data

MDPI

Article

Regression-Based Approach to Test Missing Data Mechanisms

Serguei Rouzinov [1,*] and André Berchtold [2]

[1] Statistique Vaud, 1003 Lausanne, Switzerland
[2] Institute of Social Sciences & NCCR LIVES, University of Lausanne, 1015 Lausanne, Switzerland; Andre.Berchtold@unil.ch
* Correspondence: rouzinovs@gmail.com

Abstract: Missing data occur in almost all surveys; in order to handle them correctly it is essential to know their type. Missing data are generally divided into three types (or generating mechanisms): missing completely at random, missing at random, and missing not at random. The first step to understand the type of missing data generally consists in testing whether the missing data are missing completely at random or not. Several tests have been developed for that purpose, but they have difficulties when dealing with non-continuous variables and data with a low quantity of missing data. Our approach checks whether the missing data are missing completely at random or missing at random using a regression model and a distribution test, and it can be applied to continuous and categorical data. The simulation results show that our regression-based approach tends to be more sensitive to the quantity and the type of missing data than the commonly used methods.

Keywords: distribution; Dixon test; generating mechanisms; Jamshidian and Jalal test; Little test; missing data; regression

Citation: Rouzinov, S.; Berchtold, A. Regression-Based Approach to Test Missing Data Mechanisms. *Data* **2022**, 7, 16. https://doi.org/10.3390/data7020016

Academic Editor: Khalid Belhajjame

Received: 30 November 2021
Accepted: 17 January 2022
Published: 25 January 2022

1. Introduction

Data collection is a difficult process for any study and the probability of missing information is huge. It is important that we efficiently deal with missing data (MD) to minimise the bias of estimates and optimise the estimation of the variance of parameters of interest [1]. A variety of methods exist, such as deletion, imputation, or Bayesian methods, but, in general, they are powerful only for certain types of MD [2].

Following the definitions given by Rubin [3], MD are said to be Missing Completely At Random (MCAR) when (1) the probability of the data being missing is unrelated to the variable itself and (2) there is no link between the MD and one or several auxiliary variables, whether collected or not. For instance, a researcher's typo error when collecting data is considered as MCAR data. When the amount of missing information is limited, data analysts often assume that the data are MCAR, although this type of MD is very rare in reality [2].

MD are said to be Missing At Random (MAR) when the probability of missing (1) is unrelated to the variable itself but (2) can be explained, at least partially, by another variable. For example, workers could be more or less likely to answer questions regarding their income in function of their profession.

MD are said to be Missing Not At Random (MNAR) when they can be explained (at least partially) by the missing values themselves, in addition to other variables. For example, assume that highly paid workers are less likely to report their salary. Even if this situation is intuitively understandable, it is difficult to analyse this type of MD because the information about missingness is both unrelated with the observed data and unobserved, so we cannot determine whether the (unknown) values that are missing are themselves the cause for missingness or not. Moreover, the Rubin's definition of *MNAR* includes MAR hypothesis and it is very difficult to distinguish these two mechanisms.

The type of MD is important because it helps to determine which treatment should be applied to MD. Suppose we would like to estimate the IQ score from a specific sample

composed of 12 individuals. Table 1 provides complete and incomplete data (four values were deleted for the example). The average of the complete sample is 111.75 and the average of observed values in the incomplete sample is 101.75. In this example, we see that the average of incomplete data is biased, but, of course, in real life we do not have access to the value of MD. This relatively high difference is problematic and it let us suppose that MD are not MCAR. Ignoring them is therefore not the best choice, as the observed values do not form a representative sub-sample of the population. The first step would be to formally test whether these MD are MCAR or MAR. Then, an appropriate treatment of the MD should be applied.

Table 1. Sample with 12 IQ values.

	IQ1	**IQ2**	**IQ3**	**IQ4**	**IQ5**	**IQ6**	**IQ7**	**IQ8**	**IQ9**	**IQ10**	**IQ11**	**IQ12**
Complete data	95	119	94	134	130	92	128	100	99	110	105	132
Incomplete data	95	119	94	-	-	92	-	100	99	110	105	-

The usefulness of testing MD mechanisms is multiple. First of all, it helps to detect the variables that explain the MD. It is important to identify them to understand the MD mechanism. This helps to choose the best method to handle MD and to extract the maximum amount of information. Moreover, the comprehension of the mechanism generating the MD helps to understand the data and the psychology of respondents. This information can be used in similar studies to decrease the non-response rate. Finally, when the mechanism is known, it is easier to apply a correct method to handle MD.

The usual assumption of MCAR is problematic because in the majority of cases it is false [4], and we can have a relatively large bias if the handling of MD is not appropriate [5]. A common concern with MD is how to determine whether the data are MCAR or MAR. The difference between these two mechanisms was explored in detail by Heitjan and Basu [6]. On one hand, if the data are MCAR, it is possible to use a relatively easy method to analyse the data, such as listwise deletion, if the amount of MD is small [1]. On the other hand, if the data are MAR, they should be treated, for instance, with multiple imputation [3,7], inverse probability weighting [8], or Bayesian methods [9]. Moreover, it is very useful to know which variables explain the presence of MD in order to include them in the final model to answer to the research question. Whether the MAR condition holds is avowedly impossible to test given only observed data [10,11]. In the case of *MNAR* data, it is impossible to test the mechanism on observed data as well, and thus, it has to be analysed with caution with the use of sensitivity analysis [12].

Several approaches have been developed to test the MD mechanism. Three of them are commonly used currently: the Dixon test [13], the Little test [14] along with its extensions, and the Jamshidian and Jalal test [15–17]. These approaches have several advantages and limitations, which are discussed later. Their aim is to test whether the MD are MCAR. It is important to mention that when the method for handling MD is valid for both MCAR and MAR data, then it is also interesting to test the MD mechanisms in order to find additional information, such as the the combination of variables which explain the MD. Thus, a procedure that can extract maximum information from the data is needed, because this is the key to distinguish between MCAR and MAR.

Hereafter, we propose a brief review of the literature on testing MD mechanisms. Then, we develop a complementary approach to test missingness based on regression, which can be applied to both categorical (polytomous) and continuous variables. The aim of our approach is not to test whether MD are ignorable or not, but to correctly differentiate between MCAR and MAR. This is similar to what is performed by existing tests, but one should be aware that the possibility that the MD are actually *MNAR* is left out. We assume that our data are in a restricted environment so that the amount of information does not tend to infinity. A simulation study demonstrates the behaviour of our approach, the objectives being (1) to compare our approach with the commonly used procedure to

test missingness, and (2) to check the robustness of this approach in function of the MD mechanism, the quantity of MD, the sample size, and the data distribution.

2. Existing Approaches for Testing Missing Data Mechanisms

A common concern with MD is how to determine whether the data are MCAR or MAR. Hereafter, we first review the three main tests developed in the literature: the Dixon [13], the Little [14], and Jamshidian and Jalal [15] tests, which aim to test the following hypotheses:

$$\begin{cases} H_0 : MCAR \\ H_1 : MAR \end{cases}$$

2.1. Dixon Test

One of the easiest tests to determine whether the MD are MCAR is to verify the sample group's differences [13]. Assume that X_1 and X_2 are two numerical vectors, that a certain percentage of observations in X_1 are missing, and that X_2 is a complete vector. The first step is to divide X_2 into two subsamples, one with individuals where X_1 is observed, and the other with individuals where X_1 is missing:

$$X_2 = \begin{cases} X^{2,mis} \text{ if } X_1 \text{ is missing} \\ X^{2,obs} \text{ if } X_1 \text{ is observed} \end{cases}$$

Next, the means of $X^{2,mis}$ and $X^{2,obs}$ are compared by using a two-sample t-test to check for either significant difference between means (alternative hypothesis) or no difference (null hypothesis). When the null hypothesis is rejected, the conclusion is that the data are not MCAR. This approach can be extended to the more general case with more than two variables. However, even in the case of only two variables, a mean comparison does not allow for a conclusive test, because there could be other differences between the two distributions [5]. Moreover, the mean comparison can be applied only on continuous data. To conclude, this test is inappropriate in the majority of cases and this is why other approaches have been developed.

2.2. Little Test

One of the most used methods to test MD mechanisms was developed by Little [14]. It tests whether the MD in a multivariate Gaussian distributed dataset are MCAR or not. According to this approach, the missingness can be tested by applying a likelihood ratio test asymptotically based on a chi-squared distribution. This test is summarised in the following example. Suppose a dataset $D(n \times k)$ containing n observations and k variables with a general pattern of MD (each variable can contain a certain amount of MD). Let $X_i = X_{i1}, ..., X_{ik}$ be the vector of values for observation i, $X_{obs,i} = X_{obs,i1}, ..., X_{obs,ik}$ the vector of values of observed variables in case i, and r_i the indicator for MD for observation i, such that $r_i = 1$ if X_i is missing and $r_i = 0$ if X_i is observed. This gives the number of distinct MD patterns, J, and the set S_j of MD patterns for each $j = 1, ..., J$. Furthermore, $\mu_{obs,j}$ and $\Sigma_{obs,j}$ are, respectively, the vector of means and the covariance matrix of observed variables in j. The aim of this approach is to test whether the means of each S_j are significantly identical or not. It assumes that the variance is known and uses a likelihood ratio test on the following test hypotheses:

$$\begin{cases} H_0 : (X_{obs,i}|r_i) \underset{ind}{\sim} N(\mu_{obs,j}, \Sigma_{obs,j}), & i \in S_j, 1 \leq j \leq J \\ H_1 : (X_{obs,i}|r_i) \underset{ind}{\sim} N(\gamma_{obs,j}, \Sigma_{obs,j}), & i \in S_j, 1 \leq j \leq J \end{cases}$$

where $\gamma_{obs,j}$ is the vector of means of observed variable in j, but, unlike $\mu_{obs,j}$, is distinct for each pattern j. If the true variance–covariance matrix is unknown, it can be estimated by the maximum likelihood estimator [14].

Although this test is relatively efficient for large samples, it is conservative for small ones. The Little test works well for continuous variables, but categorical data must be recoded into numerical discrete data, a transformation that can rarely be justified theoretically [18]. Two other important problems are that the Little test does not test the dispersion of MD sequence and it is not robust to a deviation from a Gaussian distribution unless the sample size is large enough. The Little test has several extensions, such as a test for repeated categorical or longitudinal data using a stratification procedure [19,20] and a Wald-type test for generalised estimating equations with MD [21].

2.3. Jamshidian and Jalal Test

Another commonly used approach for testing the MCAR hypothesis was developed by Jamshidian and Jalal [15]. The authors aim was to create a test of homoscedasticity, which is a synonym of a test of homogeneity in covariances, that works for data that do not necessarily follow a Gaussian distribution. The first step of this approach is to test the normality and homoscedasticity by using the modified Hawkins [22] test. Jamshidian and Jalal [15] modify it for being able to use it when data are incomplete with the following hypotheses:

$$\begin{cases} H_0 : \text{Normality and homogeneity of covariances} \\ H_1 : \text{Non-normality or non-homogeneity of covariances} \end{cases}$$

If the null hypothesis is accepted, then the data are considered normal and with a homogeneity of covariances and, without additional information, MD can be considered as MCAR. On the other hand, when the modified Hawkins test is rejected, data can be either: (1) non-normal and with a homogeneity of covariances; (2) normal and with a non-homogeneity of covariances; or (3) non-normal and with a non-homogeneity of covariances. The next step after the rejection of the modified Hawkins test is to decide whether the data are normally distributed or not. On one hand, if the researcher assumes that data follow a normal distribution, then there is no other choice but to reject MCAR hypothesis, because (1) and (3) are excluded. On the other hand, if it is impossible to make such an assumption, then Jamshidian and Jalal [15] developed a non-parametric pairwise variable (NPV) procedure with the following test hypothesis:

$$\begin{cases} H_0 : \text{Homogeneity of covariances} \\ H_1 : \text{Nonhomogeneity of covariances} \end{cases}$$

If the null hypothesis cannot be rejected, then the homogeneity of covariances cannot be rejected as well as the non-normality; consequently, MD are considered to be MCAR. When the test is rejected, then the MCAR hypothesis is rejected as well. However, in this last case no test conclusion can be made about normality. First, each variable is divided in two groups based on the unobserved and observed cases of the variable. The first part consists of observed values and the the second part consists of missing values. Then, the Scholz and Stephens [23] rank-based test or the Anderson and Darling [24] k-sample test is applied to the two parts of the variable to evaluate the equality of distribution between the subsamples. This is performed $K(k-1)$ times, where k is the number of variables and K is the total number of partly observed variables. Thus, the tests are performed simultaneously, which can be complicated when $k > 2$ [25]. Moreover, because this test has a relatively low power to reject Type I errors [26], Jamshidian and Jalal proposed to use the Benjamini and Hochberg [27] test to fix this issue. More details about the NPV test can be found elsewhere [16].

