A Unified Empirical Likelihood Approach for Testing MCAR and Subsequent Estimation

By SHIXIAO ZHANG

Department of Statistics and Actuarial Science, University of Waterloo

PEISONG HAN

Department of Biostatistics, School of Public Health, University of Michigan

and CHANGBAO WU

Department of Statistics and Actuarial Science, University of Waterloo

Abstract

For estimation with missing data, a crucial step is to determine if the data are missing completely at random (MCAR), in which case a complete-case analysis would suffice. Most existing tests for MCAR do not provide a method for subsequent estimation once the MCAR is rejected. In the setting of estimating means, we propose a unified approach for testing MCAR and the subsequent estimation. Upon rejecting MCAR, the same set of weights used for testing can then be used for estimation. The resulting estimators are consistent if the missingness of each response variable depends only on a set of fully observed auxiliary variables and the true outcome regression model is among the user-specified functions for deriving the weights. The proposed method is based on the calibration idea from survey sampling literature and the empirical likelihood theory.

Key words: calibration, empirical likelihood, missing completely at random, missingness mechanism

1 Introduction

Data collected from statistical studies are often incomplete. There are three widely adopted missingness mechanisms in the missing-data literature (e.g., Little and Rubin 2002): missing completely at random (MCAR) where the missingness does not depend on either the observed or the missing data, missing at random (MAR) where the missingness depends on the observed but not the missing data, and missing not at random (MNAR) where the missingness depends on both the observed and the missing data. Most existing methods for missing-data analysis are developed under the

MAR mechanism, largely due to the mathematical triviality of MCAR and complexity of MNAR. However, in cases where the data are indeed MCAR, a simple complete-case analysis would suffice without turning to other possibly complicated methods. Therefore, a crucial first step for analysis with missing data is to determine if the missingness mechanism is MCAR.

The most widely used test for MCAR mechanism was due to Little (1988). Although it was proposed in the setting of multivariate normal data, the test is asymptotically valid regardless of the distribution of the data. The basic idea behind the construction of the test is that, if the data are MCAR, the subjects with each particular missingness pattern can be viewed as a random sample from the population, and thus any significant difference between subjects with different missingness patterns provides evidence against MCAR. For longitudinal data with dropouts, Diggle (1989) proposed a nonparametric test and Ridout (1991) considered a parametric alternative by modeling the dropout mechanism. Park and Davis (1993) extended the idea of Little (1988) to the case of incomplete repeated categorical data. Chen and Little (1999) applied similar ideas and developed a test for longitudinal data with intermittent missingness using the generalized estimating equations (GEE) method (Liang and Zeger 1986). The test is carried out by testing the unbiasedness of the GEE across different missingness patterns, and thus is not equivalent to testing MCAR. Besides, this test requires the GEE model to be correctly specified. There have been some recent extensions of Little (1988)'s idea by comparing the means, the covariance matrices and/or the distributions across different missingness patterns (e.g., Kim and Bentler 2002; Jamshidian and Jalal 2010; Li and Yu 2015).

Despite the importance of determining the missingness mechanism, the ultimate task of data analysis is usually the subsequent estimation and inference. All the aforementioned works, however, treat the testing for MCAR as a stand-alone problem without providing a natural way for subsequent estimation once the MCAR mechanism is rejected. The subsequent estimation calls for some existing methods that may require an implementation that is completely different from the testing procedure itself. Our contribution in this paper is to propose a test for MCAR that also takes the subsequent estimation into account, so that an estimator of the quantity of interest with desirable properties is readily available once the MCAR is rejected. Our test does not impose any parametric assumptions on the underlying data distribution.

Our proposed unified procedure for testing and subsequent estimation is based on the calibration idea used in survey sampling literature (Deville and Särndal 1992; Wu and Sitter 2001) combined with the empirical likelihood method (Owen 1988, 2001; Qin and Lawless 1994). Under the MCAR mechanism, the complete cases are a random sample from the population, and thus the calibration weights assigned to the complete cases should be uniform with some random perturbation. Therefore, a significant deviation of the calibration weights from the uniform weights provides evidence against MCAR. Upon rejecting MCAR, the calibration weights can be readily used to construct a weighted estimator of the quantity of interest. Such an estimation approach agrees with the multiply robust estimation procedure in recent missing-data literature (e.g., Han and Wang 2013; Chan and Yam 2014; Han 2014; Han 2016a, 2016b).

For ease of methodology illustration, we take the quantities of interest to be the population means of certain response variables that are subject to missingness whereas some covariates are fully observed, a commonly encountered scenario in practice, especially in survey sampling and causal inference. The calibration weights are derived by matching the weighted average of certain user-specified functions of the covariates based on the complete cases to the unweighted average of those functions based on the whole sample. The functions may be certain moments of the covariates or regression models of the response variables on the covariates. Upon rejecting MCAR, the calibration weights lead to estimators that are the weighted average of the observed values of the response variables, and these estimators are consistent if the missingness of each response variable depends only on the covariates and the corresponding correct regression model is among the user-specified functions used for calibration.

2 A Review of Some Existing Tests for MCAR

Following the notation in Little (1988), let $\boldsymbol{Y}_i = (Y_{1i}, \cdots, Y_{pi})^{\mathrm{T}}$ denote the p-dimensional data vector we intend to collect from subject i, $i = 1, \cdots, n$, and $\boldsymbol{R}_i = (R_{1i}, \cdots, R_{pi})^{\mathrm{T}}$ the vector of missingness indicators for $\boldsymbol{Y}_i$ such that $R_{ki} = 1$ if Y_{ki} is observed and $R_{ki} = 0$ otherwise, $k = 1, \cdots, p$. Under MCAR the probability of observing Y_k given the full data vector $\boldsymbol{Y}$, $\mathbb{P}(R_k = 1 \mid \boldsymbol{Y})$, does not depend on $\boldsymbol{Y}$. Let $\pi_k \equiv \mathbb{P}(R_k = 1)$ denote this probability and assume that $\pi_k > 0$ without loss of generality. Let L denote the number of distinct missingness patterns in the data set, $\mathcal{M}_l$ the set of subjects with pattern l, $l = 1, \cdots, L$, and m_l the number of subjects in $\mathcal{M}_l$. The test statistic proposed by Little (1988) for testing MCAR is

$$D^2 = \sum_{l=1}^{L} m_l (\bar{\boldsymbol{Y}}_{\mathrm{obs},l} - \hat{\boldsymbol{\mu}}_{\mathrm{obs},l})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathrm{obs},l}^{-1} (\bar{\boldsymbol{Y}}_{\mathrm{obs},l} - \hat{\boldsymbol{\mu}}_{\mathrm{obs},l}),$$

where $\bar{\boldsymbol{Y}}_{\mathrm{obs},l}$ is the vector of sample means for the observed variables for pattern l, and $\hat{\boldsymbol{\mu}}_{\mathrm{obs},l}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{obs},l}$ are the maximum likelihood estimators of the mean vector and the covariance matrix for the observed variables for pattern l. Under MCAR, Little (1988) showed that D^2 has an χ^2-distribution with degree of freedom $\sum_{l=1}^{L} p_l - p$ for $\boldsymbol{Y}$ following a multivariate normal distribution, where p_l is the number of observed variables in pattern l, and that this result is asymptotically true for $\boldsymbol{Y}$ following other distributions. Little (1988) also raised the issue of possible heteroscedasticity of covariance matrices across different missingness patterns. For normally distributed data, Kim and Bentler (2002) proposed a method to address this issue by considering a combined test of homogeneity of means and covariance matrices with the test statistic

