Statistical Papers (2026) 67:56
https://doi.org/10.1007/s00362-026-01826-x

REGULAR ARTICLE

# Improvement of Chatterjee's correlation with missing at random data in the *Y* variable

**Fayyaz Bahari**[1]



**Abstract**
Establishing whether a variable $Y$ is a function of a variable $X$ is a fundamental problem in statistical analysis. Recently, a novel coefficient known as Chatterjee's correlation has been introduced to address this problem of testing for functional dependencies. This correlation is now widely used to assess dependencies between variables. However, statistical studies frequently encounter missing data. Failing to account for these missing values can lead to biased and misleading results. To address this issue, we consider the scenario where the variable $Y$ contains data that are missing at random. We then extend Chatterjee's correlation by incorporating inverse probability weighting. A theoretical framework is established to study the asymptotic properties of the proposed estimator. Our method demonstrates strong performance in finite-sample simulations and remains robust to the functional form of the dependency. Simulation studies confirm the excellent performance of the proposed estimator with finite sample sizes, even under varying rates of missing data. Furthermore, we demonstrate its practical utility by applying it to real-world datasets where the variable $Y$ is subject to missingness.

**Keywords** Chatterjee's correlation · Missing data · Rank-based statistics · Functional dependency · Inverse probability weights

## 1 Introduction

Recently, Chatterjee (2021) introduced a correlation coefficient designed to measure the extent to which a variable $Y$ is a function of a variable $X$. Consider a sample of size $n > 2$, denoted by the pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. To compute this coefficient, the data are first sorted by the $X$ variable, yielding the order statistics $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Let $y_{[j]}$ denote the $Y$ value concomitant to $x_{(j)}$. The reordered sample can thus be represented as $(x_{(1)}, y_{[1]}), (x_{(2)}, y_{[2]}), \ldots, (x_{(n)}, y_{[n]})$.

✉ Fayyaz Bahari
fayyaz.bahari@uma.ac.ir

[1] Department of Statistics and Applications, Faculty of Mathematical Sciences, University of Mohaghegh Ardabili, Ardabil, Iran

Published online: 21 April 2026

Assuming no ties in the $X$ observations, the rank of $y_{[j]}$ is defined as:

$$R_{[j]} = \sum_{i=1}^{n} I(y_i \leq y_{[j]}), \quad j = 1, 2, \cdots, n, \tag{1}$$

where $I(\cdot)$ is the indicator function. Chatterjee's correlation is then defined as:

$$\xi_n(X, Y) = 1 - \frac{3\sum_{j=1}^{n-1} |R_{[j+1]} - R_{[j]}|}{n^2 - 1}. \tag{2}$$

In Equations (1) and (2), we used notations that differ slightly from Chatterjee (2021) to better set the missing data formulation. As established by Chatterjee (2021), the coefficient $\xi_n(X, Y)$ possesses several key properties as follow:

1. It is asymmetric: $\xi_n(X, Y) \neq \xi_n(Y, X)$.
2. It is normalized: for sufficiently large sample sizes $n$, $\xi_n(X, Y) \in [0, 1]$.
3. For sufficiently large sample sizes, it equals zero if and only if $X$ and $Y$ are independent.
4. For sufficiently large sample sizes, it equals one if and only if $Y$ is a measurable function of $X$, i.e., $Y = f(X)$.
5. It is consistent: under the condition that $Y$ is not almost surely constant, $\xi_n(X, Y)$ converges almost surely to a population quantity $\xi(X, Y)$ as $n \to \infty$, where

$$\xi(X, Y) = \frac{\int var(E[I(Y \geq t) \mid X])d\mu(t)}{\int var(I(Y \geq t))d\mu(t)}. \tag{3}$$

Here, $\mu$ represents the probability law of $Y$.

For further details on Chatterjee's correlation, including its definition in the presence of ties, we refer the reader to the original work by Chatterjee (2021).

Since its introduction, Chatterjee's correlation has inspired numerous extensions and applications in the literature. For instance, Azadkia and Chatterjee (2021) introduced a conditional version of the coefficient, while Lin and Han (2023) proposed an estimation procedure using $m$-nearest neighbors. In a similar vein, Xia et al. (2025) developed a measure of functional dependency using a full nearest-neighbors approach, building upon the principles of Chatterjee's method. The asymptotic properties of hypothesis tests for independence based on $\xi_n$ have been studied by several authors, including (Auddy and Deb 2021; Lin 2022; Zhang 2023; Xia et al. 2025). Furthermore, the utility of Chatterjee's correlation has been demonstrated across a diverse range of statistical analyses. It has been applied to hierarchical variable clustering (Fuchs and Wang 2024), the development of prediction models (Thottolil et al. 2023), and variable selection in the general linear model (Bahari 2025).

A common challenge in statistical studies is the presence of missing data. According to the framework established by Little and Rubin (2019), simply ignoring missing values can lead to unreliable and misleading results. This issue provides the primary motivation for our work: to extend Chatterjee's correlation to handle missing data.

To address this problem methodically, we adopt the foundational missing data taxonomy introduced by Rubin (1976). This classification, formalized by Rubin (1976), is crucial for guiding appropriate analytical methods and is defined as follows:

1. *Missing Completely at Random (MCAR)*: The probability of a value being missing is independent of both observed and unobserved data.
2. *Missing at Random (MAR)*: The probability of a value being missing depends only on the observed data.
3. *Missing Not at Random (MNAR)*: The probability of a value being missing depends on the unobserved (missing) values themselves, potentially also on the observed data.

In this paper, we assume that missingness occurs only in the variable $Y$ under the MAR mechanism. Formally, we define the $i$th observation probability as:

$$\pi(x_i) = P(\delta_i = 1 \mid x_i), \quad i = 1, 2, \cdots, n, \tag{4}$$

where $\delta_i$ is the missingness indicator variable such that $\delta_i = 1$ if $y_i$ is observed and $\delta_i = 0$ if it is missing. The function $\pi(\cdot)$ represents the observing probability, which depends solely on the fully observed variable $X$, thereby satisfying the MAR assumption. For comprehensive overviews of missing data theory and inferences with missing data, we refer readers to Little and Rubin (2019); Creemers et al. (2012).
The rest of this paper is organized as follows. In Section 2, we introduce the proposed methodology and establish the asymptotic properties of our estimator. Section 3 is devoted to a simulation study that investigates the finite-sample performance of the proposed method under different missingness mechanisms and missing rates. In Section 4, we demonstrate the practical utility of our approach by applying it to real-world datasets. Finally, detailed proofs of the lemmas are provided in the Appendix.

## 2 Methodology and theoretical approaches

In this section, we assume that the response variable $Y$ is subject to missingness under the MAR mechanism. The presence of missing data implies that the true ranks $R_{[j]}$ are inestimable. To address this, we propose two estimators for $R_{[j]}$: one based on the Complete Case (CC) method, which uses only the observed data, and another using the Inverse Probability Weighting (IPW) method, which weights the observed data by the inverse of their probability of being observed.
For an index $j$ where the concomitant $y_{[j]}$ is observed, the CC and IPW estimators of the rank are defined, respectively, as:

$$R^{CC}_{[j]} = \sum_{i=1}^{n} \delta_i I(y_i \le y_{[j]}), \quad \text{if j'th data to be observed,} \tag{5}$$

$$R^{*}_{[j]} = \sum_{i=1}^{n} \frac{\delta_i}{\pi(x_i)} I(y_i \le y_{[j]}), \quad \text{if j'th data to be observed.} \tag{6}$$

Note that these estimators are only defined when the $j$th individual is observed ($\delta_{[j]} = 1$); the rank for a missing $j$th observation is considered unknown since $y_{[j]}$ is missing. Using these rank estimators, we now define the corresponding estimators for Chatterjee's correlation. The Complete Case estimator is given by:

$$\xi_n^{CC}(X, Y) = 1 - \frac{3\sum_{j=1}^{n-1} \delta_{[j+1]}\delta_{[j]}|R_{[j+1]}^{CC} - R_{[j]}^{CC}|}{m^2 - 1}, \tag{7}$$

where $m = \sum_{i=1}^{n} \delta_i$ is the number of complete cases and $\delta_{[j]}$ is the missingness indicator corresponding to the order statistic $x_{(j)}$. The IPW estimator is defined as:

$$\xi_n^{*}(X, Y) = 1 - \frac{3\sum_{j=1}^{n-1} \frac{\delta_{[j+1]}}{\pi(x_{(j+1)})}\frac{\delta_{[j]}}{\pi(x_{(j)})}|R_{[j+1]}^{*} - R_{[j]}^{*}|}{n^2 - 1}. \tag{8}$$

In Equation (8), the observing probability corresponding to $x_{(j)}$ is denoted by $\pi(x_{(j)})$, where $\pi(x_{(j)}) = P(\delta_{[j]} = 1 \mid x_{(j)})$. In both proposed methods, a pair is excluded from the summation if at least one of its concomitant $Y$ values is missing. Consequently, the correlation estimators rely solely on the estimated ranks for which the corresponding $Y$ values are observed. The IPW method compensates for the exclusion of incomplete pairs by assigning inverse probability weights to the observed data. This weighting scheme constructs a pseudo-population that is representative of the complete data, thereby correcting for the bias introduced by the missing data mechanism. Specifically, units with a lower probability of being observed receive a higher weight, effectively up-weighting them to account for similar units that are missing.
In practice, the propensity scores $\pi(x_i)$ are typically unknown and must be estimated. A consistent nonparametric estimator, based on kernel smoothing, is given by:

$$\hat{\pi}(x_i) = \frac{\sum_{j=1}^{n} \delta_j K_{h_n}(x_j - x_i)}{\sum_{j=1}^{n} K_{h_n}(x_j - x_i)}, \quad i = 1, 2, \cdots, n, \tag{9}$$

where $K_{h_n}(\cdot)$ is a kernel function with bandwidth $h_n$. Substituting this estimator into our proposed estimator, we define the feasible IPW rank estimator as:

$$R_{[j]}^{IPW} = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i)} I(y_i \le y_{[j]}), \quad \text{for } j \text{ such that } \delta_j = 1. \tag{10}$$

The corresponding feasible IPW estimator for Chatterjee's correlation is then:

$$\xi_n^{IPW}(X, Y) = 1 - \frac{3\sum_{j=1}^{n-1} \frac{\delta_{[j+1]}}{\hat{\pi}(x_{(j+1)})}\frac{\delta_{[j]}}{\hat{\pi}(x_{(j)})}|R_{[j+1]}^{IPW} - R_{[j]}^{IPW}|}{n^2 - 1}. \tag{11}$$

To obtain asymptotic properties of our proposed estimators, we consider some middle assumption which is known as regularity conditions. (see, e.g., Wang and Wang (2001); Creemers et al. (2012)).

**Regularity Conditions**

C1) The propensity score is bounded away from zero: $\inf_x \pi(x) > \zeta$ for some $\zeta > 0$.
C2) The propensity score $\pi(x)$ is twice continuously differentiable.
C3) The kernel function $K_{h_n}(\cdot)$ is a symmetric, continuous function of order 2, satisfying $\int K_{h_n}(u)\,du = 1$, $\int uK_{h_n}(u)\,du = 0$, and $\int u^2 K_{h_n}(u)\,du < \infty$.
C4) The bandwidth $h_n$ is a sequence such that, as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$, with the additional requirement that $h_n^2 + (nh_n)^{-1/2} = o(1)$.

