A Test of Missing Completely at Random for Multivariate Data With Missing Values

RODERICK J. A. LITTLE*

A common concern when faced with multivariate data with missing values is whether the missing data are missing completely at random (MCAR); that is, whether missingness depends on the variables in the data set. One way of assessing this is to compare the means of recorded values of each variable between groups defined by whether other variables in the data set are missing or not. Although informative, this procedure yields potentially many correlated statistics for testing MCAR, resulting in multiple-comparison problems. This article proposes a single global test statistic for MCAR that uses all of the available data. The asymptotic null distribution is given, and the small-sample null distribution is derived for multivariate normal data with a monotone pattern of missing data. The test reduces to a standard t test when the data are bivariate with missing data confined to a single variable. A limited simulation study of empirical sizes for the test applied to normal and nonnormal data suggests that the test is conservative for small samples.

KEY WORDS: Incomplete data; Multivariate normal distribution; Nonresponse.

1. INTRODUCTION

Many statistical analyses of data with missing values make the assumption that data are missing completely at random (MCAR), in the sense that missingness does not depend on the values of variables in the data set subject to analysis. Nevertheless, formal tests of MCAR have not received much attention. When missing values are confined to a single variable y, the standard procedure is to compare the distributions of fully observed variables for respondents and nonrespondents to y, either informally or formally, via t tests for the differences in means. In a regression setting, Simon and Simonoff (1986) considered a sensitivity analysis of deviations from the MCAR assumption (which they call missing at random) when missing values are confined to a single independent variable.

Dixon (1983) in the program BMDP8D extended the t-test approach to multivariate data with missing values on any of p variables. For each variable with missing values, the sample is split into cases with that variable observed and cases with that variable missing. The means of observed values of the other variables in the two groups are then compared by two sample t tests. Significant differences between these means are evidence that the data are not MCAR. This procedure is informative, but yields up to $(p - 1)$ t tests for each variable in the data set, or up to $p(p - 1)$ t statistics for assessing the MCAR assumption. The difficulties of simultaneous inference are considerable, since the t statistics are correlated with a complex correlation structure depending on the pattern of missing data and the correlation matrix of the y variables.

Example: Blood Chemistry Data With Values Deleted. As an illustration, I consider the Werner blood-chemistry data (Werner, Tolls, Hultin, and Mellecker 1970; see Dixon 1983, table 5.1) with values randomly deleted. The data record eight variables for $n = 188$ women. Six of the variables were selected for our purposes: age, weight, birthpill (1 = user, 2 = nonuser), cholesterol, albumin, and calcium. About 20% of the values of each of the latter four variables were randomly deleted; weight was missing for two cases in the original data set. Of the $p(p - 1) = (6)(5) = 30$ possible pairwise t statistics for testing MCAR, the five that split the sample by whether age was missing are vacuous, since age is never missing. Also, the five that split the sample by whether weight was missing are discarded, since for these the nonrespondent group consists of only two cases. A stem-and-leaf plot of the remaining 20 t statistics is shown in Figure 1; for the sample size studied they can be viewed as normal deviates. Note that the extreme t statistics (-2.3, 2.4, 2.7, 3.3) might be regarded as evidence against MCAR, although an MCAR deletion mechanism was in fact employed.

I propose a single test statistic for testing MCAR and show that its null distribution is asymptotically chi-squared. For the data in the example, the statistic takes the value 76.5 on 60 df ($P = .074$), suggesting that the evidence against random missingness is in fact weak. Before describing the test statistic, I discuss the MCAR assumption in more detail.