$$G = \sum_{l=1}^{L} \left[m_l (\bar{\boldsymbol{Y}}_{\mathrm{obs},l} - \hat{\boldsymbol{\mu}}_{\mathrm{obs},l})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathrm{obs},l}^{-1} (\bar{\boldsymbol{Y}}_{\mathrm{obs},l} - \hat{\boldsymbol{\mu}}_{\mathrm{obs},l}) + \frac{m_l - 1}{2} tr \left\{ (\boldsymbol{S}_{\mathrm{obs},l} - \hat{\boldsymbol{\Sigma}}_{\mathrm{obs},l}) \hat{\boldsymbol{\Sigma}}_{\mathrm{obs},l}^{-1} \right\}^2 \right],$$

which asymptotically follows a χ^2-distribution with degree of freedom $\sum_{l=1}^{L} p_l(p_l+3)/2 - p(p+3)/2$, where $\boldsymbol{S}_{\text{obs},l}$ is the sample covariance matrix for the observed variables for pattern l and $tr(\boldsymbol{A})$ is the trace of a matrix $\boldsymbol{A}$. Extensions without the normality assumption can be found in Jamshidian and Jalal (2010) and Li and Yu (2015). Many of the aforementioned tests rely heavily on iterative estimation procedures such as the EM algorithm, which can become computationally burdensome especially when the number of missingness patterns is not small.

3 The Proposed Method

For ease of idea illustration, we first consider the simple scenario where the missingness only occurs to one variable, denoted by Y, and a vector of auxiliary variables $\boldsymbol{X}$ is fully observed. Let R denote the missingness indicator such that $R = 1$ if Y is observed and $R = 0$ otherwise. For a random sample of size n, let $S = \{i : R_i = 1, i = 1, \ldots, n\}$ denote the set of complete cases and $n_1 = \sum_{i=1}^{n} R_i$ the number of complete cases. Under MCAR, S is a random sample from the population, and thus the sample mean of $\boldsymbol{X}$ based on the complete cases should be close to the sample mean based on the whole sample since both are consistent estimators of $E(\boldsymbol{X})$. In other words, if we assign positive weights w_i to the subjects in S so that $\sum_{i\in S} w_i \boldsymbol{X}_i = n^{-1}\sum_{j=1}^{n} \boldsymbol{X}_j$ and $\sum_{i\in S} w_i = 1$, then the w_i can be chosen to be close to the uniform weight $1/n_1$ where the deviation occurs only due to randomness. Therefore, a measure of the deviation from these w_i to $1/n_1$ provides an assessment of whether MCAR holds.

In practice, the ultimate goal is usually to estimate $E(Y)$ regardless of whether Y is MCAR. The estimation is often carried out by fitting a regression model for $E(Y \mid \boldsymbol{X})$ and then taking the sample mean of the fitted values over the whole sample. It is clear that the argument in the previous paragraph on using $\boldsymbol{X}$ to form constraints also applies to regression models viewed as functions of $\boldsymbol{X}$. Following the formulation of the empirical likelihood (EL) method (e.g., Owen, 1988; Qin and Lawless 1994), we consider the weights $\hat{w}_i$ that maximize $\prod_{i\in S} w_i$ subject to the constraints

$$w_i > 0 \quad (i \in S), \quad \sum_{i\in S} w_i = 1, \quad \sum_{i\in S} w_i \boldsymbol{h}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{h}(\boldsymbol{X}_j; \hat{\boldsymbol{\theta}}), \tag{1}$$

where $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ is a d-dimensional vector of user-specified functions of $\boldsymbol{X}$, possibly depending on some parameter $\boldsymbol{\theta}$ that is estimated by $\hat{\boldsymbol{\theta}}$. For example, $\boldsymbol{h}(\boldsymbol{X}; \boldsymbol{\theta})$ may include different moments of $\boldsymbol{X}$ and/or different regression models for $E(Y \mid \boldsymbol{X})$, and in the latter case $\boldsymbol{\theta}$ is the vector of all regression parameters. It turns out that, under MCAR, the $\hat{w}_i$ are the weights we referred to in the previous paragraph that are close to the uniform weights $1/n_1$ where the deviation occurs only due to randomness.

The constraints in (1) are constructed based on the intuition that S is a random sample from the population under MCAR. A natural question then is whether these constraints are still compatible, or in other words whether there still exist w_i satisfying (1), when Y is not MCAR. The answer is affirmative. It can be easily shown that (e.g., Han and Wang 2013)

$$E\left(w(Y,\boldsymbol{X})\left[\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})-E\{\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})\}\right]\mid R=1\right)=\boldsymbol{0},$$

where $w(Y,\boldsymbol{X})=1/\mathbb{P}(R=1\mid Y,\boldsymbol{X})$. Then the constraints in (1) are simply the data version of the above moment equality, and thus are compatible even when Y is not MCAR.

It follows from standard EL theory that the $\hat{w}_i$ that maximize $\prod_{i\in S}w_i$ subject to (1) are given by

$$\hat{w}_i=\frac{1}{n_1}\frac{1}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})}\qquad i\in S,$$

where $\hat{\boldsymbol{\rho}}$ is the Lagrange multiplier solving

$$\frac{1}{n_1}\sum_{i\in S}\frac{\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})}=\boldsymbol{0}\tag{2}$$

and $\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})=\boldsymbol{h}(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})-n^{-1}\sum_{j=1}^{n}\boldsymbol{h}(\boldsymbol{X}_j;\hat{\boldsymbol{\theta}})$. From the EL theory again, under MCAR, we have $\hat{\boldsymbol{\rho}}=O_p(n^{-1/2})$, which implies that the $\hat{w}_i$ are indeed equal to $1/n_1$ with a higher order perturbation. Now define

$$T=\frac{-2\sum\limits_{i\in S}\log(n_1\hat{w}_i)}{1-n_1/n},\tag{3}$$

which is a measure of discrepancy between the $\hat{w}_i$ and $1/n_1$. The following result shows that T can be used to test for MCAR, the proof of which is given in the Appendix.

Theorem 1. *Under H_0: Y is MCAR, the test statistic T has an asymptotic χ^2-distribution with d degrees of freedom.*

When the MCAR is rejected, the $\hat{w}_i$ can be directly used to construct an estimator $\hat{\mu}=\sum_{i\in S}\hat{w}_iY_i$ for the quantity of interest $\mu_0=E(Y)$. The following proposition states the consistency of $\hat{\mu}$.

Proposition. *Under MAR where the missingness of Y only depends on $\boldsymbol{X}$, the estimator $\hat{\mu}$ is consistent for μ_0 if $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})$ contains a correctly specified regression model for $E(Y|\boldsymbol{X})$.*

This result is easy to see. Let $a(\boldsymbol{X};\boldsymbol{\beta})$ be a correctly specified model such that $a(\boldsymbol{X};\boldsymbol{\beta}_0)=E(Y|\boldsymbol{X})$ for some $\boldsymbol{\beta}_0$, then

$$\begin{aligned}\hat{\mu}&=\sum_{i\in S}\hat{w}_i\{Y_i-a(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})\}+\frac{1}{n}\sum_{j=1}^{n}a(\boldsymbol{X}_j;\hat{\boldsymbol{\beta}})\\&\xrightarrow{p}\frac{1}{\mathbb{P}(R=1)}E\left[\frac{R\{Y-a(\boldsymbol{X};\boldsymbol{\beta}_0)\}}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)}\right]+E\{a(\boldsymbol{X};\boldsymbol{\beta}_0)\}=0+\mu_0=\mu_0,\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ that can be derived based on a complete-case analysis because $E(Y|\boldsymbol{X})=E(Y|\boldsymbol{X},R=1)$ due to MAR, $\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta})=\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})-E\{\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})\}$ and $\boldsymbol{\theta}_*$ and $\boldsymbol{\rho}_*$

are the probability limits of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\rho}}$, respectively. Therefore, the usage of the weights $\hat{w}_i$ is two-fold: they provide a test for MCAR and an estimator for μ_0, and thus make our proposed method more attractive than existing ones.

