the imputed tests. However when $\rho_{(X^c,X^m)}$ is high, the imputed versions perform better because the covariate X^c provides sufficient information to impute suitably.

Table 11 shows the behaviour of the tests for different sample sizes and $\alpha = 1\%, 5\%, 10\%$, for $\rho_{(X^c,X^m)} = 0.8$. It should be noted that the model used for the simulation of covariates X^m (4) implies that the correlation coefficient not only influences in the case of imputation, but it also does so in the case of complete data $(D_{n,C}, W_{n,C})$. The abrupt change in the power of the tests, when $\rho_{(X^c,X^m)} = 0.8$ to 0.98, is not only due to a bigger dependence but it is also due to the fact that the parameter γ in model (4) increases the range of response.

Figures 1 and 2 show the performance of statistics tests as long as the parameter a grows for Cramer-von-Mises and Kolmogorov-Smirnov statistics, respectively. The figures suggest that if we consider a suitable imputation method, the simplified test is no longer competitive.

Table 10. Percentage of times that H_0 is rejected with $n = 50$, $a = 0, 0.05, 0.1$ and $\alpha = 5\%$.

	a=0			a=0.05			a=0.1		
$\rho_{(X,Z)}$	0.64	0.80	0.98	0.64	0.80	0.98	0.64	0.80	0.98
$D_{n,C}$	4.2	5.2	4.5	4.6	6.3	96.0	6.0	10.9	99.9
$D_{n,S}$	6.1	4.1	4.5	6.1	6.5	88.4	7.2	10.7	97.8
$D_{n,I_{true}}$	4.6	4.1	5.0	5.0	5.3	93.7	5.6	9.20	99.9
$D_{n,I_{ls}}$	5.0	4.2	5.3	4.9	4.9	93.7	5.6	8.8	99.9
$D_{n,I_{np}}$	5.7	3.3	5.3	5.3	5.2	91.6	5.5	8.8	99.8
$W_{n,C}$	5.8	5.6	5.9	6.4	7.8	99.6	8.5	16.9	100
$W_{n,S}$	6.5	4.8	6.1	7.2	8.7	97.9	10.4	14.9	100
$W_{n,I_{true}}$	6.4	5.9	5.5	6.7	7.4	99.1	8.5	14.1	100
$W_{n,I_{ls}}$	6.3	6.8	5.5	7.0	7.5	98.9	8.2	14.6	100
$W_{n,I_{np}}$	7.7	6.0	6.4	7.5	7.2	97.4	8.7	11.9	100

Table 11. Percentage of times that H_0 is rejected with $a = 0$, $\rho_{(X,Z)} = 0.80$, $n = 50, 100$ and $\alpha = 1\%, 5\%, 10\%$.

	n=50			n=100		
$\rho_{(X,Z)} = 0.80$	1%	5%	10%	1%	5%	10%
$D_{n,C}$	0.5	5.2	10.9	0.4	4.8	8.9
$D_{n,S}$	0.9	4.1	11.7	0.7	3.9	9.8
$D_{n,I_{true}}$	0.8	4.1	9.9	0.3	4.6	9.3
$D_{n,I_{ls}}$	0.4	4.2	10.8	0.3	4.4	9.9
$D_{n,I_{np}}$	0.4	3.3	10.7	0.5	5.5	10.4
$W_{n,C}$	0.3	5.6	11.7	1.0	4.2	9.8
$W_{n,S}$	0.7	4.8	11.4	0.8	4.9	11.3
$W_{n,I_{true}}$	0.7	5.9	12.2	0.9	3.8	9.7
$W_{n,I_{ls}}$	0.6	6.8	11.2	0.9	4.6	9.9
$W_{n,I_{np}}$	0.9	6.0	11.9	1.0	5.0	11.9