2. FORMAL DEFINITIONS OF RANDOMLY MISSING DATA

Let $\mathbf{y}$ denote an $(n \times p)$ data matrix of n observations on p variables and $\mathbf{r}$ an $(n \times p)$ missingness indicator matrix, such that $r_{ij} = 1$ if y_{ij} is missing and 0 otherwise. A full model for the data and the missing-data mechanism specifies a distribution $f(\mathbf{y} \mid \boldsymbol{\theta})$ for $\mathbf{y}$, indexed by unknown parameters $\boldsymbol{\theta}$, and a distribution $f(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\psi})$ for $\mathbf{r}$, given $\mathbf{y}$, indexed by unknown parameters $\boldsymbol{\psi}$. Write $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$, where $\mathbf{y}_{\text{obs}}$ represents the observed values of $\mathbf{y}$ and $\mathbf{y}_{\text{mis}}$ represents the missing values. Rubin (1976) defined the missing data as MCAR if $f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \boldsymbol{\psi}) = f(\mathbf{r} \mid \boldsymbol{\psi})$ for all $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$; that is, missingness does not depend on the observed or missing values of $\mathbf{y}$. Rubin also defined a weaker condition for the missing-data mechanism, calling

* Roderick J. A. Little is Professor, Department of Biomathematics, School of Medicine, University of California, Los Angeles, CA 90024. This work was supported by National Institute of Mental Health Grant USPHS MH 37188. The author thanks the referees and associate editors for several helpful suggestions.

Journal of the American Statistical Association
December 1988, Vol. 83, No. 404, Theory and Methods

3–3.9	3
2–2.9	047
1–1.9	4
0–.9	3457888
−.9–−0	322
−1.9–−1	500
−2.9–−2	31
−3.9–−3	
−4.9–−4	

Figure 1. Distribution of Pairwise t Statistics for Werner Data With Data Deleted by an MCAR Mechanism.

the missing data missing at random (MAR) if $f(\mathbf{r} \mid \mathbf{y}_{obs}, \mathbf{y}_{mis}, \psi) = f(\mathbf{r} \mid \mathbf{y}_{obs}, \psi)$ for all $\mathbf{y}_{mis}$; that is, missingness does not depend on the missing values $\mathbf{y}_{mis}$ of $\mathbf{y}$ but may depend on observed values in the data set. Rubin showed that if the missing data are MAR and θ and ψ are distinct, then likelihood inference for θ can be based on the likelihood obtained by integrating $\mathbf{y}_{mis}$ out of the density $f(\mathbf{y}_{obs}, \mathbf{y}_{mis} \mid \theta)$, without including a term for the missing-data mechanism in the likelihood. Under these conditions, Rubin called the missing-data mechanism *ignorable* for likelihood-based inferences.

Tests for the MAR assumption occur in the literature on models for selectivity bias (see Amemiya 1985; Heckman 1976; Olsen 1980) but are highly sensitive to model misspecification (e.g., see Little 1985). The procedures in this article test the stronger MCAR assumption. Such tests are too restrictive for testing whether the missing-data mechanism is ignorable for likelihood inferences, since the latter requires MAR, not MCAR. Nevertheless, they are useful for other purposes.

First, many simple missing-data methods, including restriction to complete cases and pairwise methods, generally require the MCAR assumption (Little and Rubin 1987, chap. 3). Second, maximum likelihood (ML) estimation from data with missing values based on ignorable missing-data models do not require the MCAR assumption, but are more sensitive to model misspecification when the data are not MCAR. In particular, ML estimation for the multivariate normal model (Orchard and Woodbury 1972) provides consistent estimates of θ under mild assumptions (finite fourth moments) when the data are MCAR, but uses the multivariate normal assumptions when the data are MAR but not MCAR, and does not in general supply consistent estimates when the data are not MAR. Finally, standard errors for the parameter estimates based on the expected information matrix are valid only if the data are MCAR. Standard errors based on the observed information matrix are preferable when the data are not MCAR, since they remain valid when the data are MAR but not MCAR; however, for the multivariate normal model they require more computation, particularly for the mean parameters. Hence a test for MCAR provides guidance as to when standard errors based on the expected information matrix are adequate. [Note that even when the data are MCAR arguments can be advanced for preferring standard errors based on the observed information; see Efron and Hinkley (1978).]