Now we consider the case where $\boldsymbol{Y} = (Y_1, \cdots, Y_p)^{\mathrm{T}}$ and each component of $\boldsymbol{Y}$ is subject to missingness but the auxiliary variables $\boldsymbol{X}$ are still fully observed. Let S_k denote the set of subjects with Y_k observed and n_k the number of subjects in S_k, $k = 1, \ldots, p$. To test if Y_k is MCAR, we can directly apply the test statistic given in (3) to Y_k based on a d_k-dimensional vector of user-specified functions $\boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k)$. Let $\hat{w}_{ki}$, $i \in S_k$, denote the resulting weights for the subjects in S_k. It follows from Theorem 1 that the test statistic

$$T_k = \frac{-2 \sum_{i \in S_k} \log(n_k \hat{w}_{ki})}{1 - n_k/n}$$

asymptotically follows the χ^2-distribution with d_k degrees of freedom if Y_k is MCAR. Furthermore, using the T_k, we are able to construct a test statistic to test if $\boldsymbol{Y}$ is MCAR as shown in the following result, the proof of which is given in the Appendix.

Theorem 2. *Under H_0: $\boldsymbol{Y}$ is MCAR, the test statistic $T_{sum} = \sum_{k=1}^{p} T_k$ has asymptotically the same distribution as $\sum_{l=1}^{m} \lambda_l Q_l$, where $m = d_1 + \cdots + d_p$ and, for $l = 1, \ldots, m$, the Q_l are independent χ^2-distributed random variables with 1 degree of freedom and the λ_l are the eigenvalues of*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I}_{d_1} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1p} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{I}_{d_2} & & \vdots \\ \vdots & & \ddots & \\ \boldsymbol{\Sigma}_{1p} & \cdots & & \boldsymbol{I}_{d_p} \end{pmatrix}.$$

Here $\boldsymbol{I}_{d_k}$ is the identity matrix with dimension d_k and, for $k, r = 1, \ldots, p$ and $k \neq r$,

$$\begin{aligned} \boldsymbol{\Sigma}_{kr} &= \{\pi_k \pi_r (1 - \pi_k)(1 - \pi_r)\}^{-1/2} (\pi_{kr} - \pi_k \pi_r) \\ &\quad \times \left[E\{\boldsymbol{g}_k(\boldsymbol{\theta}_{k*}) \boldsymbol{g}_k(\boldsymbol{\theta}_{k*})^T\}\right]^{-1/2} \left[E\{\boldsymbol{g}_k(\boldsymbol{\theta}_{k*}) \boldsymbol{g}_r(\boldsymbol{\theta}_{r*})^T\}\right] \left[E\{\boldsymbol{g}_r(\boldsymbol{\theta}_{r*}) \boldsymbol{g}_r(\boldsymbol{\theta}_{r*})^T\}\right]^{-1/2}, \end{aligned}$$

$\pi_k = \mathbb{P}(R_k = 1)$, $\pi_{kr} = \mathbb{P}(R_k = 1, R_r = 1)$ and $\boldsymbol{g}_k(\boldsymbol{\theta}_k) \equiv \boldsymbol{g}_k(\boldsymbol{X}; \boldsymbol{\theta}_k) = \boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k) - E\{\boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k)\}$.

The eigenvalues λ_l are not necessarily distinct (e.g., Imhof 1961). In practice, in order to determine the critical value for the asymptotic distribution of T_{sum}, $\boldsymbol{\Sigma}_{kr}$ can be consistently estimated by replacing π_{kr} and π_k with n_{kr}/n and n_k/n, respectively, where n_{kr} is the number of subjects with Y_k and Y_r observed simultaneously, and the expectations can be estimated by sample averages. When the MCAR is rejected, the weights $\hat{w}_{ki}$ used for testing can then be used to construct an estimator for $E(Y_k)$: $\sum_{i=1}^{n} R_{ki} \hat{w}_{ki} Y_{ki}$. Following the same argument as before, such an estimator is consistent if the missingness of Y_k depends only on $\boldsymbol{X}$ and one component of $\boldsymbol{h}_k(\boldsymbol{X}; \boldsymbol{\theta}_k)$ is the correctly specified regression model for $E(Y_k \mid \boldsymbol{X})$.

The construction of constraints in (1) is flexible in the sense that, in principle, any user-specified functions of $\boldsymbol{X}$ can be considered. The use of moments of $\boldsymbol{X}$ is standard in survey sampling literature on the calibration method (e.g., Deville and Särndal 1992; Chen and Sitter 1999). The use of regression models has become popular in recent literature on calibration-based missing data analysis (e.g., Wu and Sitter 2001; Qin and Zhang 2007; Qin et al. 2008; Tan 2010; Han and Wang 2013; Chan and Yam 2014; Han 2014, 2016a, 2016b). Our extensive simulation study shows that, using moments of $\boldsymbol{X}$ tends to lead to more power for the proposed test compared to using regression models only. This makes intuitive sense because (1) holds for any functions of $\boldsymbol{X}$ whereas a regression model only represents a particular function. On the other hand, including a correctly specified regression model helps to achieve estimation consistency, as argued before in this section. Therefore, in practice we would recommend using both moments of $\boldsymbol{X}$ and regression models to construct the constraints in (1).

The power of the proposed test is also affected by the missingness mechanism of each Y_k. If the missingness mechanism does not depend on $\boldsymbol{X}$, then the proposed test has no power detecting deviation from MCAR because the constraints in (1) are all functions of $\boldsymbol{X}$. In addition, for estimation, the proposed procedure implicitly assumes a regression model of $\boldsymbol{Y}$ on $\boldsymbol{X}$. When this assumption is violated, the proposed weighted estimator will no longer be consistent.

Implementation of the proposed test is straightforward. A crucial step is to calculate $\hat{\boldsymbol{\rho}}$ by solving (2). It turns out that this $\hat{\boldsymbol{\rho}}$ can be derived by minimizing $F(\boldsymbol{\rho}) \equiv -\sum_{i\in S} \log\{1+\boldsymbol{\rho}^{\mathrm{T}}\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})\}$, which is a convex minimization problem. See Han (2014) for more discussions on the implementation and for a Newton-Raphson-type algorithm.

4 Extensions to Intermittent Missingness Patterns

We now consider the most challenging case where every variable in the data set is subject to missingness and the missingness pattern is intermittent. Without loss of generality, in this case we drop the notation $\boldsymbol{X}$ and denote the full data vector by $\boldsymbol{Y}$. We assume that there exits a subset of subjects in the sample that have $\boldsymbol{Y}$ fully observed and denote this subset by $\mathcal{M}_1$. Let m_1 be the number of subjects in $\mathcal{M}_1$. Following the notation in Section 3, we let S_k denote the set of subjects with Y_k observed and n_k the number of subjects in S_k, $k = 1, \ldots, p$. Under MCAR, any subset of subjects taken from the original sample based only on their missingness patterns form a random sample from the population. In particular, for any $k = 1, \ldots, p$, the subjects in $\mathcal{M}_1$ and those in S_k with Y_k observed form two random samples, and thus the sample mean of Y_k based on $\mathcal{M}_1$ should be close to the sample mean based on S_k. Such an intuition provides a way to construct constraints on a set of weights for the subjects in $\mathcal{M}_1$, where these weights should be close to the uniform weights under MCAR.

