Little's Test of Missing Completely at Random

Cheng Li
Northwestern University
Evanston, IL
chengli2014@u.northwestern.edu

Abstract. In missing data analysis, Little's test (Little 1988) is useful for testing the assumption of missing completely at random (MCAR) for multivariate partially observed quantitative data. I introduce the `mcartest` command that implements Little's MCAR test and its extension for testing the covariate-dependent missingness (CDM). The command also includes an option to perform the likelihood ratio test with adjustment for unequal variances. I illustrate the usage of `mcartest` through an example, and evaluate the finite sample performance of these tests in simulation studies.

Keywords: st0001, CDM, MAR, MCAR, MNAR, chi-square, mcartest, missing data, missing-value patterns, multivariate, power

1 Introduction

Statistical inference based on incomplete data typically involves certain assumptions for the missing data mechanism. The validity of these assumptions requires formal evaluation before any further analysis. For example, likelihood based inference is valid only if the missing data mechanism is ignorable (Rubin 1976), which usually relies on the missing at random assumption (MAR). MAR assumes that the missingness of the data may depend on the observed data, but is independent of the unobserved data. Therefore testing MAR is in general impossible since it requires unavailable information about the missing data. Instead, the missing completely at random assumption (MCAR) assumes that the missingness of the data is independent of both the observed and unobserved data, which is stronger than MAR and possible to test using only the observed data. When missing data mechanism depends on the unobserved data, data are missing not at random (MNAR). Although the likelihood inference only requires the MAR assumption, testing of MCAR is still of interest in real applications, since many simple missing data methods such as complete case analysis are valid only under MCAR (Chapter 3 of Little and Rubin 1987, also see the blood test example in Section 4). Also the maximum likelihood estimation for the multivariate normal model may be more sensitive to the distributional assumption when the data are not MCAR (Little 1988).

In this article, I present a new command `mcartest` that implements the chi-square test of MCAR for multivariate quantitative data proposed by Little (1988), which tests whether there exists significant difference between the

means of different missing-value patterns. The test statistic takes a similar form to the likelihood ratio statistic for multivariate normal data and is asymptotically chi-square distributed under the null hypothesis that there are no differences between the means of different missing-value patterns. Rejection of the null provides sufficient evidence to indicate that the data are not MCAR. The command also accommodates the testing of covariate-dependent missingness assumption, a straightforward extension of Little's MCAR test when covariates are present. It also allows unequal variances between different missing-value patterns.

2 Methods and formulas

2.1 MCAR, MAR, MNAR, and CDM

First I introduce the formal definitions of the four missing data mechanisms. Suppose we have an i.i.d. sequence of p-dimensional vectors $\boldsymbol{y}_i = (y_{i1}, ..., y_{ip})^\top$, $i = 1, 2, ..., n$ where n is the sample size, and $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)^\top$ is the $n \times p$ data matrix. Hereafter we are mainly interested in testing if $\boldsymbol{Y}$ is MCAR. Denote the observed entries and missing entries of $\boldsymbol{Y}$ as $\boldsymbol{Y_o}$ and $\boldsymbol{Y_m}$, respectively. In some situations, we may also have additional completely observed q-dimensional covariates $\boldsymbol{x}$. Let $\boldsymbol{X}$ be the $n \times q$ data matrix of covariate values. Let the p-dimensional vector $\boldsymbol{r}_i = (r_{i1}, ..., r_{ip})^\top$ denote the indicator of whether each component in vector $\boldsymbol{y}_i$ is observed, i.e. $r_{ik} = 1$ if y_{ik} is observed and $r_{ik} = 0$ if y_{ik} is missing for $i = 1, 2, ..., n$ and $k = 1, 2, ..., p$. The stacked matrix of $\boldsymbol{r}$ is $\boldsymbol{R} = (\boldsymbol{r}_1, ..., \boldsymbol{r}_n)^\top$. Then the MAR assumption is defined as

$$\Pr(\boldsymbol{R}|\boldsymbol{Y_m}, \boldsymbol{Y_o}, \boldsymbol{X}) = \Pr(\boldsymbol{R}|\boldsymbol{Y_o}, \boldsymbol{X}) \tag{1}$$

In other words, the distribution of the missing indicators depends only on the observed data.

The stronger assumption of MCAR is defined as

$$\Pr(\boldsymbol{R}|\boldsymbol{Y_m}, \boldsymbol{Y_o}, \boldsymbol{X}) = \Pr(\boldsymbol{R}) \tag{2}$$

which implies that the missing indicators are completely independent of both the missing data and the observed data. Note that here $\boldsymbol{R}$ is also independent of covariates $\boldsymbol{X}$, as suggested by Little (1995). This means that under the MCAR assumption, the missingness should be totally independent of any observed variables. Instead, if $\boldsymbol{R}$ only depends on covariates $\boldsymbol{X}$

$$\Pr(\boldsymbol{R}|\boldsymbol{Y_m}, \boldsymbol{Y_o}, \boldsymbol{X}) = \Pr(\boldsymbol{R}|\boldsymbol{X}) \tag{3}$$

then Little (1995) suggested that (3) be referred to as "covariate-dependent missingness" (CDM) (Chapter 17 of Fitzmaurice et al. 2009), while the term "MCAR" is reserved for (2). It is worth noting that according to the definition, CDM is a special case of MAR since covariates $\boldsymbol{x}$ are always fully observed. Finally, any missing data mechanism that does not satisfies (1) is MNAR.