Finally, the rejection of both the Hawkins and NPV tests gives evidence against the MCAR hypothesis.

2.4. Comparison of the Three Tests

These three approaches for testing MD mechanisms have various limitations. The Dixon test is a very simple and understandable approach, but it is problematic because of (1) the collinearity aspect when there are more than two variables and (2) the non-use of the dispersion within the test. The Little test is the most used approach and it is available in the majority of statistical software. However, it can accept the MCAR hypothesis as soon as there is no difference in means between two subsets, which is not always sufficient because a difference can exist in terms of distribution with no difference in means. Furthermore, it is not robust to a deviation from a Gaussian distribution unless the sample size is large enough. The Jamshidian and Jalal procedure takes into account the dispersion of the data, it does not assume data normality, and no large sample size is required. Nevertheless, its structure is relatively complicated to implement because it uses a combination of two tests and it is non-intuitive for end users. Finally, none of the three tests is really appropriate for categorical data, and they have difficulties dealing with data with a relatively low number of MD. Therefore, the objective of this article is to provide an alternative and complementary approach to these tests by (1) testing the distribution rather than the mean of the variables with MD, and (2) allowing for the approach to be adapted to any type of variables, continuous or categorical.

3. Regression-Based Approach

This section presents our regression-based (RB) approach for testing the following hypotheses:

$$\begin{cases} H_0 : MCAR \\ H_1 : MAR \end{cases}$$

First, the procedure is described for the continuous case and then the binary and categorical cases are described.

3.1. Principle

Consider a dataset with n observations and k variables. Suppose that only one variable, X_1, has MD and the other variables are fully observed. Let A be the subset of completely observed data, and let B be composed of observations missing on X_1 but fully observed on $X_2, X_3 ..., X_k$ (Figure 1). $X^A_{1,obs}$ is the observed part of X_1, and $X^B_{1,mis}$ is the missing part of X_1. The objective is to compare the distribution of $X^A_{1,obs}$ and $X^B_{1,mis}$, but since data are unobserved on $X^B_{1,mis}$, a regression model is built to explain $X^A_{1,obs}$ from A. Then, this model is used to make predictions $\widehat{X^A_{1,mis}}$ and $\widehat{X^B_{1,obs}}$ for both $X^A_{1,obs}$ and $X^B_{1,mis}$.

In practice, a regression model is defined on the fully observed data (part A of Figure 1):

$$X^A_{1,obs} = g_1(X^A_2, X^A_3, ..., X^A_k) \quad (1)$$

where $X^A_2, ..., X^A_k$ are the fully observed variables from A and $g_1(.)$ is a link function that depends on the type of X_1. Interaction terms can also be included in the model.

Partly observed	Completely observed	
$X^A_{1,obs}$	$X^A_{2,obs}, X^A_{3,obs}, \dots, X^A_{k,obs}$	A
$X^B_{1,mis}$	$X^B_{2,obs}, X^B_{3,obs}, \dots, X^B_{k,obs}$	B

Figure 1. Data structure.

The next step of the RB procedure is to compare $\widehat{X^A_{1,obs}}$ and $\widehat{X^B_{1,mis}}$ according to the data type. Regression (1) provides a set of estimated coefficients, $\widehat{\beta}_A$. $\widehat{\beta}_A$ and A are used to predict $X^A_{1,obs}$, and then $\widehat{\beta}_A$ and B are used to predict $X^B_{1,mis}$, to obtain two sets of predicted values, $\widehat{X^A_{1,obs}}$ and $\widehat{X^B_{1,mis}}$. This means that the same set of coefficients is used to predict the observed and missing data.

The difference in distribution between $\widehat{X^A_{1,obs}}$ and $\widehat{X^B_{1,mis}}$ is the key element of the RB approach to accept or reject the hypothesis that data are MCAR. If the MD are truly MCAR, their presence on the response indicator variable X_1 cannot be explained by other available variables in the dataset. However, if the data on this variable are MAR, some information obtained from the other variables should at least partially be able to explain the presence of MD. Consequently, there should be differences between A and B on variables $X_2, \ldots, X_k$, and these differences should imply a difference between $\widehat{X^A_{1,obs}}$ and $\widehat{X^B_{1,mis}}$. When the null hypothesis is accepted, it is assumed that the missing part of data on the dependent variable is MCAR.

3.2. Continuous Case

$\widehat{X^A_{1,obs}}$ and $\widehat{X^B_{1,mis}}$ are compared in function of the type of X_1. A linear regression is used to estimate Equation (1) when X_1 is numerical, and the general Kolmogorov and Smirnov (KS) test is applied [28] to compare the estimated values. However, other less usual tests could be used, because the KS test is known to be limited for different reasons, one of them being that it has problems when data are not normally distributed [29]. An interesting alternative could be the Cucconi test [30] because it is generally more powerful than other non-parametric distribution tests and the assumption about the normality of the data is not required. It is available in the R software [31]. Other alternatives, such as the Wilcoxon–Mann–Whitney procedure (a general non-parametric test, [32]) or the Kruskal–Wallis approach (for ordinal data, [33]), could also be used. We advise the use of the Kolmogorov–Smirnov test when data are normally distributed and the Kruskal–Wallis approach when data are ordinal.

Notice that the quality of the regression model used to compute the two sets of predicted subsamples is not essential for the test to perform well. What is essential is, as with the Little, the Dixon, and the Jamashidian and Jalal tests, the inclusion of all explanatory variables relevant to the missingness of data on the response indicator variable of X_1. When such explanatory variables are not sufficient to explain the variations in the dependent variable, the model fit can be low even if the auxiliary variables are associated with missingness.

3.3. Binary Case

Suppose that the variable with MD, X_1, is binary with two categories, a and b. In this case, Equation (1) becomes a logistic regression:

$$\pi_{lr} = g_2(X^A_{2,obs}, X^A_{3,obs}, ..., X^A_{k,obs}) \tag{2}$$

where $g_2(.)$ is the link function and π_{lr} is:

$$\pi_{lr} = \log \frac{p}{1-p}$$

with $p = Pr(X_{1,obs} = a)$ and $1 - p = Pr(X_{1,obs} = b)$. Then, the predicted probabilities are computed by using the predicted coefficients $\widehat{\beta_0}, ..., \widehat{\beta_k}$ from model (2):

$$\widehat{Pr}(X_{1,obs} = a) = \frac{\exp(\widehat{\beta_0} + \widehat{\beta_1} X^A_2 + ... + \widehat{\beta_k} X^A_k)}{1 + \exp(\widehat{\beta_0} + \widehat{\beta_1} X^A_2 + ... + \widehat{\beta_k} X^A_k)} \tag{3}$$

$$\widehat{Pr}(X_{1,mis} = a) = \frac{\exp(\widehat{\beta_0} + \widehat{\beta_1} X_2^B + ... + \widehat{\beta_k} X_k^B)}{1 + \exp(\widehat{\beta_0} + \widehat{\beta_1} X_2^B + ... + \widehat{\beta_k} X_k^B)} \quad (4)$$

$$\widehat{Pr}(X_{1,obs} = b) = 1 - \widehat{Pr}(X_{1,obs} = a) \quad (5)$$

$$\widehat{Pr}(X_{1,mis} = b) = 1 - \widehat{Pr}(X_{1,mis} = a) \quad (6)$$

Then, the differences between (3) and (4) and between (5) and (6) must be tested using the KS test, because the acceptance or rejection of the test for one category does not automatically imply the acceptance or rejection of the test for the other category. This procedure requires two tests (number of categories), and thus the problem of simultaneity of tests should be addressed. A general Bonferroni-type method could be used. However, this procedure is known to have a low power to reject incorrect null hypotheses [26]. To deal with this problem, we propose adjusting the *p*-values of the KS test by applying the Benjamini and Hochberg (BH) test [27]. If the hypothesis of no significant distribution differences is not rejected for both categories, then the MD are assumed to be MCAR. Otherwise, it is supposed that the MD are not MCAR.

3.4. Categorical Case

Consider the more general case, where the variable with MD, X_1, is categorical (or polytomous) with $z > 2$ modalities ($z = 1, ..., Z$). A multinomial regression model must then be used. In some cases, an ordinal regression can be considered, but it is often impossible to assume the same gaps between modalities [34], so the multinomial model is generally preferred.

After the prediction process based on the multinomial regression, the difference between predicted probabilities needs to be verified with the KS and BH tests:

$$\widehat{Pr}(X_{1,obs} = z) \overset{d}{\sim} \widehat{Pr}(X_{1,mis} = z)$$

The number of comparisons is equal to the number of categories, so that there are J KS and BH tests, because the acceptance or rejection of the first $J - 1$ categories does not automatically imply the acceptance or rejection of the test for the last category. The BH test is always conducted to avoid the problems due to multiple comparisons. If all J KS and BH tests accept the equality of distributions, then we conclude that the MD for the incomplete variable are MCAR. Otherwise, the null hypothesis is rejected.

3.5. Discussion

In contrast to the existing approaches for testing missing data mechanisms which consider the mean [13,14] or the covariance [15], we developed a regression-based approach, which can be applied to either continuous or categorical data. A related, but different, procedure was also proposed in the literature. It consists of dichotomising the variable with MD between observed and non-observed values, and to apply a logistic regression model [35,36]. However, this method has never been applied on a large scale, and it is not proposed in the commonly used statistical software or in textbooks as a standard method.

The RB method assumes that only one variable in the dataset has MD. Obviously, this situation rarely occurs in reality. One possibility is to replace the MD of independent variables in the regression model using imputation [3], but imputing data before testing to determine the type of MD is problematic because wrong imputations can modify the relationship between the cause of the MD and the MD themselves, thus limiting the probability of identifying the real MD type. A possible solution to this issue is an iterative procedure composed of two steps: an initialisation and an iterative phase. We describe this procedure in the case of single imputation, but it can be used with multiple imputation as well. During the initialisation step, the RB approach is first applied to one of the variables with MD using only the variables without MD as predictors. Then, in function of the result of the RB test, the MD are imputed with respect to their mechanism: a random draw

from observed data is used to replace MCAR data, and MAR data are imputed using the chained equation principle [37]. An additional complete variable is then available. Next, a second variable with MD is tested for missingness, using the imputed variable and the variables without MD. This procedure is repeated until the data are fully completed. Next, the RB approach is iteratively applied to each variable with MD, using all other variables as predictors in the regression model (including imputed variables). This is essential, because any change between two iterations can have an impact on the prediction of the dependent variable. If the test result for a variable with MD differs from that of the previous iteration, the MD are imputed again. At the end of each iteration, the pattern of MD is computed again. This step is repeated until no differences are found between the MD patterns of two successive iterations. It must be noted that convergence is not certain, so a maximal number of iterations has to be set. This procedure is applied to a real example in Section 5.

4. Simulation Study

We describe hereafter a simulation study demonstrating the behaviour of the RB test.

4.1. General Setting

In this section, we compare the behaviour of the Regression-based approaches with the more standard Little, Dixon, and Jamshidian and Jalal procedures (see Dixon [13], Little [14] and Jamshidian and Jalal [15]). Two sets of experiments, consisting of independent and correlated data, were designed for this comparison:

1. Experiment set 1: independent data:
   - Continuous data with a $U(0,1)$ distribution;
   - Continuous data with a $N(0,1)$ distribution;
   - Binary data with a Bernoulli distribution $B(1,0.4)$;
   - Polytomous data with a $B(1;0.1,0.3,0.6)$ distribution.
2. Experiment set 2: correlated data:
   - Continuous data with different uniform and normal distributions.

In the case of independent data, a set of 10 variables $(X_1,...,X_{10})$ with the given distribution was first randomly generated, X_1 being then modified to include a missing part, and the nine other variables being used as explanatory factors in the regression models. In the case of correlated data, after generating the initial set of 10 variables, four variables were randomly selected between X_2 and X_{10} (noted $X_{sel\_1},...,X_{sel\_4}$). Then, these four variables were transformed as follows to correlate them with the dependent variable X_1:

- $X_{sel\_1} = \zeta_{U(0,3)} X_1 + \zeta_{U(0,1)}$;
- $X_{sel\_2} = \zeta_{U(0,0.5)} X_1^2 + \zeta_{U(0,1)}$;
- $X_{sel\_3} = \frac{X_1+1}{10} + \zeta_{U(0,1)}$;
- $X_{sel\_4} = X_1^3 + \zeta_{N(0,1)}\zeta_{U(0,1)}$

where $\zeta_{U(0,1)}$ and $\zeta_{N(0,1)}$ are two random vectors generated from, respectively, $U(0,1)$ and $N(0,1)$ distributions, and $\zeta_{U(0,3)}$ and $\zeta_{U(0,0.5)}$ are two random values generated from, respectively, $U(0,3)$ and $U(0,0.5)$ distributions.