3. A TEST OF MCAR FOR MULTIVARIATE DATA

3.1 Notation

I use the following notation: $\mathbf{y}_i = (1 \times p)$ vector of values for case i, in the absence of missing data. $\mathbf{r}_i = (1 \times p)$ vector of missing-data indicators for case i. J = number of distinct missing-data patterns $\mathbf{r}_i$ in the data set. Fully observed cases, if present, count as a pattern. S_j = set of cases with missing-data pattern j ($j = 1, \ldots, J$). m_j = number of cases in S_j; $\Sigma m_j = n$. p_j = number of observed variables for cases in S_j. $\mathbf{D}_j = (p \times p_j)$ matrix indicating which variables are observed for pattern j. The matrix has one column for each variable present, consisting of $p - 1$ 0s and one 1 corresponding to the variable identified. $\mathbf{y}_{obs,i} = (1 \times p_j)$ vector of values of observed variables in case i. $\bar{\mathbf{y}}_{obs,j} \equiv m_j^{-1} \Sigma_{i \in S_j} \mathbf{y}_{obs,i} = (1 \times p_j)$ vector of means of observed variables for pattern j. μ, $\Sigma = (1 \times p)$ population mean vector and $(p \times p)$ covariance matrix of $\mathbf{y}_i$. $\hat{\mu}$, $\hat{\Sigma}$ = ML estimates of μ and Σ, assuming the $\mathbf{y}_i$ are iid normal and the missing-data mechanism is ignorable. $\tilde{\Sigma} = n\hat{\Sigma}/(n-1)$, the ML estimate of Σ with a correction for degrees of freedom. $\mu_{obs,j} \equiv \mu \mathbf{D}_j = (1 \times p_j)$ vector of means of observed variables in pattern j. $\Sigma_{obs,j} \equiv \mathbf{D}_j^T \Sigma \mathbf{D}_j = (p_j \times p_j)$ covariance matrix of observed variables in pattern j.

3.2 A Likelihood Ratio Test Statistic, Assuming Σ Is Known

To motivate the test statistic I first consider the (unrealistic) case where Σ is known. Let μ^* denote the ML estimate of μ, assuming the missing data are MAR and known Σ, and let $\mu^*_{obs,j} = \mu^* \mathbf{D}_j$. I propose the following test statistic for the MCAR assumption:

$$d_0^2 = \sum_{j=1}^{J} m_j (\bar{\mathbf{y}}_{obs,j} - \mu^*_{obs,j}) \Sigma^{-1}_{obs,j} (\bar{\mathbf{y}}_{obs,j} - \mu^*_{obs,j})^T. \quad (1)$$

Suppose that $\mathbf{y}_i$ is multivariate normal distributed with mean μ and covariance matrix Σ. If the data are MCAR, then conditional on $\mathbf{r}_i$,

$$(\mathbf{y}_{obs,i} \mid \mathbf{r}_i) \underset{ind}{\sim} N(\mu_{obs,j}, \Sigma_{obs,j}), \qquad i \in S_j, \; 1 \le j \le J. \quad (2)$$

If the data are not MCAR then the means of the observed variables can vary across the patterns, suggesting the alternative model

$$(\mathbf{y}_{obs,i} \mid \mathbf{r}_i) \underset{ind}{\sim} N(\nu_{obs,j}, \Sigma_{obs,j}), \qquad i \in S_j, \; 1 \le j \le J, \quad (3)$$

where $\{\nu_{obs,j}, j = 1, \ldots, J\}$ are $(1 \times p_j)$ vectors of mean parameters for observed variables that (unlike $\mu_{obs,j}$) are distinct for each pattern j. Note that the variances and covariances are assumed the same for each pattern; the case where they too are allowed to vary is considered in Section 4.

The statistic d_0^2 tests Model (2) against the alternative model (3), as the following lemma shows.

