Received: 26 May 2021 | Revised: 2 March 2022 | Accepted: 22 March 2022

DOI: 10.1111/biom.13666

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Score test for missing at random or not under logistic missingness models

Hairu Wang | **Zhiping Lu** | **Yukun Liu**

KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai, China

Correspondence
Yukun Liu, KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai 200062, China.
Email: ykliu@sfs.ecnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 11771144, 32030063, 71931004; the Natural Science Foundation of Shanghai, Grant/Award Number: 17ZR1409000

Abstract

Missing data are frequently encountered in various disciplines and can be divided into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Valid statistical approaches to missing data depend crucially on correct identification of the underlying missingness mechanism. Although the problem of testing whether this mechanism is MCAR or MAR has been extensively studied, there has been very little research on testing MAR versus MNAR. A critical challenge that is faced when dealing with this problem is the issue of model identification under MNAR. In this paper, under a logistic model for the missing probability, we develop two score tests for the problem of whether the missingness mechanism is MAR or MNAR under a parametric model and a semiparametric location model on the regression function. The implementation of the score tests circumvents the identification issue as it requires only parameter estimation under the null MAR assumption. Our simulations and analysis of human immunodeficiency virus data show that the score tests have well-controlled type I errors and desirable powers.

KEYWORDS
missing at random, missing not at random, score test

1 | INTRODUCTION

Missing data are frequently encountered in economic, medical, and social science disciplines. Valid statistical inferences for missing data depend crucially on correct identification of the underlying missingness mechanism, which was divided by Rubin (1976) into three categories. The missingness is called missing at random (MAR) or ignorable if it does not depend on the missing values themselves conditioning on the observed data, and it is called missing not at random (MNAR) or nonignorable otherwise. A degenerate case of MAR is missing completely at random (MCAR), where the missingness does not depend on either the observed or the missing data.

Methods for handling MAR data and MNAR data are generally different. For MAR data, both the propensity score and outcome regression models are nonparametrically identifiable, and it is therefore always tractable to conduct valid inferences (Little and Rubin, 2019; Tsiatis, 2006; Kim and Shao, 2013). Things become much more challenging when data are MNAR, as the underlying model is often not identifiable based on the observed data. A popular tool to overcome the identifiability issue is an "instrumental variable" (Wang et al., 2014) or "ancillary variable" (Miao and Tchetgen Tchetgen, 2016), which does not affect the missingness but may affect the conditional distribution of the response variable. However, an instrumental variable may not be readily available or may not be straightforward to find in practice, which complicates the identifiability

and inferences of the existing statistical approaches. See Tang and Ju (2018) and Wang and Kim (2021) for more comprehensive reviews of statistical inferences for nonignorable missing-data problems.

These discussions arguably reveal that methods for handling MAR data and MNAR data are totally different: The former are relatively easy whereas the latter are much more difficult. The research in this paper is motivated by the analysis of an human immunodeficiency virus (HIV) data, where the response variable subject to missingness was assumed to be MAR by Hammer et al. (1996) and Han et al. (2019) but to be MNAR by Liu et al. (2021) and Zhang *et al.* (2020). As correctly specifying the underlying missingness mechanism is crucial to the subsequent development of valid inference methods, this raises the hypothesis testing problem of whether the missingness mechanism is MAR or MNAR.

A relative simple counterpart of this hypothesis testing problem is whether the missingness mechanism is MCAR or MAR. Many tests for MCAR have been provided in recent decades, since the MCAR category is at the center of interest of many behavioral and social scientists confronted with missing data (Simon and Simonoff, 1986). Little (1988) constructed a test by comparing the means of recorded values of each variable between groups of different missingness patterns. Chen and Little (1999) extended Little's test to longitudinal data by comparing the means of the general estimating equations across different missingness patterns with zero, with any departure from zero then indicating rejection of the MCAR hypothesis. More extensions of Little's idea to comparisons of the means, the covariance matrices, and/or the distributions across different missingness patterns have also been investigated (see, eg, Jamshidian and Jalal, 2010; Kim and Bentler, 2002; Li and Yu, 2015). Recently, Zhang *et al.* (2019) proposed a nonparametric approach for testing MCAR based on empirical likelihood (Owen, 1988, 1990, 2001). Their approach also provides a unified procedure for estimation after the MCAR hypothesis has been rejected.

In contrast to MCAR, testing for MAR has not received much attention so far. The first contribution in this direction was the nonparametric test proposed by Breunig (2019), which was based on an integrated squared distance. Under a generalized linear regression model, Duan et al. (2020) proposed to test MAR by a quadratic form of the difference of the estimators of the regression coefficient under the MAR and MNAR assumptions, respectively. To the best of our knowledge, these are the only two formal tests for MAR. They both require the existence of an instrumental variable to guarantee identifiability, because their test statistics depend on consistent estimates under MNAR. However, as mentioned previously, the identification of an instrumental variable is usually not straightforward, and, even worse, it may not exist. Also, consistent parameter estimation itself under MNAR is rather challenging.

In this paper, we propose two score tests for MAR under a linear logistic model when a completely parametric model and a semiparametric location model, are imposed on the outcome regression model, respectively. Compared with the tests proposed by Breunig (2019) and Duan et al. (2020), the new tests have at least three advantages. The first remarkable advantage is that no identification condition is required under MNAR because the implementations of the new tests require only parameter estimation under MAR. This implies even if there is no instrument variable, the new tests are still applicable. However, without an instrumental variable, the tests of Breunig (2019) and Duan et al. (2020) may fail to work. The second advantage is that the new tests involve much easier calculations, because the underlying unknown parameters need only be estimated under MAR. As we have pointed out, identifiability is not an issue for MAR data and parameter estimation has been well studied. Third, our simulation results indicate that the two new tests are often more powerful than or at least comparable to that proposed by Duan et al. (2020).

The rest of this paper is organized as follows. Section 2 presents the model set-up, the proposed two score tests, and establishes the limiting distributions of the new tests under MAR. Their generalizations and optimalities are also discussed. In Section 3, we report a simulation study to investigate the finite-sample performance of the score tests. In Section 4, we apply the proposed score tests to analyze a HIV data. Section 5 concludes with a discussion. For clarity, all technical details are provided in the supporting information.

2 | SCORE TEST

Let Y denote an outcome that is subject to missingness and let $\mathbf{X}$ be a fully observed covariate vector whose first component is 1. We denote by D the missingness indicator of the outcome, with $D = 1$ if Y is observed and 0 otherwise. We wish to test whether the missingness mechanism is MAR or MNAR, namely, $H_0 : \text{pr}(D = 1|Y, \mathbf{X}) = \text{pr}(D = 1|\mathbf{X})$. Molenberghs et al. (2008) pointed out that every missingness not at random model has a missingness at random counterpart with equal fit. This implies that the MAR assumption is not testable against a general MNAR alternative unless some model assumptions are imposed. We assume that the nonmissingness probability or the propensity score follows a linear logistic model

$$\text{pr}(D = 1|\mathbf{X} = \mathbf{x}, Y = y) = \pi(\mathbf{x}^\top \boldsymbol{\beta} + \gamma y) \tag{1}$$

with $\pi(t) = e^t/(1+e^t)$. Under model (1), the testing problem of interest is equivalent to $H_0 : \gamma = 0$, because the missingness mechanism is MAR if $\gamma = 0$ and MNAR otherwise.

Suppose that $\{(d_i, d_i y_i, \mathbf{x}_i), i = 1, 2, \ldots, n\}$ are n independent and identically distributed observations from $(D, DY, \mathbf{X})$. Let $f(y|\mathbf{x})$ denote the conditional density function of Y given $\mathbf{X} = \mathbf{x}$. The loglikelihood based on the observed data is

$$\ell(\gamma, \boldsymbol{\beta}, f) = \sum_{i=1}^{n} \Big[d_i \{\log \pi(\mathbf{x}_i^\top \boldsymbol{\beta} + \gamma y_i) + \log f(y_i|\mathbf{x}_i)\} + (1-d_i) \log \int \{1 - \pi(\mathbf{x}_i^\top \boldsymbol{\beta} + \gamma y)\} f(y|\mathbf{x}_i) dy \Big].$$

The likelihood ratio test is the most natural and preferable for testing $\gamma = 0$. Unfortunately, Miao et al. (2016) showed that parameter identifiability is not guaranteed even when a parametric model is postulated for $f(y|\mathbf{x})$. Without parameter identifiability, consistent parameter estimation is not feasible, and therefore nor is the likelihood ratio test, under general parametric assumptions because these require consistent parameter estimation under the null and alternative hypotheses. The Wald test has the same problem.