The sample size ranged from 100 to 10,000 and the percentage of missingness on X_1 varied between 1% and 50%. MCAR and MAR mechanisms were studied.

4.2. Simulated Missing Data Mechanisms

Different MD mechanisms were simulated by deleting the observations on variable X_1, following rules defined to mimic MCAR and MAR situations. Let h denote the percentage of MD generated for X_1. The algorithms used to generate the MD are described below with examples:

- MCAR: A random vector ν of size n containing uniformly distributed data between 0 and 1 is generated. Then, all data above the $(100-h)$th percentile in ν are selected and the corresponding observations in X_1 are replaced with MD.

Example. Let $h = 20$. In this case, X_1 is missing when the random vector $\nu \sim U(0,1)$ is larger than its 80th percentile.

- MAR1: In the first MAR mechanism, the MD for X_1 are caused by only one other variable. One of the variables between X_2 and X_{10}, say, X_ℓ, is randomly chosen to become the cause of the MD for X_1. Then, all data above the $(100-h)$th percentile in X_ℓ are selected, and the corresponding observations in X_1 are replaced with MD.
  Example. Let $h = 20$. In this case, X_1 is missing when X_ℓ is larger than its 20th percentile.
- MAR2: In the second MAR mechanism, the MD for X_1 are caused by two independent variables. Two variables between X_2 and X_{10}, say, X_ℓ and X_k, are randomly chosen as the cause of the MD for X_1. Then, first, select all data above the $(100-h/2)$th percentile in X_ℓ and replace the corresponding observations in X_1 with MD. Second, do the same with X_k. Since some missing data generated from X_k could have already been generated from X_ℓ, continue to generate MD from X_k by going under the $(100-h/2)$th percentile until exactly h percent of data are replaced by MD for X_1.
  Example. Let $h = 20$. In this case, X_1 is missing when X_ℓ and X_k are larger than their 90th percentile. Since some MD generated from X_k could have already been generated from X_ℓ, then the largest values from X_k are additionally used until exactly 20% of the data for X_1 are missing.
- MAR3: The third MAR-generating mechanism is quite similar to MAR2, but it uses three different variables to generate MD. The difference with MAR2 is that it uses the second and third variables to build an interaction term (simple multiplication) and generates the second part of the MD from this interaction term rather than from the original variables. The interaction term allows to make the generation of MD more complex and to have an indirect explanation of MD.
  Example. Let $h = 20$. In this case, X_1 is missing when X_ℓ and the interaction between X_k and X_j (simple multiplication) are larger than their 90th percentile. Since some MD generated from the interaction term could have already been generated from X_ℓ, then the largest values from X_kX_j are additionally used until exactly 20% of the data for X_1 is missing.
- MAR4: The last MAR mechanism is similar to MAR1, except that the MD are caused by an interaction term built from two variables randomly selected from $X_2, ..., X_{10}$ instead of from only one randomly selected variable.
  Example. Let $h = 20$. In this case, X_1 is missing when the interaction between X_ℓ and X_k is larger than its 80th percentile.

Figures 2 and 3 explain the different MD mechanisms on a reduced example with only 4 variables (instead of 10 in the real simulations) and 20 observations. Figure 2 represents a complete dataset composed of four uncorrelated variables, X_1 to $X4$, following a $U(0,1)$ distribution, and of two interaction terms, X_3X_4 and X_2X_3. These interaction terms are a simple multiplication element by element of the generating variables, which is commonly called the Hadamard product [38]. In addition, there is a a random generated vector, ν, that does not belong to the dataset, but that is used to generate MCAR data. The desired quantity of MD is 20% (4 observations). Figure 3 represents the MD mechanisms. For MCAR, the highest four values of the random vector ν are used to assign missing values to X_1. Then, the original $X_2, ..., X_4$ complete variables are used to define the MD in the different MAR mechanisms. For MAR2 and MAR3, some of the MD can be generated identically from both independent variables. In this case, one or several additional values from the second vector have to be used in order to have exactly 20% of MD for X_1. That is why three values (instead of two) from the interaction X_3X_4 are used to generate MAR3 for X_1. Appendix A presents the *R* code for (1) the general way to simulate the MCAR and MAR1 to MAR4 mechanisms, and (2) the simulated example from Figures 2 and 3. Notice that the difference between MCAR and MAR mechanisms lies in the fact that the random vector ν, which explains the presence of MCAR data, does not belong to the dataset, so it

cannot be used in the regression model of the RB test, while vectors X_1 to X_4, which explain the presence of MAR data, are in the dataset.

random vector	observed variables					
v	X_1	X_2	X_3	X_4	X_3X_4	X_2X_3
0.684	0.472	0.908	0.786	0.911	0.715	0.713
0.873	0.302	0.178	0.277	0.881	0.244	0.049
0.690	0.110	0.276	0.100	0.624	0.062	0.028
0.116	0.572	0.405	0.518	0.980	0.043	0.210
0.195	0.791	0.946	0.218	0.980	0.213	0.206
0.461	0.638	0.198	0.511	0.789	0.403	0.101
0.789	0.311	0.133	0.925	0.428	0.395	0.123
0.591	0.693	0.508	0.290	0.884	0.257	0.148
0.374	0.675	0.072	0.072	0.069	0.028	0.029
0.141	0.758	0.575	0.892	0.804	0.717	0.513
0.096	0.841	0.768	0.730	0.654	0.478	0.560
0.703	0.240	0.312	0.794	0.643	0.511	0.248
0.078	0.680	0.054	0.072	0.942	0.068	0.004
0.235	0.070	0.639	0.691	0.239	0.165	0.441
0.960	0.647	0.630	0.596	0.239	0.535	0.375
0.795	0.102	0.823	0.788	0.363	0.286	0.648
0.387	0.956	0.777	0.375	0.192	0.072	0.292
0.976	0.748	0.551	0.172	0.195	0.034	0.095
0.095	0.150	0.032	0.805	0.641	0.516	0.026
0.582	0.308	0.251	0.685	0.245	0.168	0.172

Figure 2. A dataset of uncorrelated variables following a uniform distribution (X_1, X_2, X_3, X_4), two interaction terms (X_3X_4, X_2X_3), and an independent vector v. The sample size is 20.

MCAR		MAR1		MAR2			MAR3			MAR4	
X_1	v	X_1	X_2	X_1	X_2	X_3	X_1	X_2	X_3X_4	X_1	X_2X_3
0.472	0.684	—	**0.908**	—	**0.908**	0.786	—	**0.908**	**0.715**	—	**0.713**
—	**0.873**	0.302	0.178	0.302	0.178	0.277	0.302	0.178	0.244	0.302	0.049
0.110	0.690	0.110	0.276	0.110	0.276	0.100	0.110	0.276	0.062	0.110	0.028
0.572	0.116	0.572	0.405	0.572	0.405	0.518	0.572	0.405	0.043	0.572	0.210
0.791	0.195	—	**0.946**	—	**0.946**	0.218	—	**0.946**	0.213	0.791	0.206
0.638	0.461	0.638	0.198	0.638	0.198	0.511	0.638	0.198	0.403	0.638	0.101
0.311	0.789	0.311	0.133	—	0.133	**0.925**	0.311	0.133	0.395	0.311	0.123
0.693	0.591	0.693	0.508	0.693	0.508	0.290	0.693	0.508	0.257	0.693	0.148
0.675	0.374	0.675	0.072	0.675	0.072	0.072	0.675	0.072	0.028	0.675	0.029
0.758	0.141	0.758	0.575	—	0.575	**0.892**	—	0.575	**0.717**	—	**0.513**
0.841	0.096	0.841	0.768	0.841	0.768	0.730	0.841	0.768	0.478	—	**0.560**
0.240	0.703	0.240	0.312	0.240	0.312	0.794	0.240	0.312	0.511	0.240	0.248
0.680	0.078	0.680	0.054	0.680	0.054	0.072	0.680	0.054	0.068	0.680	0.004
0.070	0.235	0.070	0.639	0.070	0.639	0.691	0.070	0.639	0.165	0.070	0.441
—	**0.960**	0.647	0.630	0.647	0.630	0.596	—	0.630	**0.535**	0.647	0.375
—	**0.795**	—	**0.823**	0.102	0.823	0.788	0.102	0.823	0.286	—	**0.648**
0.956	0.387	—	**0.777**	0.956	0.777	0.375	0.956	0.777	0.072	0.956	0.292
—	**0.976**	0.748	0.551	0.748	0.551	0.172	0.748	0.551	0.034	0.748	0.095
0.150	0.095	0.150	0.032	0.150	0.032	0.805	0.150	0.032	0.516	0.150	0.026
0.308	0.582	0.308	0.251	0.308	0.251	0.685	0.308	0.251	0.168	0.308	0.172

Figure 3. An example of the five missing data generating mechanisms with 20% of missing data on X_1 in each case. Red font: top 20% values.

4.3. Simulation Procedure

The main purpose of these simulations is to compare the behaviour of the RB approach with the Dixon, Jamshidian and Jalal, and Little tests. The four approaches were applied to

continuous data, and the RB and Little approaches were also applied to categorical data for different MD mechanisms, sample sizes, and percentages of generated MD. For MAR mechanisms, all tests are performed using the nine variables $X_2, \ldots, X_{10}$ as explanatory variables in the regression model. Moreover, because the MAR4 mechanism uses only an interaction term, the situation with this interaction variable added to the dataset is also tested (*MAR4i*). Each situation is replicated 1000 times and the percentage of acceptance of the null hypothesis (MCAR) by each approach is analysed.

Because the data for this simulation study were randomly generated, there is no significant difference between the variables $X_2, \ldots, X_{10}$ in part A and in part B of each dataset (see Figure 1). If there was a significant difference between them, the concerned variable would have been a natural candidate for the cause of the MD for X_1. This trivial case being excluded, the reason for rejecting the null hypothesis of the test can actually be found in the MAR mechanisms described above.

The *RBtest* [39] package of the open-source statistical software R [40] was used for all computations. The Type I error was set to 5%.

4.4. Experiment Set 1: Independent Data

4.4.1. Uniform Distribution

For continuous dependent variables, two types of dataset were simulated, the only difference being the data distribution:

- $U(0,1)$ with a sample size of 1000;
- $N(0,1)$ with a sample size of 1000.

Table 2 summarises the simulation results for MCAR and MAR generating mechanisms when the sample size is 1000 and the data follow a $U(0,1)$ distribution. For this setting, all tests accept the MCAR hypothesis when the MD are really MCAR with a probability close to the degree of confidence of the test (95%). The conclusions diverge between tests for MAR mechanisms. Both the Dixon and Little tests reject MCAR in the majority of cases, while the Jamshidian and Jalal procedure tends to accept MCAR in the presence of a large quantity of MD. In addition, Jamshidian and Jalal accepts MCAR when MAR1, MAR4, and MAR4i are used and for a relatively low quantity of MD as well. This unexpected behaviour is a sign that the Jamshidian and Jalal is not appropriate for all types of MD. There is no definitive explanation for this unexpected behaviour, although the following elements should be considered:

1. The Jamshidian and Jalal procedure uses a combination of two tests: the modified Hawkins test and a non-parametric distribution test. However, there is no adjustment for the total error, which is problematic for simultaneous tests [27];
2. Its procedure is such that when data follow a multivariate normal distribution, the test rejects the null hypothesis more often. Thus, it is impossible to compare the application of the test on different types of data;
3. The construction of the Hawkins test is such that whatever the distribution, when the sample size is relatively small the test fails to reject the null hypothesis (lack of statistical power). This means that when there is a relatively small quantity of MD, it is too easy for the modified Hawkins test to accept the null hypothesis;
4. It is known that non-parametric tests are generally less powerful than parametric ones [41], and the Jamshidian and Jalal procedure uses a non-parametric test to check for missingness.

The same behaviour was also observed for other sets of simulated data. All these considerations lead us to consider that the Jamshidian and Jalal test sometimes has a poor reliability.

Finally, in presence of MAR data the probability of the RB approach to correctly identify this MD mechanism is positively correlated with the percentage of MD. The higher the percentage of MD, the higher the probability of rejecting the MCAR case. This behaviour

is not surprising, since the more information (here: missing data), the easier it should be to identify the correct type of MD.

Table 2. MCAR and MAR mechanisms, $U(0,1)$, $n = 1000$. The percentage of acceptance of the MCAR hypothesis is provided for Jamshidian and Jalal (JJ), Dixon (D), Little (L), and regression-based (RB) approaches.