Under the MAR mechanism, the CC estimator $\xi_n^{CC}(X, Y)$ is generally biased, as will be illustrated in our simulation studies. For a comprehensive discussion of the bias inherent in complete case analysis under MAR, see Little and Rubin (2019); Little (1992). The CC method is included here primarily to benchmark the performance of our main method, the IPW estimator. Consequently, we focus our theoretical analysis on the IPW estimator $\xi_n^{IPW}(X, Y)$. To establish its asymptotic properties, we first present two lemmas concerning the propensity score estimator and the rank process, followed by our main theorems.

**Lemma 1** *Under the regularity conditions, the following stochastic approximations hold:*

*a)* $R_{[j]}^{IPW} = R_{[j]}^* + O_P(nh_n^2 + \sqrt{\frac{n}{h_n}})$,
*b)* $\xi_n^{IPW}(X, Y) = \xi_n^*(X, Y) + O_P(h_n^2 + \frac{1}{\sqrt{nh_n}})$.

The proof of Lemma 1 is provided in Appendix A.1. This lemma characterizes the error introduced by estimating the propensity score.

**Lemma 2** *Under the regularity conditions, the following stochastic approximations hold:*

*a)* $R_{[j]}^* = R_{[j]} + O_P(\sqrt{n})$,
*b)* $\xi_n^*(X, Y) = \xi_n(X, Y) + O_P(\frac{1}{\sqrt{n}})$.

Te proof of Lemma 2 is provided in Appendix A.2. This lemma establishes the asymptotic relationship between the infeasible IPW estimator (with known propensity scores) and the full data estimator.

**Theorem 1** *Under the regularity conditions, the proposed estimator has the following properties:*

*a)* $\xi_n^{IPW}(X, Y) = \xi_n(X, Y) + O_P(h_n^2 + \frac{1}{\sqrt{nh_n}})$,
*b)* $\xi_n^{IPW}(X, Y) \xrightarrow{P} \xi(X, Y)$ *as* $n \to \infty$, *provided* $h_n \to 0$ *and* $nh_n \to \infty$.

***Proof*** Part (a) follows directly by combining Lemma 1(b) and Lemma 2(b). Part (b) follows from part (a) and the consistency of the full data estimator $\xi_n(X, Y)$ established in Chatterjee (2021), under the given bandwidth conditions. □

Theorem 1 establishes the asymptotic equivalence between the proposed IPW estimator, which uses estimated propensity scores, and the original Chatterjee's correlation

computed on the full data. This result guarantees that our method inherits the desirable asymptotic properties of the full data estimator despite the presence of missing values. As discussed, the CC method provides a biased estimator, and our aim is to derive the details of the IPW method. The following remarks and theorem provide further insights into the properties of the IPW estimator.

***Remark 1*** By Theorem 1(a), taking expectations and using Theorem 1.1 of Chatterjee (2021), it follows that

$$E(\xi_n^{IPW}(X,Y)) = E(\xi_n(X,Y)) + O(h_n^2 + \frac{1}{\sqrt{nh_n}}).$$

Therefore, $E(\xi_n^{IPW}(X,Y)) \to \xi(X,Y)$ as $n \to \infty$ and $nh_n \to \infty$. In particular, under the independence assumption, where $E(\xi_n(X,Y)) \to 0$ as $n \to \infty$, the expectation of the IPW estimator also converges to zero, provided that $h_n \to 0$ and $nh_n \to \infty$.

**Theorem 2** *Suppose $X$ and $Y$ are independent and $Y$ is a continuous random variable. Then, under the regularity conditions,*

$$\sqrt{n}\xi_n^{IPW}(X,Y) \xrightarrow{d} N(0, \frac{2}{5}) \quad as \quad n \to \infty.$$

***Proof*** By Slutsky's theorem, convergence in probability implies convergence in distribution. Applying Theorem 1 together with Theorem 2.1 of Chatterjee (2021), the result follows. □

An immediate consequence of Theorem 2 is that it can be used to construct a test for the independence of $X$ and $Y$. Also, the following remark discusses possible extensions when some of the assumptions of Theorem 2 are relaxed.

***Remark 2*** Theorem 2 is established under the assumptions that $Y$ is continuous and that $X$ and $Y$ are independent. If the continuity assumption on $Y$ is relaxed, similar asymptotic results can be obtained by combining Theorem 1 with Theorem 2.2 of Chatterjee (2021). Furthermore, by applying Theorem 1 together with Theorem 1.1 of Lin (2022), analogous conclusions to those of Theorem 2 can be derived in more general settings. For brevity, we omit the technical details, particularly those related to variance estimation in these cases, and leave a thorough investigation for future work.

## 3 Simulation study

In this section, we evaluate the performance of our proposed methods under various missing data scenarios. Each Monte Carlo simulation was run for $t = 5000$ iterations to compute the results presented in the tables and to generate the figures. To compare the different methods, we employ the following performance metrics:

- The mean of the estimators across the $t$ Monte Carlo replications, denoted by $\bar{\xi}$.

- The Intraclass Correlation Coefficient (ICC) for absolute agreement, as defined by Koo and Li (2016), which measures the consistency between the proposed estimators and the full data Chatterjee's correlation.
- The Mean Squared Bias (MSB) relative to the full data Chatterjee's estimator, defined for method $i$ as:

$$MSB_i = \frac{1}{t}\sum_{j=1}^{t}(\xi^i_{n_j}(X,Y) - \xi_{n_j}(X,Y))^2, \quad i \in \{CC, IPW\}, \tag{12}$$

where $\xi_{n_j}(X,Y)$ is the estimate from the $j$th full dataset and $\xi^i_{n_j}(X,Y)$ is the corresponding estimate from method $i$ applied to the incomplete data.

We consider three main simulation studies (Studies 1–3) to comprehensively evaluate the proposed methods. These studies investigate the impact of various factors, including sample size, missing data mechanism (MCAR, MAR), missing rate, and the underlying functional form of the dependency between $X$ and $Y$. In Study 4, we employ visualization techniques to provide deeper insight into the behavior of the estimators and to further illustrate the effect of the functional relationship on the analysis. Moreover, in the final scenario, we use Theorem 2 to compare the IPW method with the Full method for testing the independence hypothesis through Monte Carlo simulation studies.
The R codes used for the simulation studies and the real data analysis are available at the following link: https://drive.google.com/file/d/1BFY5WmL0CHWOFPi61AP8cW5Y2vSrG38t/view?usp=sharing.

### 3.1 Study 1: A model with strong dependency

We begin with a model characterized by a strong, nonlinear dependency:

**Model 1:** $Y = X^3 + \epsilon$,

where $X \sim N(0,1)$ and $\epsilon \sim N(0, 0.5)$. To evaluate the performance of our estimators under diverse missing data scenarios, we generate missing values in $Y$ according to the following four mechanisms:

1) $\pi_1(x) = \dfrac{1}{1 + 0.4 \mid X|e^{-x^2}}$ (MAR, moderate complexity).
2) $\pi_2(x) = 0.8$ (MCAR, 20% missing rate).
3) $\pi_3(x) = \dfrac{1}{1 + e^{-1.5-0.5x}}$ (MAR, logistic).
4) $\pi_4(x) = \dfrac{1}{1 + \sin^2(2\pi x)}$ (MAR, periodic).

The simulation results for this study, including different sample sizes $n$ and the resulting missing rates under each mechanism, are presented in Table 1.

The consistency of Chatterjee's correlation ensures that the full data estimator converges to the true population value as $n$ increases. Consequently, comparing the proposed methods (CC and IPW) against this full data benchmark is meaningful across

**Table 1** Simulation results for Study 1: Performance of the Full, Complete Case (CC), and Inverse Probability Weighting (IPW) estimators of Chatterjee's correlation for different sample sizes ($n$), missingness mechanisms (MM), and missing rates (MR), showing the mean estimate ($\bar{\xi}$), Intraclass Correlation Coefficient (ICC), and Mean Squared Bias (MSB) based on 5000 Monte Carlo replications

| | | | n=50 | | | n=100 | | | n=500 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | MM | MR | $\bar{\xi}$ | MSB | ICC | $\bar{\xi}$ | MSB | ICC | $\bar{\xi}$ | MSB | ICC |
| Full | – | – | 0.6488 | – | – | 0.6751 | – | – | 0.6920 | – | – |
| CC | $\pi_1$ | 0.09 | 0.6854 | 0.0024 | 0.9090 | 0.7080 | 0.0016 | 0.9157 | 0.7232 | 0.0011 | 0.9181 |
| IPW | | | 0.6541 | 0.0008 | 0.9373 | 0.6776 | 0.0003 | 0.9463 | 0.6930 | 0.0001 | 0.9499 |
| CC | $\pi_2$ | 0.20 | 0.7204 | 0.0075 | 0.8015 | 0.7403 | 0.0054 | 0.8169 | 0.7436 | 0.0040 | 0.8166 |
| IPW | | | 0.6632 | 0.0022 | 0.8531 | 0.6828 | 0.0009 | 0.8718 | 0.6938 | 0.0002 | 0.8826 |
| CC | $\pi_3$ | 0.20 | 0.7128 | 0.0063 | 0.8170 | 0.7323 | 0.0043 | 0.8301 | 0.7559 | 0.0031 | 0.8297 |
| IPW | | | 0.6637 | 0.0020 | 0.8626 | 0.6828 | 0.0009 | 0.8797 | 0.6941 | 0.0002 | 0.8887 |
| CC | $\pi_4$ | 0.30 | 0.7327 | 0.0099 | 0.7686 | 0.7489 | 0.0069 | 0.7747 | 0.7600 | 0.0049 | 0.7747 |
| IPW | | | 0.6697 | 0.0038 | 0.7531 | 0.6879 | 0.0017 | 0.7726 | 0.7024 | 0.0004 | 0.7900 |

different sample sizes. The results in Table 1 lead to several key conclusions. Overall, the IPW method demonstrates strong performance, substantially outperforming the CC method. As the rate of missing data increases, the accuracy of the CC estimator deteriorates rapidly. In contrast, the IPW estimator maintains its accuracy remarkably well, showing little sensitivity to the missing rate. Furthermore, both proposed methods exhibit the same convergence behavior as the full data estimator; specifically, the IPW estimates converge toward the true value as the sample size increases, as theoretically expected. By comparing the $\bar{\xi}$ column, it is evident that the IPW method gives estimates closer to the Full method than the CC method. This is further supported by the MSB criterion, where the MSB of the IPW method is smaller than that of the CC method in nearly all scenarios, indicating that IPW estimates are generally closer to the full-data case. The ICC values are largely invariant to sample size but decrease with higher missing rates. A comparison of ICC values confirms that the IPW method maintains significantly higher agreement with the full data estimates than the CC method. To isolate the effect of the missing mechanism, we compared mechanisms (2) (MCAR) and (3) (MAR) at comparable missing rates. The IPW method performs robustly under both mechanisms, which is theoretically justified as MCAR is a special case of MAR. In general, the IPW method provides excellent results. However, its accuracy in small samples can be affected when the missing rate is very high. Finally, we caution against the use of the CC method in any missing data scenario, as it consistently introduces substantial bias.