2.2 Test of MCAR

In Little's test of MCAR (Little 1988), the data $\boldsymbol{y}_i$, $(i = 1, 2, ..., n)$ are modeled as p-dimensional multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with part of the components in $\boldsymbol{y}_i$s missing. When the normality is not satisfied, Little's test still works in the asymptotic sense for quantitative random vectors $\boldsymbol{y}_i$s, but is not suitable for categorical variables (Little 1988). We suppose that there are a total of J missing-value patterns among all $\boldsymbol{y}_i$s. For each pattern j, let $\boldsymbol{o}_j$ and $\boldsymbol{m}_j$ be the index sets of the observed components and the missing components, respectively, and $p_j = |\boldsymbol{o}_j|$ is the number of observed components in pattern j. Furthermore, let $\boldsymbol{\mu}_{\boldsymbol{o}_j}$ and $\boldsymbol{\Sigma}_{\boldsymbol{o}_j}$ be the $p_j \times 1$-dimensional mean vector and the $p_j \times p_j$ covariance matrix of only the observed components for jth missing pattern, and $\overline{\boldsymbol{y}}_{\boldsymbol{o}_j}$ $(p_j \times 1)$ is the observed sample average for the jth missing pattern. Finally, let $\boldsymbol{I}_j \subseteq \{1, 2, ..., n\}$ be the index set of pattern j in the sample, and $n_j = |\boldsymbol{I}_j|$, then $\sum_{j=1}^{J} n_j = n$.

The Little's χ^2 test statistic for MCAR takes the following form

$$d_0^2 = \sum_{j=1}^{J} n_j (\overline{\boldsymbol{y}}_{\boldsymbol{o}_j} - \boldsymbol{\mu}_{\boldsymbol{o}_j})^\top \boldsymbol{\Sigma}_{\boldsymbol{o}_j}^{-1} (\overline{\boldsymbol{y}}_{\boldsymbol{o}_j} - \boldsymbol{\mu}_{\boldsymbol{o}_j}) \tag{4}$$

The idea is that if the data are MCAR, then conditional on the missing indicator $\boldsymbol{r}_i$, the following null hypothesis holds

$$H_0 : \quad \boldsymbol{y}_{o,i} | \boldsymbol{r}_i \sim N(\boldsymbol{\mu}_{\boldsymbol{o}_j}, \boldsymbol{\Sigma}_{\boldsymbol{o}_j}) \qquad \text{if} \quad i \in \boldsymbol{I}_j, 1 \le j \le J \tag{5}$$

where $\boldsymbol{\mu}_{\boldsymbol{o}_j}$ is a subvector of the mean vector $\boldsymbol{\mu}$.

Instead, if (5) is not true, then conditional on the missing indicator $\boldsymbol{r}_i$, the means of the observed $\boldsymbol{y}$'s are expected to vary across different patterns, which implies

$$H_1 : \quad \boldsymbol{y}_{o,i} | \boldsymbol{r}_i \sim N(\boldsymbol{\nu}_{\boldsymbol{o}_j}, \boldsymbol{\Sigma}_{\boldsymbol{o}_j}) \qquad \text{if} \quad i \in \boldsymbol{I}_j, 1 \le j \le J \tag{6}$$

where $\boldsymbol{\nu}_{\boldsymbol{o}_j}, j = 1, 2..., J$ are mean vectors of each pattern j and can be distinct. Rejecting (5) is sufficient for rejecting the MCAR assumption (2), but not necessary.

Little (1988) proved that the statistic (4) is the likelihood ratio statistic for testing (5) against (6). If the normality assumption holds, then d_0^2 follows χ^2 distribution with df $= \sum_{j=1}^{J} p_j - p$. If $\boldsymbol{y}_i$'s are not multivariate normal but has the same mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then by the multivariate central limit theorem (see for example, part (c) of the lemma in Little 1988), under the null assumption of MCAR, d_0^2 follows the same χ^2 distribution asymptotically.

In practice, since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually unknown, Little (1988) proposed to replace them with the unbiased estimators $\widehat{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}} = n\widehat{\boldsymbol{\Sigma}}/(n-1)$, where $\widehat{\boldsymbol{\mu}}$

and $\widehat{\boldsymbol{\Sigma}}$ are the maximum likelihood estimators based on the null hypothesis (5). Thus $\boldsymbol{\Sigma}_{\boldsymbol{o}_j}$ in (4) is replaced by the submatrix $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{o}_j}$ of $\widetilde{\boldsymbol{\Sigma}}$, which gives

$$d^2 = \sum_{j=1}^{J} n_j (\overline{\boldsymbol{y}}_{\boldsymbol{o}_j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{o}_j})^\top \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{o}_j}^{-1} (\overline{\boldsymbol{y}}_{\boldsymbol{o}_j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{o}_j}) \tag{7}$$

Asymptotically, d^2 follows χ^2 distribution with degrees of freedom $\mathrm{df} = \sum_{j=1}^{J} p_j - p$, and (5) is rejected if $d^2 > \chi^2_{\mathrm{df}}(1-\alpha)$ where α is the significance level. $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ can be obtained from EM algorithm using the observed data $\boldsymbol{Y}_{\boldsymbol{o}}$ (Little and Rubin 1987, Schafer 1997).