% of MD	MCAR				MAR1				MAR2			
	JJ	D	L	RB	JJ	D	L	RB	JJ	D	L	RB
50%	95.4	95.0	95.6	95.3	90.6	0	0	8.5	0	0	0	11.8
45%	94.7	94.5	93.2	94.6	43.8	0	0	9.7	1.3	0	0	12.3
40%	95.4	95.6	96.0	94.7	4.2	0	0	10.8	4.8	0	0	16.6
35%	94.9	94.9	94.5	96.0	0.1	0	0	9.8	12.0	0	0	18.5
30%	95.5	95.2	94.8	96.3	0	0	0	15.5	20.3	0	0	21.0
25%	95.0	95.9	95.3	96.4	0	0	0	17.9	32.0	0	0	20.8
20%	94.4	94.5	94.5	94.4	0	0	0	19.5	44.3	0	0	26.2
15%	92.8	95.1	94.4	96.5	0.6	0	0	22.6	44.3	0	0	26.2
10%	95.1	95.1	95.3	94.6	4.0	0	0	25.7	63.2	0	0	34.4
5%	93.1	95.3	95.4	95.6	28.8	0	0	41.3	75.8	0	0	50.8
4%	94.5	94.9	94.8	95.4	38.2	0	0	40.9	78.5	0	0	58.0
3%	95.3	94.4	95.6	96.1	51.6	0	0	46.7	81.5	0.6	0	63.3
2%	93.1	94.8	95.6	95.2	61.8	0	0	57.9	82.7	11.8	0.3	72.5
1%	92.2	90.9	95.6	95.7	75.1	0	0	70.1	81.1	62.5	19.0	82.8
% of MD	**MAR3**				**MAR4**				**MAR4i**			
	JJ	**D**	**L**	**RB**	**JJ**	**D**	**L**	**RB**	**JJ**	**D**	**L**	**RB**
50%	0	0	0	14.2	9.6	0	0	12.2	95.1	0	0	15.0
45%	0	0	0	14.2	9.0	0	0	12.0	79.5	0	0	11.9
40%	0.1	0	0	14.2	0	0	0	11.9	36.5	0	0	12.7
35%	0.5	0	0	20.4	0	0	0	13.1	8.6	0	0	12.9
30%	2.7	0	0	20.9	0	0	0	14.2	1.7	0	0	11.1
25%	7.0	0	0	20.8	0	0	0	15.2	0.1	0	0	10.4
20%	16.1	0	0	27.0	0	0	0	16.9	0	0	0	12.3
15%	31.0	0	0	29.5	0	0	0	19.3	0	0	0	13.6
10%	46.6	0	0	36.0	0	0	0	23.9	0	0	0	14.2
5%	64.5	0	0	54.5	0.2	0	0	32.2	0	0	0	20.5
4%	67.6	0.2	0	55.8	0.4	0	0	32.7	0	0	0	19.0
3%	72.0	0.8	0	61.9	3.6	0	0	41.7	0.4	0	0	27.6
2%	80.3	14.3	0.5	72.3	12.0	0	0	45.6	1.7	0	0	32.5
1%	78.3	61.8	13.5	79.5	41.9	0	0	56.1	10.6	0	0	37.4

4.4.2. Normal Distribution

Table 3 summarises the simulation results for MCAR and MAR generating mechanisms when the sample size is 1000 and the data follow a $N(0,1)$ distribution. In this setting, all approaches accept MCAR when the MD are truly MCAR. In the case of MAR mechanisms, both the Dixon and Little tests reject MCAR when MD are MAR1, MAR2, MAR3, and MAR4i. On the other hand, the MAR4 has a specific behaviour. This special behaviour is due to the absence of the interaction variable when computing these tests (note that when the interaction is added (*MAR4i.*) this problem disappears). This difference for the MAR4 case is specific to $N(0,1)$ data as compared to $U(0,1)$ data. To understand this difference between distributions, scatter plots, and Pearson correlation tests were performed between X_1, the complete variables that form the interaction and the interaction itself. For normal data, Figure 4 shows that there are no significant correlations between the variables that form the interaction (X_2 and X_3) and the interaction (X_2X_3). This means that when the interaction is not added to the model (MAR4 case), the presence of X_2 and X_3 in the model may not impact on the presence of MD for X_1 (even if they are explained by the interaction). Consequently, there is no other choice but to accept the MCAR hypothesis. The RB approach rejects the MCAR hypothesis less often when the number of MD is relatively

low, except for MAR4. The results of the Jamshidian and Jalal procedure show, once again, unexpected behaviour, as explained in the previous section.

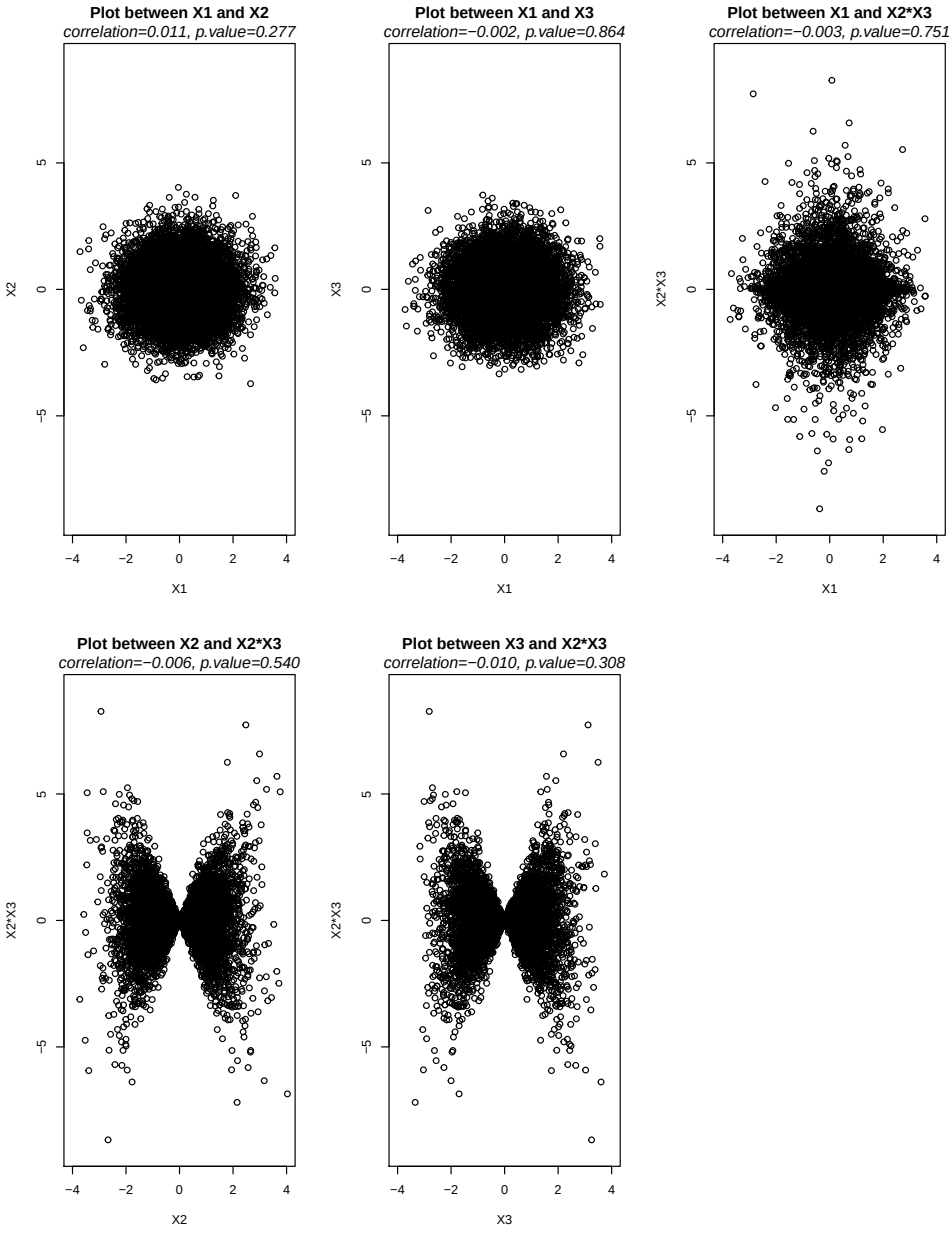

Figure 4. $N(0,1)$ distribution for MAR4 missing data mechanism.

Table 3. MCAR and MAR mechanisms, $N(0,1)$, $n = 1000$. The percentage of acceptance of the MCAR hypothesis is provided for the Jamshidian and Jalal (JJ), Dixon (D), Little (L), and regression-based (RB) approaches.

% of MD	MCAR				MAR1				MAR2			
	JJ	D	L	RB	JJ	D	L	RB	JJ	D	L	RB
50%	95.6	94.4	94.9	94.0	94.7	0	0	12.6	4.7	0	0	11.8
45%	96.1	95.0	95.0	95.3	90.3	0	0	12.6	6.8	0	0	14.0
40%	95.6	94.9	95.4	96.2	81.5	0	0	12.3	8.0	0	0	12.5
35%	94.0	94.4	96.1	94.6	71.8	0	0	13.2	7.8	0	0	13.6
30%	95.0	95.0	94.7	95.1	60.9	0	0	14.6	9.0	0	0	15.8
25%	96.3	93.9	93.9	95.8	48.5	0	0	16.7	10.3	0	0	15.3
20%	94.5	95.0	95.1	96.2	45.9	0	0	17.9	9.4	0	0	19.2
15%	95.2	94.6	95.5	94.4	48.4	0	0	18.9	11.3	0	0	20.7
10%	95.1	95.8	95.2	95.5	54.6	0	0	22.5	11.0	0	0	25.4
5%	94.4	95.3	95.2	94.9	70.2	0	0	29.7	15.8	0	0	34.5
4%	92.8	96.2	95.8	96.5	77.1	0	0	32.1	16.8	0	0	39.0
3%	94.7	95.7	96.1	96.4	79.9	0	0	36.8	19.6	0	0	43.6
2%	94.3	95.3	95.3	97.2	85.4	0	0	43.6	27.1	2.1	0	52.0
1%	94.3	94.2	95.7	95.4	87.5	0	0	51.2	37.3	64.1	4.6	62.9
% of MD	**MAR3**				**MAR4**				**MAR4i**			
	JJ	**D**	**L**	**RB**	**JJ**	**D**	**L**	**RB**	**JJ**	**D**	**L**	**RB**
50%	0.2	0	0	16.6	95.2	93.8	94.7	64.8	95.1	0	0	15.0
45%	0	0	0	18.1	82.4	95.3	96.3	61.7	80.6	0	0	14.5
40%	0	0	0	19.7	42.7	94.5	94.7	61.4	36.7	0	0	13.6
35%	0	0	0	19.0	12.5	95.4	94.4	59.2	9.7	0	0	12.6
30%	0	0	0	18.8	1.8	95.5	93.5	62.2	0.8	0	0	11.9
25%	0	0	0	18.6	0.3	95.6	91.1	61.6	0	0	0	10.6
20%	0	0	0	24.6	0.1	94.4	88.7	61.1	0	0	0	11.4
15%	0	0	0	27.9	0	95.3	85.4	62.2	0	0	0	13.6
10%	0	0	0	33.9	0	95.0	85.4	69.2	0	0	0	15.1
5%	0	0	0	43.8	0	95.2	79.5	71.3	0	0	0	19.9
4%	0	0	0	46.0	0	94.8	78.6	74.8	0	0	0	19.6
3%	0.3	0.8	0	53.6	0.1	94.3	75.2	78.6	0.1	0	0	24.2
2%	1.4	14.3	0	63.0	0.7	95.3	71.6	77.1	1.8	0	0	28.3
1%	9.5	79.3	2.4	70.9	7.8	95.3	66.0	83.8	15.4	0	0	39.5

4.4.3. Comparison of the Continuous Results

In this subsection, we summarise and graphically compare the simulation results on all continuous data (Figure 5). The left side of the figure represents the results for data following a $U(0,1)$ distribution, while the right side represents data following a $N(0,1)$ distribution. Only the MCAR and MAR1 generated MD mechanisms are presented here.

The MCAR hypothesis is widely accepted by all approaches when data are truly MCAR. For the MAR1 case, the main comparison results are that (1) the Little and Dixon tests always reject the MCAR hypothesis and (2) the Jamshidian and Jalal and RB procedures accept the MCAR hypothesis more often when data follow the $N(0,1)$ distribution.

When there is a small number of MD, it should be more difficult to find a powerful argument to reject the null hypothesis. Thus, the results are not expected to be correct in 100% of cases. Figure 5 shows that the decrease in the percentage of MD has no impact on both Little and Dixon tests. This result is also explained by the strong MAR1 assumption, where MD are completely explained by one variable. The RB approach gives more weight to MCAR data and balances the results in function of the percentage of MD.