### 3.2 Study 2: A model with no functional dependency

This study examines the performance of the estimators under the scenario of independence between $X$ and $Y$. Consider the following model:

**Model 2:** $Y = 1 + \epsilon$,

where $X \sim U(1, 3)$ and $\epsilon \sim N(0, 0.5)$. In this model, $Y$ is independent of $X$, and thus the true value of Chatterjee's correlation is 0. This allows for a direct comparison of the estimates against the known null value. Missingness in $Y$ is generated according to the following mechanisms:

5) $\pi_5(x) = \dfrac{1}{1 + e^{-1.5-0.5x}}$ (MAR, logistic).
6) $\pi_6(x) = 0.8$ (MCAR, 20% missing rate).
7) $\pi_7(x) = \dfrac{1}{1 + \frac{x}{8}}$ (MAR, decreasing).
8) $\pi_8(x) = \dfrac{1}{1 + 0.3\sqrt{x}}$ (MAR, decreasing).

The simulation results for Study 2 are presented in Table 2 for different sample sizes and the resulting missing rates under each mechanism.

Since Model 2 implies independence between $X$ and $Y$, the true value of Chatterjee's correlation is 0, and a consistent estimator should converge to this value as $n \to \infty$. The relative performance of the proposed methods in this setting is consistent with the findings from Study 1. The IPW estimator continues to demonstrate excellent performance, with $\xi_n^{IPW}(X, Y)$ converging toward 0 as the sample size increases. While the

**Table 2** Simulation results for Study 2: Performance of the Full, Complete Case (CC), and Inverse Probability Weighting (IPW) estimators of Chatterjee's correlation for different sample sizes ($n$), missingness mechanisms (MM), and missing rates (MR), showing the mean estimate ($\bar{\xi}$), Intraclass Correlation Coefficient (ICC), and Mean Squared Bias (MSB) based on 5000 Monte Carlo replications

| | | | n=50 | | | n=100 | | | n=500 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | MM | MR | $\bar{\xi}$ | MSB | ICC | $\bar{\xi}$ | MSB | ICC | $\bar{\xi}$ | MSB | ICC |
| Full | – | – | 0.0056 | – | – | 0.0003 | – | – | 0.0002 | – | – |
| CC | $\pi_5$ | 0.08 | 0.0802 | 0.0100 | 0.7868 | 0.0784 | 0.0080 | 0.7997 | 0.0734 | 0.0065 | 0.8032 |
| IPW | | | 0.0197 | 0.0034 | 0.8212 | 0.0125 | 0.0017 | 0.8355 | 0.0055 | 0.0003 | 0.8567 |
| CC | $\pi_6$ | 0.20 | 0.2056 | 0.0515 | 0.5517 | 0.2018 | 0.0455 | 0.5679 | 0.2013 | 0.0413 | 0.5884 |
| IPW | | | 0.0550 | 0.0138 | 0.5608 | 0.0357 | 0.0064 | 0.5879 | 0.0152 | 0.0011 | 0.6369 |
| CC | $\pi_7$ | 0.20 | 0.2004 | 0.0489 | 0.5657 | 0.1968 | 0.0433 | 0.5858 | 0.1957 | 0.0390 | 0.5968 |
| IPW | | | 0.0551 | 0.0138 | 0.5649 | 0.0359 | 0.0065 | 0.5898 | 0.0150 | 0.0011 | 0.6329 |
| CC | $\pi_8$ | 0.30 | 0.3000 | 0.1027 | 0.4333 | 0.2965 | 0.0948 | 0.4483 | 0.2936 | 0.0873 | 0.4701 |
| IPW | | | 0.0903 | 0.0283 | 0.3936 | 0.0602 | 0.0139 | 0.4057 | 0.0237 | 0.0023 | 0.4726 |

accuracy of the IPW estimator decreases slightly with higher missing rates, this effect is modest. In stark contrast, the CC estimator exhibits severe bias. Its estimates are consistently and substantially inflated above 0, and this bias *increases* with the missing rate. For instance, with $n = 500$ and a 30% missing rate, $\xi_n^{CC}(X, Y) = 0.2936$, a value far from the true value of 0. Under the same conditions, the IPW estimator yields $\xi_n^{IPW}(X, Y) = 0.0237$, demonstrating its remarkable accuracy and robustness even under the hypothesis of independence.

### 3.3 Study 3: A model with moderate dependency

We now examine a scenario with a moderate, nonlinear functional dependency:

**Model 3:** $Y = 1 + 2\log(1 + X) + \epsilon$,

where $X \sim \text{Beta}(3, 2)$ and $\epsilon \sim N(0, 0.5)$. Missingness in $Y$ is simulated through the following four mechanisms:

9) $\pi_9(x) = \dfrac{1}{1 + e^{-1-1.5x-x^2}}$ (MAR, complex logistic).

10) $\pi_{10}(x) = \dfrac{1}{1 + 0.4 \mid X \mid}$ (MAR, V-shaped).

11) $\pi_{11}(x) = \dfrac{1}{1 + x^3}$ (MAR, polynomial decay).

12) $\pi_{12}(x) = \dfrac{1}{1 + 0.25\mid \sin(\pi x) + \cos(\pi x)\mid}$ (MAR, oscillatory).

The simulation results for Study 3, including different sample sizes and the resulting missing rates, are presented in Table 3.

For the full data case with $n = 500$, the estimated Chatterjee's correlation is 0.3227, confirming a moderate functional dependency of $Y$ on $X$, as specified by the model. The results from this study reinforce the patterns observed in Studies 1 and 2. The IPW method consistently and substantially outperforms the CC method across all evaluated conditions. It provides estimates that are both more accurate and more stable, demonstrating its robustness as a general-purpose method for estimating functional dependency in the presence of missing data.

### 3.4 Study 4: Visualization across functional forms

This study visually compares the performance of the estimators across four distinct functional relationships, designed to cover a range of dependency structures:

**Model 4:** $Y = \rho(1 + X) + (1 - \rho)\epsilon$ (Linear).
**Model 5:** $Y = \rho \sin(X) + (1 - \rho)\epsilon$ (Sinusoidal).
**Model 6:** $Y = \rho(1 + \log(1 + X^2)) + (1 - \rho)\epsilon$ (Logarithmic).
**Model 7:** $Y = \rho(1 + e^{-1-X}) + (1 - \rho)\epsilon$ (Exponential).

In all models, $X \sim N(0, 1)$ and $\epsilon \sim N(0, 0.5)$. The parameter $\rho \in [0, 1]$ controls the noise of the models, allowing the true functional dependency to vary smoothly from pure noise ($\rho = 0$) to a deterministic relationship ($\rho = 1$). Missingness in $Y$

**Table 3** Simulation results for Study 3: Performance of the Full, Complete Case (CC), and Inverse Probability Weighting (IPW) estimators of Chatterjee's correlation for different sample sizes ($n$), missingness mechanisms (MM), and missing rates (MR), showing the mean estimate ($\bar{\xi}$), Intraclass Correlation Coefficient (ICC), and Mean Squared Bias (MSB) based on 5000 Monte Carlo replications

| | | | n=50 | | | n=100 | | | n=500 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | MM | MR | $\bar{\xi}$ | MSB | ICC | $\bar{\xi}$ | MSB | ICC | $\bar{\xi}$ | MSB | ICC |
| Full | | | 0.3036 | – | – | 0.3143 | – | – | 0.3227 | – | – |
| CC | $\pi_9$ | 0.10 | 0.3667 | 0.0064 | 0.8699 | 0.3771 | 0.0052 | 0.8740 | 0.3831 | 0.0039 | 0.8728 |
| IPW | | | 0.3202 | 0.0024 | 0.8900 | 0.3246 | 0.0011 | 0.9032 | 0.3257 | 0.0002 | 0.9088 |
| CC | $\pi_{10}$ | 0.19 | 0.4408 | 0.0237 | 0.7433 | 0.4515 | 0.0212 | 0.7533 | 0.4561 | 0.0183 | 0.7504 |
| IPW | | | 0.3319 | 0.0062 | 0.7598 | 0.3311 | 0.0028 | 0.7811 | 0.3266 | 0.0005 | 0.7976 |
| CC | $\pi_{11}$ | 0.20 | 0.4373 | 0.0229 | 0.7387 | 0.4472 | 0.0202 | 0.7464 | 0.4535 | 0.0176 | 0.7385 |
| IPW | | | 0.3435 | 0.0081 | 0.7221 | 0.3400 | 0.0038 | 0.7345 | 0.3318 | 0.0007 | 0.7464 |
| CC | $\pi_{12}$ | 0.25 | 0.4802 | 0.0376 | 0.6725 | 0.4901 | 0.0341 | 0.6826 | 0.4924 | 0.0294 | 0.6895 |
| IPW | | | 0.3451 | 0.0095 | 0.6834 | 0.3393 | 0.0043 | 0.7031 | 0.3284 | 0.0007 | 0.7322 |

are generated using mechanisms $\pi_1(x)$ and $\pi_4(x)$ from Study 1, resulting in approximately 10% and 30% missing data. We estimated Chatterjee's correlation using the Full, CC, and IPW methods for different values of $\rho$. The results are displayed in Figs. 1, 2, 3 and 4 for Models 4–7, respectively. In all figures, the Full data estimates are represented by solid curves, the IPW estimates by dashed curves, and the CC estimates by dotted curves. The simulations were conducted for sample sizes $n = 50$, $n = 100$ and $n = 500$ under both missingness mechanisms to examine their interacting effects. The figures corroborate the conclusions from Studies 1–3. The IPW estimator (dashed curves) consistently tracks the full data estimates (solid curves) closely across all functional forms, sample sizes, and missingness mechanisms. In contrast, the CC estimator (dotted curves) exhibits significant bias, particularly in scenarios with low to moderate dependency. This bias in the CC method diminishes as the true dependency strength ($\rho$) increases, with its estimates gradually shifting toward the full data curve. The effect of a higher missing rate is visually apparent for both methods, but is substantially more pronounced for the CC estimator. A comparison across Figs. 1, 2, 3 and 4 reveals two key insights: first, the superior performance of the IPW method is robust to the underlying functional form of the dependency; and second, the sample size and missing rate remain the primary factors influencing the precision of the estimates.

### 3.5 Study 5: Testing independence

In this section, we use simulation studies to evaluate the performance of Theorem 2 for testing independence. To this end, we consider the following models:

**Model 8:** $Y = \rho(1 + X^3) + (1 - \rho)\epsilon$,
**Model 9:** $Y = \rho(1 + \sin(X)) + (1 - \rho)\epsilon$.

In all models, $X \sim N(0, 1)$, $\epsilon \sim N(0, 0.25)$ and $\rho \in [0, 1]$. To generate missing values in the response variable $Y$, we consider the following missingness mechanisms:

13) $\pi_{14}(x) = \dfrac{1}{1 + e^{-2-0.5x}}$, (MAR, logistic).