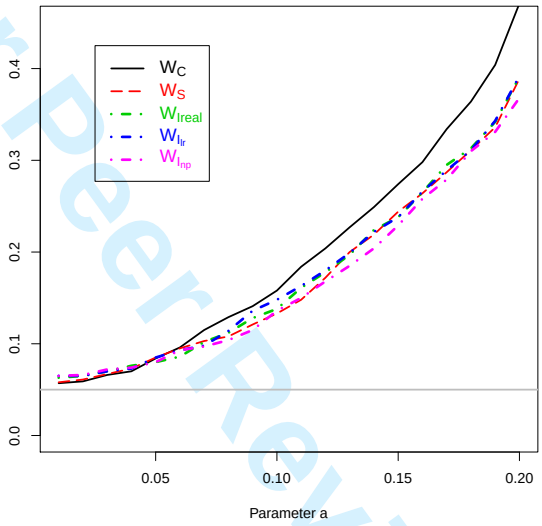

Figure 1. Estimated Power for Cramer-von-Mises statistics for missing data in the covariate context with $\rho_{(X,Z)} = 0.8$ and $n = 50$.

6. Real data analysis

This section applies the proposed tests on two real situations.

6.1. *Missing data in the response variable: Wind Energy data*

To illustrate a situation with missing data in the response variable, we consider Wind Energy data. The data set consists of 219 daily observations of daily Electric Power Production (MW in Megawatts) and daily average of Wind Velocity (Wv in m/s). Our interest is to check the linear model of the Electric Power versus Wind Velocity. The number of missing observations in the Electric Power variable is 8 (about 3.8%). Figure 3 plots daily electric power against daily average wind velocity with estimates from a linear regression with the complete cases.

A simple linear regression model proceeds and the null hypothesis is:

$$H_0 : E\,(MW|Wv) = \beta_0 + \beta_1 Wv$$

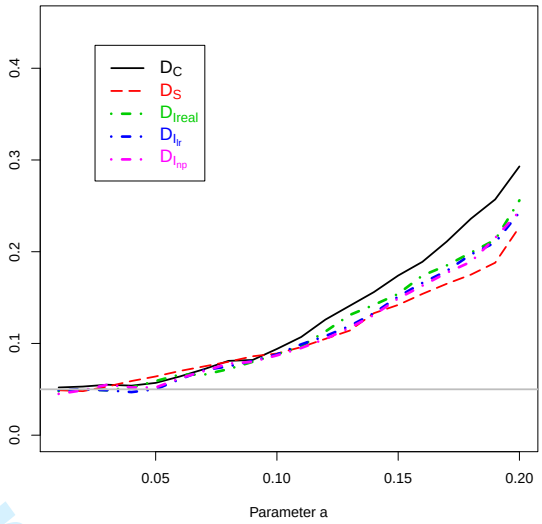

Figure 2. Estimated Power for Kolmogorov-Smirnov statistics for missing data in the covariate context with $\rho_{(X,Z)} = 0.8$ and $n = 50$.

for some β_0 and β_1.

We applied our test statistics, just like we did in the simulation study (see Table 1), to these data. All p-values were less than 0.0001, except for the simplified process ($R_{n,S,R}$) with 4% and 1.44% for KS and CvM, respectively. The null hypothesis can therefore be rejected in all the cases. This is consistent with graphic evidence and expert opinions in this field. Figure 4 shows the approximate bootstrap density estimation of Kolmogorov-Smirnov statistic tests.

6.2. *Missing data in the covariate*

In this subsection, we apply the statistical test to check whether the relationship between annual Salary (S in €) and the covariables "Duration of employment" (D in days) and "Age" (A in years) is linear (Data are provided by the Galician Institute of Statistics (IGE) in Spain and obtained from the "*Muestra continua de vidas laborales (MCVL) 2011*" elaborated by the Ministry of Employment and Social Security in Spain):

$$H_0 : E\left[S/A, D\right] = \beta_0 + \beta_1 A + \beta_2 D$$

Now the missing observations are in covariate D. They are imputed under a linear model with missing observations in the response, given by:

$$E\left[D/A\right] = \gamma_0 + \gamma_1 A. \tag{5}$$

This data contains a total of 37405 observations with 3409 (9.11%) observations missing in the covariate D. In the previous section, we observed that the correlation between the variable with missing observations and the covariate (used to impute) is of great importance to the performance of the imputed test. In this case, the linear correlation between D and A is 0.377. We used test statistics based on the sub-complete sample and the least squares imputed sample.