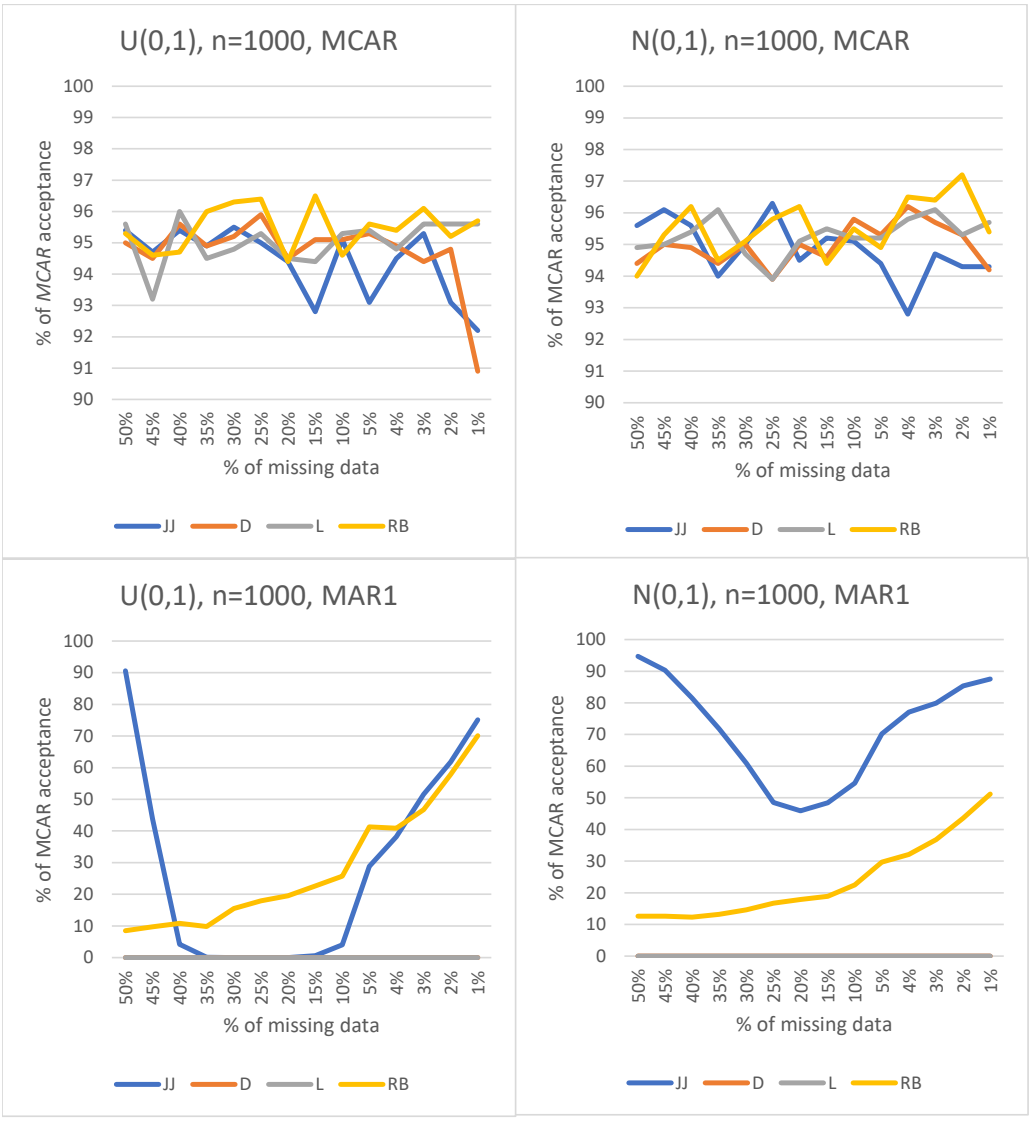

Figure 5. MCAR and MAR1 mechanisms, $U(0,1)$ and $N(0,1)$, $n = 1000$. The percentage of acceptance of the MCAR hypothesis is provided for the Dixon (D), Jamshidian and Jalal (JJ), Little (L), and regression-based (RB) approaches. For MAR1, the Dixon and Little tests achieve exactly the same results.

4.4.4. Binary Distribution

For data with a dependant binary variable, X_1 has two modalities, $[a]$ with probability $p = 0.3$ and $[b]$ with probability $1 - p = 0.7$. The sample size takes different values in the following set: $n = [100; 250; 500; 1000; 2000; 10,000]$. From 1% to 50% of MD are generated for the dependant incomplete variable X_1 either purely by random (MCAR case) or by using the MAR1 mechanism.

Notice that even if it is used for comparison purpose, the Little test is not really able to test missingness when data are categorical; therefore, categorical data are first transformed into numerical values. The transformed data could be tested by the Jamshidian and Jalal and the Dixon tests as well, but since these approaches were not developed for categorical data, there is no sense in applying them to the categorical case. This argument is also valid for the Little test but, because it is nevertheless commonly used for categorical data in the literature, it was important to compare it with the RB approach.

The results are summarised as percentage of cases in which the null hypothesis, that is the MCAR hypothesis, is accepted. Results are given in Table A1 and Figure 6 (Appendix B). The figure contains five graphics:

- (a) Represents the results of the RB approach for MAR1 data. It shows how this approach behaves as a function of the quantity of MD and the sample size;
- (b) Represents the difference between the RB approach and the Little test when the sample size is equal to 100;
- (c)–(e) are the same as (b), except that the sample size increases respectively to 250, 500, and 1000.

Figure 6a shows that whatever the sample size, the RB approach rejects the MCAR hypothesis more often when the quantity of MD increases. Moreover, the results of the RB approach evolve with the sample size. Moreover, the larger the sample size, the easier it is to reject the MCAR hypothesis. Figure 6b–e show that the Little test and the RB approach tend to be closer in terms of rejection of the MCAR hypothesis (while the data are truly MAR) with an increase in the sample size. Furthermore, the Little test always rejects the MCAR mechanism when the sample size is at least 1000, even when the sample includes a very small percentage of MAR data. However, it may not always be correct to reject MCAR when the percentage of MD is very small, because the information set on which to base the decision is then very small. The relatively strong MAR1 assumption, where MD are completely explained by one variable, could partially explain this result of the Little test.

4.4.5. Multinomial Distribution

For the multinomial case, we consider three modalities $\{a; b; c\}$ for variable X_1 with the following probability of occurrence:

$$X_1 = \begin{cases} a \text{ with } p = 0.1 \\ b \text{ with } p = 0.3 \\ c \text{ with } p = 0.6 \end{cases}$$

For this set of simulations, the sample size is constant and equal to 10,000. Five MD mechanisms are used: MCAR, MAR1, MAR2, MAR3, MAR4, and MAR4i (same algorithms as before).

Table A2 (Appendix B) summarises the results for the Little and RB tests by giving the percentage of acceptance of the *MCAR* hypothesis. Both tests lead to similar conclusions. They accept the MCAR hypothesis when the data are MCAR in 95% of cases on average and when the MD are MAR, they reject MCAR in the large majority of cases. However, when the total amount of MD is less than 3%, the probability of acceptance of the MCAR hypothesis in the RB approach is larger than 5% when the MD are MAR.

4.5. Experiment Set 2: Correlated Data

Real data are often correlated. It is then important to verify how the tests of MD mechanisms are influenced by the level of inter-variable correlation. Therefore, we simulated one set of correlated continuous data, as explained in Section 4.1.

We applied the RB, Little, and Dixon approaches to verify the impact of the quality of the linear regression models on the results. The regression models for the dependent variable X_1 used the nine other variables as explanatory variables, including variables $X_{sel\_1}, \ldots, X_{sel\_4}$. Tables A3–A5 present the results for, respectively, the Dixon, Little, and RB approaches with different ranges of R^2 (Appendix C). The R square measure is very helpful to evaluate the quality of a regression model. We used it because we wanted to explore the relationship between the quality of the regression model and the accuracy of the RB test. Table A6 (Appendix C) shows the sample size for each range of the R^2, given that there are 1000 replications for each percentage of MD and, thus, there are 14,000 datasets. The results show that whatever the level of the R^2, the three approaches work well for MCAR data because they accept the MCAR hypothesis in 95% of the cases in average. However, for MAR data, the RB approach tends to accept the MCAR hypothesis much more often than when data are independent (whatever the R^2), while the Dixon and Little tests always accept the MAR hypothesis. As already noted in the case of independent data,

the behaviour of these latter two tests is somewhat strange, since it implies a near-infinite power of the tests, even in presence of a very low amount of information. In contrast, even if the RB test proves sometimes wrong in detecting correctly MAR data, at least its behaviour is related to the R^2 of the model, the probability of obtaining the true result being inversely related to the degree of dependence between data.

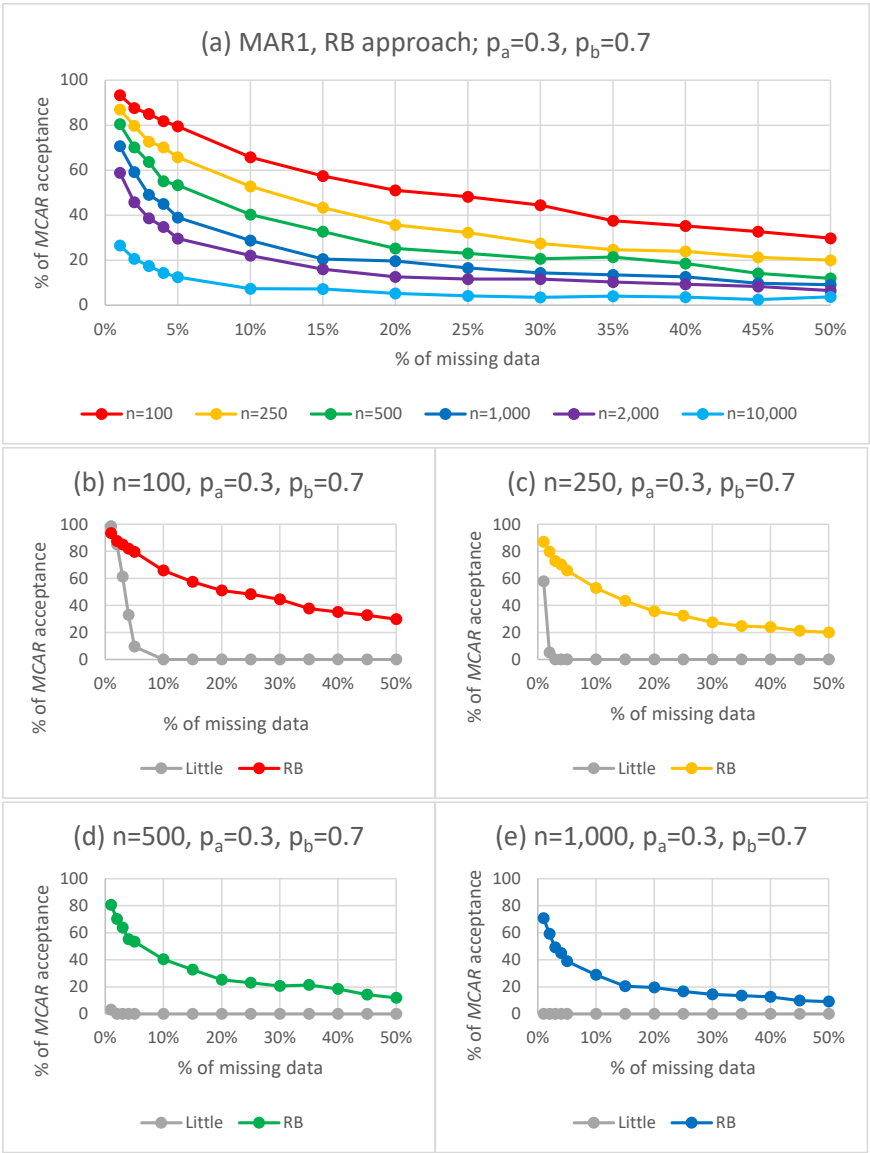

Figure 6. MAR1 mechanism, $B(1, 0.3)$, different sample sizes. The percentage of acceptance of the MCAR hypothesis is provided for the Little and RB approaches.

4.6. Discussion

For continuous data the results are relatively similar whether the data follow a $N(0, 1)$ or a $U(0, 1)$ distribution, except for one of the MAR mechanisms (MAR4, $N(0, 1)$). The percentage of non-observed data is an important factor: the RB and Jamshidian and Jalal approaches tend to accept the MCAR hypothesis more often as compared to the Little and Dixon procedures when there is a relatively large quantity of MAR data.

The RB approach is less conservative than the Little and Dixon tests, while the Jamshidian and Jalal approach seems to be more similar to the RB procedure. However, as explained in Section 4.4.1, the Jamshidian and Jalal approach has an unexpected behaviour, mainly due to its principle based on a combination of two tests. For very small percentages of MD, the MCAR hypothesis is more often accepted by the RB approach.