2.3 Test of CDM

A natural extension of Little's test of MCAR is to test the CDM assumption (3) of $\boldsymbol{y}_i$ conditional on $\boldsymbol{x}_i$ when covariates $\boldsymbol{x}_i$'s are present. For simplicity, we assume that $\boldsymbol{x}_i$ contains the constant term 1 as one of its components. If $\boldsymbol{y}$ depends linearly on $\boldsymbol{x}$, then the model becomes

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{B}$ is a $p \times q$ matrix of coefficients and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. Under homoscedasticity assumption, $\boldsymbol{\Sigma}$ does not depend on $\boldsymbol{x}$. Compared with the model without covariates, we need to replace every unconditional mean of $\boldsymbol{y}$ with the conditional mean of $\boldsymbol{y}$ given $\boldsymbol{x}$, and test whether the coefficient matrix $\boldsymbol{B}$ varies among different missing patterns. The χ^2 test statistic (4) now becomes

$$\begin{aligned} d_0^2 &= \sum_{j=1}^{J} \sum_{i \in \boldsymbol{I}_j} (\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j} \boldsymbol{x}_i - \boldsymbol{B}_{\boldsymbol{o}_j} \boldsymbol{x}_i)^\top \boldsymbol{\Sigma}_{\boldsymbol{o}_j}^{-1} (\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j} \boldsymbol{x}_i - \boldsymbol{B}_{\boldsymbol{o}_j} \boldsymbol{x}_i) \\ &= \sum_{j=1}^{J} \sum_{i \in \boldsymbol{I}_j} \boldsymbol{x}_i^\top (\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j} - \boldsymbol{B}_{\boldsymbol{o}_j})^\top \boldsymbol{\Sigma}_{\boldsymbol{o}_j}^{-1} (\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j} - \boldsymbol{B}_{\boldsymbol{o}_j}) \boldsymbol{x}_i \end{aligned} \tag{8}$$

where $\boldsymbol{B}_{\boldsymbol{o}_j}$ is a $p_j \times q$ submatrix of $\boldsymbol{B}$, whose rows correspond to the jth missing pattern, and $\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j}$ is the OLS estimator of $\boldsymbol{B}_{\boldsymbol{o}_j}$ using the observed data from pattern j. It is straightforward to see that d_0^2 in (4) is a special case of d_0^2 in (8) when $\boldsymbol{x}$ only contains the constant component 1.

Accordingly, we are now testing the null hypothesis

$$H_0: \quad \boldsymbol{y}_{o,i} | \boldsymbol{r}_i, \boldsymbol{x}_i \sim N(\boldsymbol{B}_{\boldsymbol{o}_j} \boldsymbol{x}_i, \boldsymbol{\Sigma}_{\boldsymbol{o}_j}) \qquad \text{if} \quad i \in \boldsymbol{I}_j, 1 \le j \le J \tag{9}$$

versus

$$H_1: \quad \boldsymbol{y}_{o,i} | \boldsymbol{r}_i, \boldsymbol{x}_i \sim N(\boldsymbol{D}_{\boldsymbol{o}_j} \boldsymbol{x}_i, \boldsymbol{\Sigma}_{\boldsymbol{o}_j}) \qquad \text{if} \quad i \in \boldsymbol{I}_j, 1 \le j \le J \tag{10}$$

where under H_1, the CDM assumption does not hold and $\boldsymbol{y}_{\boldsymbol{o}_j} = \boldsymbol{D}_{\boldsymbol{o}_j} \boldsymbol{x} + \boldsymbol{\varepsilon}$ for pattern j, with $\boldsymbol{D}_{\boldsymbol{o}_j}$ potentially different among all patterns, but the error terms

still sharing the same multivariate distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$.

In practice, we replace $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ in (8) with unbiased estimators $\widehat{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{\Sigma}} = n\widehat{\boldsymbol{\Sigma}}/(n-q)$ where $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the maximum likelihood estimators using all data under H_0, and calculate

$$d^2 = \sum_{j=1}^{J} \sum_{i \in \boldsymbol{I}_j} \boldsymbol{x}_i^\top (\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j} - \widehat{\boldsymbol{B}}_{\boldsymbol{o}_j})^\top \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{o}_j}^{-1} (\widetilde{\boldsymbol{B}}_{\boldsymbol{o}_j} - \widehat{\boldsymbol{B}}_{\boldsymbol{o}_j}) \boldsymbol{x}_i \tag{11}$$

which asymptotically follows χ^2 distribution with degrees of freedom df $= q(\sum_{j=1}^{J} p_j - p)$, and (9) is rejected if $d^2 > \chi^2_{\text{df}}(1-\alpha)$ where α is the significance level. Again, when there are no covariates, and $\boldsymbol{x}$ only contains the constant component 1 with $q = 1$, then df $= \sum_{j=1}^{J} p_j - p$, which coincides with the degrees of freedom in the test of MCAR.