The score test was introduced by Rao (1948) as an alternative to the likelihood ratio test and Wald test. The most significant advantage of the score statistic is that it depends only on estimates of parameters under H_0; in other words, it automatically circumvents the notorious identifiability issue under MNAR. This motivates us to consider testing $\gamma = 0$ by a score test. Let ∇_γ denote the partial differential operator with respect to γ. The score function with respect to γ at $\gamma = 0$ is

$$\nabla_\gamma \ell(\gamma, \boldsymbol{\beta}, f)|_{\gamma=0} = \sum_{i=1}^{n} \big[d_i \{1 - \pi(\mathbf{x}_i^\top \boldsymbol{\beta})\} y_i - (1-d_i) \pi(\mathbf{x}_i^\top \boldsymbol{\beta}) \mu(\mathbf{x}_i) \big],$$

which depends on the unknown parameters $\boldsymbol{\beta}$ and $\mu(\mathbf{x}) = \int y f(y|\mathbf{x}) dy$.

The score test statistic is constructed by replacing $\boldsymbol{\beta}$ and $\mu(\cdot)$ with their estimators under the null hypothesis $H_0 : \gamma = 0$. The likelihood function under H_0 becomes

$$\ell_0(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \Big[d_i \log \pi(\mathbf{x}_i^\top \boldsymbol{\beta}) + d_i \log f(y_i|\mathbf{x}_i) + (1-d_i) \log\{1 - \pi(\mathbf{x}_i^\top \boldsymbol{\beta})\} \Big].$$

In this situation, a natural estimator for $\boldsymbol{\beta}$ is the maximum likelihood estimator $\widehat{\boldsymbol{\beta}} = \arg\max_\beta \ell_1(\boldsymbol{\beta})$, where $\ell_1(\boldsymbol{\beta}) = \sum_{i=1}^n [d_i \log \pi(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1-d_i) \log\{1 - \pi(\mathbf{x}_i^\top \boldsymbol{\beta})\}]$ is the likelihood function of $\boldsymbol{\beta}$ under the null hypothesis. Estimation of $\mu(\cdot)$ depends on model assumptions on $f(y|\mathbf{x}) = \text{pr}(Y = y|\mathbf{X} = \mathbf{x})$ or $f(y|\mathbf{x}, D = 1) = \text{pr}(Y = y|\mathbf{X} = \mathbf{x}, D = 1)$ as they are the same under MAR, although the latter is checkable with available data but the former is not. To finish the construction of the score test, we consider postulating either a fully parametric or semiparametric model on $f(y|\mathbf{x}, D = 1)$.

2.1 | Score test under a parametric model on $f(y|\mathbf{x}, D = 1)$

Under a fully parametric model $f(y|\mathbf{x}, \boldsymbol{\xi})$ on $f(y|\mathbf{x}, D = 1)$, we estimate $\boldsymbol{\xi}$ by $\widehat{\boldsymbol{\xi}} = \arg\max_\xi \ell_2(\boldsymbol{\xi})$, where $\ell_2(\boldsymbol{\xi}) = d_i \log f(y_i|\mathbf{x}_i, \boldsymbol{\xi})$ is the likelihood function of $\boldsymbol{\xi}$ under H_0. Because $f(y|\mathbf{x}) = f(y|\mathbf{x}, D = 1) = f(y|\mathbf{x}, \boldsymbol{\xi})$ under H_0 or equivalently when the missingness mechanism is MAR, the score test statistic is then

$$S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}) = \sum_{i=1}^{n} \Big[d_i \{1 - \pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})\} y_i - (1-d_i) \pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}) \int y f(y|\mathbf{x}_i, \widehat{\boldsymbol{\xi}}) dy \Big].$$

To calculate the p-value of a score test statistic, we need to determine the sampling distribution of $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ under H_0. The exact form of this sampling distribution is in general intractable. A more practical solution is to approximate it by its null limiting distribution under H_0 or MAR.

Let $\boldsymbol{\beta}_0$ and $\boldsymbol{\xi}_0$ be the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, respectively. Our theoretical results on $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ are built on the following regularity conditions on $\mathbf{X}$ and $f(y|\mathbf{x}, \boldsymbol{\xi})$:

Condition (C1) $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and $\mathbf{A} = \mathbb{E}[\pi(\mathbf{X}^\top \boldsymbol{\beta}_0)\{1 - \pi(\mathbf{X}^\top \boldsymbol{\beta}_0)\}\mathbf{X}\mathbf{X}^\top]$ is of full rank.

Condition (C2) (i) The parameter space Ω of $\boldsymbol{\xi}$ is independent of $(y, \mathbf{x})$ and compact. (ii) The true value $\boldsymbol{\xi}_0$ of $\boldsymbol{\xi}$ is an interior point of Ω. (iii) $\boldsymbol{\xi}$ is identifiable, that is, $\mathbb{E}\{\int |f(y|\mathbf{X}, \boldsymbol{\xi}) - f(y|\mathbf{X}, \boldsymbol{\xi}')| dy\} > 0$ for any different elements $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ in Ω. (iv) $\mathbb{E}\{\sup_{\xi \in \Omega} |\log f(Y|\mathbf{X}, \boldsymbol{\xi})|\} < \infty$. (v) $f(y|\mathbf{x}, \boldsymbol{\xi})$ is continuous in $\boldsymbol{\xi}$ for almost all $(y, \mathbf{x})$.

Condition (C3) (i) $f(y|x, \boldsymbol{\xi})$ is twice differentiable with respect to $\boldsymbol{\xi}$ for almost all $(y, \mathbf{x})$, and $\nabla_{\xi\xi^\top} f(y|\mathbf{x}, \boldsymbol{\xi})$ is continuous at $\boldsymbol{\xi}_0$. (ii) $\mathbf{B} = \mathbb{E}[\pi(\mathbf{X}^\top \boldsymbol{\beta}_0)\{\nabla_\xi \log f(Y|\mathbf{X}, \boldsymbol{\xi}_0)\}^{\otimes 2}]$ is positive definite. (iii) There exist a $\delta > 0$ and positive functions $M_1(\mathbf{x})$ and $M_2(y, \mathbf{x})$ such that $\mathbb{E}\{M_1(\mathbf{X})\} < \infty$ and $\mathbb{E}\{M_2(Y, \mathbf{X})\} < \infty$, and $\|\mathbf{x}\| \int |t| \{f(t|\mathbf{x}, \boldsymbol{\xi}) + \|\nabla_\xi f(t|\mathbf{x}, \boldsymbol{\xi})\|\} dt \le M_1(\mathbf{x})$ and

$\|\nabla_{\xi\xi^\top} \log f(y|\mathbf{x},\boldsymbol{\xi})\| \le M_2(y,\mathbf{x})$ for all $\boldsymbol{\xi}$ satisfying $\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\| \le \delta$.

Under Condition (C1), the limit function of $\ell_1(\boldsymbol{\beta})/n$ is well defined. Conditions (i) and (ii) in Condition (C2) are trivial. Condition (iii) guarantees that $\boldsymbol{\xi}_0$ is a unique maximizer of the likelihood $\ell_2(\boldsymbol{\xi})$. Condition (iv) provides an envelope for $\{\log f(y|\mathbf{x},\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Omega\}$. These conditions plus the continuity condition of Condition (C2)(v) are sufficient for the consistency of $\widehat{\boldsymbol{\xi}}$. Under Condition (C3), the loglikelihood $\ell_2(\boldsymbol{\xi})$ can be approximated by a quadratic form of $\boldsymbol{\xi}$. The asymptotic normality of $\widehat{\boldsymbol{\xi}}$ follows immediately. Condition (C3) also guarantees that the matrices defined in Theorem 1 are well defined.