14) $\pi_{15}(x) = \dfrac{1}{1 + 0.25\log(0.5 + |x|)}$ (MAR, U-shaped).

Under mechanism 13, approximately 13% of the observations in $Y$ are missing, whereas under mechanism 14, approximately 6% of the observations in $Y$ are missing. For Models 8 and 9, when $\rho = 0$, the variables $X$ and $Y$ are independent. As $\rho$ increases, the dependence between $X$ and $Y$ becomes stronger. In particular, when $\rho = 1$, $Y$ is a deterministic function of $X$. We compute the empirical power of the independence test for different values of $\rho$ at the significance level $\alpha = 0.05$. As expected, the empirical power increases toward one as $\rho$ approaches one. The results of this study are presented in Figs. 5 and 6. Figures 5 and 6 display the empirical power as a function of the noise parameter $\rho$. In both figures, the power for rejecting the independence hypothesis is an increasing function of $\rho$. When $\rho$ is close to zero, the probability of rejecting independence is low, whereas for values of $\rho$ greater than approximately 0.5, the empirical power is close to one. A comparison of the two figures highlights the effect of the missing rate. Under a lower missing rate, the

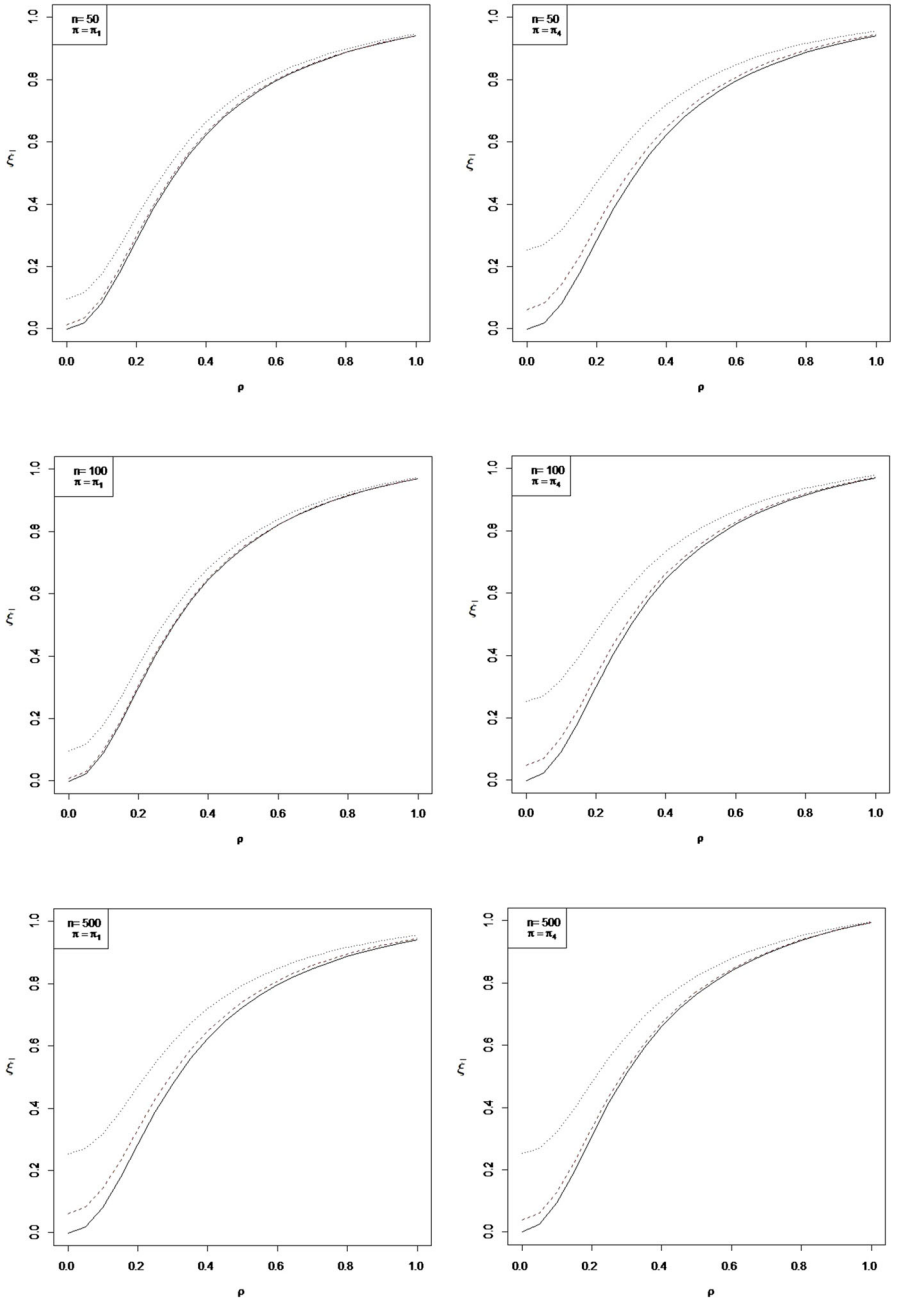


**Fig. 1** Estimated functional dependency for Model 4 (Linear) across different signal strengths $\rho$. Results are shown for the Full data (solid line), Complete Case (CC; dotted line), and Inverse Probability Weighting (IPW; dashed line) estimators. The panels display results for all combinations of sample size ($n = 50, 100, 500$) and missingness mechanism ($\pi_1$ or $\pi_4$)

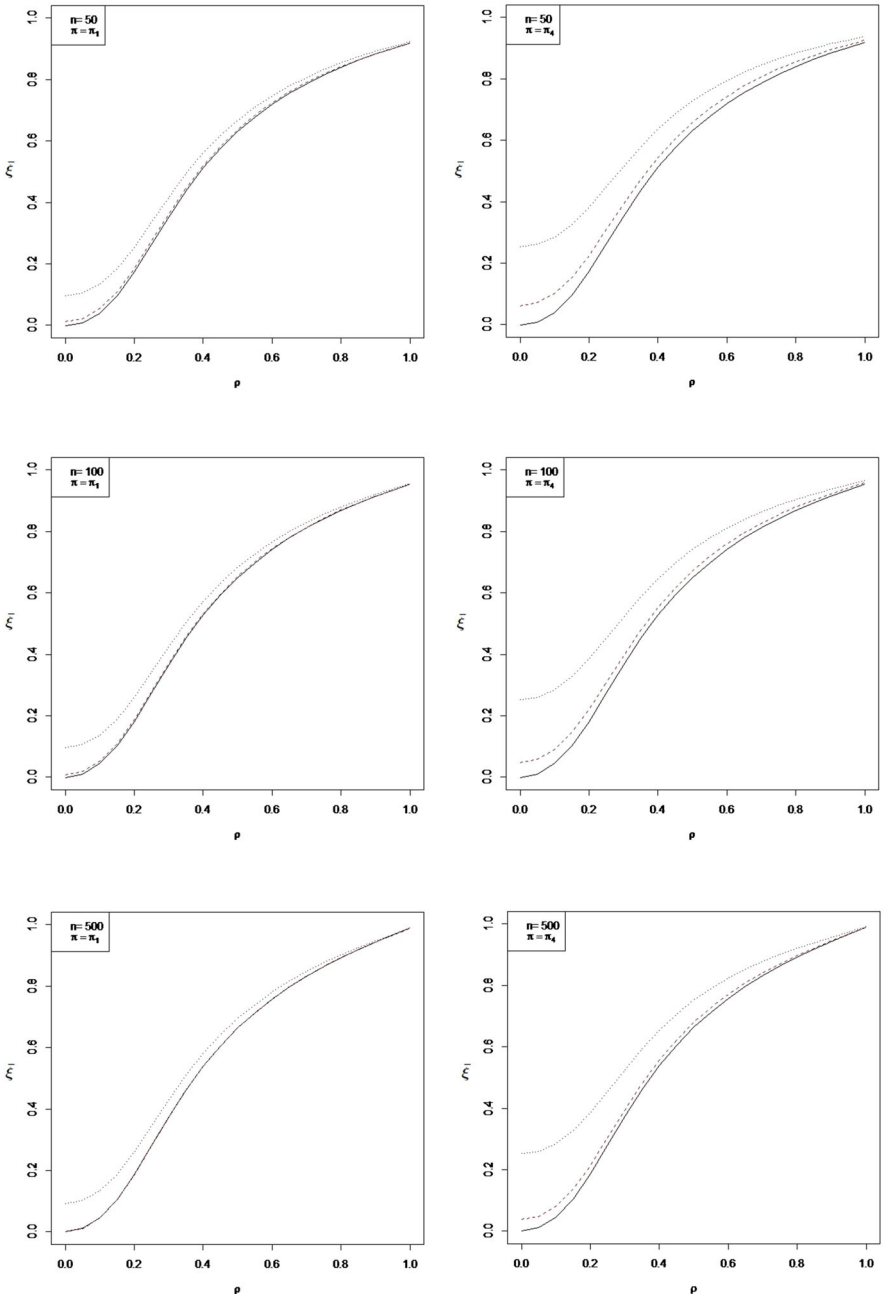


**Fig. 2** Estimated functional dependency for Model 5 (Sinusoidal) across different signal strengths $\rho$. Results are shown for the Full data (solid line), Complete Case (CC; dotted line), and Inverse Probability Weighting (IPW; dashed line) estimators. The panels display results for all combinations of sample size ($n = 50, 100, 500$) and missingness mechanism ($\pi_1$ or $\pi_4$)