2.4 Adjustment for unequal variances

As Little (1988) pointed out, one important limitation of d^2 in (7) and (11) is that the covariance matrix of observed $\boldsymbol{y}$ (or observed $\boldsymbol{y}$ conditional on $\boldsymbol{x}$) is still the same for all missing-value patterns even in the alternative hypotheses (6) and (10). This assumption may not be satisfied in general, especially when the number of missing patterns is large. Therefore, we can relax this limitation on covariance matrices and replace the alternative hypothesis with

$$H_1: \quad \boldsymbol{y}_{o,i} | \boldsymbol{r}_i, \boldsymbol{x}_i \sim N(\boldsymbol{D}_{\boldsymbol{o}_j} \boldsymbol{x}_i, \boldsymbol{\Gamma}_{\boldsymbol{o}_j}) \qquad \text{if} \quad i \in \boldsymbol{I}_j, 1 \le j \le J \tag{12}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{o}_j}$ contains distinct parameters for each missing pattern j. To test (9) against (12), we can derive the following likelihood ratio statistic as in Little (1988)

$$d^2_{\text{aug}} = d^2 + \sum_{j=1}^{J} n_j \Big\{ \text{tr}(\boldsymbol{S}_{\boldsymbol{o}_j} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{o}_j}^{-1}) - p_j - \log|\boldsymbol{S}_{\boldsymbol{o}_j}| + \log|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{o}_j}| \Big\} \tag{13}$$

where d^2 is the same as in (7) without covariates or (11) with covariates, $\boldsymbol{S}_{\boldsymbol{o}_j}$ is the estimated covariance matrix of residuals from the regression of observed $\boldsymbol{y}_{\boldsymbol{o}_j}$ on $\boldsymbol{x}$ in pattern j, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{o}_j}$ is the same as in (7). aug stands for "augmented" since more parameters need to be estimated for covariance matrices in the new test. Asymptotically, d^2_{aug} follows χ^2 distribution with degrees of freedom df $= q(\sum_{j=1}^{J} p_j - p) + \sum_{j=1}^{J} \frac{p_j(p_j+1)}{2} - \frac{p(p+1)}{2}$, and (5) or (9) is rejected if $d^2_{\text{aug}} > \chi^2_{\text{df}}(1-\alpha)$ where α is the significance level. This augmented test using d^2_{aug} tends to have higher power than the test using d^2 for large sample sizes, especially when the covariance structures of different missing-value patterns vary a lot, as shown later in our simulation results in Section 5. On the other hand, d^2_{aug} may not be applicable if some patterns have too small sample sizes such that $n_j < p_j + q$, since $\boldsymbol{S}_{\boldsymbol{o}_j}$ will then be singular and hence $\log|\boldsymbol{S}_{\boldsymbol{o}_j}|$ in the expression of d^2_{aug} cannot be computed.

3 The mcartest command

3.1 Description

`mcartest` performs Little's chi-square test for the MCAR assumption, and accommodates arbitrary missing-value patterns. *depvars* contains a list of variables with missing values to be tested. *depvars* requires at least two variables. *indepvars* contains a list of covariates. When *indepvars* are specified, `mcartest` tests the CDM assumption for *depvars* conditional on *indepvars* (see Little 1995). The test statistic uses multivariate normal estimates from the EM algorithm (see [MI] **mi impute mvn**). The `unequal` option performs Little's augmented chi-square test which allows unequal variances between missing-value patterns. See Little (1988) for details.

3.2 Syntax

Test for MCAR

`mcartest` *depvars* [*if*] [*in*] [`,` `noconstant` `unequal` `emoutput` *em_options*]

Test for CDM

`mcartest` *depvars* `=` *indepvars* [*if*] [*in*] [`,` `noconstant` `unequal` `emoutput` *em_options*]

3.3 Options

`noconstant` suppresses constant term.

`unequal` specifies to allow unequal variances between missing-value patterns. By default, the test assumes equal variances between different missing-value patterns.

`emoutput` specifies to display intermediate output from EM estimation.

em_options specifies the options in EM algorithm. See [MI] **mi impute mvn** in StataCorp (2011) for details.

3.4 Saved results

Scalars

`r(N)`	number of observations used in the test, excluding observations with all values missing
`r(N_S_em)`	number of unique missing-value patterns
`r(chi2)`	Little's chi-square statistic
`r(df)`	the chi-square degrees of freedom
`r(p)`	the chi-square p-value

4 Example

I illustrate the usage of `mcartest` command through an example. The fictional dataset used here is the blood test results in a study of obesity that contains 371 observations and 11 variables: cholesterol level, triglycerides level, diastolic blood pressure, systolic blood pressure, age, gender, height, weight, exercise

time in a week, alcohol, and smoking. Suppose the variables of interest are the first four, coded as `chol`, `trig`, `diasbp`, and `sysbp`, and the other seven are used as auxiliary variables, coded as `age`, `female`, `height`, `weight`, `exercise`, `alcohol`, and `smoking`. Descriptions of these variables are shown in Table 1.