For the categorical and independent data experiment, only the RB and Little approaches were used, because the other methods are not appropriate for this data type. As shown by the simulations, the Little and RB approaches tend to behave similarly, except that the Little test is less sensible to a variation in the percentage of MD than the RB approach. In the case of a binary incomplete variable and MAR data, neither approach is recommended when $n = 100$; in a conservative approach, a minimum sample size of $n = 1000$ seems necessary. In the case of a multinomial incomplete variable with three categories and MAR data, there are no important differences between the Little and RB approaches, whatever the MAR mechanism. The RB approach works well even if the number of modalities increases, however the size of each subsample of a category should be at least 100 to achieve plausible results.

When data are correlated, the results for MCAR data are generally very similar to the ones obtained on the independent data: the MCAR hypothesis is accepted in 95% of the situations on average. In the case of $U(0,1)$ data and MAR mechanisms, the different ranges of the R^2 show that the RB approach tends to give too much importance to MCAR data when data are truly MAR (as compared to the Dixon and Little tests). Meanwhile, the Little test always rejects correctly the MCAR hypothesis for both independent and correlated data.

5. Application on Real Data

We apply the four approaches—RB (use of the two step procedure explained in Section 3.5), Dixon, Little, and Jamshidian–Jalal—to test MD mechanisms on a set of variables (completely and partially observed) of the first wave of the Professional Path Survey [42]. The main purpose of this survey was to understand the professional transitions, career pathways, personal experiences, and well-being of employed and unemployed middle-aged adults (25–55 years) living in the French and German speaking regions of Switzerland. The sample size is 1902 (excluding the unit non-responses). The selected variables are listed in Table 4.

For the RB approach, the maximum number of iterations is set to 50. Moreover, given that there is a certain amount of uncertainty in the iterative step of the RB approach (mainly due to the high amount of variables and the multiple imputation by the chained equation method), we replicate the RB procedure 100 times (number of simulations). This gives more credibility to the RB approach when compared to the other tests.

The Dixon test is applied $(K-1)K$ times, K being the number of variables with at least one missing value. Thus, only variables with at least one missing value are tested by the Dixon approach. Two special cases require discussion. First, the test cannot be computed when all MD on one variable are also missing on the second variable. In such a case, we suppose that MD are MAR for not taking the risk of a bad replacement. Second, when both variables have a certain amount of MD in common, in addition to MD that are differently distributed over the two variables, the t-test is applied. This is problematic when the remaining quantity of observed data is relatively low, because it is known that the power of the t-test decreases with the sample size [43].

Table 4 summarizes the results. The variables are listed in the first column. The second column gives the absolute frequency of MD per variable. The third column represents the number of times the MCAR hypothesis is accepted by the RB approach over the 100 simulations. The results of the Dixon procedure are in the fourth column, where the first value is the number of testing variables leading to the acceptance of the MCAR hypothesis and the second value is the number of testing variables leading to the rejection of the MCAR hypothesis. The fifth and sixth columns provide the p-values of the Jamshidian and Jalal and Little procedures, which test missingness using the overall information, and, thus, only one number per test is provided. Note that for the Jamshidian and Jalal procedure, the p-value shown in Table 4 corresponds to the last step of the procedure where the homogeneity of covariances is tested. Of course, in this case, the null hypothesis in the first step of the procedure was rejected.

Table 4. Results of the RB, Dixon, Jamshidian and Jalal, and Little tests applied on the first wave of the Professional Path Survey data. The first column lists the variables. The second column gives the number of missing data per variable. The third column provides the number of times the MCAR hypothesis is accepted by the RB approach over the 100 simulations. The fourth column represents the number of times the missing data are considered as MCAR or MAR for the variable of interest by the Dixon test. The fifth and sixth columns represent the *p*-values of the Jamshidian and Jalal and Little procedures (overall tests).

Variables	Number of MD	RB	D	JJ	L
Children	103	0	6; 6		
Income	62	100	8; 4		
General health	20	82	8; 4		
Education	13	98	8; 4		
Household	9	58	0; 12		
Self-rated health	9	99	6; 6		
Marital status	7	90	5; 7		
Agreeableness	4	99	0; 12		
Conscientiousness	4	100	0; 12		
Extraversion	4	100	0; 12		
Neuroticism	4	100	0; 12		
Number of jobs	4	100	8; 4		
Openness	4	100	0; 12		
Age	0	-	-		
Benefits	0	-	-		
Gender	0	-	-		
Nationality	0	-	-		
Work rate	0	-	-		
Overall				0.086	0

When the RB approach is applied, there is no ambiguity for self-rated health, number of jobs, income, neuroticism, extraversion, openness, agreeableness, conscientiousness, and education; the MCAR hypothesis is accepted in at least 95% of cases. The conclusion is the opposite for Children, because the MCAR hypothesis is always rejected, and, thus, the MD are supposed to be MAR. Marital status is supposed to be MCAR, because the MCAR hypothesis is accepted in 90% of cases. Results are more contrasted for the two remaining variables; the MCAR hypothesis is rejected in 42% of cases for Household and in 18% of cases for General health. A first naive explanation could be that only the number of MD affects the result, that is there is a higher probability of accepting the MCAR hypothesis when this number is relatively low. For instance, Education has a relatively low number of MD (13) and the RB approach accepts the MCAR hypothesis in 98% of cases. However, this hypothesis does not hold for Household, which contains only 9 missing values and for which the RB approach approximately equally accepts and rejects the MCAR hypothesis over the 100 simulations. Both variables (Household and Education) have a similar and relatively low quantity of MD; however, the difference in the distributions between the acceptance and rejection of the MCAR hypothesis is important.

For the Dixon approach, the results are more contrasted with respect to the RB procedure. For instance the Dixon test always rejects the MCAR hypothesis for the four variables related to personality, while the RB approach accepts MCAR in the large majority of cases. The Little test suggests that these MD are globally not MCAR, but without doing any distinction between variables, while the Jamshidian and Jalal globally accepts the MCAR hypothesis. Such results provide a contrasting conclusion about the type of MD.

6. Conclusions

There are few methods in the literature to test missingness. The most commonly used ones, the Dixon [13], Little [14] and Jamshidian and Jalal [15] methods, are all limited in different ways. This article has described an innovative approach to test MCAR versus MAR data. This method is adapted both to continuous and categorical data. Demonstrating that MD are not MCAR is important because in that case it is generally recommended to use multiple imputation to handle MD, while it is very common in practice to simply delete non-observed data. Moreover, knowing the MD mechanism can help to detect some response patterns and to better understand the data, as well as the psychology of respondents.

Our results show that no test can be universally applied to correctly detect the correct MD mechanism. All tests have difficulties when data are correlated. A limitation of the our approach, but similar to other tests, is that it is not designed to test the *MNAR* hypothesis. However, in the longitudinal context, the RB approach could be extended to detect non-ignorable MD by combining it with the drawn indicator family of methods [44]. Furthermore, simulations of MAR mechanisms with a weaker link between MD and there explanation would be necessary.

The method developed in this article is not meant to replace existing testing procedures, but to offer an additional tool for scientists and data analysts to check whether their MD are completely randomly distributed over their dataset or not, and to offer them an additional point of view on missingness.

Author Contributions: Conceptualisation, S.R. and A.B.; software development, S.R.; formal analysis, S.R.; interpretation of results, S.R. and A.B.; writing—original draft preparation, S.R.; writing—review and editing, S.R. and A.B.; supervision, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This publication benefited from the support of the Swiss National Centre of Competence in Research LIVES—Overcoming vulnerability: Life course perspectives, which is financed by the Swiss National Science Foundation(grant number: 51NF40-160590). The authors are grateful to the Swiss National Science Foundation for its financial assistance.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data from the simulation study can be generated using the R code provided in the appendix. Data from the Professional Path Survey can be downloaded from here: https://www.swissubase.ch/fr/catalogue/studies/12734/17161/overview, accessed on 30 November 2021.

Acknowledgments: Most of this work was completed while Serguei Rouzinov was Ph.D. student within the NCCR LIVES.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. R Code for Generated Missing Data Mechanisms

```
m<-10  # number of variables.
n<-10000   # sample size.
mis<-20  # percentage of missing data on the dependent variable

set.seed(40)

############## MCAR and MAR mechanisms ##############

#----------------#
#----- MCAR -----#
#----------------#

X<-matrix(NA,ncol=m,nrow=n)
for (i in 1:m){
X[,i]<-runif(n,0,1)
}
X<-data.frame(X)

X_MCAR<-X
aa<-runif(n,0,1)
X_MCAR[which(aa<=sort(aa)[mis*(n/100)]),1]=NA
```

```
#----------------#
#----- MAR1 -----#
#----------------#

X<-matrix(NA,ncol=m,nrow=n)
for (i in 1:m){
X[,i]<-runif(n,0,1)
}
X<-data.frame(X)

X_MAR1<-X
selection<-sample(c(2:ncol(X_MAR1)),1)
x_s<-X_MAR1[,selection] # one sampled independent variable.

q_MAR_1<-quantile(x_s,seq(0,1,0.01))[100-mis+1]
X_MAR1[which(x_s>q_MAR_1),1]<-NA


#----------------#
#----- MAR2 -----#
#----------------#

X<-matrix(NA,ncol=m,nrow=n)
for (i in 1:m){
X[,i]<-runif(n,0,1)
}
X<-data.frame(X)

X_MAR2<-X

selection<-sample(c(2:ncol(X_MAR2)),2)
x_s<-X_MAR2[,selection] # two sampled independent variables.

q_MAR_1<-quantile(x_s[,1],seq(0,1,0.005))[mis+1]

X_MAR2[which(x_s[,1]<q_MAR_1),1]=NA
X_MAR2[which(is.na(X_MAR2[,1])==F)
[which(x_s[which(is.na(X_MAR2[,1])==F),2]<
quantile(x_s[which(is.na(X_MAR2[,1])==F),2],
(mis/2)/(100-(mis/2))))],1]=NA

#----------------#
#----- MAR3 -----#
#----------------#

X<-matrix(NA,ncol=m,nrow=n)
for (i in 1:m){
X[,i]<-runif(n,0,1)
}
X<-data.frame(X)

X_MAR3<-X
selection<-sample(c(2:ncol(X_MAR3)),3)
x_s<-x_s_1<-X_MAR3[,selection] # three sampled independent variables.
x_s<-cbind.data.frame(x_s[,1],x_s[,2]*x_s[,3])

q_MAR_1<-quantile(x_s[,1],seq(0,1,0.005))[mis+1]

X_MAR3[which(x_s[,1]<q_MAR_1),1]=NA
X_MAR3[which(is.na(X_MAR3[,1])==F)
[which(x_s[which(is.na(X_MAR3[,1])==F),2]<
quantile(x_s[which(is.na(X_MAR3[,1])==F),2],
(mis/2)/(100-(mis/2))))],1]=NA

#----------------#
#----- MAR4 -----#
#----------------#

X<-matrix(NA,ncol=m,nrow=n)
for (i in 1:m){
X[,i]<-runif(n,0,1)
}
X<-data.frame(X)

X_MAR4<-X
x_s<-X_MAR4[,sample(c(2:ncol(X_MAR4)),2)] # two sampled independent variables.
x_s<-x_s[,1]*x_s[,2]

q_MAR_1<-quantile(x_s,seq(0,1,0.01))[100-mis+1]
X_MAR4[which(x_s>q_MAR_1),1]=NA


#########################################################
#-------------------------------------------------------#
#----- Simulated example of missing data mechanisms -----#
#-------------------------------------------------------#
```

```
############################################################


set.seed(40)
# 5 variables with a uniform distribution and a sample size of 20
D<-matrix(NA,ncol=5,nrow=20)
for (i in 1:5){
D[,i]<-runif(20,0,1)
}

colnames(D)=c("v","X1","X2","X3","X4") # names of columns
D<-data.frame(D)

D$X3X4<-D$X3*D$X4 #interaction term between X3 and X4
D$X2X3<-D$X2*D$X3 #interaction term between X2 and X3

D<-round(D,digit=3) # 3 digits

# Check for the highest 20% of values for each variable

D_s<-D # sorted data.frame
for (i in 1:ncol(D)){
D_s[,i]<-sort(D[,i])
}
D_s[17:20,c(1,3:7)] # the 4 highest values of each
# variable and interactions,
# except X1
```

Appendix B. Simulation Results of Experiment Set 1

Table A1. MAR1 and MCAR mechanisms, binary data with $p = 0.3$. The percentage of acceptance of the MCAR hypothesis is provided for the Little (L) and regression-based (RB) approaches.