Lemma. (a) d_0^2 is the likelihood ratio statistic for testing Model (2) against the alternative (3). (b) Under (2), d_0^2 has a chi-squared distribution with $f = \Sigma p_j - p$ df. (c) If the data are MCAR and $\mathbf{y}_i$ has any distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, d_0^2 is asymptotically chi-squared with f df. That is, for large samples the assumption of normality in (2) can be relaxed.

Proof. The log-likelihood under (3) (to within a constant) is

$$l(\boldsymbol{\nu} \mid \mathbf{y}_{\text{obs}}) = -\frac{1}{2}\sum_{j=1}^{J} m_j(\bar{\mathbf{y}}_{\text{obs},j} - \boldsymbol{\nu}_{\text{obs},j})\boldsymbol{\Sigma}_{\text{obs},j}^{-1}(\bar{\mathbf{y}}_{\text{obs},j} - \boldsymbol{\nu}_{\text{obs},j})^T, \quad (4)$$

where $\boldsymbol{\nu}$ denotes the set of means for all of the patterns. Substituting ML estimates $\hat{\boldsymbol{\nu}}_{\text{obs},j} = \bar{\mathbf{y}}_{\text{obs},j}$ yields $l(\hat{\boldsymbol{\nu}} \mid \mathbf{y}_{\text{obs}}) = 0$. Substituting ML estimates of the means under Model (2) yields $l(\boldsymbol{\mu}^* \mid \mathbf{y}_{\text{obs}}) = -d_0^2/2$. Hence the likelihood ratio test statistic is $-2[l(\boldsymbol{\mu}^* \mid \mathbf{y}_{\text{obs}}) - l(\hat{\boldsymbol{\nu}} \mid \mathbf{y}_{\text{obs}})] = d_0^2$, proving (a).

To prove (b), concatenate the pattern mean vectors $\{\bar{\mathbf{y}}_{\text{obs},j}, j = 1, \ldots, J\}$ into a single $1 \times \Sigma p_j$ vector, and note that under Model (2) this vector is normal with mean $\boldsymbol{\mu}\mathbf{X}$ and known covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\Sigma})$, where $\mathbf{X}$ is a known $p \times \Sigma p_j$ matrix of 0s and 1s, and $\boldsymbol{\Gamma}$ is a known matrix, since $\boldsymbol{\Sigma}$ is assumed known. It is easily seen that $\boldsymbol{\mu}^*$ is the generalized least squares estimate of $\boldsymbol{\mu}$ and d_0^2 is the residual sum of squares. Hence by standard least squares theory, under (2) the distribution of d_0^2 conditional on the response pattern $(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ is chi-squared with $f = \Sigma p_j - p$ df.

Finally, to prove (c) let the number of observations for each observed pattern tend to infinity, ignoring patterns that do not appear in the n observed cases. The pattern mean vectors $\bar{\mathbf{y}}_{\text{obs},j}$ tend to normality by the central limit theorem, so d_0^2 tends to a chi-squared deviate with f df, by the proof of (b).

3.3 A Test Statistic When $\boldsymbol{\Sigma}$ Is Unknown

When $\boldsymbol{\Sigma}$ is not known, I propose replacing $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}$ in (1) with $\hat{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ from the multivariate normal ML algorithm, yielding the test statistic

$$d^2 = \sum_{j=1}^{J} m_j(\bar{\mathbf{y}}_{\text{obs},j} - \hat{\boldsymbol{\mu}}_{\text{obs},j})\tilde{\boldsymbol{\Sigma}}_{\text{obs},j}^{-1}(\bar{\mathbf{y}}_{\text{obs},j} - \hat{\boldsymbol{\mu}}_{\text{obs},j})^T. \quad (5)$$

Suppose that the observed data contain information on all pairs of variables so that all of the means, variances, and covariances are estimable. If the data are MCAR and the distribution of $\mathbf{y}_i$ has finite fourth moments, $\tilde{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$. Hence under these conditions, d^2, like d_0^2, is asymptotically chi-squared distributed with f df. This result follows from the multivariate analog of a theorem of Cramer (1946, sec. 20.6). Thus a large-sample test of the MCAR assumption compares d^2 with a chi-squared distribution with f df, rejecting when d^2 is large.