More formally, let w_i be the weights on the subjects in $\mathcal{M}_1$. We consider the $\hat{w}_i$ that maximize

$\prod_{i\in\mathcal{M}_1} w_i$ subject to the following constraints on w_i:

$$w_i > 0, \qquad \sum_{i\in\mathcal{M}_1} w_i = 1, \qquad \sum_{i\in\mathcal{M}_1} w_i Y_{ki} = \bar{Y}_k \text{ for } k \in \mathcal{K}, \tag{4}$$

where $\bar{Y}_k = n_k^{-1}\sum_{i\in S_k} Y_{ki}$ and $\mathcal{K} = \{k^* :\ 1 \le k^* \le p \text{ and } n_{k^*} > m_1\}$. Suppose that $\mathcal{K} = \{k_1, \ldots, k_d\}$ with $d \le p$. We then have

$$\hat{w}_i = \frac{1}{m_1}\frac{1}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i}, \qquad i \in \mathcal{M}_1,$$

where $\hat{\boldsymbol{\rho}}$ solves

$$\frac{1}{m_1}\sum_{i\in\mathcal{M}_1}\frac{\hat{\boldsymbol{g}}_i}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i} = \mathbf{0} \tag{5}$$

and $\hat{\boldsymbol{g}}_i = (Y_{k_1 i} - \bar{Y}_{k_1}, \cdots, Y_{k_d i} - \bar{Y}_{k_d})^{\mathrm{T}}$. A large deviation from the $\hat{w}_i$ to $1/m_1$ will provide evidence against MCAR. More specifically, we define the test statistic as

$$T_{\mathrm{INT}} = -2\sum_{i\in\mathcal{M}_1}\log\left(m_1\hat{w}_i\right),$$

where the subscript "INT" denotes intermittent missingness patterns. The following result gives the asymptotic distribution of T_{INT} and can be used to test if $\boldsymbol{Y}$ is MCAR. The proof is given in the Appendix.

Theorem 3. *Under H_0: $\boldsymbol{Y}$ is MCAR, the test statistic T_{INT} has asymptotically the same distribution as $\sum_{l=1}^{d}\gamma_l Q_l$, where the Q_l are independent χ^2-distributed random variables with 1 degree of freedom and the γ_l are the eigenvalues of $\{E(\boldsymbol{g}^*\boldsymbol{g}^{*T})\}^{-1}\boldsymbol{V}$. Here $\boldsymbol{g}^* = (Y_{k_1} - \mu_{k_1}, \ldots, Y_{k_d} - \mu_{k_d})^T$, $\mu_{k_r} = E(Y_{k_r})$ for $r = 1, \ldots, d$, $\boldsymbol{V} = (v_{rs})_{r,s=1,\ldots,d}$,*

$$\begin{aligned} v_{rr} &= \left(1 - \frac{\pi_c}{\pi_{k_r}}\right) E(Y_{k_r} - \mu_{k_r})^2, \\ v_{rs} &= \left(1 - \frac{\pi_c}{\pi_{k_r}} - \frac{\pi_c}{\pi_{k_s}} + \frac{\pi_c \pi_{k_s k_r}}{\pi_{k_s}\pi_{k_r}}\right) E\{(Y_{k_r} - \mu_{k_r})(Y_{k_s} - \mu_{k_s})\}, \quad r \ne s, \end{aligned}$$

$\pi_c = \mathbb{P}(R_c = 1)$, $\pi_{k_s k_r} = \mathbb{P}(R_{k_s} = 1, R_{k_r} = 1)$ and R_c is the indicator indicating if a subject is in $\mathcal{M}_1$

For implementation, the quantities needed in Theorem 3 are estimated as follows: $\mu_{k_r} \simeq n_{k_r}^{-1}\sum_{i\in S_{k_r}} Y_{k_r i}$, $E(\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}}) \simeq m_1^{-1}\sum_{i\in\mathcal{M}_1}\hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i^{\mathrm{T}}$, $\pi_c \simeq m_1/n$, $\pi_k \simeq n_k/n$, $\pi_{k_s k_r} \simeq n_{k_s k_r}/n$,

$$E(Y_{k_r} - \mu_{k_r})^2 \simeq n_{k_r}^{-1}\sum_{i\in S_{k_r}}(Y_{k_r i} - n_{k_r}^{-1}\sum_{j\in S_{k_r}} Y_{k_r j})^2,$$

$$E\{(Y_{k_r} - \mu_{k_r})(Y_{k_s} - \mu_{k_s})\} \simeq n_{k_s k_r}^{-1}\sum_{i\in S_{k_s k_r}}\{(Y_{k_s i} - n_{k_s}^{-1}\sum_{j\in S_{k_s}} Y_{k_s j})(Y_{k_r i} - n_{k_r}^{-1}\sum_{j\in S_{k_r}} Y_{k_r j})\},$$

where $S_{k_sk_r}$ is the set of subjects with both Y_{k_s} and Y_{k_r} observed and $n_{k_sk_r}$ is the number of subjects in $S_{k_sk_r}$.

Unlike (1) in Section 3 where $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})$ can include both moments of $\boldsymbol{X}$ and regression models for $E(Y \mid \boldsymbol{X})$, for the constraints in (4) we only used moments of $\boldsymbol{Y}$. In principle, regression models for one component of $\boldsymbol{Y}$ conditional on other components can also be included in (4). However, the implementation becomes impractical due to the complexity of intermittent missingness patterns. When MCAR is rejected by the test in Theorem 3, estimators constructed using the calibration weights $\hat{w}_i$ are not consistent in general. For example, $E(Y_k)$ may be estimated by $\sum_{i\in\mathcal{M}_1}\hat{w}_iY_{ki}$, which is simply $\bar{Y}_k = n_k^{-1}\sum_{i\in S_k}Y_{ki}$ from (4) and is not a consistent estimator of $E(Y_k)$ unless the missingness of Y_k does not depend on any other components of $\boldsymbol{Y}$. In this case, similar to all existing methods, some specific model assumptions on both the missingness mechanism and/or the data distribution are needed to obtain consistent estimators for the quantities of interest.

5 Simulation Studies

5.1 Simulation Study 1

For the scenario considered in Section 3, we use a simulation setup mimicing the one in Chen and Little (1999) to study the type I error of the proposed test under MCAR and the power under different missingness mechanisms. Three covariates are independently generated as $X_1 \sim$ Uniform$(-1, 1)$, $X_2 \sim N(0, 1)$ and $X_3 \sim$ Bernoulli(0.5). Given the covariates, $\tilde{Y}_1$ and $\tilde{Y}_2$ are independently generated from $N(X_1+2X_2+3X_3, 1)$. The two response variables are then generated as $Y_1 = \tilde{Y}_1$ and $Y_2 = U\tilde{Y}_1 + (1-U)\tilde{Y}_2$ where $U \sim$ Bernoulli$\{(1 + X_1)/2\}$.