Theorem 1. *Assume Conditions (C1)–(C3) and that $H_0 : \gamma = 0$ is true. As n goes to infinity, $n^{-1/2}S_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}}) \xrightarrow{d} \mathcal{N}(0,\sigma_1^2)$, where $\sigma_1^2 = A_2 + B_2 - \mathbf{A}_1^\top\mathbf{A}^{-1}\mathbf{A}_1 - \mathbf{B}_1^\top\mathbf{B}^{-1}\mathbf{B}_1$, and*

$$\mathbf{A}_1 = \mathbb{E}[\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}\mathbf{X}Y],$$

$$A_2 = \mathbb{E}\{\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}^2Y^2\},$$

$$B_2 = \mathbb{E}[\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}\{\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}^2\{\int yf(y|\mathbf{X},\boldsymbol{\xi}_0)dy\}^2],$$

$$\mathbf{B}_1 = \mathbb{E}\left[\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\int y\nabla_\xi f(y|\mathbf{X},\boldsymbol{\xi}_0)dy\right].$$

An estimator for the asymptotic variance σ_1^2 is $\widehat{\sigma}_1^2 = \widehat{A}_2 + \widehat{B}_2 - \widehat{\mathbf{A}}_1^\top\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{A}}_1 - \widehat{\mathbf{B}}_1^\top\widehat{\mathbf{B}}^{-1}\widehat{\mathbf{B}}_1$, where $\widehat{\mathbf{A}},\widehat{\mathbf{A}}_1,\widehat{A}_2,\widehat{\mathbf{B}},\widehat{\mathbf{B}}_1$, and $\widehat{B}_2$ are all moment estimates with $\boldsymbol{\beta}_0$ and $\boldsymbol{\xi}_0$ replaced by their maximum likelihood estimates. For example, the moment estimate $\widehat{\mathbf{A}} = (1/n)\sum_{i=1}^n \pi(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})\{1-\pi(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})\}\mathbf{x}_i\mathbf{x}_i^\top$. The consistency of $\widehat{\sigma}_1^2$ follows from the consistency of $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}})$ and the continuity of $\pi(\mathbf{x}^\top\boldsymbol{\beta})$ and of $f(y|\mathbf{x},\boldsymbol{\xi})$. Formally, the proposed score test rejects H_0 if $|S_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}})|/(\sqrt{n}\widehat{\sigma}_1)$ is too large, and its p-value is approximately $2-2\Phi\{|S_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}})|/(\sqrt{n}\widehat{\sigma}_1)\}$, where $\Phi(\cdot)$ is the standard normal distribution function.

2.2 | Score test under a semiparametric location model on $f(y|\mathbf{x}, D=1)$

The score function depends on $f(y|\mathbf{x}, D=1)$ under MAR through the conditional mean $\mu(\mathbf{x}) = \int yf(y|\mathbf{x},D=1)dy$. Instead of imposing a fully parametric conditional density model, it is sufficient to assume a parametric model $\mu(\mathbf{x},\boldsymbol{\theta})$ for $\mu(\mathbf{x})$, where $\boldsymbol{\theta}$ is an unknown parameter. This model assumption is equivalent to a location model on the completely observed data $\{(\mathbf{x}_i,y_i) : d_i = 1\}$, namely, $y_i = \mu(\mathbf{x}_i,\boldsymbol{\theta}) + \varepsilon_i$, where ε_i satisfies $\mathbb{E}(\varepsilon_i|\mathbf{X}_i = \mathbf{x}_i, D_i = 1) = 0$. We estimate $\boldsymbol{\theta}$ by the least square estimator

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n d_i\{y_i - \mu(\mathbf{x}_i,\boldsymbol{\theta})\}^2.$$

Given the estimators $\widehat{\boldsymbol{\beta}}$ and $\mu(\mathbf{x},\widehat{\boldsymbol{\theta}})$ of $\boldsymbol{\beta}$ and $\mu(\mathbf{x})$, the score test statistic under the location model on $f(y|\mathbf{x})$ is

$$S_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^n \left[d_i\{1-\pi(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})\}y_i - (1-d_i)\pi(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})\mu(\mathbf{x}_i,\widehat{\boldsymbol{\theta}})\right].$$

Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. To establish the limiting distribution of $S_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})$, we impose the following regularity conditions on $\mu(\mathbf{x},\boldsymbol{\theta})$:

Condition (D1) (i) $\mathbb{E}(Y|\mathbf{X}) = \mu(\mathbf{X},\boldsymbol{\theta}_0)$ holds for all $\mathbf{X}$. (ii) The parameter space $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}$ is compact, and $\boldsymbol{\theta}_0$ is an interior in $\boldsymbol{\Theta}$. (iii) $\boldsymbol{\theta}_0$ is the unique minimizer of $Ł_*(\boldsymbol{\theta}) = \mathbb{E}[\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\{Y-\mu(\mathbf{X},\boldsymbol{\theta})\}^2]$.

Condition (D2) (i) $\mu(\mathbf{x},\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$. (ii) There exists $M_3(\mathbf{x},y)$ such that $\mathbb{E}\{M_3(\mathbf{X},Y)\} < \infty$ and $y^2 + \{\mu(\mathbf{x},\boldsymbol{\theta})\}^2 \le M_3(\mathbf{x},y)$ holds for all $\boldsymbol{\theta}$. (iii) There exists $M_4(\mathbf{x},y)$ such that $\{|y| + |\mu(\mathbf{x},\theta)|\}\|\nabla_{\theta\theta^\top}\mu(\mathbf{x},\boldsymbol{\theta})\| + \|\nabla_\theta\mu(\mathbf{x},\boldsymbol{\theta})\|^2 \le M_4(\mathbf{x},y)$ holds for all $\boldsymbol{\theta}$ and $\mathbb{E}\{M_4(\mathbf{X},Y)\} < \infty$. (iv) $C_1 = \mathbb{E}[\{\nabla_\theta\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^{\otimes 2}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)]$ is positive definite and $C_2 = \mathbb{E}[\{Y-\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^2\{\nabla_\theta\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^{\otimes 2}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)]$ is well defined.

Conditions (D1) and (D2) are the analogues of Conditions (C2) and (C3) on the conditional mean model $\mu(\mathbf{X},\boldsymbol{\theta})$.

Theorem 2. *Assume Conditions (C1), (D1), and (D2) and that $H_0 : \gamma = 0$ is true. As n goes to infinity, $n^{-1/2}S_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}(0,\sigma_2^2)$, where $\sigma_2^2 = A_2 + B_4 - \mathbf{A}_1^\top\mathbf{A}^{-1}\mathbf{A}_1 + \mathbf{B}_3^\top\mathbf{C}_1^{-1}\mathbf{C}_2\mathbf{C}_1^{-1}\mathbf{B}_3 - 2\mathbf{B}_3^\top\mathbf{C}_1^{-1}\mathbf{C}_3$ and*

$$\mathbf{B}_3 = \mathbb{E}[\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\nabla_\theta\mu(\mathbf{X},\boldsymbol{\theta}_0)],$$

$$B_4 = \mathbb{E}[\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}\{\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}^2\{\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^2],$$

$$\mathbf{C}_1 = \mathbb{E}[\{\nabla_\theta\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^{\otimes 2}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)],$$

$$\mathbf{C}_2 = \mathbb{E}[\{Y-\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^2\{\nabla_\theta\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^{\otimes 2}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)],$$

$$\mathbf{C}_3 = \mathbb{E}[\{1-\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\}\pi(\mathbf{X}^\top\boldsymbol{\beta}_0)\{Y-\mu(\mathbf{X},\boldsymbol{\theta}_0)\}^2\nabla_\theta\mu(\mathbf{X},\boldsymbol{\theta}_0)].$$

A consistent estimator for the asymptotic variance σ_2^2 is

$$\widehat{\sigma}_2^2 = \widehat{A}_2 + \widehat{B}_4 - \widehat{\mathbf{A}}_1^\top\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{A}}_1 + \widehat{\mathbf{B}}_3^\top\widehat{\mathbf{C}}_1^{-1}\widehat{\mathbf{C}}_2\widehat{\mathbf{C}}_1^{-1}\widehat{\mathbf{B}}_3 - 2\widehat{\mathbf{B}}_3^\top\widehat{\mathbf{C}}_1^{-1}\widehat{\mathbf{C}}_3,$$

where $\widehat{\mathbf{A}},\widehat{\mathbf{A}}_1,\widehat{A}_2,\widehat{\mathbf{B}}_3,\widehat{B}_4,\widehat{\mathbf{C}}_1,\widehat{\mathbf{C}}_2$, and $\widehat{\mathbf{C}}_3$ are all moment estimates with $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ replaced by their maximum likelihood estimates. We reject H_0 if $|S_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})|/(\sqrt{n}\widehat{\sigma}_2)$

is large enough. Its p-value is approximately $2-2\Phi\{|S_2(\widehat{\boldsymbol{\beta}},\widehat{\theta})|/(\sqrt{n}\widehat{\sigma}_2)\}$. The result in Theorem 2 and the variance estimator $\widehat{\sigma}_2^2$ allow the error ε_i given $D_i=1$ and $\mathbf{X}_i=\mathbf{x}$ to depend on $\mathbf{x}$, or to have a heterogeneous variance. If we assume that $\mathbf{x}_i$ and ε_i are conditionally independent given $D_i=1$, then $\mathbf{C}_3=\mathbf{B}_3\times\mathbb{V}\mathrm{ar}(\varepsilon_i|D_i=1)$ and $\mathbf{C}_2=\mathbf{C}_1\times\mathbb{V}\mathrm{ar}(\varepsilon_i|D_i=1)$ and σ_2^2 reduces to $\sigma_2^2=A_2+B_4-\mathbf{A}_1^\top\mathbf{A}^{-1}\mathbf{A}_1-\mathbf{B}_3^\top\mathbf{C}_1^{-1}\mathbf{B}_3\times\mathbb{V}\mathrm{ar}(\varepsilon_i|D_i=1)$.