% of MD	n = 100				n = 250				n = 500				n = 1000				n = 2000				n = 10,000			
	MCAR		MAR1		MCAR		MAR1		MCAR		MAR1		MCAR		MAR1		MCAR		MAR1		MCAR		MAR1	
	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB
50%	96.2	93.9	0	29.8	94.7	94.5	0	20.0	93.9	95.3	0	11.9	95.6	95.9	0	9.1	95.6	94.7	0	6.5	95.2	95.7	0	3.7
45%	96.3	93.4	0	32.7	95.4	94.4	0	21.3	94.0	94.6	0	14.2	96.0	96.3	0	9.8	95.4	95.2	0	8.3	96.1	94.2	0	2.5
40%	95.9	94.6	0	35.2	93.8	93.7	0	23.9	94.6	95.3	0	18.5	96.1	95.2	0	12.6	95.6	94.8	0	9.3	95.8	95.4	0	3.6
35%	96.0	93.5	0	37.6	95.1	95.4	0	24.7	93.7	95.8	0	21.4	95.6	95.4	0	13.5	95.5	94.5	0	10.3	96.1	94.8	0	4.0
30%	96.3	95.8	0	44.5	95.0	96.4	0	27.4	94.6	96.6	0	20.6	96.1	95.6	0	14.4	94.5	94.6	0	11.6	95.3	95.1	0	3.5
25%	95.5	96.4	0	48.2	95.6	96.1	0	32.3	94.5	95.3	0	23.0	96.9	96.5	0	16.6	94.4	94.3	0	11.6	96.5	94.7	0	4.2
20%	95.0	95.7	0	51.1	95.4	95.4	0	35.7	93.4	94.8	0	25.2	95.2	94.8	0	19.6	94.1	95.6	0	12.6	95.4	94.7	0	5.3
15%	95.1	95.3	0	57.4	95.2	94.6	0	43.4	94.8	96.1	0	32.7	95.9	96.3	0	20.5	95.6	94.1	0	16.0	95.6	95.1	0	7.2
10%	96.2	94.6	0	65.8	95.7	95.4	0	52.8	94.9	95.6	0	40.3	96.2	95.8	0	28.8	96.3	95.3	0	22.1	95.4	95.3	0	7.3
5%	96.5	94.5	9.6	79.5	95.5	95.0	0	65.8	94.4	95.5	0	53.4	96.2	96.3	0	39.0	95.5	95.6	0	29.6	95.1	95.6	0	12.5
4%	96.5	94.7	32.9	81.8	95.5	95.4	0	70.2	95.3	94.4	0	55.1	95.3	97.1	0	45.0	95.4	95.1	0	34.8	95.6	95.2	0	14.4
3%	97.3	94.2	61.3	85.0	95.6	95.3	0	72.7	95.2	95.6	0	63.7	96.1	96.3	0	49.1	95.0	94.7	0	38.6	96.3	93.4	0	17.4
2%	97.4	93.3	85.3	87.6	95.4	95.0	5.1	79.7	96.1	94.6	0	70.2	96.3	96.0	0	59.2	95.3	94.5	0	45.8	95.7	95.6	0	20.6
1%	99.2	93.9	98.5	93.4	97.7	95.0	57.8	87.0	96.6	95.3	3.2	80.5	97.2	95.6	0	70.7	94.0	96.8	0	58.9	95.1	95.8	0	26.6

Table A2. MCAR and MAR mechanisms, $B(1, 0.1, 0.3, 0.6)$, $n = 10,000$. The percentage of acceptance of the MCAR hypothesis is provided for the Little (L) and regression-based (RB) approaches.

	n = 10,000											
% of MD	MCAR		MAR1		MAR2		MAR3		MAR4		MAR4i	
	L	RB	L	RB	L	RB	L	RB	L	RB	L	RB
50%	95.2	94.5	0	0	0	0	0	0	0	0.1	0	0
45%	96.1	95.0	0	0	0	0	0	0.1	0	0.2	0	0
40%	95.8	94.1	0	0.3	0	0.1	0	0	0	0.2	0	0.1
35%	96.1	95.7	0	0.5	0	0	0	0.3	0	0.2	0	0
30%	95.3	96.2	0	0.3	0	0.4	0	0.1	0	0.1	0	0
25%	96.5	95.3	0	0.2	0	0.2	0	0.1	0	0.3	0	0
20%	95.4	96.1	0	0.4	0	0.4	0	0.3	0	0.1	0	0
15%	95.6	95.3	0	0.2	0	0.5	0	0.6	0	0.5	0	0
10%	95.4	95.9	0	0.8	0	1.3	0	1.0	0	0.3	0	0
5%	95.1	96.1	0	2.3	0	2.8	0	3.0	0	0.5	0	0.2
4%	95.6	96.2	0	1.8	0	3.4	0	5.0	0	0.9	0	0.6
3%	96.3	96.0	0	2.8	0	4.3	0	4.7	0	2.0	0	0.2
2%	95.7	96.5	0	4.6	0	7.0	0	8.8	0	2.5	0	0.9
1%	95.1	95.8	0	9.2	0	15.1	0	17.5	0	3.9	0	2.1

Appendix C. Simulation Results of Experiment Set 2

Table A3. Dixon test, MCAR and MAR mechanisms, $U(0, 1)$, n = 10,000, $R^2 \in [0,1[$.

% of MD	$R^2 \in [0, 0.05[$						$R^2 \in [0, 0.1[$					
	MCAR	MAR1	MAR2	MAR3	MAR4	MAR4i	MCAR	MAR1	MAR2	MAR3	MAR4	MAR4i
50%	94.1	0	0	0	0	0	97.4	0	0	0	0	0
45%	100.0	0	0	0	0	0	100.0	0	0	0	0	0
40%	92.9	0	0	0	0	0	97.1	0	0	0	0	0
35%	76.9	0	0	0	0	0	87.9	0	0	0	0	0
30%	86.4	0	0	0	0	0	87.2	0	0	0	0	0
25%	85.7	0	0	0	0	0	91.2	0	0	0	0	0
20%	96.2	0	0	0	0	0	96.4	0	0	0	0	0
15%	94.4	0	0	0	0	0	94.9	0	0	0	0	0
10%	87.5	0	0	0	0	0	92.1	0	0	0	0	0
5%	100.0	0	0	0	0	0	100.0	0	0	0	0	0
4%	100.0	0	0	0	0	0	97.3	0	0	0	0	0
3%	100.0	0	0	0	0	0	100.0	0	0	0	0	0
2%	100.0	0	0	0	0	0	100.0	0	0	0	0	0
1%	85.7	0	0	0	0	0	90.0	0	0	0	0	0
% of MD	$R^2 \in [0, 0.25[$						$R^2 \in [0.25, 0.5[$					
	MCAR	MAR1	MAR2	MAR3	MAR4	MAR4i	MCAR	MAR1	MAR2	MAR3	MAR4	MAR4i
50%	97.5	0	0	0	0	0	94.9	0	0	0	0	0
45%	96.7	0	0	0	0	0	93.8	0	0	0	0	0
40%	94.2	0	0	0	0	0	93.1	0	0	0	0	0
35%	92.6	0	0	0	0	0	94.6	0	0	0	0	0
30%	92.9	0	0	0	0	0	96.0	0	0	0	0	0
25%	92.1	0	0	0	0	0	95.5	0	0	0	0	0
20%	95.1	0	0	0	0	0	95.1	0	0	0	0	0
15%	96.2	0	0	0	0	0	91.1	0	0	0	0	0
10%	93.9	0	0	0	0	0	94.3	0	0	0	0	0
5%	96.6	0	0	0	0	0	95.7	0	0	0	0	0
4%	93.9	0	0	0	0	0	93.9	0	0	0	0	0
3%	96.0	0	0	0	0	0	94.9	0	0	0	0	0
2%	97.0	0	0	0	0	0	93.8	0	0	0	0	0
1%	92.3	0	0	0	0	0	92.5	0	0	0	0	0

Table A3. *Cont.*

% of MD	$R^2 \in [0.5, 0.75[$						$R^2 \in [0.75, 1[$					
	MCAR	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**	**MCAR**	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**
50%	97.5	0	0	0	0	0	94.5	0	0	0	0	0
45%	97.0	0	0	0	0	0	95.8	0	0	0	0	0
40%	93.6	0	0	0	0	0	95.1	0	0	0	0	0
35%	96.7	0	0	0	0	0	94.8	0	0	0	0	0
30%	94.6	0	0	0	0	0	98.6	0	0	0	0	0
25%	95.6	0	0	0	0	0	95.9	0	0	0	0	0
20%	97.3	0	0	0	0	0	95.8	0	0	0	0	0
15%	95.2	0	0	0	0	0	96.3	0	0	0	0	0
10%	94.5	0	0	0	0	0	95.9	0	0	0	0	0
5%	93.7	0	0	0	0	0	95.6	0	0	0	0	0
4%	91.9	0	0	0	0	0	91.9	0	0	0	0	0
3%	94.4	0	0	0	0	0	94.0	0	0	0	0	0
2%	97.9	0	0	0	0	0	93.5	0	0	0	0	0
1%	94.7	0	0	0	0	0	94.5	0	0	0	0	0

Table A4. Little test, MCAR and MAR mechanisms, $U(0, 1)$, n = 10,000, $R^2 \in [0,1[$.

% of MD	$R^2 \in [0, 0.05[$						$R^2 \in [0, 0.1[$					
	MCAR	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**	**MCAR**	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**
50%	88.2	0	0	0	0	0	94.7	0	0	0	0	0
45%	95.2	0	0	0	0	0	97.7	0	0	0	0	0
40%	92.9	0	0	0	0	0	97.1	0	0	0	0	0
35%	92.3	0	0	0	0	0	93.9	0	0	0	0	0
30%	90.9	0	0	0	0	0	89.7	0	0	0	0	0
25%	85.7	0	0	0	0	0	91.2	0	0	0	0	0
20%	96.2	0	0	0	0	0	96.4	0	0	0	0	0
15%	88.9	0	0	0	0	0	92.3	0	0	0	0	0
10%	93.8	0	0	0	0	0	89.5	0	0	0	0	0
5%	200.0	0	0	0	0	0	100.0	0	0	0	0	0
4%	92.3	0	0	0	0	0	89.2	0	0	0	0	0
3%	94.1	0	0	0	0	0	91.2	0	0	0	0	0
2%	91.7	0	0	0	0	0	96.2	0	0	0	0	0
1%	92.9	0	0	0	0	0	90.0	0	0	0	0	0
% of MD	$R^2 \in [0, 0.25[$						$R^2 \in [0.25, 0.5[$					
	MCAR	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**	**MCAR**	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**
50%	96.2	0	0	0	0	0	93.9	0	0	0	0	0
45%	93.9	0	0	0	0	0	94.2	0	0	0	0	0
40%	97.9	0	0	0	0	0	90.3	0	0	0	0	0
35%	93.2	0	0	0	0	0	94.6	0	0	0	0	0
30%	93.5	0	0	0	0	0	95.7	0	0	0	0	0
25%	95.8	0	0	0	0	0	96.1	0	0	0	0	0
20%	97.5	0	0	0	0	0	94.7	0	0	0	0	0
15%	94.5	0	0	0	0	0	92.3	0	0	0	0	0
10%	92.4	0	0	0	0	0	94.7	0	0	0	0	0
5%	96.6	0	0	0	0	0	95.1	0	0	0	0	0
4%	93.9	0	0	0	0	0	96.5	0	0	0	0	0
3%	96.6	0	0	0	0	0	94.9	0	0	0	0	0
2%	97.6	0	0	0	0	0	93.8	0	0	0	0	0
1%	94.7	0	0	0	0	0	92.5	0	0	0	0	0

Table A4. *Cont.*

% of MD	$R^2 \in [0.5, 0.75[$						$R^2 \in [0.75, 1[$					
	MCAR	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**	**MCAR**	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**
50%	95.1	0	0	0	0	0	94.8	0	0	0	0	0
45%	97.6	0	0	0	0	0	95.5	0	0	0	0	0
40%	93.1	0	0	0	0	0	94.0	0	0	0	0	0
35%	96.1	0	0	0	0	0	96.5	0	0	0	0	0
30%	92.6	0	0	0	0	0	96.3	0	0	0	0	0
25%	95.6	0	0	0	0	0	95.3	0	0	0	0	0
20%	97.8	0	0	0	0	0	93.3	0	0	0	0	0
15%	94.0	0	0	0	0	0	96.6	0	0	0	0	0
10%	95.8	0	0	0	0	0	95.9	0	0	0	0	0
5%	96.2	0	0	0	0	0	94.7	0	0	0	0	0
4%	97.1	0	0	0	0	0	91.3	0	0	0	0	0
3%	97.0	0	0	0	0	0	94.3	0	0	0	0	0
2%	94.2	0	0	0	0	0	93.8	0	0	0	0	0
1%	98.8	0	0	0	0	0	95.4	0	0	0	0	0

Table A5. RB test, MCAR and MAR mechanisms, $U(0, 1)$, n = 10,000, $R^2 \in [0,1[$.