Table 1: Descriptions of the variables

Name	Type	Description
`chol`	Continuous	Cholesterol level
`trig`	Continuous	Triglycerides level
`diasbp`	Continuous	Diastolic blood pressure
`sysbp`	Continuous	Systolic blood pressure
`age`	Categorical	1 if 21-30, 2 if 31-40, 3 if 41-50, 4 if above 50
`female`	Categorical	1 if female, 0 if male
`height`	Continuous	Height in inches
`weight`	Continuous	Weight in lbs
`exercise`	Discrete	Exercise in hours per week
`alcohol`	Categorical	1 if drinking alcohol, 0 if not
`smoking`	Categorical	1 if smoking, 0 if not

After loading the data, we can check the missing-value patterns by using the `misstable` command.

```
. use bloodtest, clear
(fictional blood test data)
. misstable summarize
```

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
chol	90		281	265	187.73	224.57
trig	70		301	280	103.22	136.21
diasbp	34		337	24	66	90
sysbp	73		298	32	106	138

```
. misstable pattern, freq
  Missing-value patterns
    (1 means complete)
```

	Pattern			
Frequency	1	2	3	4
122	1	1	1	1
72	1	1	1	0
70	1	0	1	1
55	1	1	0	1
34	0	1	1	1
18	1	1	0	0
371				

```
Variables are  (1) diasbp  (2) trig  (3) sysbp  (4) chol
```

The results suggest that the dataset contains missing values in the first four variables, but all the other variables are completely observed. 122 observations out of the 371 in total are complete, while over 2/3 of the observations contain missing values, with six missing-value patterns in total that are not monotone.

Now we can determine if the data are missing completely at random using the `mcartest` command. Suppose in the beginning we don't include any of the auxiliary variables in the analysis and only apply Little's MCAR test to `chol`, `trig`, `diasbp`, and `sysbp`. We try both the regular MCAR test and the test with unequal variances.

```
. mcartest chol trig diasbp sysbp, emoutput nolog
Expectation-maximization estimation      Number obs           =       371
                                         Number missing       =       267
                                         Number patterns      =         6
Prior: uniform                           Obs per pattern: min =        18
                                                          avg =  61.83333
                                                          max =       122

Observed log likelihood = -2623.2645 at iteration 17
```

	chol	trig	diasbp	sysbp
Coef				
_cons	206.2264	120.5829	78.8161	121.196
Sigma				
chol	41.91012	22.33289	3.762825	3.48862
trig	22.33289	42.08035	6.622086	10.69249
diasbp	3.762825	6.622086	18.45518	14.37273
sysbp	3.48862	10.69249	14.37273	35.92427

```
Little´s MCAR test

Number of obs       = 371
Chi-square distance = 25.7412
Degrees of freedom  = 14
Prob > chi-square   = 0.0279
```

We specified the `emoutput` option to display the EM estimates and also suppressed the log using *em_options*' `nolog` option. If the EM algorithm does not converge, `mcartest` will generate a warning message in blue, similar to what `mi impute mvn` does. EM has converged in this test. The regular Little's MCAR test gives a χ^2 distance of 25.74 with degrees of freedom 14 and p-value 0.0279. The test provides evidence that the missing data in the four variables of interest are not MCAR under significance level 0.05.

We can also specify the `unequal` option to run the test with unequal variances.

(*Continued on next page*)

```
. mcartest chol trig diasbp sysbp, unequal

Little´s MCAR test with unequal variances

Number of obs       = 371
Chi-square distance = 56.7101
Degrees of freedom  = 41
Prob > chi-square   = 0.0522
```

This test gives a χ^2 distance of 56.71 with degrees of freedom 41 and p-value 0.0522. The p-value is only slightly larger than 0.05, indicating that although the evidence against MCAR is not strong, the power of the test could be possibly low. Both tests cast doubts on the MCAR assumption.

Next we add auxiliary variables as covariates into the test and test the CDM assumption. Note that `age` is grouped into 4 brackets and `female` has two groups, so we use the factor variables `i.age` and `i.female` in the test. We also specify the `emoutput` option to display the EM estimates of the linear regression coefficients.

```
. mcartest chol trig diasbp sysbp = weight i.age i.female, emoutput nolog
Expectation-maximization estimation      Number obs           =       371
                                         Number missing       =       267
                                         Number patterns      =         6
Prior: uniform                           Obs per pattern: min =        18
                                                          avg =  61.83333
                                                          max =       122

Observed log likelihood = -2477.8319 at iteration 24
```

	chol	trig	diasbp	sysbp
Coef				
weight	.0898433	.1155952	.0035606	.0315919
1b.age	0	0	0	0
2.age	-.0790635	-.598354	.0120911	-.6006885
3.age	-.3147961	-.6971391	-.4392923	-1.07614
4.age	-2.220313	-2.172395	.4254206	-.582046
0b.female	0	0	0	0
1.female	2.10565	-4.386112	-4.315367	-2.971464
_cons	191.5976	103.5614	79.32499	117.3274
Sigma				
chol	38.04902	15.04927	2.537881	1.435059
trig	15.04927	21.60197	-.5490975	1.695223
diasbp	2.537881	-.5490975	14.83308	10.89443
sysbp	1.435059	1.695223	10.89443	32.07185

```
Little´s CDM test

Number of obs       = 371
Chi-square distance = 89.4992
Degrees of freedom  = 84
Prob > chi-square   = 0.3204
```