2.3 | Local power

To study the asymptotic power of the proposed score tests, we consider a series of local alternatives $H_a:\gamma=n^{-1/2}\gamma_0$, where γ_0 is fixed. This local alternative tends to the null hypothesis at a root-n rate as n goes to infinity. A test for H_0 is root-n consistent if it can detect the local alternative as n goes to infinity for any fixed γ_0. We expect that both of the proposed score tests have root-n consistency, which is a desirable property of a nice test for MAR.

Theorem 3. *Assume Condition (C1) and that the alternative $H_a:\gamma=n^{-1/2}\gamma_0$ is true. Let $\boldsymbol{\beta}_0,\boldsymbol{\xi}_0$, and θ_0 be the true values of $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and θ, respectively. (i) If Conditions (C2) and (C3) are satisfied, then, as n goes to infinity, $n^{-1/2}S_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}})\xrightarrow{d}\mathcal{N}(\gamma_0\sigma_1^2,\sigma_1^2)$, where σ_1^2 is defined in Theorem 1. (ii) If Conditions (D1) and (D2) are satisfied, then, as n goes to infinity, $n^{-1/2}S_2(\widehat{\boldsymbol{\beta}},\widehat{\theta})\xrightarrow{d}\mathcal{N}(\gamma_0\delta,\sigma_2^2)$, where $\delta=A_2+B_4-\mathbf{A}_1^\top\mathbf{A}^{-1}\mathbf{A}_1-\mathbf{B}_3^\top\mathbf{C}_1^{-1}\mathbf{C}_3$ and σ_2^2 is defined in Theorem 2.*

Under the propensity model (1), a nonzero γ characterizes the departure of the true missingness mechanism from the null hypothesis. Theorem 3 indicates that if for some fixed γ_0 the alternative H_a is true, then both of the score test statistics converge in distribution to nondegenerate distributions with nonzero location parameters. Because the absolute values of the location parameters are increasing functions of $|\gamma_0|$, the powers of both score tests tend to 1 as γ_0 goes to infinity, which means that both of them are root-n consistent.

Remark 1. A nice property of the score tests is that their implementations are free from the identifiability issue that is inevitable in nonignorable missing data problems, and therefore instrument variables are not needed. Even so theoretically their performances still suffer from the identifiability issue. Molenberghs et al. (2008) pointed out that every MNAR model, fitted to a set of incomplete data, can be replaced by an MAR version, which produces exactly the same fit to the observed data. Under the logistic regression model for the propensity score, if the model is not identifiable, then we can find an MAR model with equal fit, and therefore we cannot distinguish MAR and MNAR based on the observed data. By Theorem 3, the score tests have nontrivial powers if $\sigma_1\neq 0$ or $\delta\neq 0$, where σ_1 is defined in Theorem 1 and δ is defined in Theorem 3. We believe that in certain situations where the identifiability issue is present under alternatives, we should have $\sigma_1=0$ or $\delta=0$. In such cases, the score tests fail to work.

Remark 2. As pointed out by an anonymous referee, when the score tests reject the null hypothesis, it may well be the case that the missingness mechanism is MAR while the assumed models are not correctly specified. To circumvent such a dilemma, desirable tests should be robust to model misspecification as much as possible. Our score tests are built on the parametric missingness model (1) and a parametric/semiparametric outcome model. The correctness of the missingness model (1) is untestable because we have no direct data from it. To alleviate the risk of model misspecification, we may conduct a goodness-of-fit test for the reduced logistic missingness model $\pi(\mathbf{x}^\top\boldsymbol{\beta})$ (which is $\pi(\mathbf{x}^\top\boldsymbol{\beta}+\gamma y)$ under MNAR) when the MAR mechanism is not rejected. A rejection of the model $\pi(\mathbf{x}^\top\boldsymbol{\beta})$ provides evidence that the assumed model $\pi(\mathbf{x}^\top\boldsymbol{\beta}+\gamma y)$ may be questionable and we need to consider an alternative model for the missingness mechanism. For the conditional distribution $\mathrm{pr}(y|\mathbf{x},D=1)$ or the conditional mean $\mathbb{E}(Y|\mathbf{x},D=1)$, we can conduct goodness-of-fit tests for them based on $\{(\mathbf{x}_i,y_i):d_i=1,1\le i\le n\}$ before the score tests are applied to test the missingness mechanism. Goodness-of-fit tests for general parametric regression models and logistic regression models have been well studied in the literature; see, for example, Fan and Huang (2001) and Qin and Zhang (1997).

2.4 | Generalization and optimality

As one referee pointed out, a general test statistic leveraging the conditional independence of D and Y given $\mathbf{X}$ under MAR can be constructed as

$$T_g=\sum_{i=1}^{n}\left[\frac{d_i}{\widehat{q}(\mathbf{x}_i)}g(\mathbf{x}_i,y_i)-\frac{1-d_i}{1-\widehat{q}(\mathbf{x}_i)}\widehat{\mathbb{E}}\{g(\mathbf{x}_i,y_i)\mid\mathbf{x}_i\}\right],$$

where g is a user-specific function, and $\widehat{q}(\mathbf{x})$ and $\widehat{\mathbb{E}}\{g(\mathbf{X},Y)\mid\mathbf{X}\}$ are consistent estimates of the nonmissingness probability $\mathrm{pr}(D=1|\mathbf{X}=\mathbf{x})$ and $\mathbb{E}\{g(\mathbf{X},Y)\mid\mathbf{X}\}$, respectively, under MAR. The proposed score test corresponds to T_g with $g(\mathbf{x},y)=\widehat{q}(\mathbf{x})\{1-\widehat{q}(\mathbf{x})\}y$. The test based on T_g is usually a valid test for MAR without any model assumptions. However, its implementation necessitates nonparametric estimates of $\mathrm{pr}(D=1|\mathbf{X}=\mathbf{x})$ and

$\mathbb{E}\{g(\mathbf{X}, Y) \mid \mathbf{X}\}$, which often suffer from bandwidth selection and possible curse of dimensionality. Our score test, which is built on some parametric models, trades off model assumptions and practical performance: It is free of bandwidth selection and curse of dimensionality at the expense of possible model misspecification. An appealing property of our two tests is that they are the most powerful among two classes of tests based on T_g if our model assumptions are correct.

We first consider the score test $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ under the logistic nonmissingness probability model (1) and the regression model $f(y|\mathbf{x}, \xi)$ for $\text{pr}(Y = y|\mathbf{X} = \mathbf{x})$. Here $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ under MAR. For any function g, the general test statistic T_g becomes

$$T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}) = \sum_{i=1}^{n} \left\{ \frac{d_i}{\pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})} g(\mathbf{x}_i, y_i) - \frac{(1-d_i)}{1-\pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})} \int g(\mathbf{x}_i, y) f(y|\mathbf{x}_i, \widehat{\boldsymbol{\xi}}) dy \right\}.$$

Condition (C4) The function $g(\mathbf{x}, y)$ is a completely deterministic function of $(\mathbf{x}, y)$, and there exist a $\delta > 0$ and positive function $M_5(\mathbf{x})$ such that $\mathbb{E}\{M_5(\mathbf{X})\} < \infty$, and $\|\mathbf{x}\| \int |g(\mathbf{x}, t)|\{f(t|\mathbf{x}, \boldsymbol{\xi}) + \|\nabla_{\boldsymbol{\xi}} f(t|\mathbf{x}, \boldsymbol{\xi})\|\} dt \le M_5(\mathbf{x})$ for all $\boldsymbol{\xi}$ satisfying $\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| \le \delta$.