% of MD	$R^2 \in [0, 0.05[$						$R^2 \in [0, 0.1[$					
	MCAR	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**	**MCAR**	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**
50%	94.1	5.6	4.2	11.5	10.7	13.0	89.5	11.4	4.8	15.6	10.5	10.9
45%	90.5	2.3	20.8	4.5	4.3	23.8	93.2	6.0	20.0	16.1	2.4	14.6
40%	92.9	6.1	6.7	11.1	3.4	11.1	94.1	10.7	11.5	12.5	5.7	12.5
35%	92.3	7.7	10.5	7.1	10.0	5.0	97.0	13.5	26.5	21.1	10.0	17.0
30%	86.4	5.9	10.0	10.5	0.0	7.7	92.3	9.1	18.4	11.8	7.7	16.7
25%	100.0	0.0	15.0	10.0	13.0	20.0	94.1	12.3	21.6	15.9	18.3	28.6
20%	100.0	17.4	20.0	12.5	3.8	23.5	96.4	24.6	25.0	10.5	8.2	25.0
15%	100.0	19.0	16.7	37.5	12.9	14.3	97.4	38.0	32.5	33.3	20.4	24.4
10%	81.2	48.0	29.4	20.0	23.4	20.0	89.5	48.1	28.6	30.0	21.7	28.0
5%	86.7	55.0	46.7	33.3	31.2	40.0	90.9	39.0	37.8	34.1	40.9	51.3
4%	100.0	50.0	20.0	71.4	26.9	23.8	97.3	60.0	36.4	61.8	29.5	28.9
3%	100.0	50.0	41.2	50.0	39.1	25.0	100.0	52.2	38.9	31.9	30.2	35.4
2%	75.0	54.5	22.2	22.2	44.0	56.5	88.5	60.5	18.4	33.3	43.2	56.0
1%	100.0	33.3	21.4	35.7	9.1	43.5	100.0	40.5	21.6	46.2	29.7	52.3
% of MD	$R^2 \in [0, 0.25[$						$R^2 \in [0.25, 0.5[$					
	MCAR	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**	**MCAR**	**MAR1**	**MAR2**	**MAR3**	**MAR4**	**MAR4i**
50%	96.2	18.1	17.0	14.9	15.7	10.0	95.5	32.0	25.2	15.7	25.5	20.4
45%	94.5	18.3	22.1	17.0	12.8	7.5	95.5	32.9	32.2	18.9	27.7	17.7
40%	93.7	22.4	17.7	16.6	14.3	14.4	95.8	36.1	26.8	22.0	28.2	26.1
35%	93.8	27.3	22.0	20.8	14.6	14.3	94.6	43.4	27.8	19.4	37.8	28.7
30%	92.9	23.0	27.6	19.9	15.9	19.1	94.6	44.8	33.0	25.2	30.8	28.2
25%	96.4	25.3	33.1	21.5	21.2	23.6	95.2	44.8	31.6	23.2	30.7	30.2
20%	95.6	36.6	29.3	25.8	19.6	30.2	95.4	56.9	40.0	26.5	33.0	36.1
15%	94.0	38.9	33.7	25.1	22.8	34.3	91.4	58.5	35.4	28.3	34.2	37.5
10%	93.9	56.0	32.3	34.1	24.0	34.5	92.5	58.2	38.2	26.6	35.4	38.1
5%	96.0	48.8	40.3	33.9	30.9	42.7	93.6	63.8	44.3	35.3	34.2	44.4
4%	93.3	55.6	37.2	37.9	32.6	35.5	97.1	58.2	35.6	33.1	38.3	46.1
3%	97.7	58.4	34.8	35.4	29.1	45.2	92.8	56.4	41.7	35.8	37.5	46.7
2%	94.6	55.2	35.5	39.8	34.9	50.5	95.2	66.9	43.4	36.6	41.0	47.5
1%	94.1	56.9	36.7	43.6	38.3	47.6	95.2	64.1	36.9	47.7	43.6	52.4

Table A5. *Cont.*

% of MD	$R^2 \in [0.5, 0.75[$						$R^2 \in [0.75, 1[$					
	MCAR	MAR1	MAR2	MAR3	MAR4	MAR4i	MCAR	MAR1	MAR2	MAR3	MAR4	MAR4i
50%	91.7	29.0	22.5	11.9	20.3	22.1	95.4	58.0	36.2	17.0	30.3	28.4
45%	97.6	40.9	22.4	14.6	28.3	20.0	93.6	57.8	30.3	15.5	31.3	25.7
40%	95.4	43.4	24.3	10.3	25.1	24.4	95.1	58.5	30.8	15.5	30.3	32.5
35%	94.4	42.6	24.7	12.6	28.4	25.3	93.9	62.1	36.1	13.9	30.8	28.9
30%	95.0	45.9	24.4	13.7	22.3	28.0	94.0	63.1	30.2	16.7	33.4	28.6
25%	95.6	49.2	28.9	11.5	31.8	25.0	93.7	58.3	30.0	22.3	32.4	28.5
20%	98.4	48.9	28.1	14.4	27.2	29.3	94.2	55.6	27.8	14.8	26.9	29.7
15%	94.0	58.1	29.1	20.5	22.8	32.4	96.9	59.9	30.8	18.4	29.8	30.7
10%	92.7	54.5	25.8	18.4	24.3	32.4	96.9	55.7	30.2	21.9	31.1	34.3
5%	96.9	54.9	30.2	20.8	22.7	34.6	92.3	55.2	37.9	20.4	36.4	36.1
4%	96.0	55.6	25.0	21.9	33.0	40.2	96.4	57.2	26.7	24.3	36.4	38.3
3%	95.4	57.2	25.9	17.9	31.1	42.1	93.7	58.6	27.6	22.3	33.9	42.6
2%	94.8	57.0	31.8	26.5	30.0	37.4	96.6	57.3	31.0	25.5	31.6	34.3
1%	98.2	61.4	32.1	23.1	25.0	43.4	95.4	56.8	27.9	28.7	30.6	44.6

Table A6. Sample sizes for the examples of Tables A3–A5.

% of MD	$R^2 \in [0, 0.05[$	$R^2 \in [0, 0.1[$	$R^2 \in [0, 0.25[$	$R^2 \in [0.25, 0.5[$	$R^2 \in [0.5, 0.75[$	$R^2 \in [0.75, 1[$
50%	17	38	157	311	204	328
45%	21	44	181	292	168	359
40%	14	34	190	289	173	348
35%	13	33	176	299	180	345
30%	22	39	170	277	202	351
25%	21	34	165	310	206	319
20%	26	55	203	285	182	330
15%	18	39	183	326	168	323
10%	16	38	197	318	165	320
5%	15	33	176	327	159	338
4%	13	37	180	312	173	335
3%	17	34	176	292	197	335
2%	12	26	166	291	191	352
1%	14	30	169	333	169	329

References

1. Allison, P.D. *Missing Data*; Sage Publications: Thousand Oaks, CA, USA, 2001.
2. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef] [PubMed]
3. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: Hoboken, NJ, USA, 2004; Volume 81.
4. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
5. Enders, C.K. *Applied Missing Data Analysis*; Guilford Press: New York, NY, USA, 2010.
6. Heitjan, D.F.; Basu, S. Distinguishing "missing at random" and "missing completely at random". *Am. Stat.* **1996**, *50*, 207–213.
7. Van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **2007**, *16*, 219–242. [CrossRef] [PubMed]
8. Seaman, S.R.; White, I.R. Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **2013**, *22*, 278–295. [CrossRef] [PubMed]
9. Huang, L.; Chen, M.H.; Ibrahim, J.G. Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics* **2005**, *61*, 767–780. [CrossRef] [PubMed]
10. Fitzmaurice, G. Missing data: Implications for analysis. *Nutrition* **2008**, *24*, 200–202. [CrossRef] [PubMed]
11. Dong, Y.; Peng, C.Y.J. Principled missing data methods for researchers. *SpringerPlus* **2013**, *2*, 222. [CrossRef] [PubMed]
12. Diggle, P.; Kenward, M.G. Informative drop-out in longitudinal data analysis. *Appl. Stat.* **1994**, *43*, 49–93. [CrossRef]
13. Dixon, W.J. *BMDP Statistical Software Manual: To Accompany the 1990 Software Release*; University of California Press: Oakland, CA, USA, 1990; Volume 1.
14. Little, R.J. A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* **1988**, *83*, 1198–1202. [CrossRef]
15. Jamshidian, M.; Jalal, S. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika* **2010**, *75*, 649–674. [CrossRef] [PubMed]
16. Jamshidian, M.; Yuan, K.H. Examining missing data mechanisms via homogeneity of parameters, homogeneity of distributions, and multivariate normality. *Wiley Interdiscip. Rev. Comput. Stat.* **2014**, *6*, 56–73. [CrossRef]
17. Jamshidian, M.; Jalal, S.J.; Jansen, C. MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). *J. Stat. Softw.* **2014**, *56*, 1–31. [CrossRef]

18. Powers, D.A.; Xie, Y. *Statistical Methods for Categorical Data Analysis*; Academic Press: Cambridge, MA, USA, 2000.
19. Park, T.; Davis, C.S. A test of the missing data mechanism for repeated categorical data. *Biometrics* **1993**, *49*, 631–638. [CrossRef]
20. Park, T.; Lee, S.Y. A test of missing completely at random for longitudinal data with missing observations. *Stat. Med.* **1997**, *16*, 1859–1871. [CrossRef]
21. Chen, H.Y.; Little, R.J. A test of missing completely at random for generalised estimating equations with missing data. *Biometrika* **1999**, *86*, 1–13. [CrossRef]
22. Hawkins, D.M. A new test for multivariate normality and homoscedasticity. *Technometrics* **1981**, *23*, 105–110. [CrossRef]
23. Scholz, F.W.; Stephens, M.A. K-sample Anderson-Darling tests. *J. Am. Stat. Assoc.* **1987**, *82*, 918–924.
24. Anderson, T.W.; Darling, D.A. A test of goodness of fit. *J. Am. Stat. Assoc.* **1954**, *49*, 765–769. [CrossRef]
25. Gabriel, K.R. Simultaneous test procedures–some theory of multiple comparisons. *Ann. Math. Stat.* **1969**, *40*, 224–250. [CrossRef]
26. Westfall, P.H.; Young, S.S. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*; John Wiley & Sons: Thousand Oaks, CA, USA, 1993.
27. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
28. Noé, M. The calculation of distributions of two-sided Kolmogorov-Smirnov type statistics. *Ann. Math. Stat.* **1972**, *43*, 58–64. [CrossRef]
29. Baumgartner, W.; Weiß, P.; Schindler, H. A nonparametric test for the general two-sample problem. *Biometrics* **1998**, *54*, 1129–1135. [CrossRef]
30. Marozzi, M. Some notes on the location–scale Cucconi test. *J. Nonparametr. Stat.* **2009**, *21*, 629–647. [CrossRef]
31. Marozzi, M. The multisample Cucconi test. *Stat. Methods Appl.* **2014**, *23*, 209–227. [CrossRef]
32. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
33. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
34. Bartholomew, D.J.; Steele, F.; Galbraith, J.; Moustaki, I. *Analysis of Multivariate Social Science Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2008.
35. Tshering, S.; Okazaki, T.; Endo, S. A Method to Identify Missing Data Mechanism in Incomplete Dataset. *IJCSNS* **2013**, *13*, 14.
36. Diggle, P.J. Testing for random dropouts in repeated measurement data. *Biometrics* **1989**, *45*, 1255–1258. [CrossRef]
37. Van Buuren, S.; Oudshoorn, C. *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*; Technical Report; TNO: Den Haag, The Netherlands, 2000.
38. Styan, G.P. Hadamard products and multivariate statistical analysis. *Linear Algebra Its Appl.* **1973**, *6*, 217–240. [CrossRef]
39. Rouzinov, S.; Berchtold, A. RBtest: Regression-Based Approach for Testing the Type of Missing Data. 2019. Available online: http://CRAN.R-project.org/package=RBtest (accessed on 24 January 2020).
40. R Core Team. R: A Language and Environment for Statistical Computing. 2015. Available online: https://www.R-project.org/ (accessed on 1 April 2019).
41. Tanizaki, H. Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. *J. Appl. Stat.* **1997**, *24*, 603–632. [CrossRef]
42. Maggiori, C.; Rossier, J.; Krings, F.; Johnston, C.S.; Massoudi, K. Career Pathways and Professional Transitions: Preliminary Results from the First Wave of a 7-Year Longitudinal Study. In *Surveying Human Vulnerabilities across the Life Course*; Oris, M., Roberts, C., Joye, D., Stähli, M.E., Eds.; Springer: Oxford, UK, 2016; pp. 131–157.
43. Cohen, J.; Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*; Routledge: Abingdon-on-Thames, UK, 2013.
44. Van Buuren, S. *Flexible Imputation of Missing Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.