This CDM test gives a χ^2 distance of 89.50 with degrees of freedom 84 and p-value 0.3204. We find that for this dataset, adding `age`, `female`, and `weight`

as covariates can pass the CDM test. The EM outputs in the table give the EM estimates of the multivariate linear regression of `chol`, `trig`, `diasbp`, and `sysbp` on `weight`, `age` and `female`, including the regression coefficients (`Coef`) and the covariance matrix of the errors (`Sigma`). As comparison, we also run the test with all the seven auxiliary variables as covariates.

```
. mcartest chol trig diasbp sysbp = weight height exercise i.age i.female i.alc
> ohol i.smoking

Little´s CDM test

Number of obs       = 371
Chi-square distance = 141.1465
Degrees of freedom  = 140
Prob > chi-square   = 0.4569
```

This CDM test gives a χ^2 distance of 141.15 with degrees of freedom 140 and p-value 0.4569. Both CDM tests are highly nonsignificant, which implies that although `chol`, `trig`, `diasbp`, and `sysbp` are not MCAR, the missing data mechanism can be reasonably viewed as CDM given the auxiliary variables `age`, `female`, and `weight`. Therefore, for this dataset, any analysis of `chol`, `trig`, `diasbp`, and `sysbp` using only the 122 completely observed samples without adjusting the effect of the auxiliary variables is not valid since the MCAR assumption is violated. The means of these four variables are significantly different in the 122 completely observed samples and in the other samples that contain missing values. On the other hand, the plausible CDM assumption implies that the means of these four variables change linearly with the auxiliary variables. For example, the mean level of the cholesterol level changes from case to case with linear dependence on the subject's weight, age and gender, and the linear regression coefficients are displayed in the foregoing output of EM estimates. Since CDM is a special case of MAR as we mentioned in Section 2.1, this example also implies that simple methods such as complete case analysis do not necessarily work under the more general MAR assumption.

As suggested by Little (1995), since in real applications no information about the covariates is known beforehand, it seems preferable to include all possible covariates in the model. However, including more covariates will increase the chi-square degrees of freedom considerably, as can be seen in this example, which could make the estimation less efficient and the test less powerful. Therefore we need to balance between the limited sample size and the number of covariates, and choose the appropriate MCAR or CDM assumptions for testing.

5 Simulation study

In this section, I evaluate the performance of Little's chi-square test of MCAR and CDM through simulation studies. In general, when the true missing-data mechanism is MCAR, the empirical rejection probability of Little's test of MCAR fits well with the nominal significance level, with a stable performance even for small samples, different proportions of missing values, and different

numbers of variables with missing values, as was found in Little (1988), Kim and Bentler (2002) and confirmed by my own simulations that are not included here. However, for Little's test of CDM, the natural extension of MCAR test, it remains unclear whether increasing the number of covariates has an impact on its finite sample performance. I explored this by simulating the following model

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \boldsymbol{B} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

where $\boldsymbol{B}$ is a $p \times (q-1)$ matrix of all 1's, $x_1, x_2, ..., x_{q-1}$ are independent $N(0,1)$ variables, and the error terms follow

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\}$$

y_1 is missing completely at random with probability 0.5 and y_2 is always completely observed, yielding 2 missing-value patterns. $\boldsymbol{y} = (y_1, y_2)^\top$ is tested for CDM with auxiliary variables (covariates) $\boldsymbol{x} = (x_1, ..., x_{q-1})^\top$. The number of covariates $q-1$ (constant term not included) varies among 0, 1, 2, 5, 10, and 20, and the sample size increases from 100, 250, 500 to 1000. For each scenario, 10,000 Monte Carlo replications are used. Under the null hypothesis (9), d^2 in (11) asymptotically follows χ^2 distribution with df $= q$. At significance level $\alpha = 0.05$, I report the empirical rejection probability of the CDM test in Table 2. The Monte Carlo standard errors are displayed in the parentheses right after each rejection rate.

Table 2: Empirical rejection rates of the CDM test with $\alpha = 0.05$

Covariates	χ^2 df	Sample size			
		100	250	500	1000
0	1	0.051 (0.002)	0.043 (0.002)	0.050 (0.002)	0.048 (0.002)
1	2	0.051 (0.002)	0.052 (0.002)	0.050 (0.002)	0.052 (0.002)
2	3	0.044 (0.002)	0.049 (0.002)	0.049 (0.002)	0.048 (0.002)
5	6	0.045 (0.002)	0.049 (0.002)	0.050 (0.002)	0.051 (0.002)
10	11	0.036 (0.002)	0.045 (0.002)	0.046 (0.002)	0.047 (0.002)
20	21	0.023 (0.001)	0.039 (0.002)	0.045 (0.002)	0.046 (0.002)

Table 2 shows that in this model, when the number of covariates is small, the empirical rejection rate of Little's CDM test is sufficiently close to the nominal level 0.05 with a sample size of 100 or 250. However, as the number of covariates increases to 10 and 20, the empirical rejection rate is much lower than the nominal level 0.05 when the sample size is 100 or 250. Therefore in small samples, the CDM test tends to be more conservative when the number of covariates is large.