Theorem 4. *Suppose that Conditions (C1)-(C4) are satisfied.*

(a) When $H_0 : \gamma = 0$ is true, then as n goes to infinity, $n^{-1/2} T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}) \xrightarrow{d} \mathcal{N}(0, \sigma_3^2)$, where $\sigma_3^2 = A_6 + B_6 - \mathbf{A}_5^\top \mathbf{A}^{-1} \mathbf{A}_5 - \mathbf{B}_5^\top \mathbf{B}^{-1} \mathbf{B}_5$, and

$$\mathbf{A}_5 = \mathbb{E}\{g(\mathbf{X}, Y)\mathbf{X}\}, \quad \mathbf{B}_5 = \mathbb{E}\left\{ \int g(\mathbf{X}, y) \nabla_{\boldsymbol{\xi}} f(y|\mathbf{X}, \boldsymbol{\xi}_0) dy \right\},$$

$$A_6 = \mathbb{E}\left\{ \frac{g^2(\mathbf{X}, Y)}{\pi(\mathbf{X}^\top \boldsymbol{\beta}_0)} \right\}, \quad B_6 = \mathbb{E}\left[\frac{\mathbb{E}^2\{g(\mathbf{X}, Y)|\mathbf{X}\}}{1-\pi(\mathbf{X}^\top \boldsymbol{\beta}_0)} \right].$$

(b) When $H_a : \gamma = n^{-1/2}\gamma_0$ is true, then as n goes to infinity, $n^{-1/2} T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}) \xrightarrow{d} \mathcal{N}(\gamma_0 \delta_g, \sigma_3^2)$, where $\delta_g = A_7 + B_7 - \mathbf{A}_5 \mathbf{A}^{-1} \mathbf{A}_1 - \mathbf{B}_5 \mathbf{B}^{-1} \mathbf{B}_1$ and

$$A_7 = \mathbb{E}[\{1-\pi(\mathbf{X}^\top \boldsymbol{\beta}_0)\} g(\mathbf{X}, Y) Y],$$

$$B_7 = \mathbb{E}\left[\pi(\mathbf{X}^\top \boldsymbol{\beta}_0) Y \int g(\mathbf{X}, y) f(y|\mathbf{X}, \boldsymbol{\xi}_0) dy \right].$$

Result (a) of Theorem 4 suggests that a reasonable test is to reject H_0 if $|T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})|/(\sqrt{n}\widehat{\sigma}_3)$ is large enough, where $\widehat{\sigma}_3$ is a consistent estimator of σ_3. The p-value of this test ($T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ for short) is approximately $2 - 2\Phi\{|T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})|/(\sqrt{n}\widehat{\sigma}_3)\}$. Hereafter let $H(t) = \Phi(t - z_{1-\frac{\alpha}{2}}) + 1 - \Phi(t + z_{1-\frac{\alpha}{2}})$. It follows from result (b) of Theorem 4 that at the significance level $\alpha \in (0, 1)$ and under $H_a : \gamma = n^{-1/2}\gamma_0$, the local power of $T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ is $H(\gamma_0 \delta_g / \sigma_3)$. By Theorems 1 and 3, at the same α significance level, the local power of our score test $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ is $H(\gamma_0 \sigma_1)$.

Lemma 1. *If the σ_1 in Theorem 1 and the σ_3 in Theorem 4 are well defined, then $H(\gamma_0 \delta_g / \sigma_3) \le H(\gamma_0 \sigma_1)$ for any constant γ_0 and any g satisfying Condition (C4).*

Lemma 1 indicates that our score test $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ is no less powerful than $T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ for any non random function g satisfying Condition (C4). Let $g(\mathbf{x}, y) = \pi(\mathbf{x}^\top \widehat{\boldsymbol{\beta}})\{1 - \pi(\mathbf{x}^\top \widehat{\boldsymbol{\beta}})\} h(\mathbf{x}, y)$ and $\widetilde{g}(\mathbf{x}, y) = \pi(\mathbf{x}^\top \boldsymbol{\beta}_0)\{1 - \pi(\mathbf{x}^\top \boldsymbol{\beta}_0)\} h(\mathbf{x}, y)$. Under either H_0 or H_a,

$$\begin{aligned} & T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}) - T_{\widetilde{g}}^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}) \\ &= \sum_{i=1}^{n} \left\{ \frac{d_i}{\pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})} h(\mathbf{x}_i, y_i) - \frac{(1-d_i)}{1-\pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})} \int h(\mathbf{x}_i, y) f(y|\mathbf{x}_i, \widehat{\boldsymbol{\xi}}) dy \right\} \\ &\quad \times [\pi(x_i^\top \widehat{\boldsymbol{\beta}})\{1 - \pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})\} - \pi(\mathbf{x}_i^\top \boldsymbol{\beta}_0)\{1 - \pi(\mathbf{x}_i^\top \boldsymbol{\beta}_0)\}] \\ &= o_p(n^{1/2}) \end{aligned} \tag{2}$$

for h satisfying Condition (C4), which implies that $n^{-1/2} T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ has the same limiting distributions as $n^{-1/2} T_{\widetilde{g}}^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ under both H_0 and H_a. It also indicates that our score test $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ is asymptotically no less powerful than $T_{\widetilde{g}}^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ for any h satisfying Condition (C4). In summary, our score test $S_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ is the most powerful among all the general tests $T_g^{(1)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}})$ with $g(\mathbf{x}, y) = \pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})\{1 - \pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})\} h(\mathbf{x}, y)$ or $g(\mathbf{x}, y) = h(\mathbf{x}, y)$ for h satisfying Condition (C4).

Similarly our second score test $S_2(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ is the most powerful among all the general tests

$$T_h^{(2)}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \left\{ \frac{d_i}{\pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})} h(\mathbf{x}_i) y_i - \frac{(1-d_i)}{1-\pi(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})} h(\mathbf{x}_i) \mu(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}) \right\}$$

with $h(\mathbf{x}) = r(\mathbf{x})$ or $h(\mathbf{x}) = \pi(\mathbf{x}^\top \widehat{\boldsymbol{\beta}})\{1 - \pi(\mathbf{x}^\top \widehat{\boldsymbol{\beta}})\} r(\mathbf{x})$, where $r(\mathbf{x})$ is a nonrandom function satisfying mild conditions.

To save space, we postpone the details to Section 1.1 of the supporting information.

3 | SIMULATION

We conduct simulations to evaluate the finite-sample performance of the proposed score tests. Specifically, we compare the following three tests: (1) S1, the proposed score test under a parametric model on $\text{pr}(y|x)$; (2) S2, the proposed score test under a semiparametric location model on $\text{pr}(y|x)$; and (3) DUAN, the test proposed by Duan et al. (2020). We generate data from two examples. In Example 1, which comes from Duan et al. (2020), an instrumental variable is present, whereas there is no instrumental variable in Example 2. All our simulation results are calculated based on 5000 simulated samples and the significance level is set to 5%.

Example 1. Let (Y, U, Z) follow a multivariate normal distribution such that $(Y|U, Z) \sim \mathcal{N}(1 + U + b_z Z, 1)$, $(U \mid Z) \sim \mathcal{N}(1 - Z, 1)$, and $Z \sim \mathcal{N}(0, 1)$. The missingness indicator D of Y follows a Bernoulli distribution with success probability $\text{pr}(D = 1 \mid Y, U, Z) = \Phi(c_0 + c_1 w(Y) + c_2 U)$, where $\Phi(\cdot)$ is the standard normal distribution function. We consider three choices for $w(y)$, namely, $y, 0.4y^2$, and $2.5I(y > 1)$, two choices for b_z, namely, 0.5 and 1, four choices for c_2, namely, 0, 0.25, 0.5, and 0.75, and 11 choices for c_1, namely, $0.05 \times k$, for $k = 0, 1, \ldots, 10$. For each (c_1, c_2), we choose an appropriate value of c_0 so that the overall nonmissingness probability is about 20%. Details of the parameter settings are given in Table 1.

The DUAN test requires the existence of an instrumental variable, and is applicable under the settings of Example 1 because the variable Z is an instrumental variable. For data generated from this example, we model $\text{pr}(Y = y|X = x)$ by $f(y|x, \xi) = (2\pi)^{-1/2} \exp\{-(y - x^\top \xi)^2/2\}$ in the construction of S1, and model $\mathbb{E}(Y|X = x)$ by $\mu(x, \theta) = x^\top \theta$ in the construction of S2. To save space, we report the simulated rejection rates of S1, S2, and DUAN for $n = 1000$ in Tables S1 and S2 in the supporting information, and display the power (vs c_1) lines of S2 and DUAN in Figures 1 and 2, corresponding to $b_z = 0.5$ and 1, respectively. The rejection rates corresponding to the DUAN test are directly copied from tables 3 and 4 in Duan et al. (2020), which were calculated based on 1000 simulated samples. Although the true missingness indicator is generated from a probit model, we model it by the logistic model (1) in the constructions of the proposed two score tests.