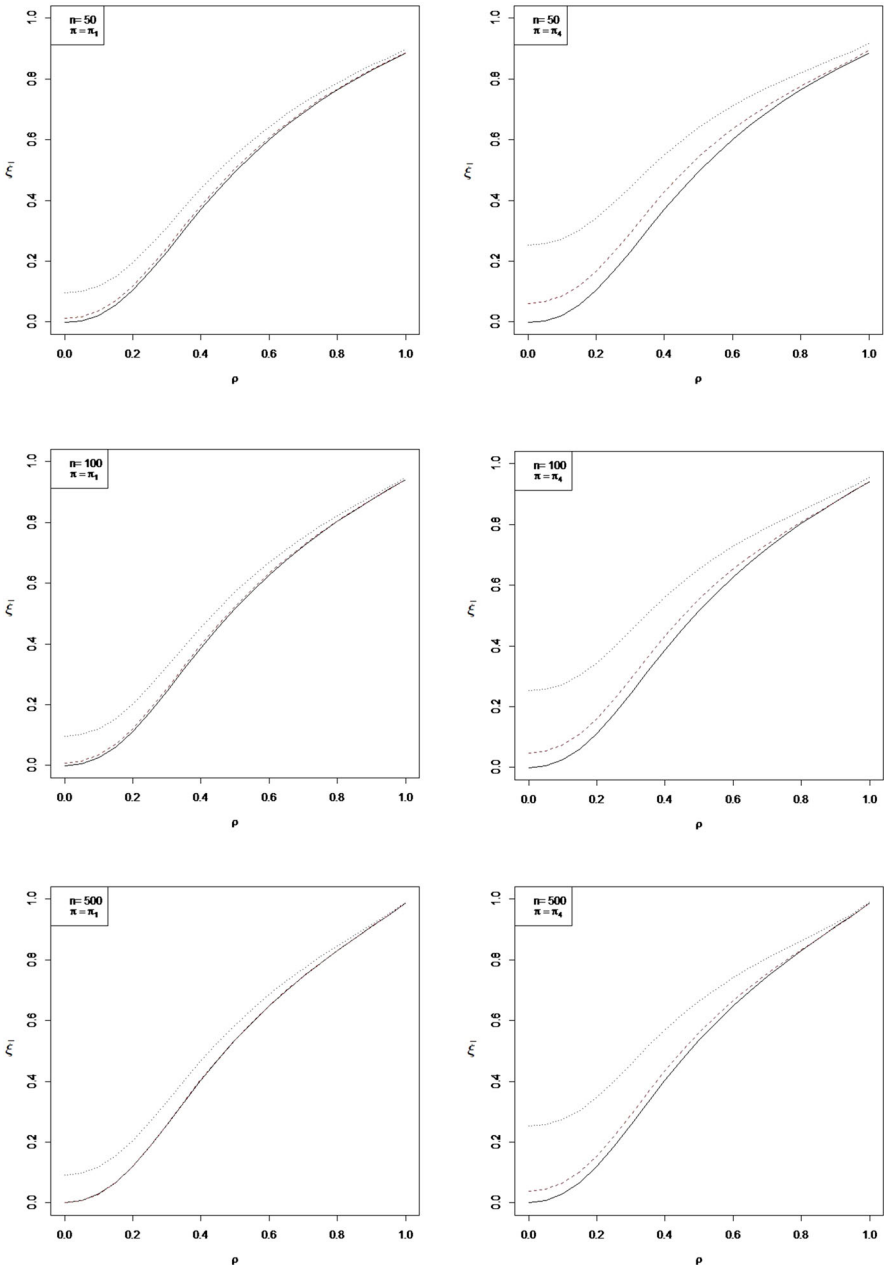


**Fig. 3** Estimated functional dependency for Model 6 (Logarithmic) across different signal strengths $\rho$. Results are shown for the Full data (solid line), Complete Case (CC; dotted line), and Inverse Probability Weighting (IPW; dashed line) estimators. The panels display results for all combinations of sample size ($n = 50, 100, 500$) and missingness mechanism ($\pi_1$ or $\pi_4$)

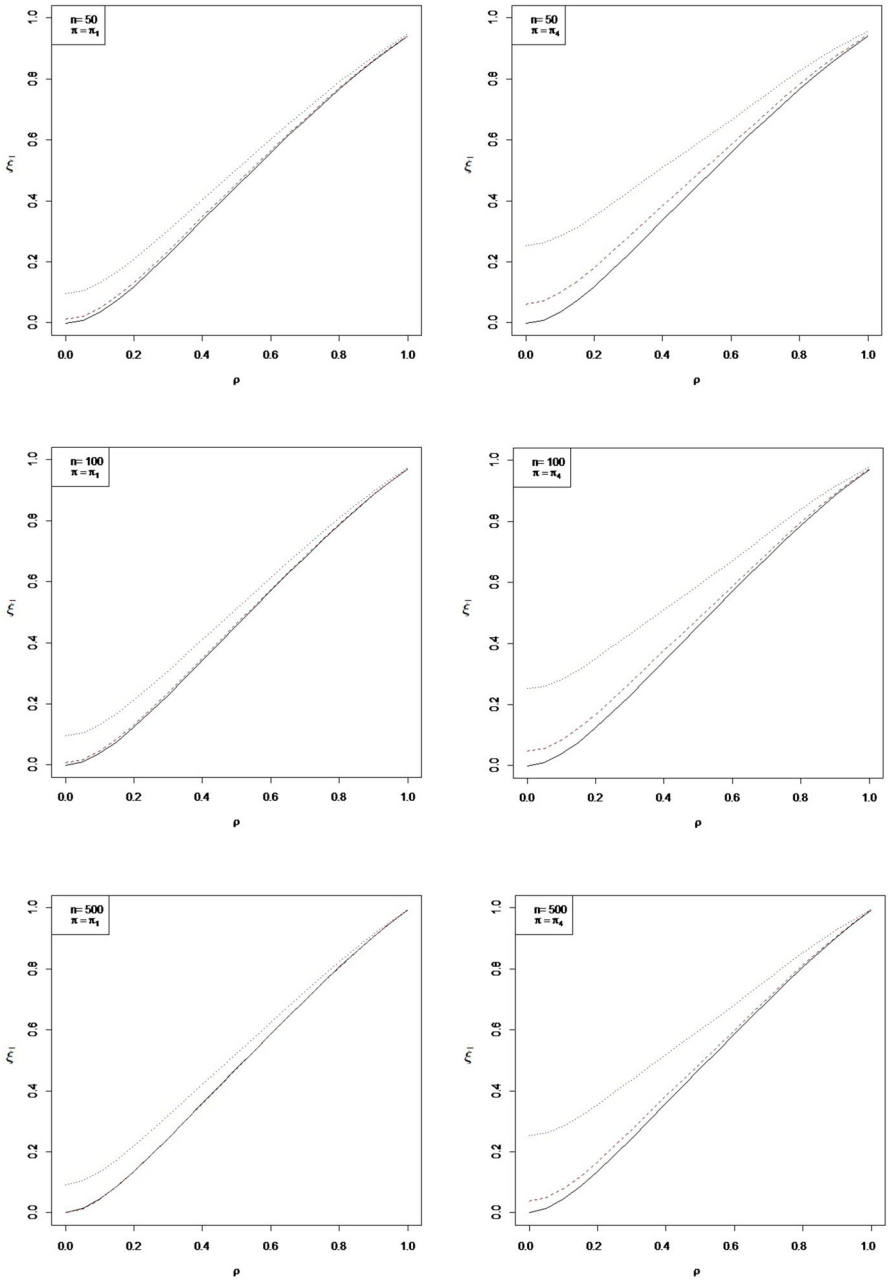


**Fig. 4** Estimated functional dependency for Model 7 (Exponential) across different signal strengths $\rho$. Results are shown for the Full data (solid line), Complete Case (CC; dotted line), and Inverse Probability Weighting (IPW; dashed line) estimators. The panels display results for all combinations of sample size ($n = 50, 100, 500$) and missingness mechanism ($\pi_1$ or $\pi_4$)

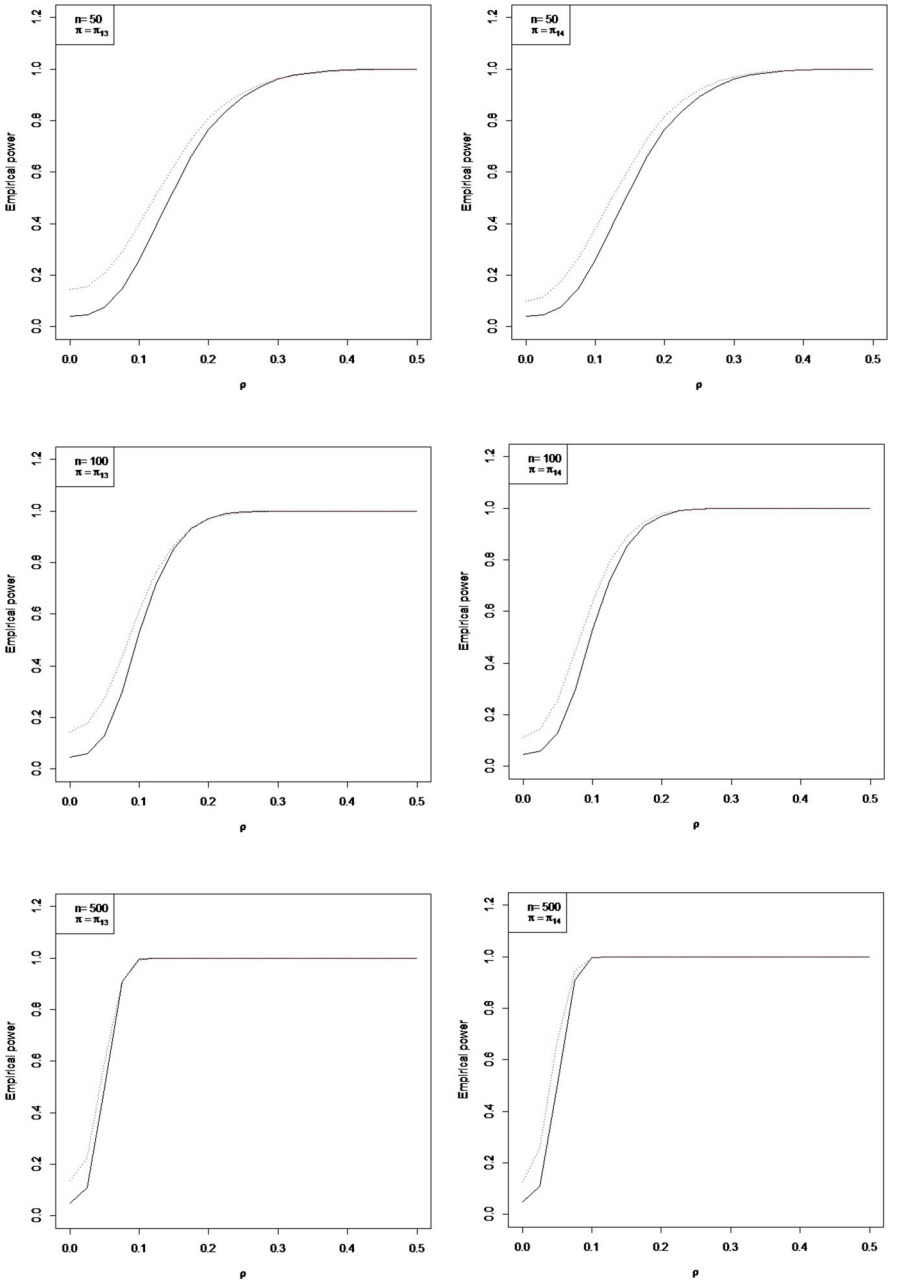


**Fig. 5** Empirical power of the independence test versus different values of the noise parameter $\rho$ for Model 8, comparing the Full (solid curve) and IPW (dotted curve) methods