The computation of $\hat{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ is iterative, but the EM algorithm is easy to program with the aid of the SWEEP operator and is available in current software (see the BMDPAM program of Dixon 1983). The additional calculation for (5) is trivial, since $\tilde{\boldsymbol{\Sigma}}_{\text{obs},i}^{-1}$ is a submatrix of the swept covariance matrix computed by EM (see Little and Rubin 1987).

3.4 The Test Statistic for Monotone Missing Data

The small-sample null distribution of d^2 is extremely complex for a general pattern of missing data, but simplifies for particular missing-data patterns. Consider first the special case of $p = 2$ variables Y_1 and Y_2, where Y_1 is observed for all n cases and Y_2 is observed for $n_2 < n$ cases, say $i = 1, \ldots, n_2$. There are $J = 2$ patterns: Pattern 1 denotes cases with Y_1 and Y_2 present and pattern 2 denotes cases with only Y_1 present. Then, $\mathbf{y}_{\text{obs},i} = (y_{i1}, y_{i2})$ for $i = 1, \ldots, n_2$ and $\mathbf{y}_{\text{obs},i} = y_{i1}$ for $i = n_2 + 1, \ldots, n$; $\bar{\mathbf{y}}_{\text{obs},1} = (\bar{y}_1, \bar{y}_2)$, the sample means of Y_1 and Y_2 based on the first n_2 cases; $\bar{\mathbf{y}}_{\text{obs},2} = \bar{y}_1^*$, the sample mean of Y_1 based on the last $n - n_2$ cases; $\tilde{\boldsymbol{\Sigma}}_{\text{obs},1} = \tilde{\boldsymbol{\Sigma}}$; and $\tilde{\boldsymbol{\Sigma}}_{\text{obs},2} = \tilde{\sigma}_{11}$. Thus (5) becomes

$$d^2 = n_2\begin{pmatrix}\bar{y}_1 - \hat{\mu}_1 \\ \bar{y}_2 - \hat{\mu}_2\end{pmatrix}^T \tilde{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix}\bar{y}_1 - \hat{\mu}_1 \\ \bar{y}_2 - \hat{\mu}_2\end{pmatrix} + (n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2/\tilde{\sigma}_{11}, \quad (6)$$

which can be rewritten as

$$\frac{n_2(\bar{y}_1 - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}} + \frac{n_2[\bar{y}_2 - \hat{\mu}_2 - \hat{\beta}_{21.1}(\bar{y}_1 - \hat{\mu}_1)]^2}{\tilde{\sigma}_{22.1}} + \frac{(n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}}. \quad (7)$$

Explicit ML estimates of the parameters are available for this problem (see Anderson 1957; Little and Rubin 1987, chap. 6). Substituting these in (7) yields, after a little algebra,

$$d^2 = [n_2(\bar{y}_1 - \hat{\mu}_1)^2 + (n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2]/\tilde{\sigma}_{11} = \text{SSB}_1/\text{MST}_1 = (n - 1)F/(n - 2 + F),$$

where SSB_1, MST_1, and F are, respectively, the between-groups sum of squares, the total mean square, and the F statistic from the analysis of variance (ANOVA) of Y_1 on the missing-data pattern. Since there are just two patterns here, $F = t^2$, where t is the t statistic for comparing pattern means discussed in Section 1. Hence the test based on d^2 is equivalent to the t test. Under the null hypothesis of MCAR and assuming that the values of Y_1 are normal, F has an F distribution with 1 and $n - 2$ df.