The coefficient c_1 quantifies the departure of the true missingness mechanism from the null hypothesis. When $c_1 = 0$, the null hypothesis holds and the results reported are all type I errors. We see that all three tests have desirable controls on their type I errors. As c_1 increases, the true missingness mechanism departs more and more from the null hypothesis and, as expected, all tests have increasing powers. The proposed two score tests are more powerful than DUAN in most situations. When $b_z = 0.5$, their power gains against DUAN can be greater than 25%; see the case with $c_1 = 0.4, c_2 = 0.75$, and $w(y) = y$. As b_z increases from 0.5 to 1, the power gain can be as large as 41%; see the case with $b_z = 1, c_1 = 0.25, c_2 = 0.75$, and $w(y) = 0.4y^2$. These observations show that the proposed score tests have obvious advantages over the DUAN test. Meanwhile, the two score tests S1 and S2 have almost the same powers in all cases, although S2 requires much weaker model assumptions. We also conduct simulations for $n = 2000$, and the simulation results, provided in the supporting information, are similar.

From Figures 1 and 2, we see that the power lines of S1 coincide with those of S2 and hence are omitted. It is clear that the power lines of S2 always lie above those of the DUAN test, or S2 is uniformly more powerful, except for two scenarios where $b_z = 0.5, c_2 = 0$, and $w(y) = 0.4y^2$ or $2.5I(y > 1)$. In the two exceptional cases, compared with DUAN, S2 is more powerful for small c_1 and becomes less powerful for large c_1. As c_1 quantifies the departure of the true missingness mechanism from the null hypothesis, a possible explanation for this phenomenon is that the score test is usually most powerful for "local" alternatives, but may be suboptimal when the alternative is not very local.

Example 2. Let $X \sim \mathcal{N}(0, 1)$, $(Y|X = x) \sim \mathcal{N}(\xi_1 x + \xi_2 x^2, e^{\xi_3 + \xi_4 x})$ with $\xi = (\xi_1, \ldots, \xi_4) = (-1, 1, 0.5, 0)$ or $(1, 1, 0.5, 1)$, and $\text{pr}(D = 1|x, y) = \pi(\beta_0 + \beta_1 x + \gamma y)$. We consider eight choices of (β_0, β_1), namely, $(0.85, 0)$, $(0.6, 0.25)$, $(0.4, 0.5)$, $(0.1, 1)$ in the case of $\xi = (-1, 1, 0.5, 0)$ and $(0.85, 0)$, $(0.7, 0.25)$, $(0.5, 0.5)$, $(0.2, 1)$ in the case of $\xi = (1, 1, 0.5, 1)$. These settings are chosen such that the missingness rates are about 20%-30%. The parameter γ is set to 0, 0.05, ..., and 0.25, respectively.

Example 2 is designed to represent the case where no instrument is present, and therefore the DUAN test is not applicable. The choices of $\xi_4 = 0$ and 1 correspond to a homogeneous variance and a heterogeneous variance, respectively. We take $f(y|x, \xi)$ and $\mu(x, \theta)$ in the constructions of S1 and S2 to be the density functions of $\mathcal{N}(\xi_1 x + \xi_2 x^2, e^{\xi_3 + \xi_4 x})$ and $\theta_1 x + \theta_2 x^2$, respectively, where $\theta = (\theta_1, \theta_2)^\top$. Table 2 presents the simulated rejection rates of the S1 and S2 tests when data are generated from Example 2 and the sample size $n = 1000$. The results corresponding to $\gamma = 0$ are type I errors, and the type I errors of both S1 and S2 are under control. As γ increases from 0 to 0.25, both tests have desirable and increasing

TABLE 1 Details of the parameter settings in Example 1

			$b_z = 0.5\, c_1$									
$w(y)$	c_2	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
y	0	0.84	0.75	0.66	0.57	0.48	0.39	0.32	0.24	0.18	0.12	0.06
	0.25	0.65	0.57	0.49	0.41	0.33	0.25	0.17	0.13	0.08	0.03	-0.04
	0.50	0.55	0.47	0.39	0.31	0.23	0.15	0.07	0.04	0.02	-0.02	-0.06
	0.75	0.5	0.42	0.34	0.26	0.2	0.16	0.1	0.04	-0.02	-0.08	-0.14
$0.4y^2$	0	0.84	0.74	0.64	0.54	0.44	0.4	0.34	0.28	0.2	0.15	0.1
	0.25	0.65	0.57	0.49	0.41	0.33	0.25	0.19	0.15	0.11	0.07	0.03
	0.50	0.55	0.47	0.39	0.31	0.25	0.2	0.14	0.10	0.06	0.02	-0.02
	0.75	0.5	0.42	0.34	0.26	0.18	0.1	0.07	0.04	0	-0.04	-0.08
$2.5I(y > 1)$	0	0.84	0.75	0.66	0.57	0.48	0.39	0.3	0.21	0.14	0.08	0.02
	0.25	0.65	0.57	0.49	0.41	0.33	0.25	0.17	0.09	0.01	-0.07	-0.15
	0.50	0.55	0.47	0.39	0.31	0.23	0.15	0.07	-0.01	-0.09	-0.17	-0.25
	0.75	0.5	0.42	0.34	0.26	0.18	0.1	0.02	-0.06	-0.14	-0.22	-0.3
			$b_z = 1\, c_1$									
$w(y)$	c_2	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
y	0	0.84	0.75	0.66	0.57	0.48	0.39	0.30	0.21	0.12	0.03	-0.06
	0.25	0.65	0.57	0.49	0.41	0.33	0.25	0.17	0.09	0.01	-0.07	-0.15
	0.50	0.55	0.47	0.39	0.31	0.23	0.15	0.07	-0.01	-0.09	-0.17	-0.25
	0.75	0.5	0.42	0.34	0.26	0.18	0.1	0.02	-0.06	-0.14	-0.22	-0.3
$0.4y^2$	0	0.84	0.74	0.64	0.54	0.44	0.4	0.34	0.28	0.2	0.15	0.1
	0.25	0.65	0.57	0.49	0.41	0.33	0.25	0.17	0.12	0.08	0.04	0
	0.50	0.55	0.47	0.39	0.31	0.23	0.15	0.07	0.04	0.02	-0.02	-0.06
	0.75	0.5	0.42	0.34	0.26	0.18	0.1	0.07	0.04	0	-0.04	-0.08
$2.5I(y > 1)$	0	0.84	0.75	0.66	0.57	0.48	0.39	0.3	0.21	0.14	0.08	0.02
	0.25	0.65	0.57	0.49	0.41	0.33	0.25	0.17	0.09	0.01	-0.07	-0.15
	0.50	0.55	0.47	0.39	0.31	0.23	0.15	0.07	-0.01	-0.09	-0.17	-0.25
	0.75	0.5	0.42	0.34	0.26	0.18	0.1	0.02	-0.06	-0.14	-0.22	-0.3

powers whether the variance is homogeneous or heterogeneous. Again, the results for both tests are all nearly equal to each other in all cases. As S1 requires stronger model assumptions, it may be more risky for model misspecification than S2. Hence, we would recommend using S2 rather than S1 for testing whether the missingness mechanism is ignorable missing or nonignorable missing.

We have also studied how robust the proposed tests are when the outcome model $f(y|x)$ or the mean function $\mu(x)$ is misspecified. Our general finding is that when these models are misspecified, the score tests have controllable or slightly inflated type I errors and desirable power trend at the price of certain power loses. See Example A and the corresponding discussion in the supporting information.

4 | APPLICATION TO HIV DATA

For illustration, we analyze HIV data from AIDS Clinical Trials Group Protocol 175 (Hammer et al., 1996; Han et al., 2019; Liu et al., 2021). These data are available from the `R` package `speff2trial` and consist of various measurements of $n = 2139$ HIV-infected patients. The patients were randomly divided into four arms according to the regimen of treatment they received: (I) zidovudine monotherapy, (II) zidovudine + didanosine, (III) zidovudine + zalcitabine, and (IV) didanosine monotherapy. Important measurements from the patients include CD4 cell count at baseline (cd40), CD4 cell count at 20±5 weeks (cd420), CD4 cell count at 96 ±5 weeks (cd496), CD8 cell count at 20±5 weeks (cd820), and arm number (arms). The effectiveness of an HIV treatment can be assessed by monitoring the CD4 cell counts of HIV-positive patients: An increase in such counts is an indication of improvement in the patients' health. The typical problem of interest is to estimate the mean of the CD4 cell counts in each arm after the patients were treated for about 96 weeks.