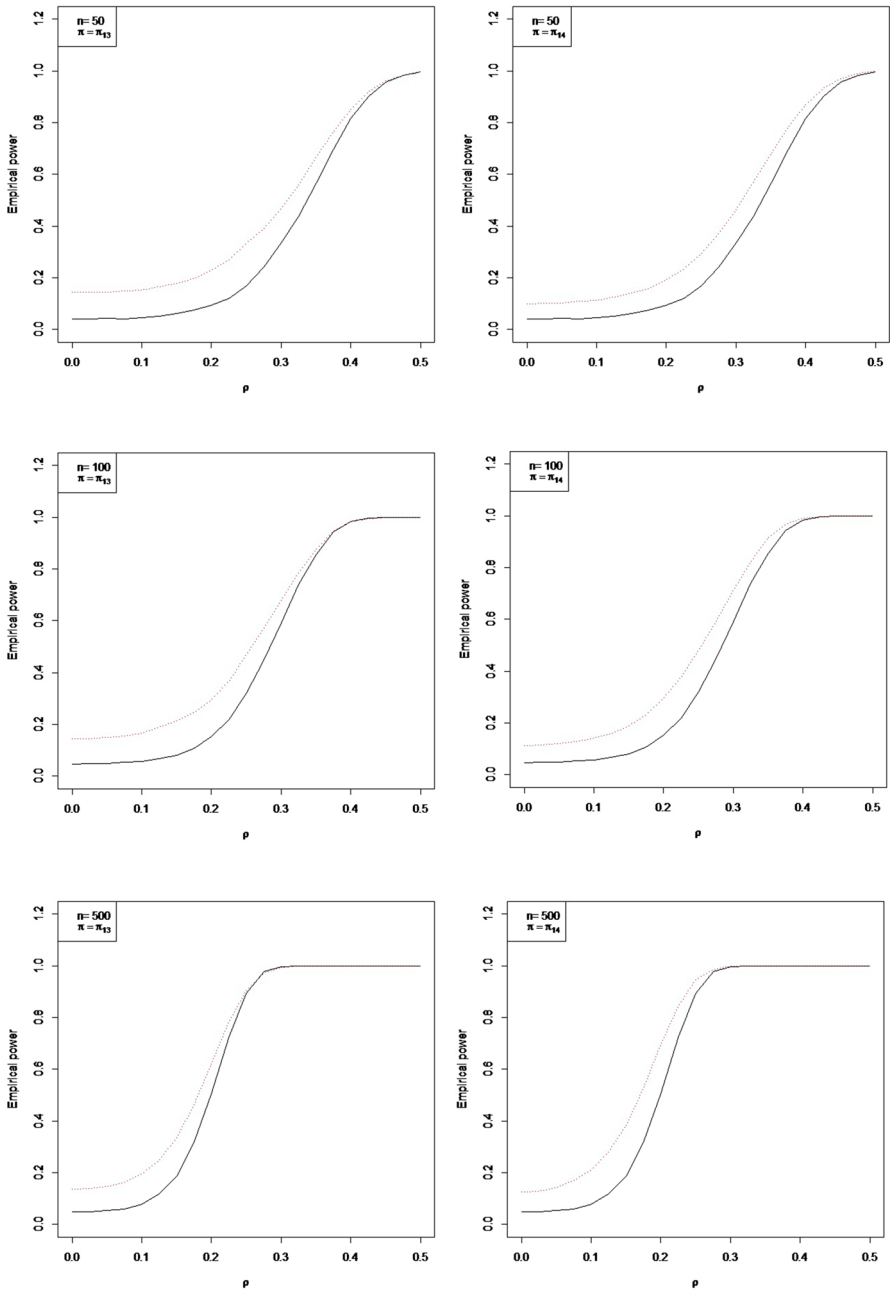


**Fig. 6** Empirical power of the independence test versus different values of the noise parameter $\rho$ for Model 9, comparing the Full (solid curve) and IPW (dotted curve) methods

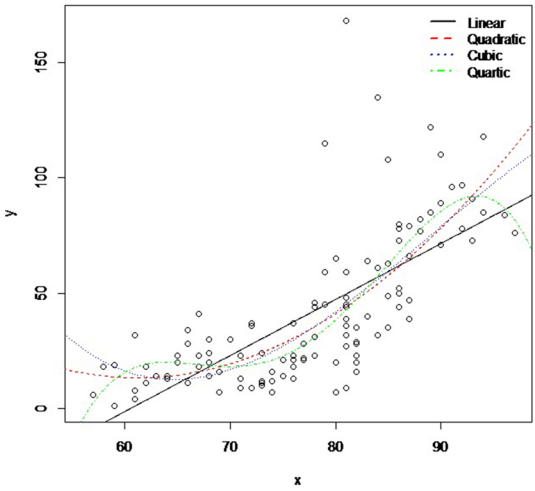


**Fig. 7** Scatter plot of Ozone ($Y$) versus Temperature ($X$) with fitted regression models for the `airquality` dataset

empirical power of the IPW method is closer to that of the Full method. Overall, the proposed IPW estimator of Chatterjee's correlation in the presence of missing data demonstrates strong performance in testing independence. Also, as the sample size increases, the power to reject the independence hypothesis increases. In summary, the empirical power approaches one in all cases, across different models, different sample sizes, missingness mechanisms, and missing rates, for moderate values of the noise parameter

## 4 Real data studies

To illustrate the practical utility of our proposed method, we apply it to two real-world datasets where the variable $Y$ is subject to missingness. For each dataset, we estimate Chatterjee's correlation using the CC and the IPW approaches, demonstrating the performance of our method in realistic settings.

### 4.1 Air quality dataset

We first analyze the `airquality` dataset, available in the R software environment, which provides daily air quality measurements in New York City from May to September 1973. This dataset includes six variables; for our analysis, we treat `Ozone` as the variable $Y$ and `Temp` (temperature) as the variable $X$. The dataset contains 153 observations, with the Ozone variable having 37 missing values, corresponding to a missing rate of approximately 24%. The variables $X$ and $Y$, obtained from the `airquality` dataset, are shown in Fig. 7.

From Fig. 7, it appears that the relationship between $X$ and $Y$ is not linear. Therefore, using classical correlation measures such as Pearson's correlation to assess the dependence of $Y$ on $X$ is not appropriate. In this case, applying Chatterjee's correlation

**Table 4** Chatterjee's correlation estimates for Ozone ($Y$) versus Temperature ($X$) in the `airquality` dataset

| Method | $\sigma^2$ | $\bar{\xi}$ | MSE | 95% Empirical CI |
|---|---|---|---|---|
| CC | 0.01 | 0.7028 | 0.0004 | (0.6899, 0.7414) |
| IPW | | 0.6172 | 0.0007 | (0.6005, 0.6701) |
| CC | 0.1 | 0.6997 | 0.0004 | (0.6005, 0.6701) |
| IPW | | 0.6137 | 0.0008 | (0.5947, 0.6668) |
| CC | 0.25 | 0.6911 | 0.0005 | (0.6760 ,0.7351) |
| IPW | | 0.6009 | 0.0009 | (0.5808, 0.6604) |
| CC | 0.50 | 0.6853 | 0.0006 | (0.6698, 0.7307) |
| IPW | | 0.5924 | 0.0011 | (0.5711, 0.6549) |
| CC | 0.75 | 0.6718 | 0.0006 | (0.6649, 0.7280) |
| IPW | | 0.5874 | 0.0011 | (0.5647, 0.6501) |
| CC | 1.00 | 0.6783 | 0.0006 | (0.6615, 0.7258) |
| IPW | | 0.5874 | 0.0011 | (0.5597, 0.6471) |
| CC | 1.50 | 0.6739 | 0.0007 | (0.6566, 0.7242) |
| IPW | | 0.5763 | 0.0013 | (0.5527, 0.6446) |
| CC | 2.00 | 0.6711 | 0.0007 | (0.6531, 0.7246) |
| IPW | | 0.5725 | 0.0014 | (0.5481, 0.6455) |

Estimates from CC and IPW methods are shown, with the $X$ variable jittered using $N(0, \sigma^2)$ noise to break ties. The mean of estimates ($\bar{\xi}$), mean squared error (MSE), and 95% empirical confidence intervals (CI) for different values of $\sigma^2$ are reported based on 5000 Monte Carlo replications

provides a more meaningful measure of dependence. Since Chatterjee's correlation requires the $X$ variable to be continuous without ties, and `Temp` contains repeated values, we introduced a small jittering noise to break the ties. Specifically, we added independent noise $\epsilon \sim N(0, \sigma^2)$ to the $X$ values. To ensure the stability of our results against the random jitter, we repeated the estimation process 5000 times for different noise levels $\sigma^2$ and report the average estimates. The results of this analysis are presented in Table 4.

Table 4 presents the mean, mean squared error (MSE), and 95% empirical confidence interval of the estimators across 5000 Monte Carlo replications for different values of the jittering parameter $\sigma^2$. The results demonstrate that the estimated correlations are robust to the specific choice of $\sigma^2$, with only minor variations observed as the dispersion of the noise increases. Based on the consistent performance of the IPW method established in our simulation studies, we recommend its estimate for interpreting the relationship between Ozone and Temperature. The IPW method indicates a moderate functional dependency, with a Chatterjee's correlation of approximately 0.61.

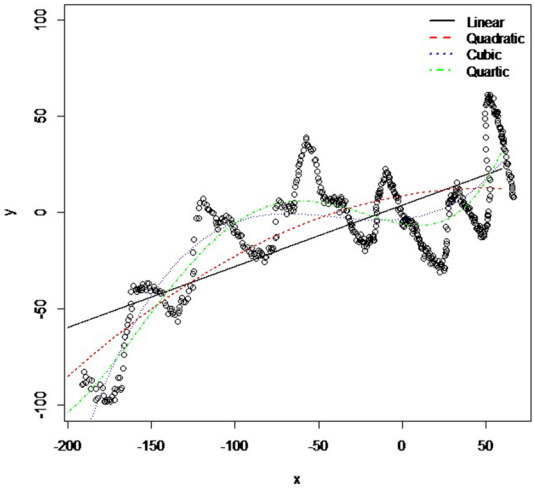


**Fig. 8** Scatter plot of number of unemployed individuals ($Y$) versus personal consumption expenditures ($X$) with fitted regression models for the `unemploy` dataset

### 4.2 Economics dataset

We further validate our method using the `economics` dataset from the `ggplot2` package in R, which contains US economic time series data from 1967 to 2015. This dataset comprises 7 variables; for our analysis, we use the third column, personal consumption expenditures (`pce`), as the $X$ variable, and the seventh column, number of unemployed individuals (`unemploy`), as the $Y$ variable. Both variables are fully observed in the original dataset. The variables $X$ and $Y$, obtained from the `unemploy` dataset, are shown in Fig. 8.

From Fig. 8, it is clear that the relationship between $X$ and $Y$ is nonlinear. Therefore, applying Chatterjee's correlation provides a more meaningful measure of dependence compared to classical correlation measures. To evaluate the performance of our estimators, we artificially introduced missing data into the $Y$ variable according to the following six mechanisms:

15) $\pi_{15}(x) = 0.85$ (MCAR, 15% missing).
16) $\pi_{16}(x) = \dfrac{1}{1+e^{-2-0.01x}}$ (MAR, logistic).
17) $\pi_{17}(x) = \dfrac{1}{1+0.0025 \mid X \mid}$ (MAR, V-shaped).
18) $\pi_{18}(x) = 1 - 0.5\sin^2(2\pi x)$ (MAR, periodic).
19) $\pi_{19}(x) = 1 - 0.5\cos^2(2\pi x)$ (MAR, periodic).
20) $\pi_{20}(x) = 1 + \log(1 + 0.01 \mid X \mid)$ (MAR, logarithmic).

The simulation results, based on 5,000 Monte Carlo replications for each mechanism, are summarized in Table 5.