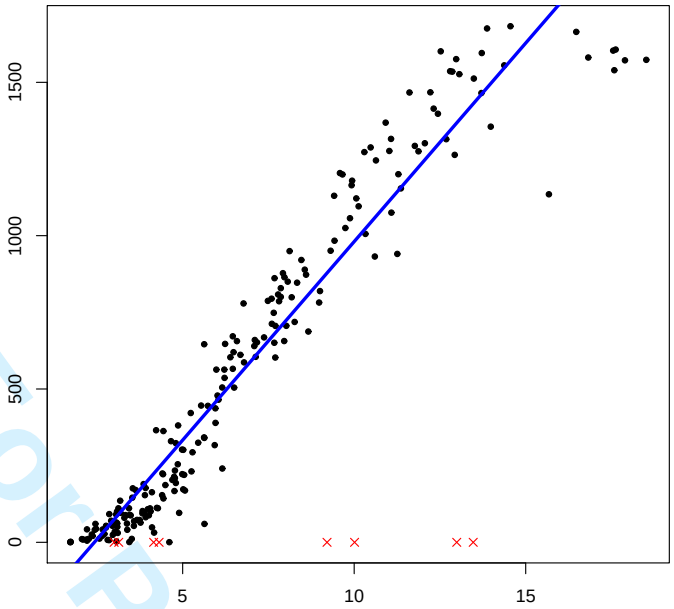

Figure 3. Daily electric power (MW) against daily average wind velocity (m/s). The solid straight line is the linear-regression fit to the complete case analysis. Solid circles are the 211 complete data cases and points with $\times$ correspond to missing data in the response variable.

We used software R code along with npsp package [25] (to apply the binning techniques) to obtain the results. Linear binning was used with $(40, 25)$ bins on each dimension.

All bootstrap p-values obtained with $B = 1000$, are less than 0.0001. Therefore, the null hypothesis of linearity can be rejected for all the cases.

7. Conclusions

This paper uses a simulation study to compare the behavior of several statistical tests based on functionals of empirical process that are used to test a parametric regression model with missing responses or missing covariates. In the case of missing data in the response variable, our results show that empirical processes based on imputed samples perform better in most cases.

In the case of the scenario with missing observations in the covariates, the results are different. If the correlation coefficient between the covariates is high, the test based on the imputed data performs better because the observed covariates provide more information about the missing observations. In the other case, when the correlation is not high, the imputed version is not competitive.

The asymptotic behaviour of these tests merits further research. The analysis of the effect of the covariance between the covariates and the missing data model in the asymptotic distribution of the tests could represent a future line of work.

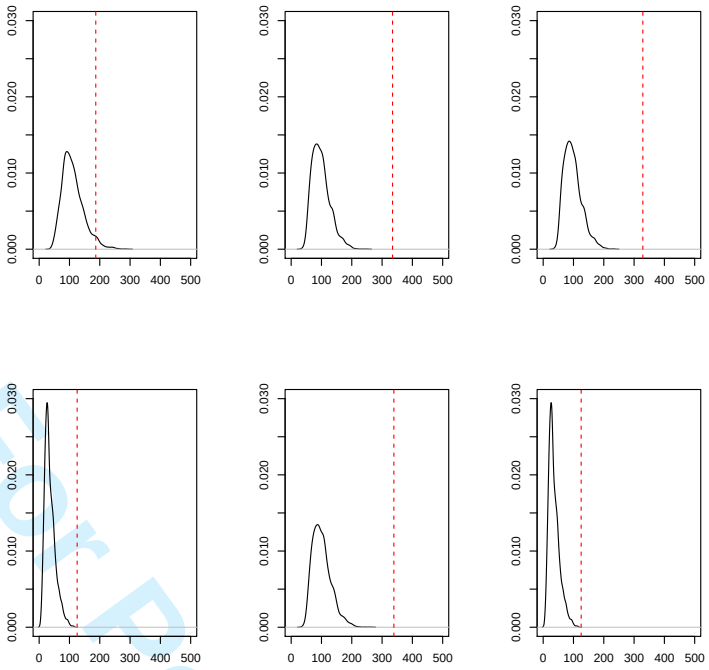

Figure 4. Bootstrap distribution of statistics test based on Kolmogorov-Smirnov. From top to bottom and left to right $D_{n,S,R}, D_{n,S}, D_{n,I,r}, D_{n,I}, D_{n,A}, D_{n,Sm}$.