We follow steps similar to those in Chen and Little (1999) to create missing values. First, each subject is classified into one of two sets with probabilities p^s and $1 - p^s$, respectively. Then, in the first set, Y_2 is fully observed while Y_1 is missing with probability p_1^s; in the second set, Y_1 is fully observed while Y_2 is missing with probability p_2^s. The dependence of p^s, p_1^s and p_2^s on $\boldsymbol{X}$ and/or $\boldsymbol{Y}$ determines the missingness mechanism. Table 1 gives a list of some specific combinations of (p^s, p_1^s, p_2^s) we use in the simulation study, where the parameters α_1 and α_2 take different values corresponding to different degrees of departure from MCAR ($\alpha_1 = 0$ and $\alpha_2 = 0$). The missingness mechanism that each specific combination corresponds to is also given. To distinguish different combinations and make them easier to be referred to in Tables 2, 3 and 4, each specific combination, except the one corresponding to MCAR, is assigned a code in the form of "letter-number", where "a" and "b" correspond to $p^s = 0.5$ and $p^s = (1 + X_1)/2$ and "1", "2" and "3" correspond to MAR with missingness depending only on $\boldsymbol{X}$, MAR with missingness depending on the observed response and MNAR, respectively.

Since the correct regression models for $E(Y_1|\boldsymbol{X})$ and $E(Y_2|\boldsymbol{X})$ are linear models with regressors X_1, X_2 and X_3, including both the first moment of $\boldsymbol{X}$ and those linear regression models in $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})$

results in collinearity. Therefore, we simply take $\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{X}$. We compare the proposed test with the ones in Little (1988) and Chen and Little (1999). Simulation results are summarized based on 1000 replications with sample size $n = 100$ and 200 for each replication, and the significance level is set at 5%.

Table 2 contains results on the type I error under MCAR and the power under different missingness mechanisms. The overall performance of the proposed test is quite close to that of Little (1988), and both are better than the test of Chen and Little (1999). As pointed out by Chen and Little (1999), their test actually tests the unbiasedness of a set of generalized estimating equations rather than the MCAR mechanism, and thus the performance depends on the specific form of the estimating equations and does not always agree with the theoretical behaviour of a test for MCAR.

Tables 3 and 4 show the performance of the weighted estimators of $E(Y_1)$ and $E(Y_2)$ based on the calibration weights that were used to construct the test statistic, with sample size $n =$ 100 and 200, respectively. Under MCAR, both the proposed estimator $\hat{\mu}_k$ and the complete-case average estimator $\hat{\mu}_{k\mathrm{cc}}$ have negligible bias, $k = 1, 2$. The estimator $\hat{\mu}_{k\mathrm{cc}}$ loses consistency when the missingness mechanism is no longer MCAR, demonstrated by its non-negligible relative bias in those cases. On the contrary, the proposed estimator $\hat{\mu}_k$ is still consistent in cases a-1 and b-1 where the missingness depends only on the fully observed covariates. Surprisingly, for the other cases a-2, a-3, b-2 and b-3, although $\hat{\mu}_k$ is theoretically not consistent, its relative bias is very small compared to that of $\hat{\mu}_{k\mathrm{cc}}$. This observation that calibration-based estimators have relatively small bias even if their theoretical consistency cannot be formally shown has also been noted in Han (2014, 2016a) and demonstrates the superiority of these estimators.

5.2 Simulation Study 2

For the scenario of intermittent missingness considered in Section 4, we use a simulation setup similar to that in Little (1988). Random variables $\tilde{Y}_1$, $\tilde{Y}_2$, $\tilde{Y}_3$ and $\tilde{Y}_4$ are generated as

$$\begin{aligned}
\tilde{Y}_1 &= Z_1\sqrt{1/q},\\
\tilde{Y}_2 &= Z_1\sqrt{0.9/q} + Z_2\sqrt{0.1/q},\\
\tilde{Y}_3 &= Z_1\sqrt{0.2/q} + Z_2\sqrt{0.1/q} + Z_3\sqrt{0.7/q},\\
\tilde{Y}_4 &= -Z_1\sqrt{0.6/q} + Z_2\sqrt{0.25/q} + Z_3\sqrt{0.1/q} + Z_4\sqrt{0.05/q},
\end{aligned}$$

where $(Z_1, Z_2, Z_3, Z_4)^{\mathrm{T}} \sim N(0, \boldsymbol{I})$. Three different distributions for the final responses Y_1, Y_2, Y_3 and Y_4 are considered: multivariate normal distribution by setting $q = 1$ and $\boldsymbol{Y} = \tilde{\boldsymbol{Y}}$, lognormal distribution by setting $q = 1$ and $\boldsymbol{Y} = \exp(\tilde{\boldsymbol{Y}})$, and multivariate t-distribution with 3 degrees of freedom by setting $q \sim \chi^2(3)$ and $\boldsymbol{Y} = \tilde{\boldsymbol{Y}}$. The missingness mechanism is set to be MCAR with 70% of the subjects being complete cases, i.e., with the pattern $(1,1,1,1)$ for $\boldsymbol{R} = (R_1, R_2, R_3, R_4)$, and 5% for each of the six patterns $(1,1,1,0)$, $(1,1,0,0)$, $(1,1,0,1)$, $(1,0,0,1)$, $(1,0,1,1)$ and $(1,0,1,0)$.

Therefore, Y_1 is always observed but each of Y_2, Y_3 and Y_4 is observed only in four different patterns.

For this simulation setup, let w_i be the weights on the subjects in $\mathcal{M}_1$, i.e., the subjects with pattern $(1,1,1,1)$. The calibration constraints in (4) now become

$$\begin{aligned} w_i > 0, \quad \sum_{i\in\mathcal{M}_1} w_i = 1, \\ \sum_{i\in\mathcal{M}_1} w_i Y_{1i} = \frac{1}{n}\sum_{j=1}^{n} Y_{1j}, \\ \sum_{i\in\mathcal{M}_1} w_i Y_{2i} = \frac{1}{0.85n}\sum_{j\in S_2} Y_{2j}, \\ \sum_{i\in\mathcal{M}_1} w_i Y_{3i} = \frac{1}{0.85n}\sum_{j\in S_3} Y_{3j}, \\ \sum_{i\in\mathcal{M}_1} w_i Y_{4i} = \frac{1}{0.85n}\sum_{j\in S_4} Y_{4j}. \end{aligned}$$

Table 5 contains simulation results on type I error summarized based on 1000 replications, with the test of Little (1988) included as a comparison. While the comparison is inconclusive with $n = 100$, it seems to become clear as n increases to 200, 500 and 800. Under the latter three sample sizes, when the data are normally distributed, both tests have type I error close to the nominal level. When the data distribution is skewed as in the lognormal case, Little (1988)'s test tends to have type I error larger than the nominal level when the sample size is not large enough, whereas the proposed test has type I error closer to the nominal level. For the t-distribution case, the proposed test also has type I error closer to the nominal level. The better overall performance of the proposed test is partially due to the nature of the empirical likelihood method that it does not require assumptions of a specific data distribution. Similar to Little (1988), power analysis is not included here.

6 Data Application

As an application of the proposed method, we consider data collected from 2002 New York City Social Indicators Survey. This survey was conducted by School of Social Work at Columbia University to study the household demographics of a representative sample from New York City. Detailed information can be found in the Social Indicators Survey Codebook, downloadable from `http://www.stat.columbia.edu/~gelman/arm/examples/sis/`, along with the data set.