We take cd496 as a response variable Y and we take cd40, cd420, and cd820 as covariates X_1, X_2, and X_3, respectively. Owing to the end of the trial or loss to follow-

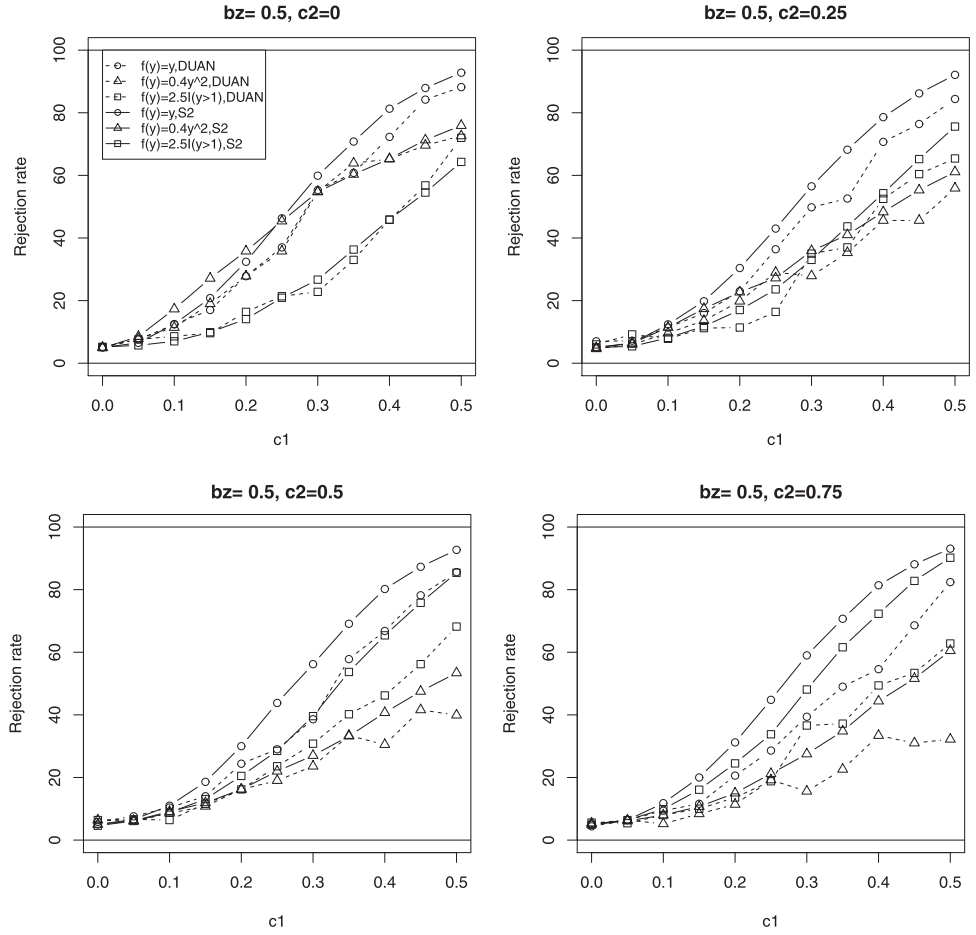

FIGURE 1 Plots of rejection rates when $b_z = 0.5$ for the S2 test (solid lines) and the DUAN test (dotted lines): $w(y) = y$ (circles); $w(y) = 0.5y^2$ (triangles); $w(y) = 2.5I(y > 1)$ (squares)

TABLE 2 Empirical rejection rates (%) of the S1 and S2 tests based on 5000 simulated samples of size $n = 1000$ from Example 2

			γ					
ξ	β_1	Test	0	0.05	0.1	0.15	0.2	0.25
$(-1, 1, 0.5, 0)$	0	S1	4.7	17.4	50.9	81.7	95.6	99.2
		S2	4.8	17.3	50.8	81.7	95.5	99.2
	0.25	S1	4.9	17.4	50.6	81.7	96.4	99.4
		S2	4.9	17.2	50.4	81.6	96.4	99.4
	0.5	S1	5.4	16.9	47.6	79.3	94.8	99.1
		S2	5.4	16.8	47.3	79.0	94.7	99.1
	1	S1	4.7	13.0	35.8	66.0	86.3	97.3
		S2	5.1	12.8	35.3	65.4	86.1	97.1
$(1, 1, 0.5, 1)$	0	S1	4.6	14.4	37.4	60.9	77.7	87.4
		S2	5.0	13.0	36.3	60.4	77.5	88.4
	0.25	S1	5.2	13.6	35.1	57.4	75.9	86.6
		S2	5.3	13.4	34.7	57.6	76.5	87.4
	0.5	S1	4.7	14.2	33.6	55.3	73.7	86.4
		S2	4.7	14.1	34.3	56.2	74.6	87.0
	1	S1	4.7	11.0	28.1	47.2	66.0	79.9
		S2	4.6	11.1	27.5	46.8	65.2	79.4

TABLE 3 AIC and BIC of the candidate models. The best candidate model is highlighted

Covariates	AIC	BIC
X_1	2823.678	2835.014
X_2	**2811.279**	**2822.615**
X_3	2828.592	2839.928
X_1, X_2	2813.229	2830.233
X_1, X_3	2825.541	2842.545
X_2, X_3	2813.139	2830.143
X_1, X_2, X_3	2815.077	2837.749

up, 39.66% of the patients' responses were missing. These data have been analyzed under the MAR (Hammer et al., 1996; Han et al., 2019) and MNAR (Liu et al., 2021; Zhang *et al.*, 2020) assumptions. We may wonder which missingness mechanism is more credible. Let $\mathbf{X} = (1, X_1, X_2, X_3)$, and suppose the missingness indicator D given $\mathbf{X}$ and Y follows the linear logistic model (1). Table 3 presents the Akaike information criterion (AIC) and Bayesian

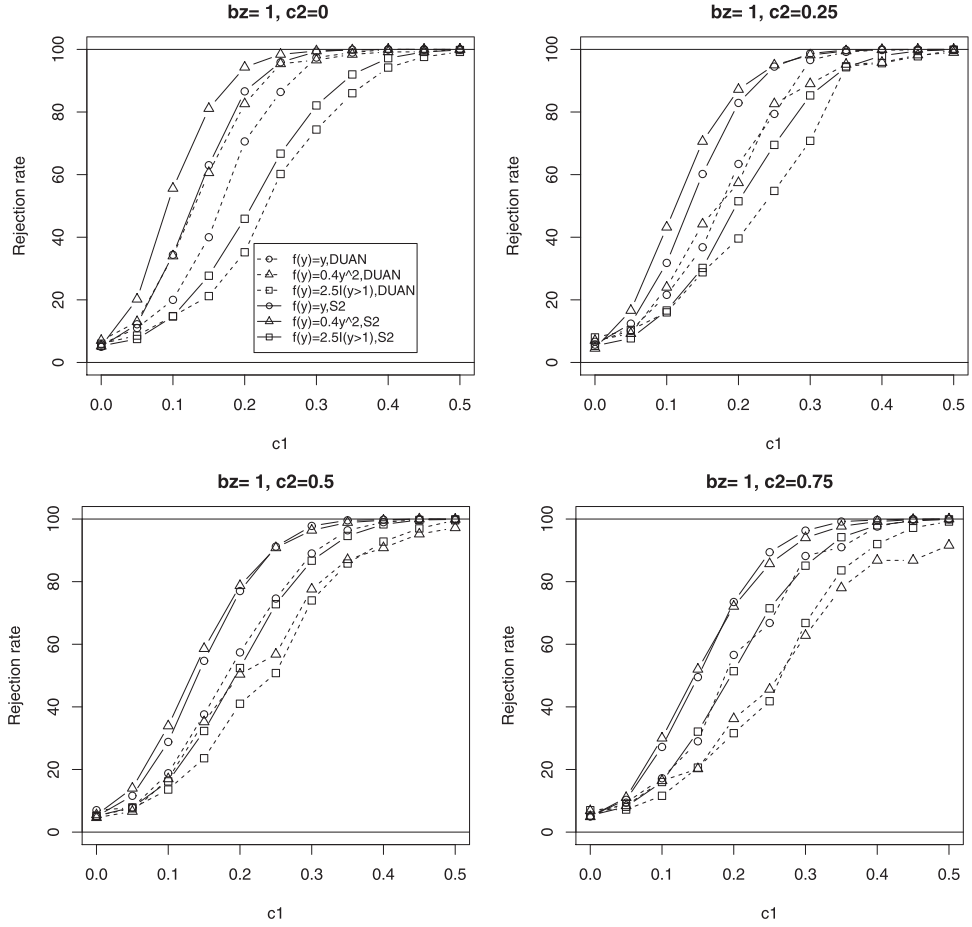

FIGURE 2 Plots of rejection rates when $b_z = 1$ for the S2 test (solid lines) and the DUAN test (dotted lines): $w(y) = y$ (circles); $w(y) = 0.5y^2$ (triangles); $w(y) = 2.5I(y > 1)$ (squares)

information criterion (BIC) of all logistic candidate models under $\gamma = 0$. The candidate model with only covariate X_2 has the smallest AIC and BIC. To some extent, this indicates that the logistic model with only covariate X_2 and possibly y is the most appropriate for the nonmissingness probability, as assumed hereafter. In addition, we choose $f(y|\mathbf{x}, \xi)$ to be the normal density with mean $\mu(\mathbf{x}, \xi) = \xi_1 + \xi_2 x_1 + \xi_3 x_2 + \xi_4 x_3 + \xi_5 x_2^2$ and variance $\sigma(\mathbf{x}, \xi) = \xi_6$, where $\xi = (\xi_1, \ldots, \xi_6)^\top$. We apply the proposed two score tests to test whether the missingness of cd496 depends on itself.