More generally, suppose that the data can be arranged in a *monotone* pattern, where variable Y_q is more observed than Y_{q-1} for $q = 1, \ldots, p - 1$ (Rubin 1974). Then, if n_q is the number of cases for which Y_q is observed, $n = n_1 \geq n_2 \geq \cdots \geq n_p$. A generalization of the previous analysis yields

$$\begin{aligned} d^2 &= \text{SSB}_1/\text{MST}_1 + \text{SSB}_{2.1}/\text{MST}_{2.1} + \cdots \\ &\quad + \text{SSB}_{p-1.12\cdots p-2}/\text{MST}_{p-1.12\cdots p-2} \\ &= \sum_{q=1}^{p-1} (n_q - 1)(k_q - 1)F_{q.12\cdots q-1} \\ &\quad \div \{n_a - k_a + (k_a - 1)F_{a.12\cdots a-1}\}, \end{aligned} \quad (8)$$

where n_q is the number of cases with Y_q observed; k_q is the number of patterns with Y_q observed; SSB_1, MST_1, and F_1 are, respectively, the between-groups sum of squares, total mean square, and F statistic from the ANOVA of Y_1 on all k_1 patterns; $SSB_{2.1}$, $MST_{2.1}$, and $F_{2.1}$ are the between-groups sum of squares, total mean square, and F statistic from the analysis of covariance of Y_2 on all k_2 patterns with Y_2 observed, adjusting for Y_1; and the remaining terms are defined similarly. Under normality and MCAR, each of the contributions in (8) is independent, so the small-sample null distribution of d^2 is a sum of functions of independent F statistics. In large samples, these functions become chi-squared distributed, and d^2 has the asymptotic chi-squared distribution discussed in Section 3.3.

3.5 Simulation Study

A limited simulation study was conducted to examine the empirical size of tests based on d^2, for incomplete multivariate normal, skewed (lognormal), and long-tailed (multivariate t_3) data sets with $n = 20$, 40, or 80 observations. To generate the multivariate normal and multivariate t data sets, observations i on $p = 4$ variables Y_1, Y_2, Y_3, and Y_4 were generated from independent standard normal deviates Z_1, Z_2, Z_3, and Z_4, using

$$y_{i1} = z_{i1}/\sqrt{q_i},$$

$$y_{i2} = z_{i1}\sqrt{.9/q_i} + z_{i2}\sqrt{.1/q_i},$$

$$y_{i3} = z_{i1}\sqrt{.2/q_i} + z_{i2}\sqrt{.1/q_i} + z_{i3}\sqrt{.7/q_i},$$

and

$$y_{i4} = -z_{i1}\sqrt{.6/q_i} + z_{i2}\sqrt{.25/q_i} + z_{i3}\sqrt{.1/q_i} + z_{i4}\sqrt{.05/q_i}.$$

For the multivariate normal data sets $q_i = 1$ for all i; for the multivariate t_3 data sets q_i equals a chi-squared deviate with 3 df. The resulting data sets all have mean vector (0, 0, 0, 0) and variances (1, 1, 1, 1). The correlations are $\rho_{12} = .9487$, $\rho_{13} = .4472$, $\rho_{23} = .5243$, $\rho_{14} = -.7746$, $\rho_{24} = -.5767$, and $\rho_{34} = .0763$, thus encompassing a range of values. The lognormal data sets were obtained by exponentiating the values (y_{ij}) generated for the multivariate normal case.

Missing data were then created in the data set by an MCAR mechanism such that for every data set exactly 40% of the cases were complete (i.e., had the pattern $r_i = 0000$), 10% of the cases had Y_4 missing (0001), and 10% of the cases had each of the patterns 0011, 0010, 0110, 0100, and 0101. Note that all cases had at least two variables present. For each of the nine problems generated by the combinations of distribution and sample size, $N = 1{,}000$ incomplete data sets were generated using the GGNPM and GGUBS subroutines in the IMSL library. The same set of random numbers was used for each of the nine problems to sharpen comparisons between problems. For each data set the MCAR test statistic was calculated; for the chosen missing-data pattern it has 15 df. Acceptance or rejection of the MCAR hypothesis was recorded for the 20%, 10%, 5%, and 1% nominal levels.