We focus on subjects who worked in 2001, with either a regular or an odd job. Our main interest is to estimate the population mean of annual income (*N09_d*) and total assets (not including home) (*N33*). Three auxiliary variables are considered: age (*age*) with a range from 18 to 80, number of months worked altogether in 2001 with a range from 1 to 12 (*N05*), and number of hours worked per week with a range from 1 to 97 (*N06*). Our analysis is based on $n = 1049$ subjects for whom these auxiliary variables are available. For the two variables of interest, *N09_d* and *N33*, values "do not know" and "refused" are also treated as missing data in our analysis. In total, there are 378 (36%) subjects with *N09_d* missing and 479 (46%) subjects with *N33* missing.

We use the first moment of the auxiliary variables to construct the calibration constraints, and this is equivalent to fitting a linear regression of the responses on the auxiliary variables with main effects. For estimation, in addition to our proposed calibration-based estimator (CAL), we also calculate the inverse probability weighted (IPW) estimator (Horvitz and Thompson 1952), the augmented IPW (AIPW) estimator (Robins et al. 1994) and the average of the complete cases (CC). For the IPW and AIPW estimators, the missingness probability is modeled by a logistic regression, and for the AIPW estimator, the response is modeled by a linear regression, both including main effects of the three auxiliary variables. Standard errors for all estimators are calculated based on 1000 bootstrap samples.

Table 6 contains results of our analysis. For testing MCAR, both the individual tests and the overall test are conducted, together with Little's (1998) test. All these tests reject MCAR. For estimation, the estimated values and standard errors of our proposed estimator are very close to those of the IPW and AIPW estimators. The complete-case analysis produces quite different results, indicating its bias in estimation. Our proposed estimator is calculated based on the same weights that were used for testing MCAR. If one were to use existing methods, however, one would need to apply Little's (1998) test first and then calculate the IPW/AIPW estimator, with completely different implementations for testing and for estimation.

7 Concluding Remarks

Ascertaining the missingness mechanism is always a crucial step in missing data analysis. While the MAR is in general not testable, the MCAR is. Under MCAR, data analysis becomes fairly easy since a complete case analysis would be sufficient. We have proposed a nonparametric approach based on the empirical likelihood method to test MCAR. The proposed approach not only provides an alternative to existing tests, but more importantly, for the commonly seen scenarios with the presence of fully observed covariates, it leads to a unified procedure for estimation after the MCAR is rejected with little extra effort beyond the calculation of the test statistic. Existing tests, on the contrary, focus exclusively on testing, and the estimation after MCAR is rejected has to invoke possibly completely different procedures.

In this paper we considered estimating the population mean of certain response variables that are subject to missingness. Extensions to estimating parameters defined through estimating equations can be made. Since the missingness mechanism does not depend on the model for parameter estimation, a simple extension is to directly apply the proposed test when the parameters of interest are defined through estimating equations. The resulting weights can then be used to weight the estimating equations for estimation. But estimators derived in this way may not be consistent under MAR because the calibration constraints in this paper are constructed to ensure consistency under MAR when estimating population means. A more complex extension leading to consistency under MAR is to follow the idea in Han (2014) and construct calibration constraints using the estimating functions rather than the moments of fully observed variables. A detailed account of this extension is beyond the scope of this paper and is of interest for future research.

Numerical performance of the proposed procedure could be jeopardized if the number of constraints gets too large. This is in particular an issue when the dimension of the fully observed covariates is high. In this case, the functions used for calibration constraints need to be carefully chosen. One possible solution would be to use moments of those covariates that are considered more relevant in explaining the missingness mechanism, instead of moments of all the covariates, combined with some selected regression models, to construct the calibration constraints. More investigation in the case of high dimensional covariates is needed, both theoretically and numerically.

Acknowledgment

We wish to thank the Editor, an Associate Editor and two reviewers for their valuable comments that have helped improve the quality of this work. Financial support for this research was partially provided by Natural Sciences and Engineering Research Council of Canada (NSERC) to the second and the third authors.

References

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statist. Sci.*, 29(3):380–396.

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

Chen, J. and Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9:385–406.

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87(418):376–382.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, 45(4):1255–1258.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *J. Amer. Statist. Assoc.*, 109(507):1159–1173.

Han, P. (2016a). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scand. J. Stat.*, 43:246–260.

Han, P. (2016b). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103(3):683–700.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47(260):663–685.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3-4):419–426.

Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.

Kim, K. H. and Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67(4):609–624.

Li, J. and Yu, Y. (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika*, 80(3):707–726.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika*, 73:13–22.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *J. Amer. Statist. Assoc.*, 83(404):1198–1202.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, Inc. New York, 2 edition.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. (2001). *Empirical likelihood*. Chapman & Hall/CRC Press, New York.

Park, T. and Davis, C. S. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49(2):631–638.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, 22(1):300–325.

Qin, J., Shao, J., and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *J. Amer. Statist. Assoc.*, 103(482):797–810.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 69(1):101–122.

Ridout, M. S. (1991). Testing for random dropouts in repeated measurement data (reader reaction). *Biometrics*, 47(4):1619–1621.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89(427):846–866.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Wu, C. and Lu, W. (2016). Calibration weighting methods for complex surveys. *Int. Stat. Rev.*, 84(1):79–98.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.*, 96:185–193.

Corresponding author

PEISONG HAN, *Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.*
Email: peisong@umich.edu

Appendix

Proof of Theorem 1. Let $\boldsymbol{\theta}_*$ denote the probability limit of $\hat{\boldsymbol{\theta}}$ and $\pi_0 = \mathbb{P}(R = 1)$. A Taylor expansion of (2) at $(\boldsymbol{\rho} = \mathbf{0}, \boldsymbol{\theta}_*)$ yields

$$\begin{aligned} \mathbf{0} &= \frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) - \left\{\frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\hat{\boldsymbol{\rho}} \\ &\quad + \left[\frac{1}{n}\sum_{i=1}^{n} R_i \left\{\frac{\partial \boldsymbol{h}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}} - \frac{1}{n}\sum_{j=1}^{n}\frac{\partial \boldsymbol{h}(\boldsymbol{X}_j;\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}}\right\}\right](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} R_i \hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) - \pi_0 E\left\{\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\hat{\boldsymbol{\rho}} + o_p(n^{-1/2}), \end{aligned}$$

where $\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta}) - E\{\boldsymbol{h}(\boldsymbol{X};\boldsymbol{\theta})\}$. This implies

$$n^{1/2}\hat{\boldsymbol{\rho}} = \left[\pi_0 E\left\{\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\right]^{-1} n^{-1/2}\sum_{i=1}^{n} R_i\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) + o_p(1).$$

On the other hand, simple calculations show that

$$n^{-1/2}\sum_{i=1}^{n} R_i\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) = n^{-1/2}\sum_{i=1}^{n}(R_i - \pi_0)\boldsymbol{g}(\boldsymbol{X}_i;\boldsymbol{\theta}_*) + o_p(1),$$

and thus

$$n^{1/2}\hat{\boldsymbol{\rho}} \xrightarrow{d} N\left(\boldsymbol{0}, \frac{1-\pi_0}{\pi_0}\left[E\left\{\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}_*)^{\mathrm{T}}\right\}\right]^{-1}\right).$$

A Taylor expansion of (3) at $(\boldsymbol{\rho} = \boldsymbol{0}, \boldsymbol{\theta}_*)$ gives

$$\begin{aligned} T &= (1-\frac{n_1}{n})^{-1}\left[2\left\{n^{-1/2}\sum_{i=1}^{n} R_i\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\right\}^{\mathrm{T}} n^{1/2}\hat{\boldsymbol{\rho}}\right. \\ &\quad \left. -n^{1/2}\hat{\boldsymbol{\rho}}^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^{n} R_i\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)^{\mathrm{T}}\right\} n^{1/2}\hat{\boldsymbol{\rho}}\right] + o_p(1) \\ &= (1-\frac{n_1}{n})^{-1} n^{1/2}\hat{\boldsymbol{\rho}}^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^{n} R_i\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)\hat{\boldsymbol{g}}(\boldsymbol{X}_i;\boldsymbol{\theta}_*)^{\mathrm{T}}\right\} n^{1/2}\hat{\boldsymbol{\rho}} + o_p(1) \xrightarrow{d} \chi^2_d. \end{aligned}$$