References

[1] González-Manteiga W, Crujeiras RM. An updated review of Goodness-of-Fit tests for regression models. *Test*. 2013;22(3):361-411.
[2] Stute W. Nonparametric model checks for regression. Ann Stat 1997;25:613-641.
[3] Stute W, González-Manteiga W, Presedo-Quindimil M. Bootstrap approximations in model checks for regression. J Am Stat Assoc. 1998;93(441):141-149.
[4] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2th Ed. New York: Wiley; 2002.
[5] Zhou Y, Wan A, Wang X. Estimating Equations Inference with missing data. J Am Stat Assoc 2008;103:1187-1199.
[6] González-Manteiga W, Pérez-González A. Goodness-of-fit tests for linear regression models with missing response data. Can J Stat 2006;34(1):149-170.
[7] Li X. Lack-of-fit testing of a regression model with response missing at random. J Stat Plan Inference. 2012;142:155-170.
[8] Sun Z, Wang Q, Dai P. Model checking for partially linear models with missing responses at random. J Multivar Anal 2009;100:636-651.
[9] Xu W, Guo X, Zhu L. Goodness-of fitting for partial linear model with missing response at random. J Nonparametr Stat 2012;24(1):103-118.
[10] Sun Z, Wang Q. Checking the adequacy of a general linear model with responses missing at random. J Stat Plan Inference 2009;139:3588-3604.
[11] Koul H, Mller U, Schick A. The Transfer Principle: A tool for complete case analysis. Ann Stat 2012;40(6):3031-3049.
[12] Zhu H, Ibrahim J, Shi X. Diagnostic measures for Generalized linear models with missing covariate. Scand Stat Theory Appl 2009;36:686-712.
[13] Guo X, Xu W. Goodness-of-fit test for general linear models with covariates missed at random. J Stat Plan Inference 2012;142:2047-2058.
[14] Escanciano J. A consistent diagnostic test for regression models using projections. Econ

Theory 2006;22:1030-1051.
[15] Lavergne P, Patilea V. Breaking the curse of dimensionality in nonparametric testing. J Econometrics 2008;143:103-122.
[16] Stute W, Xu W, Zhu L. Model diagnosis for parametric regression in high-dimensional spaces. Biometrika 2008;95:451-467.
[17] Zhu H, Ibrahim J, Shi X. Diagnostic Measures for Generalized Linear Models With Missing Covariates. Scand J Stat 2009;36:686712.
[18] Bravo F. Partially linear varying coefficient models with missing at random responses. Ann Inst Stat Math 2013:65:721762.
[19] Guo X, Xu W, Zhu L. Model checking for parametric regressions with response missing at random. Ann Inst Stat Math 2015;67(2):229-259.
[20] Niu C, Guo X, Xu W, Zhu L. Empirical likelihood inference in linear regression with nonignorable missing response. Comput Stat Data Anal 2014;79:91112.
[21] Xu W, Guo X. Checking the adequacy of partial linear models with missing covariates at random. Ann Inst Stat Math. 2013;65(3):473-490.
[22] Wand MP. Fast Computation of Multivariate Kernel Estimators. J Comput Graph Stat. 1994;3(4):433-445.
[23] Ruppert D, Wand MP, Hssjer UH. Local Polynomial Variance function estimation. Technometrics. 1997;39(3):262-273.
[24] Wand MP, Jones MC. Kernel Smoothing. London: Chapman & Hall; 1995.
[25] Fernandez-Casal R. npsp: Nonparametric spatial (geo)statistics (R package version 0.3-6) [computer software]. 2014. Available from: http://CRAN.R-project.org/package=npsp