As pointed out in Remark 2, we first conduct a goodness-of-fit test for the assumed outcome model. The p-values of the Fan and Huang (2001) test (based on their $T_{AN,1}$) for the assumed normal model are 0.107, 0.999, 0.905, and 0.005, respectively, showing that the normal model is appropriate under regimens I-III, but is inappropriate under regimen IV. Under regimens IV, when $f(y|\mathbf{x}, \xi)$ is chosen to be the normal density with mean $\mu(\mathbf{x}, \xi) = \xi_1 + \xi_2 x_1 + \xi_3 x_2 + \xi_4 x_3 + \xi_5 x_1 x_2 + \xi_6 x_2^2 + \xi_7 x_1 x_2^2$ and variance $\sigma(\mathbf{x}, \xi) = \xi_8$, where $\xi = (\xi_1, \ldots, \xi_8)^\top$, the Fan and Huang (2001) test produces supportive evidence for this model (p-value = 1).

The p-values of the proposed two score tests are reported in Table 4. None of the results are significant at the 5% level. Meanwhile the Qin and Zhang (1997) test provides no evidence (the p-values corresponding to the four regimens are 0.352, 0.935, 0.268, and 0.977, respectively) against the assumed logistic nonmissingness model. In other words, they all support the MAR mechanism in the four regimens. At the same time, both the tests have very close p-values. Table 4 also presents their p-values if we remove the covariate X_2 from the propensity score model. The p-values for regimens II and IV are seemingly unchanged and insignificant. Again the Qin and Zhang (1997) test supports the assumed logistic model (the p-values are 0.356 and 0.296, respectively). However, those for regimens I and III become much less, and much smaller than the 5% significance level. These results indicate that the MNAR mechanism seems more reasonable than MAR if the propensity score depends potentially on y. A possible explanation for the insignificant result in the presence of X_2 is that X_2 and

TABLE 4 *p*-Values of the proposed score tests S1 and S2 under the four regimens of treatment based on the HIV data

Treatment regimen	I	II	III	IV	IV under new model
X_2 appears in the logistic model					
S1	0.3263	0.3558	0.4490	0.4060	0.3996
S2	0.1291	0.4548	0.3730	0.2265	0.2131
X_2 does not appear in the logistic model					
S1	0.0065	0.3731	0.0104	0.2081	0.2108
S2	0.0003	0.3389	0.0006	0.1584	0.1615

Y stand for CD4 cell counts at 20 ± 5 weeks and at 96 ± 5 weeks, respectively, and they are highly correlated.

5 | DISCUSSION

Valid data analyses of missing data rely on a correctly specified missingness mechanism. The problem of testing whether the missingness mechanism is MCAR or MAR is relatively easy to solve and has been extensively studied. However, it is much more challenging to test whether the mechanism is MAR or not, because parameters may no longer be identifiable under the alternative hypothesis. Numerically we avoid this thorny issue by using a score test, which is constructed under the null hypothesis, namely, the MAR mechanism. The underlying parameters are usually identifiable based on MAR data. This is one of the nice properties of a score test (Rao, 2005). A score test is also invariant under transformation of parameters. Transformation of parameters may simplify parameter estimation without affecting the value of the statistic. We derive two score tests, S1 and S2, when the conditional density of Y given $\mathbf{X}$ is modeled by a completely parametric model and a semiparametric location model, respectively. Our numerical results indicate that these tests generally have nearly the same performance (type I error and power), but S2 is preferable because it requires weaker model assumptions.

When the score tests reject the null hypothesis, it may be either that the missingness mechanism is MNAR or that the missingness mechanism is MAR but the assumed models are not correctly specified. As a remedial action to the proposed score tests, when the null hypothesis is rejected, we suggest conducting follow-up goodness-of-fit tests for the logistic missingness model and the observed outcome model under MAR. If none of these models is rejected, we shall claim that the missingness mechanism is MNAR. Otherwise, we need to consider alternative models for the missingness and the observed regression. Obviously this is a multiple testing procedure, which may have an out-of-control type I error. Theoretically we need to study large-sample properties of the multiple testing procedure and make its type I error under control. We leave this interesting problem for future research.

ACKNOWLEDGMENTS

The authors thank the editor, associate editor, and two referees for constructive comments and suggestions that led to significant improvements of this paper. This research was supported by the Natural Science Foundation of Shanghai (17ZR1409000), the National Natural Science Foundation of China (71931004, 12171157, 11771144, 32030063), and the 111 Project (B14019).

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available from the `R` package `speff2trial` at https://cran.r-project.org/web/packages/speff2trial/index.html.

ORCID

Yukun Liu https://orcid.org/0000-0002-9743-9276

REFERENCES

Breunig, C. (2019) Testing missing at random using instrumental variables. *Journal of Business and Economic Statistics* 37, 223–234.

Chen, H. Y. and Little, R. J. (1999) A test of missing completely at random for generalized estimating equations. *Biometrika* 86, 1–13.

Duan, R., Liang, C. J., Shaw, P., Tang, C. Y. and Chen, Y. (2020) Missing at random or not: a semiparametric testing approach. arXiv: 2003.11181.

Fan, J. and Huang, L. (2001) Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association* 96, 640–652.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H. et al. (1996) A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* 335, 1081–1090.

Han, P., Kong, L., Zhao, J. and Zhou, X. (2019) A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society B* 81, 305–333.

Jamshidian, M. and Jalal, S. (2010) Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika* 75, 649–674.

Kim, K. H. and Bentler, P. M. (2002) Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika* 67, 609–623.

Kim, J. K. and Shao, J. (2013) *Statistical Methods for Handling Incomplete Data*. New York: CRC Press.

Li, J. and Yu, Y. (2015) A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika* 80, 707–726.

Little, R. J. (1988) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198–1202.

Little, R. J. and Rubin, D. B. (2019) *Statistical Analysis with Missing Data*, 3rd edition. Hoboken, NJ: Wiley.

Liu, Y., Li, P. and Qin, J. (2022) Full semiparametric likelihood based inference for non-ignorable missing data. *Statistica Sinica* 32(1), 271–292.

Miao, W., Ding, P. and Geng, Z. (2016) Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* 111, 1673–1683.

Miao, W. and Tchetgen Tchetgen, E. J. (2016) On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* 103, 475–442.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M.G. (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B* 70, 371–388.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.

Owen, A. B. (1990) Empirical likelihood ratio confidence regions. *Annals of Statistics* 18, 90–120.

Owen, A. B. (2001) *Empirical Likelihood*. New York: CRC Press.

Qin, J. and Zhang, B. (1997) A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* 84(3), 609–618.

Rao, C. R. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44, 50–57.

Rao, C. R. (2005) Score test: historical review and recent developments. In: N. Balakrishnan, N. Kannan and H. N. Nagaraja (Eds.) *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*. Boston: Birkhäuser, pp. 3–20.

Rubin, D. B. (1976) Inference and missing data (with discussion). *Biometrika* 63, 581–592.

Simon, G. A. and Simonoff, J. S. (1986) Diagnostic plots for missing data in least squares regression. *Journal of the American Statistical Association* 81, 501–509.

Tang, N. and Ju, Y. (2018) Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2, 105–133.

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.

Wang, H. and Kim, J. K. (2021) Statistical inference after kernel ridge regression imputation under item nonresponse. arXiv: 2102.00058v1.

Wang, S., Shao, J. and Kim, J. K. (2014) An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* 24, 1097–116.

Zhang, S., Han, P. and Wu, C. (2019) A unified empirical likelihood approach for testing MCAR and subsequent estimation. *Scandinavian Journal of Statistics*, 46, 272–288.

Zhang, T., Wang, L. and Azen, S. P. (2020) Smoothed empirical likelihood inference and variable selection for quantile regression with nonignorable missing response. *Computational Statistics and Data Analysis*, 144, 106888.

SUPPORTING INFORMATION

Web Appendices referenced in Sections 2 and 3, proofs of all theorems and lemmas, and the R code, are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Wang, H., Lu, Z. and Liu, Y. (2023) Score test for missing at random or not under logistic missingness models. *Biometrics*, 79, 1268–1279. https://doi.org/10.1111/biom.13666