It is also of interest to compare the performance of Little's MCAR test statistic d^2 with that of the augmented test statistic d^2_{aug} when the covariance matrices vary among different missing-value patterns. I simulated the following simple model without covariates

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\}$$

where y_2 always remains complete through all observations, and y_1 is missing with probability 0.5 based on the missing mechanisms below. In principle, we can compare both the rejection probabilities when the null hypothesis (5) or (9) is satisfied by the true model and the power of these tests when the null is violated. The alternative hypothesis could be either (10) or (12), and will be covered by the 5 cases below. In the following, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, and $\Phi^{-1}(\cdot)$ is its inverse.

1. (MCAR) y_1 is missing completely at random with probability 0.5.
2. (MAR) y_1 is missing if and only if $\Phi^{-1}(0.1) \leq y_2 \leq 0$ or $y_2 \geq \Phi^{-1}(0.9)$.
3. (MAR) y_1 is missing if and only if $|y_2| \geq \Phi^{-1}(0.75)$.
4. (MNAR) y_1 is missing if and only if $\Phi^{-1}(0.2) \leq y_1 \leq 0$ or $y_1 \geq \Phi^{-1}(0.8)$.
5. (MNAR) y_1 is missing if and only if $|y_1| \geq \Phi^{-1}(0.75)$.

Note that y_1 is missing with probability 0.5 in all 5 cases, yielding 2 missing-value patterns, and we always test the full vector of $\boldsymbol{y} = (y_1, y_2)^\top$. Therefore the true missing data mechanism of Case 1 corresponds to MCAR. Case 2 and Case 3 are MAR. Case 4 and Case 5 are MNAR. The covariance structures of 2 missing-value patterns are the same in Cases 1, 2, and 4 by symmetry, and different in Cases 3 and 5. Under the null hypothesis (5), d^2 in (7) asymptotically follows χ^2 distribution with df = 1, and d^2_{aug} in (13) asymptotically follows χ^2 distribution with df = 2. I report the empirical rejection rates of both tests at significance level $\alpha = 0.05$ using sample sizes 100, 250, 500, and 1000 based on 10,000 Monte Carlo replications for each of the 5 missing data mechanisms. The results are summarized in Table 3. The Monte Carlo standard errors are displayed in the parentheses right after each rejection rate.

We can compare the results from d^2 and d^2_{aug} in Table 3. In Case 1 where y_1 is MCAR, the empirical rejection rates for both d^2 and d^2_{aug} are close to the nominal level. In Case 2 (MAR) and Case 4 (MNAR), these two tests also behave similarly, though the power of d^2 seems to be slightly higher than d^2_{aug}. This is not surprising because in the true model, covariance matrices of the two missing patterns are exactly the same, and d^2_{aug} is less efficient since it estimates two covariance matrices separately. However, in either Case 3 (MAR) where y_1 is missing if $|y_2| \geq \Phi^{-1}(0.75)$, or in Case 5 (MNAR) where y_1 is missing if $|y_1| \geq \Phi^{-1}(0.75)$, the missing data and the observed data have the same mean

Table 3: Empirical rejection rates when $\alpha = 0.05$ for d^2 and d^2_{aug}

Missingness of y_1	Test stat	Sample size			
		100	250	500	1000
Case 1 (MCAR)	d^2	0.051 (0.002)	0.043 (0.002)	0.050 (0.002)	0.048 (0.002)
	d^2_{aug}	0.053 (0.002)	0.048 (0.002)	0.050 (0.002)	0.050 (0.002)
Case 2 (MAR)	d^2	0.182 (0.004)	0.346 (0.005)	0.566 (0.005)	0.851 (0.004)
	d^2_{aug}	0.184 (0.004)	0.303 (0.005)	0.490 (0.005)	0.780 (0.004)
Case 3 (MAR)	d^2	0.052 (0.002)	0.051 (0.002)	0.051 (0.002)	0.050 (0.002)
	d^2_{aug}	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Case 4 (MNAR)	d^2	0.363 (0.005)	0.728 (0.004)	0.953 (0.002)	0.999 (0.000)
	d^2_{aug}	0.292 (0.005)	0.626 (0.005)	0.916 (0.003)	0.998 (0.000)
Case 5 (MNAR)	d^2	0.050 (0.002)	0.053 (0.002)	0.048 (0.002)	0.052 (0.002)
	d^2_{aug}	0.261 (0.004)	0.572 (0.005)	0.882 (0.003)	0.996 (0.001)

zero, but different variances. As a result, the empirical rejection rates from d^2 are very low, indicating weak power of Little's test in these two situations. The power of d^2 does not improve significantly even if we increase the sample size to 1000. Instead, after adjustment for unequal variances, d^2_{aug} has much higher power, and the power increases to 1 as the sample size increases from 100 to 1000. This implies that d^2 may not be reliable when the difference between missing-value patterns does not lie in their means, while d^2_{aug} can overcome this weakness when the covariance structure varies significantly across different missing-value patterns.