The original economics dataset is complete, providing a full data benchmark estimate of Chatterjee's correlation of $\xi_n = 0.8760$ (Table 5), indicating a very strong functional dependency between unemployment and personal consumption expenditures. After artificially introducing missingness into the unemployment variable ($Y$)

**Table 5** Chatterjee's correlation estimates for unemployment ($Y$) versus personal consumption expenditures ($X$) in the `economics` dataset

| Method | MM | MR | $\bar{\xi}$ | MSE | 95% Empirical CI |
|---|---|---|---|---|---|
| Full | - | - | 0.8760 | - | - |
| CC | $\pi_{15}$ | 0.15 | 0.8940 | 0.00005 | (0.8893, 0.9081) |
| IPW | | | 0.8768 | 0.00004 | (0.8725, 0.8902) |
| CC | $\pi_{16}$ | 0.18 | 0.8824 | 0.00004 | (0.8779, 0.8954) |
| IPW | | | 0.8772 | 0.00003 | (0.8735, 0.8812) |
| CC | $\pi_{17}$ | 0.12 | 0.8840 | 0.00004 | (0.8796, 0.8971) |
| IPW | | | 0.8767 | 0.00003 | (0.8730, 0.8880) |
| CC | $\pi_{18}$ | 0.26 | 0.9040 | 0.00005 | (0.8994, 0.9179) |
| IPW | | | 0.8746 | 0.00007 | (0.8693, 0.8887) |
| CC | $\pi_{19}$ | 0.24 | 0.9057 | 0.00005 | (0.9009, 0.9195) |
| IPW | | | 0.8772 | 0.00006 | (0.8719, 0.8927) |
| CC | $\pi_{20}$ | 0.28 | 0.9005 | 0.00008 | (0.8842, 0.9181) |
| IPW | | | 0.8781 | 0.00011 | (0.8710, 0.8987) |

The mean of estimates ($\bar{\xi}$), mean squared error (MSE), and 95% empirical confidence intervals (CI) for different values of $\sigma^2$ are shown for the Full, CC, and IPW estimators under six different missing data mechanisms, based on 5000 Monte Carlo replications

under various missing mechanisms, the results in Table 5 demonstrate the robustness of the proposed IPW estimator. While a minor decrease in accuracy is observed with higher missing rates, the IPW estimates remain consistently close to the full data benchmark across all missingness scenarios. In contrast, the CC estimator shows substantially greater deviation and sensitivity to the missing data mechanism and rate. In addition, the estimated Chatterjee's correlation lies within the empirical 95% confidence interval obtained using the IPW method in all cases. However, this is not true for the empirical 95% confidence interval produced by the CC method.

## 5 Discussion and Conclusion

This paper has addressed the estimation of Chatterjee's correlation in the presence of missing data, operating under the MAR assumption. Our simulation studies demonstrate that the proposed IPW method performs robustly not only under MAR but also under MCAR mechanisms, as evidenced in Studies 1 and 2. This robustness is valuable in practice, as the true missingness mechanism is often unknown and the MNAR mechanism is not testable. Following the guidance of Little and Rubin (2019), employing a MAR-based methodology provides a principled and often satisfactory approach, allowing for valid inference even when the true mechanism is MNAR. Inference under a strict MNAR mechanism is particularly challenging because the mechanism is inherently untestable from the observed data alone, and the propensity score depends on the unobserved values themselves. While methods exist for MNAR data, such as the exponential tilting approach proposed by Bahari et al. (2021) for regression analy-

sis, they require strong assumptions about the nature of the missingness. Therefore, the MAR framework remains a practical and widely adopted foundation for handling missing data. The propensity scores required for the IPW estimator are unknown in practice and must be estimated. While we employed a kernel smoothing approach with a normal kernel and cross-validated bandwidth, alternative methods, such as parametric logistic regression, could also be used. The consistency of the final estimator relies on the consistent estimation of these propensity scores. Based on our comprehensive simulation results and real-data applications, we strongly recommend the use of the IPW method over the CC method, as the latter introduces significant bias across all scenarios considered. While the IPW method presented here provides a reliable solution, future research could explore more sophisticated approaches, such as augmented inverse probability weighting (AIPW), which may offer improved efficiency and robustness. Extending Chatterjee's correlation using such doubly robust methods may be our next study.

## A Appendix: Proofs of Lemmas

This appendix provides detailed proofs of Lemmas 1 and 2. The proof of Lemma 1 is presented in the following subsection, followed by the proof of Lemma 2 in the subsequent subsection. For notational simplicity in the proofs that follow, we adopt these conventions:

$$\begin{aligned}
&\pi(x_i) \equiv \pi_i, && \hat{\pi}(x_i) \equiv \hat{\pi}_i,\\
&\pi(x_{(j)}) \equiv \pi_{[j]}, && \hat{\pi}(x_{(j)}) \equiv \hat{\pi}_{[j]},\\
&\xi_n(X,Y) \equiv \xi_n, && \xi_n^*(X,Y) \equiv \xi_n^*,\\
&\xi_n^{CC}(X,Y) \equiv \xi_n^{CC}, && \xi_n^{IPW}(X,Y) \equiv \xi_n^{IPW},\\
&h_n \equiv h.
\end{aligned}$$

### A.1 Proof of Lemma 1.

For any index $j$ where $y_{[j]}$ is observed (i.e., where $\delta_j = 1$), we have:

$$R_{[j]}^{IPW} = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} I(y_i \le y_{[j]})$$

Applying a first-order Taylor expansion to the propensity score $\hat{\pi}_i$ around $\pi_i$, we obtain:

$$\begin{aligned}
R_{[j]}^{IPW} = &\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} I(y_i \le y_{[j]})\\
&+ \sum_{i=1}^{n} \frac{\delta_i}{\pi_i^2} (\pi_i - \hat{\pi}_i) I(y_i \le y_{[j]})
\end{aligned}$$

$$+\sum_{i=1}^{n}\delta_i I(y_i \le y_{[j]})O_P((\pi_i-\hat{\pi}_i)^2)$$
$$:= A + B + C. \tag{A.1}$$

We now analyze the asymptotic behavior of the terms $B$ and $C$ separately.

$$C = \sum_{i=1}^{n}\delta_i I(y_i \le y_{[j]})O_P((\pi_i-\hat{\pi}_i)^2)$$

By conditions (C1) and (C2) and standard results in kernel estimation (Wang and Wang 2001), we have $|\hat{\pi}_i - \pi_i| = O_P(h^2 + \frac{1}{\sqrt{nh}})$. Consequently, $|\hat{\pi}_i - \pi_i|^2 = O_P(h^4 + \frac{1}{nh})$. On the other hand, $|\delta_i I(y_i \le y_{[j]})| \le 1$ for all $i, j$, the term $C$ satisfies:

$$C = \sum_{i=1}^{n} O_P(h^4 + \frac{1}{nh}),$$

Since $h^4 = o(h^2)$ and $\frac{1}{nh} = o(\frac{1}{\sqrt{nh}})$ as $h \to 0$ and $n \to \infty$, we conclude that:

$$C = \sum_{i=1}^{n} O_P(h^2 + \frac{1}{\sqrt{nh}}) = O_P(nh^2 + \sqrt{\frac{n}{h}}). \tag{A.2}$$

For B statement we have:

$$B = \sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}(\pi_i - \hat{\pi}_i)I(y_i \le y_{[j]}),$$

by applying $|\hat{\pi}_i - \pi_i| = O_P(h^2 + \frac{1}{\sqrt{nh}})$, we can write,

$$B = \sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2} I(y_i \le y_{[j]})O_P(h^2 + \frac{1}{\sqrt{nh}}),$$

Furthermore, by condition (C1) which states $\inf_x \pi(x) > \zeta > 0$, the term $\frac{\delta_i}{\pi_i^2} I(y_i \le y_{[j]})$ is bounded from above. Therefore,

$$B = O_P(h^2 + \frac{1}{\sqrt{nh}}). \tag{A.3}$$

Substituting the results from Equations (A.2) and (A.3) into Equation (A.1) yields:

$$R_{[j]}^{IPW} = \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} I(y_i \le y_{[j]}) + O_P(h^2 + \frac{1}{\sqrt{nh}}) + O_P(nh^2 + \sqrt{\frac{n}{h}})$$
$$= \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} I(y_i \le y_{[j]}) + O_P(nh^2 + \sqrt{\frac{n}{h}}) \quad (A.4)$$

This completes the proof of part (a). Using this result, we now prove part (b). From part (a), we have:

$$|R_{[j+1]}^{IPW} - R_{[j]}^{IPW}| = |R_{[j+1]}^{*} - R_{[j]}^{*}| + O_P(nh^2 + \sqrt{\frac{n}{h}}). \quad (A.5)$$

Now, by applying the result from Equation (A.5) and performing a first-order Taylor expansion of $\hat{\pi}_i$ around $\pi_i$, we obtain:

$$\frac{\delta_{[j]}}{\hat{\pi}_{[j]}} \frac{\delta_{[j+1]}}{\hat{\pi}_{[j+1]}} |R_{[j+1]}^{IPW} - R_{[j]}^{IPW}| = [\frac{\delta_{[j]}}{\pi_{[j]}} \frac{\delta_{[j+1]}}{\pi_{[j+1]}} + \frac{\delta_{[j]}}{\pi_{[j]}^2} \frac{\delta_{[j+1]}}{\pi_{[j+1]}} (\pi_{[j]} - \hat{\pi}_{[j]})$$
$$+ \frac{\delta_{[j]}}{\pi_{[j]}} \frac{\delta_{[j+1]}}{\pi_{[j+1]}^2} (\pi_{[j+1]} - \hat{\pi}_{[j+1]})$$
$$+ O_P(h^4 + \frac{1}{nh})] \times [|R_{[j+1]}^{*} - R_{[j]}^{*}|$$
$$+ O_P(nh^2 + \sqrt{\frac{n}{h}})].$$

Expanding the product of the two bracketed expressions yields:

$$\frac{\delta_{[j]}}{\hat{\pi}_{[j]}} \frac{\delta_{[j+1]}}{\hat{\pi}_{[j+1]}} |R_{[j+1]}^{IPW} - R_{[j]}^{IPW}| = \frac{\delta_{[j]}}{\pi_{[j]}} \frac{\delta_{[j+1]}}{\pi_{[j+1]}} |R_{[j+1]}^{*} - R_{[j]}^{*}|$$
$$+ \frac{\delta_{[j]}}{\pi_{[j]}^2} \frac{\delta_{[j+1]}}{\pi_{[j+1]}} (\pi_{[j]} - \hat{\pi}_{[j]}) |R_{[j+1]}^{*} - R_{[j]}^{*}|$$
$$+ \frac{\delta_{[j]}}{\pi_{[j]}} \frac{\delta_{[j+1]}}{\pi_{[j+1]}^2} (\pi_{[j+1]} - \hat{\pi}_{[j+1]}) |R_{[j+1]}^{*} - R_{[j]}^{*}|$$
$$+ |R_{[j+1]}^{*} - R_{[j]}^{*}| O_P(h^4 + \frac{1}{nh})$$
$$+ \frac{\delta_{[j]}}{\pi_{[j]}} \frac{\delta_{[j+1]}}{\pi_{[j+1]}} O_P(nh^2 + \sqrt{\frac{n}{h}}) + O_P(nh^2 + \sqrt{\frac{n}{h}}) \quad (A.6)$$