Proof of Theorem 2. Some calculations show that $T_{\text{sum}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + o_p(1)$, where

$$\boldsymbol{W} = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{W}_{1i}^{\mathrm{T}}, \ldots, \boldsymbol{W}_{pi}^{\mathrm{T}})^{\mathrm{T}}$$

and

$$\boldsymbol{W}_{ki} = \{\pi_k(1-\pi_k)\}^{-1/2}\left[E\{\boldsymbol{g}_k(\boldsymbol{\theta}_{k*})\boldsymbol{g}_k(\boldsymbol{\theta}_{k*})^{\mathrm{T}}\}\right]^{-1/2}(R_{ki} - \pi_k)\boldsymbol{g}_{ki}(\boldsymbol{\theta}_{k*}).$$

It is easy to check that $\mathrm{Var}(\boldsymbol{W}_k) = \boldsymbol{I}_{d_k}$ and $\mathrm{Cov}(\boldsymbol{W}_k, \boldsymbol{W}_r) = \boldsymbol{\Sigma}_{kr}$. Therefore we have $\boldsymbol{W} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and thus the desired result follows (e.g., Imhof 1961).

Proof of Theorem 3. A Taylor expansion of (5) at $\boldsymbol{\rho}^* = \boldsymbol{0}$ yields

$$n^{1/2}\hat{\boldsymbol{\rho}} = \{E(R_c\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1} n^{-1/2}\sum_{i=1}^{n} R_{ci}\hat{\boldsymbol{g}}_i + o_p(1).$$

Some calculations show that

$$n^{-1/2}\sum_{i=1}^{n} R_{ci}\hat{\boldsymbol{g}}_i = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\varphi}_i + o_p(1) \equiv n^{-1/2}\sum_{i=1}^{n}(\varphi_{k_1 i},\ldots,\varphi_{k_d i})^{\mathrm{T}} + o_p(1),$$

where $\varphi_{k_r} = (R_c - R_{k_r}\pi_c/\pi_{k_r})(Y_{k_r} - \mu_{k_r})$ for $r = 1,\ldots,d$. It is easy to see that $E(\boldsymbol{\varphi}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varphi}) = \pi_c\boldsymbol{V}$. Therefore

$$n^{1/2}\hat{\boldsymbol{\rho}} \xrightarrow{d} N\left(\mathbf{0}, \pi_c^{-1}\{E(\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1}\boldsymbol{V}\{E(\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}^{-1}\right).$$

A Taylor expansion of T_{INT} at $\boldsymbol{\rho}^* = \mathbf{0}$ gives

$$T_{\mathrm{INT}} = n^{1/2}\hat{\boldsymbol{\rho}}^{\mathrm{T}}\{E(R_c\boldsymbol{g}^*\boldsymbol{g}^{*\mathrm{T}})\}n^{1/2}\hat{\boldsymbol{\rho}} + o_p(1).$$

The desired result then follows.

Table 1: The combinations of (p^s, p_1^s, p_2^s) used in Simulation Study 1

p^s	p_1^s	p_2^s	Mechanism	code
0.5	$\{1+\exp(0.5)\}^{-1}$	$\{1+\exp(0.5)\}^{-1}$	MCAR	
0.5	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 X_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 X_2)\}^{-1}$	MAR	a-1
0.5	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_1)\}^{-1}$	MAR	a-2
0.5	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_1)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_2)\}^{-1}$	MNAR	a-3
$(1+X_1)/2$	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 X_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 X_2)\}^{-1}$	MAR	b-1
$(1+X_1)/2$	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_2)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_1)\}^{-1}$	MAR	b-2
$(1+X_1)/2$	$\{1+\exp(0.5-\alpha_1/2+\alpha_1 Y_1)\}^{-1}$	$\{1+\exp(0.5-\alpha_2/2+\alpha_2 Y_2)\}^{-1}$	MNAR	b-3

Table 2: Results on Type I error under MCAR and power under different missingness mechanisms for Simulation Study 1 based on 1000 replications. The significance level is set to be 5%. The numbers are percentages.

		$n = 100$						$n = 200$					
		Little	C&L	T_{sum}	Little	C&L	T_{sum}	Little	C&L	T_{sum}	Little	C&L	T_{sum}
α_1	α_2	(a) $p^s = 0.5$			(b) $p^s = (1 + X_1)/2$			(a) $p^s = 0.5$			(b) $p^s = (1 + X_1)/2$		
		MCAR						MCAR					
0	0	4.3	30	5.7	–	–	–	3.7	18.3	4.1	–	–	–
		a-1 MAR			b-1 MAR			a-1 MAR			b-1 MAR		
0.3	-0.3	6.7	31.6	13.9	78.9	33.9	90.6	15.6	16.6	23.2	99	18.6	99.8
0.6	-0.3	15.8	29.6	25	86.8	31.7	95.1	35.1	17.8	47.5	99.7	18.3	99.8
0.3	0.3	11.6	28.8	12.5	84.1	29.3	92.9	23.3	15.5	20.1	99.6	16.1	99.8
0.6	0.3	25.5	26.7	23.6	91.5	27.3	96.9	57.1	16.3	49.9	99.9	17.6	99.9
		a-2 MAR			b-2 MAR			a-2 MAR			b-2 MAR		
0.3	-0.3	45.2	39.1	55.7	98.7	38.5	99.5	82.1	27.4	86.3	100	27.8	100
0.6	-0.3	79.2	44.5	83	99.8	44.8	99.9	99.1	32.6	99.2	100	40	100
0.3	0.3	67.8	44.3	58.6	96.9	45.9	97.3	97.1	33.4	93.6	100	30.7	99.9
0.6	0.3	93.8	49.3	89.8	99.8	50.7	99.8	100	41.2	99.9	100	40.3	100
		a-3 MNAR			b-3 MNAR			a-3 MNAR			b-3 MNAR		
0.3	-0.3	39.1	35.2	55.8	98.7	35	99.4	77.6	21.9	87.1	100	21.6	100
0.6	-0.3	72.2	39	85.1	99.4	35.7	99.7	97.7	25.9	98.6	100	24.7	100
0.3	0.3	63.1	40.5	59.8	96.4	44	97.7	95.7	25.7	93.6	100	25.5	100
0.6	0.3	91.7	44.2	89.3	99.7	44.6	99.5	99.9	30.6	99.9	100	27.2	100

Little: the test in Little (1988). C&L: the test in Chen and Little (1999). T_{sum}: our proposed test.

Table 3: Results on estimation of $E(Y_1) = E(Y_2) = 1.5$ using the calibration weights for Simulation Study 1 based on $n = 100$ and 1000 replications. The numbers have been multiplied by 100.