Although the augmented test for unequal variances has better power in some situations, such as Case 3 and Case 5 of the model above, it may be too conservative with small sample sizes and complicated missing-value patterns. In the extreme case, according to the formula (13), d^2_{aug} cannot be computed when some missing-value patterns contain too few observations. In the following we simulate the same example from Little (1988) and compare the finite sample performance of d^2 and d^2_{aug} with more complicated missing-value patterns. Little (1988) considered a multivariate normal model with 4 variables $\boldsymbol{y} = (y_1, y_2, y_3, y_4)^\top$, generated by

$$\begin{aligned} y_1 &= z_1 \\ y_2 &= z_1\sqrt{0.9} + z_2\sqrt{0.1} \\ y_3 &= z_1\sqrt{0.2} + z_2\sqrt{0.1} + z_3\sqrt{0.7} \\ y_4 &= -z_1\sqrt{0.6} + z_2\sqrt{0.25} + z_3\sqrt{0.1} + z_4\sqrt{0.05} \end{aligned}$$

where z_1, z_2, z_3, z_4 are independent standard normal random variables. We only observe y_1, y_2, y_3, y_4 but not z_1, z_2, z_3, z_4, and the missing data mechanism of y_1, y_2, y_3, y_4 is MCAR. For $\boldsymbol{y} = (y_1, y_2, y_3, y_4)^\top$, Little (1988) considered 7 missing-value patterns in total, which can be represented by the missing indicator vector $\boldsymbol{r} = 1111, 1110, 1100, 1101, 1001, 1011, 1010$. For example, $\boldsymbol{r} = 1110$

means that y_1, y_2, y_3 are observed and y_4 is missing. The proportions of the 7 missing-value patterns in the sample are 0.4, 0.1, 0.1, 0.1, 0.1, 0.1 and 0.1 respectively. We examine the empirical rejection rates of d^2 and d^2_{aug} with the sample size ranging from 100, 250, 500, 1000 to 2000, based on 10,000 Monte Carlo replications. The results are summarized in Table 4 and the Monte Carlo standard errors are displayed in the parentheses.

Table 4: Empirical rejection rates when $\alpha = 0.05$ for d^2 and d^2_{aug}

Test Stat	Sample Size				
	100	250	500	1000	2000
d^2	0.043 (0.002)	0.047 (0.002)	0.054 (0.002)	0.051 (0.002)	0.049 (0.002)
d^2_{aug}	0.213 (0.004)	0.096 (0.003)	0.070 (0.003)	0.060 (0.002)	0.053 (0.002)

Given these 7 missing-value patterns, the chi-square degrees of freedom for d^2 and d^2_{aug} are 15 and 42 respectively. The results in Table 4 suggest that with too many parameters in the covariance matrices to estimate, the empirical rejection rates for d^2_{aug} are too conservative and only get close to the nominal level 0.05 when the sample size is 2000. In comparison, d^2 has already achieved acceptable accuracy when the sample size is 250. This implies that d^2_{aug} may not perform as well as d^2 in small samples when the missing-value patterns become more complicated. Moreover, as pointed out in Little (1988), d^2_{aug} may be sensitive to departure from the normality assumption as d^2_{aug} involves the comparison of variances, while simulation results in Little (1988) suggest that d^2 is relatively robust to non-normality of the data. Therefore the augmented test works best for nearly multivariate normal data when the covariance structure differs significantly among missing-value patterns and a sufficient number of observations are available in each pattern.

6 Conclusion

In this article, I presented the `mcartest` command that implements Little's chi-square test of the MCAR assumption or the CDM assumption. The methodology is mainly based on Little (1988) and can be extended to testing the CDM assumption when covariates are included in the test. The command also allows adjustment for unequal variances via the `unequal` option. I demonstrated how to use this command and the caveats of choosing covariates through an example. Finally I examined the performance of the MCAR/CDM test, compared the strengths and weaknesses of the regular test and the test with unequal variances by simulation and provided some suggestions for how to use them in practice.

7 Acknowledgements

This work was done during my internship at StataCorp in the summer of 2012. I am grateful to Yulia Marchenko for her guidance and support. I also thank the reviewer for helpful comments that have substantially improved the article.

8 References

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs. 2009. *Handbooks of Modern Statistical Methods: Longitudinal Data Analysis*. London: Chapman & Hall.

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

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

———. 1995. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90: 1112–1121.

Little, R. J. A., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63: 581–592.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

StataCorp. 2011. *Stata Multiple-Imputation Reference Manual Release 12*. College Station, TX: Stata Press.

About the author

Cheng Li is a PhD candidate in statistics at Northwestern University. His research is currently focused on Bayesian methods for high dimensional problems.