The final expression is obtained through standard algebraic manipulation and simplification of the remaining product. Note that $|R_{[j+1]}^{*} - R_{[j]}^{*}| = O(n)$ and $(\hat{\pi}_i - \pi_i) =$

$O_P(h^2 + \frac{1}{\sqrt{nh}})$. Therefore, under the assumed regularity conditions, it follows that:

$$\begin{aligned}
&\frac{\delta_{[j]}}{\pi^2_{[j]}}\frac{\delta_{[j+1]}}{\pi_{[j+1]}}(\pi_{[j]} - \hat{\pi}_{[j]})|R^*_{[j+1]} - R^*_{[j]}| = O_P(nh^2 + \sqrt{\frac{n}{h}}),\\
&\frac{\delta_{[j]}}{\pi_{[j]}}\frac{\delta_{[j+1]}}{\pi^2_{[j+1]}}(\pi_{[j+1]} - \hat{\pi}_{[j+1]})|R^*_{[j+1]} - R^*_{[j]}| = O_P(nh^2 + \sqrt{\frac{n}{h}}),\\
&|R^*_{[j+1]} - R^*_{[j]}|O_P(h^4 + \frac{1}{nh}) = O_P(nh^4 + \frac{1}{h}),\\
&\frac{\delta_{[j]}}{\pi_{[j]}}\frac{\delta_{[j+1]}}{\pi_{[j+1]}}O_P(nh^2 + \sqrt{\frac{n}{h}}) = O_P(nh^2 + \sqrt{\frac{n}{h}}).
\end{aligned}$$

Summing both sides of the above equation from $j = 1$ to $n - 1$ yields:

$$\begin{aligned}
\sum_{j=1}^{n-1}\frac{\delta_{[j]}}{\hat{\pi}_{[j]}}\frac{\delta_{[j+1]}}{\hat{\pi}_{[j+1]}}|R^{IPW}_{[j+1]} - R^{IPW}_{[j]}| &= \sum_{j=1}^{n-1}\frac{\delta_{[j]}}{\pi_{[j]}}\frac{\delta_{[j+1]}}{\pi_{[j+1]}}|R^*_{[j+1]} - R^*_{[j]}| + (n-1)O_P(nh^2 + \sqrt{\frac{n}{h}})\\
&\quad + (n-1)O_P(nh^2 + \sqrt{\frac{n}{h}}) + (n-1)O_P(nh^4 + \frac{1}{h})\\
&\quad + (n-1)O_P(nh^2 + \sqrt{\frac{n}{h}}) + (n-1)O_P(nh^2 + \sqrt{\frac{n}{h}})\\
&= \sum_{j=1}^{n-1}\frac{\delta_{[j]}}{\pi_{[j]}}\frac{\delta_{[j+1]}}{\pi_{[j+1]}}|R^*_{[j+1]} - R^*_{[j]}| + O_P(n^2h^2 + n\sqrt{\frac{n}{h}}). \quad \text{(A.7)}
\end{aligned}$$

Substituting this result into the expression for $\xi_n^{IPW}$ and simplifying yields:

$$\begin{aligned}
\xi_n^{IPW} &= 1 - \frac{3\sum_{j=1}^{n-1}\frac{\delta_{[j+1]}}{\pi_{[j+1]}}\frac{\delta_{[j]}}{\pi_{[j]}}|R^*_{[j+1]} - Rj^*|}{n^2 - 1} + O_P(h^2 + \frac{1}{\sqrt{nh}})\\
&= \xi_n^* + O_P(h^2 + \frac{1}{\sqrt{nh}}). \quad \text{(A.8)}
\end{aligned}$$

The Equation (A.8) establishes the result in part (b) of Lemma 1, thereby completing the proof.

### A.2 Proof of Lemma 2.

To prove part (a), we analyze the expectation and variance of the estimated ranks for observed indices $j$ (where $\delta_j = 1$). We proceed as follows:

$$
\begin{aligned}
E(R^*_{[j]} \mid X, Y) &= E(\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} I(y_i \le y_{[j]}) \mid X, Y) \\
&= \sum_{i=1}^{n} E(\frac{\delta_i}{\pi_i} I(y_i \le y_{[j]}) \mid X, Y) \\
&= \sum_{i=1}^{n} I(y_i \le y_{[j]}) E(\frac{\delta_i}{\pi_i} \mid X, Y) \\
&= R_{[j]}, \qquad \text{(A.9)}
\end{aligned}
$$

In Equation (A.9), under the MAR assumption, we have $E(\frac{\delta_i}{\pi_i} \mid X, Y) = 1$. Furthermore,

$$
\begin{aligned}
var(R^*_{[j]} \mid X, Y) &= var(\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} I(y_i \le y_{[j]}) \mid X, Y) \\
&= \sum_{i=1}^{n} var(\frac{\delta_i}{\pi_i} I(y_i \le y_{[j]}) \mid X, Y) \\
&= \sum_{i=1}^{n} \frac{I(y_i \le y_{[j]})}{\pi_i^2} var(\frac{\delta_i}{\pi_i} \mid X, Y) \\
&= \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i} I(y_i \le y_{[j]}) \le \frac{1-\zeta}{\zeta} \sum_{i=1}^{n} I(y_i \le y_{[j]}) = Mn. \qquad \text{(A.10)}
\end{aligned}
$$

In Equation (A.10), the first equality follows from the independence of the $\delta_i$'s. The fourth equality utilizes the identity $var(\delta_i \mid X, Y) = \pi_i(1-\pi_i)$. The subsequent inequality is a consequence of the first regularity condition (C1). The final equality holds because the summation is asymptotically of order $n$, where $M$ is a positive constant. Therefore, we obtain the following expression for the variance:

$$
var(R^*_{[j]} \mid X, Y) = O(n). \qquad \text{(A.11)}
$$

Thus, by combining the results from Equations (A.9) and (A.11) and applying Chebyshev's inequality, we conclude that:

$$
R^*_{[j]} = R_{[j]} + O_P(\sqrt{n})
$$

This establishes part (a) of Lemma 2. To prove part (b), we decompose the rank estimator as $R^*_{[j]} = R_{[j]} + e_{[j]}$, where the error term is defined as $e_{[j]} = \sum_{i=1}^{n} (\frac{\delta_i}{\pi_i} - 1) I(y_i \le y_{[j]})$. From part (a),

$$
E(e_{[j]} \mid X, Y) = 0,
$$

and, by following the same reasoning applied in the derivation of Equation (A.10), we conclude that:

$$var(e_{[j]} \mid X, Y) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i} I(y_i \leq y_{[j]}) = O(n).$$

Thus, $e_{[j]} = O_P(\sqrt{n})$. Substituting this into the decomposition $R^*_{[j]} = R_{[j]} + e_{[j]}$, we obtain:

$$|R^*_{[j+1]} - R^*_{[j]}| = |R_{[j+1]} - R_{[j]} - (e_{[j+1]} - e_{[j]})|.$$

Thus, $||R^*_{[j+1]} - R^*_{[j]}| - |R_{[j+1]} - R_{[j]}|| \leq |e_{[j+1]} - e_{[j]}| = O_p(\sqrt{n})$. This implies that,

$$|R^*_{[j+1]} - R^*_{[j]}| = |R_{[j+1]} - R_{[j]}| + O_p(\sqrt{n}).$$

Therefore, we obtain the following result:

$$\begin{aligned} E(\xi^*_n \mid X, Y) &= 1 - \frac{3\sum_{j=1}^{n-1}[|R_{[j+1]} - R_{[j]}| + O_p(\sqrt{n})]}{n^2-1} \\ &= \xi_n + \frac{nO(\sqrt{n})}{n^2-1} \\ &= \xi_n + O(\frac{1}{\sqrt{n}}). \end{aligned} \tag{A.12}$$

Furthermore,

$$\begin{aligned} var(\xi^*_n \mid X, Y) &= var(1 - \frac{3\sum_{j=1}^{n-1}[|R_{[j+1]} - R_{[j]}| + O_p(\sqrt{n})]}{n^2-1} \mid X, Y) \\ &= (\frac{3}{n^2-1})^2 \sum_{j=1}^{n-1} var(\frac{\delta_{[j+1]}}{\pi_{[j+1]}} \frac{\delta_{[j]}}{\pi_{[j]}} |R_{[j+1]} - R_{[j]}| \mid X, Y) \\ &= (\frac{3}{n^2-1})^2 \sum_{j=1}^{n-1} (\frac{1-\pi_{[j+1]}\pi_{[j]}}{\pi_{[j+1]}\pi_{[j]}}) |R_{[j+1]} - R_{[j]}|^2 \\ &\leq (\frac{3}{n^2-1})^2 . M^* n^3, \end{aligned} \tag{A.13}$$

where $M^*$ is a bounded positive constant. Therefore, we conclude that:

$$var(\xi^*_n \mid X, Y) = O(\frac{1}{n}). \tag{A.14}$$

Then, by combining the results from Equations (A.12) and (A.14) and applying Chebyshev's inequality, we establish part (b) of Lemma 2, which completes the proof.

## Declarations

**Conflict of interests** We declare that there are no conflicts of interest in this article.

## References

Auddy A, Deb N, Nandy S (2021) Exact detection thresholds and minimax optimality of Chatterjee's correlation coefficient. arXiv preprint arXiv:2104.15140
Azadkia M, Chatterjee S (2021) A simple measure of conditional dependence. Ann Stat 49(6):3070–3102
Bahari F (2025) A new algorithm for variable selection in general linear models. Jap J Stat Data Sci. https://doi.org/10.1007/s42081-025-00303-3
Bahari F, Parsi S, Ganjali M (2021) Empirical likelihood inference in general linear model with missing values in response and covariates by MNAR mechanism. Stat Pap 62(2):591–622
Chatterjee S (2021) A new coefficient of correlation. J Am Stat Assoc 116(536):2009–2022
Creemers A, Aerts M, Hens N, Molenberghs G (2012) A nonparametric approach to weighted estimating equations for regression analysis with missing covariates. Comput Stat Data Anal 56(1):100–113
Fuchs S, Wang Y (2024) Hierarchical variable clustering based on the predictive strength between random vectors. Int J Approximate Reasoning 170:109185
Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163
Lin Z, Han F (2023) On boosting the power of Chatterjee's rank correlation. Biometrika 110(2):283–299
Lin Z, Han F (2022) Limit theorems of Chatterjee's rank correlation. arXiv preprint arXiv:2204.08031
Little RJ (1992) Regression with missing X's: a review. J Am Stat Assoc 87(420):1227–1237
Little RJ, Rubin DB (2019) Statistical analysis with missing data. John Wiley & Sons
Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592
Thottolil R, Kumar U, Chakraborty T (2023) Prediction of transportation index for urban patterns in small and medium-sized Indian cities using hybrid RidgeGAN model. Sci Rep 13(1):21863
Wang S, Wang CY (2001) A note on kernel assisted estimators in missing covariate regression. Stat Probabil Lett 55(4):439–449
Xia L, Cao R, Du J, Chen X (2025) The improved correlation coefficient of Chatterjee. J Nonparametric Stat 37(2):265–281
Zhang Q (2023) On the asymptotic null distribution of the symmetrized Chatterjee's correlation coefficient. Stat Probabil Lett 194:109759