Table 1 shows the empirical sizes of the test for each problem. For example, 20.2 in the table indicates that for this problem the null hypothesis of MCAR was rejected at the 20% level in 202 out of 1,000 data sets. Superscripts a and b indicate that the empirical size differs significantly from the nominal size at, respectively, the 1% and 5% levels of significance.

The empirical sizes do not differ significantly from nominal levels for the data sets with 80 observations. For the smaller sample sizes the test appears overly conservative, particularly at the lower nominal levels. An encouraging feature of the results is the relatively small impact of nonnormality (in the form of long tails or skewness) on the empirical sizes, suggesting a fair degree of robustness for the method. This reinforces the fact that asymptotically the test does not require normality.

These results on size should be treated as suggestive rather than definitive, given the modest scope of the simulation study. Power calculations are not included, since the power depends greatly on what departures from the MCAR assumption are contemplated. For example, in the bivariate monotone case of Section 3.4, power may be high if missingness of Y_2 depends on the fully observed variable Y_1. On the other hand, if missingness of Y_2 depends on Y_2, then the test statistic only has good power if Y_1 and Y_2 are highly correlated. For a general pattern of missing data, the global nature of the test leads to a loss of power relative to others that test specific alternative hypotheses, such as an alternative that specifies that missingness is a function of a particular variable. In most circumstances such specific alternative hypotheses are hard to formulate with much certainty, so this loss of power may be tolerated in the interests of achieving a single global test statistic. Since power may be low, it is prudent to keep in mind that accepting the null hypothesis of MCAR does not imply its correctness.

4. DISCUSSION

I conclude by discussing some limitations of the proposed test and outlining some alternative procedures. Close relatives of the test statistic d^2 in Equation (5) are d_1^2, where $\tilde{\boldsymbol{\Sigma}}$ is replaced by the ML estimate of $\boldsymbol{\Sigma}$ under the alternative model (3), and d_2^2, the likelihood ratio test

Table 1. Percent Empirical Sizes for a Test of the MCAR Assumption, From N = 1,000 Simulated Data Sets

Sample size	Distribution	Nominal level of test: 20%	10%	5%	1%
80	Normal	20.2	10.9	4.9	.5
	Lognormal	18.9	8.8	3.7	.7
	t on 3 df	21.2	11.2	5.5	1.0
40	Normal	21.2	9.5	2.5[a]	.2[b]
	Lognormal	18.8	8.9	3.2[b]	.3[b]
	t on 3 df	20.8	9.6	4.1	.8
20	Normal	20.3	6.8[a]	2.8[a]	.3[b]
	Lognormal	21.1	8.3	3.5[b]	.5
	t on 3 df	21.6	8.1[b]	2.0[a]	.3[b]
Standard errors		1.27	.95	.69	.315

[a] 1% level of significance.
[b] 5% level of significance.

statistic obtained by subtracting twice the maximized log-likelihoods for Models (3) and (2). These statistics have the same asymptotic null distribution as d^2. They both require ML estimates to be calculated under both (2) and (3), and hence involve a bit more computation.

We have seen that the test based on d^2 is valid asymptotically without normal assumptions; however, it seems most appropriate when the variables are quantitative. If they are categorical, with data forming a contingency table with supplemental margins, more appropriate large-sample tests of the MCAR assumption can be based on the chi-squared statistics in Fuchs [1982, eqs. (4.1) and (4.2)], with estimates of the cell probabilities based on the saturated model.