		Estimation of $E(Y_1)$				Estimation of $E(Y_2)$			
		$\hat{\mu}_1$		$\hat{\mu}_{1\mathrm{cc}}$		$\hat{\mu}_2$		$\hat{\mu}_{2\mathrm{cc}}$	
α_1	α_2	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE
MCAR									
0	0	-1	28	0	31	0	28	0	31
a-1 MAR									
0.3	-0.3	-1	28	5	32	0	28	-6	31
0.6	-0.3	-1	28	12	36	0	28	-6	31
0.3	0.3	-1	28	5	32	0	28	6	33
0.6	0.3	-1	28	12	36	0	28	6	33
a-2 MAR									
0.3	-0.3	1	28	16	38	-2	28	-20	43
0.6	-0.3	1	28	25	48	-2	28	-20	43
0.3	0.3	1	28	16	38	1	28	16	38
0.6	0.3	1	28	25	48	1	28	16	39
a-3 MNAR									
0.3	-0.3	2	28	18	40	-3	29	-22	45
0.6	-0.3	3	28	27	50	-3	29	-22	45
0.3	0.3	2	28	18	40	2	28	17	40
0.6	0.3	3	28	27	50	2	28	17	40
b-1 MAR									
0.3	-0.3	0	28	12	36	0	28	-10	33
0.6	-0.3	-1	28	18	42	0	28	-10	33
0.3	0.3	0	28	12	36	0	28	0	30
0.6	0.3	-1	28	18	42	0	28	0	30
b-2 MAR									
0.3	-0.3	1	28	21	44	-3	28	-28	52
0.6	-0.3	1	28	31	54	-3	28	-28	52
0.3	0.3	1	28	21	44	1	28	12	35
0.6	0.3	1	28	31	54	1	28	12	35
b-3 MNAR									
0.3	-0.3	2	28	23	45	-4	28	-29	53
0.6	-0.3	4	28	33	58	-4	29	-29	53
0.3	0.3	2	28	23	46	2	28	12	35
0.6	0.3	4	28	33	58	2	28	12	35

$\hat{\mu}_k$ and $\hat{\mu}_{k\mathrm{cc}}$: estimators of $E(Y_k)$ based on our proposed procedure and based on complete-case analysis, respectively, $k = 1, 2$. rBias: relative bias $1000^{-1} \sum_{b=1}^{1000} \{\hat{\mu}_{kb} - E(Y_k)\}/E(Y_k)$, where $\hat{\mu}_{kb}$ is the estimate of $E(Y_k)$ from the bth replication. RMSE: root mean square error.

Table 4: Results on estimation of $E(Y_1) = E(Y_2) = 1.5$ using the calibration weights for Simulation Study 1 based on $n = 200$ and 1000 replications. The numbers have been multiplied by 100.

		Estimation of $E(Y_1)$				Estimation of $E(Y_2)$			
		$\hat{\mu}_1$		$\hat{\mu}_{1\text{cc}}$		$\hat{\mu}_2$		$\hat{\mu}_{2\text{cc}}$	
α_1	α_2	rBias	RMSE	rBias	RMSE	rBias	RMSE	rBias	RMSE
				MCAR					
0	0	0	19	0	21	0	20	0	21
				a-1 MAR					
0.3	-0.3	0	19	6	23	0	20	-5	22
0.6	-0.3	0	19	12	28	0	20	-5	22
0.3	0.3	0	19	6	23	0	19	7	23
0.6	0.3	0	19	12	28	0	19	7	23
				a-2 MAR					
0.3	-0.3	1	19	17	33	-1	20	-20	37
0.6	-0.3	2	19	26	44	-1	20	-20	37
0.3	0.3	1	19	17	33	2	19	17	32
0.6	0.3	2	19	26	44	2	19	17	32
				a-3 MNAR					
0.3	-0.3	2	19	18	34	-3	20	-21	38
0.6	-0.3	4	20	28	46	-3	20	-21	38
0.3	0.3	2	19	18	34	3	20	18	34
0.6	0.3	4	20	28	46	3	20	18	34
				b-1 MAR					
0.3	-0.3	0	19	12	28	0	20	-10	26
0.6	-0.3	0	19	19	35	0	20	-10	26
0.3	0.3	0	19	12	28	0	20	0	22
0.6	0.3	0	19	19	35	0	20	0	22
				b-2 MAR					
0.3	-0.3	1	19	21	38	-2	20	-27	46
0.6	-0.3	1	19	31	51	-2	20	-27	46
0.3	0.3	1	19	21	38	2	20	12	27
0.6	0.3	1	19	31	51	2	20	12	27
				b-3 MNAR					
0.3	-0.3	3	20	23	40	-3	20	-28	48
0.6	-0.3	5	20	34	54	-3	20	-28	48
0.3	0.3	3	20	23	40	3	20	13	28
0.6	0.3	5	20	34	54	3	20	13	28

$\hat{\mu}_k$ and $\hat{\mu}_{k\text{cc}}$: estimators of $E(Y_k)$ based on our proposed procedure and based on complete-case analysis, respectively, $k = 1, 2$. rBias: relative bias $1000^{-1} \sum_{b=1}^{1000} \{\hat{\mu}_{kb} - E(Y_k)\}/E(Y_k)$, where $\hat{\mu}_{kb}$ is the estimate of $E(Y_k)$ from the bth replication. RMSE: root mean square error.

Table 5: Results on Type I error under MCAR for Simulation Study 2 based on 1000 replications. The numbers are percentages.

		significance level							
		1%		5%		10%		20%	
Distribution	n	Little	T_{INT}	Little	T_{INT}	Little	T_{INT}	Little	T_{INT}
Normal	100	1	3.5	4.6	10.2	10.6	15.4	20.3	25.7
	200	0.9	1	5.3	5.9	9.6	10.3	19	20
	500	0.7	0.8	5.2	4.4	9.8	9	19.9	19.2
	800	0.9	1.2	5	5.8	9.6	10.6	18.3	21.1
Lognormal	100	3.3	1.4	10	5.7	16.3	12.7	25.4	25.2
	200	3.6	0.8	9.6	4.3	14.8	9.7	23.4	22.4
	500	2.7	0.5	7.5	2.8	14.3	7.9	21.9	19.2
	800	2.2	1	5.2	4.5	10.3	10.1	20.2	21.2
t on 3 df	100	2.9	3.2	7.6	7.9	12.1	12.7	21.9	21.7
	200	3.1	2	8.3	6.8	12.5	10.9	21.4	19.6
	500	2.4	0.8	7.1	3.9	12.6	8.5	22.8	18.6
	800	2.2	1.2	7.1	4.7	12.1	10.1	21.4	20.5

Little: the test in Little (1988). T_{INT}: our proposed test.

Table 6: Results of the analysis of the 2002 New York City Social Indicators Survey ($n = 1049$). The estimates and standard errors are in hundreds

					Subsequent Estimation			
Testing MCAR					N09_d		N33	
Test	Value	DF	p-value	Estimator	Estimate	S.E.	Estimate	S.E.
$T_{\mathrm{N09\_d}}$	49.03	3	<0.0001	CAL	498.90	35.03	1425.63	330.31
T_{N33}	14.69	3	0.0021	CC	521.81	36.80	1358.24	313.12
T_{sum}	63.72	–	<0.0001	IPW	499.00	35.00	1426.61	329.19
Little	87.62	11	<0.0001	AIPW	498.97	35.06	1426.30	330.49

$T_{\mathrm{N09\_d}}$ and T_{N33}: our proposed individual test for *N09_d* and *N33* respectively. T_{sum}: our proposed overall test. Little: the test in Little (1988).

Value: value of corresponding test statistic. DF: degrees of freedom of the asymptotic χ^2-distribution.

CAL: our proposed calibration-based estimator. CC: the average of the complete cases. IPW: inverse probability weighted estimator. AIPW: augmented inverse probability weighted estimator. S.E.: bootstrap standard error.