An important limitation of d^2, d_1^2, and d_2^2 is that they are derived from an alternative hypothesis (3) that allows missingness to affect the means but constrains the variances and covariances to be the same for all patterns. A referee proposed relaxing this limitation. In particular, consider the alternative model

$$(\mathbf{y}_{\text{obs},i} \mid \mathbf{r}_i) \underset{\text{ind}}{\sim} N(\boldsymbol{\nu}_{\text{obs},j}, \boldsymbol{\Gamma}_{\text{obs},j}), \qquad i \in S_j,\ 1 \le j \le J, \quad (9)$$

where now the covariance matrices $\{\boldsymbol{\Gamma}_{\text{obs},j}\}$, like the means $\{\boldsymbol{\nu}_{\text{obs},j}\}$, contain distinct parameters for each pattern j. A standard calculation yields the following likelihood ratio test statistic for (9) relative to (2):

$$d^2_{\text{aug}} = d^2 + \sum_{j=1}^{J} m_j[\text{tr}(\mathbf{S}_{\text{obs},j}\hat{\boldsymbol{\Sigma}}^{-1}_{\text{obs},j}) - p_j - \ln|\mathbf{S}^{-1}_{\text{obs},j}| + \ln|\hat{\boldsymbol{\Sigma}}^{-1}_{\text{obs},j}|], \quad (10)$$

where $\mathbf{S}_{\text{obs},j}$ is the sample covariance matrix of the observations with pattern j and aug is augmented. Under (2) d^2_{aug} is asymptotically chi-squared with $\Sigma_j p_j(p_j + 3)/2 - p(p + 3)/2$ df; this can be a large number (for the data in the simulation study it is 42), raising concerns about the power of the test. Also, data for patterns j with $m_j \le p_j$ need to be discarded, since for those patterns $\mathbf{S}_{\text{obs},j}$ is singular; thus $\mathbf{S}^{-1}_{\text{obs},j}$ in (10) cannot be computed. By analogy with Bartlett's test for comparing dispersions, I expect the test to be sensitive to departures from the normality assumption, and even under normality the asymptotic null distribution seems unlikely to be reliable unless the sample size is large.

[*Received July 1986. Revised March 1988.*]

REFERENCES

Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61.

Anderson, T. W. (1957), "Maximum Likelihood Estimators for the Multivariate Normal Distribution When Some Observations Are Missing," *Journal of the American Statistical Association*, 52, 200–203.

Cramer, H. (1946), *Mathematical Methods in Statistics*, Princeton, NJ: Princeton University Press.

Dixon, W. J. (ed.) (1983), *BMDP Statistical Software*, Berkeley: University of California Press.

Efron, B., and Hinkley, D. V. (1978), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Information," *Biometrika*, 65, 457–487.

Fuchs, C. (1982), "Maximum Likelihood Estimation and Model Selection in Contingency Tables With Missing Data," *Journal of the American Statistical Association*, 77, 270–278.

Heckman, J. D. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Little, R. J. A. (1985), "A Note About Models for Selectivity Bias," *Econometrica*, 53, 1469–1474.

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

Olsen, R. J. (1980), "A Least Squares Correction for Selectivity Bias," *Econometrica*, 48, 1815–1820.

Orchard, T., and Woodbury, M. A. (1972), "Missing Information Principle: Theory and Applications," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), Berkeley: University of California Press, pp. 697–715.

Rubin, D. B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

——— (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Simon, G. A., and Simonoff, J. S. (1986), "Diagnostic Plots for Missing Data in Least Squares Regression," *Journal of the American Statistical Association*, 81, 501–509.

Werner, M., Tolls, R., Hultin, J., and Mellecker, J. (1970), "Sex and Age Dependence of Serum Calcium, Inorganic Phosphorus, Total Protein, and Albumin in a Large Ambulatory Population," in *Fifth Technical International Congress on Automation, Advances in Automated Analysis* (Vol. 2), Mount Kisco, NY: Futura Publishing, pp. 59–65.