MathSoft

Research Report
No. 94

A Graphical Representation of Little's Test for MCAR

Tim Hesterberg

Revision 1, May 7, 1999

Acknowledgments: This work was supported by NIH 2R44CA65147-02.

MathSoft, Inc.
1700 Westlake Ave. N, Suite 500
Seattle, WA 98109–9891, USA
Tel: (206) 283–8802
FAX: (206) 283–6310

E-mail: `timh@statsci.com`
Web: `www.statsci.com/Hesterberg`

A Graphical Representation of Little's Test for MCAR

Tim C. Hesterberg

Abstract

We describe a graphical representation of Little's (1988) test for MCAR (missing completely at random). The test statistic has an asymptotic χ^2 distribution. Terms in the statistic are represented graphically as rectangles, whose widths, areas, and heights are chosen to accurately indicate the contributions of the terms to the final statistic.

Key Words: Missing data.

1 Introduction

This document describes a graphical representation of the test for MCAR (missing completely at random) described in Little (1988). There are p variables (columns), J unique (row) patterns of missing values (out of 2^p possible patterns), m_j observations (rows) with the j'th missingness pattern, p_j is the number of nonmissing observations in pattern j, $\hat{\mu}$ and $\hat{\Sigma}$ are the maximum likelihood estimates of the parameters of a p-dimensional multivariate normal distribution based on the available data, $\hat{\mu}_{\mathrm{obs},j}$ and $\hat{\Sigma}^{-1}_{\mathrm{obs},j}$ are the subsets of the parameters corresponding to nonmissing observations for pattern j (a vector of length p_j and matrix of dimension $p_j \times p_j$, respectively), and $\overline{y}_{\mathrm{obs},j}$ the p_j-dimensional sample average of observed data in pattern j.

The test statistic is

$$
\begin{aligned}
d^2 &= \sum_{j=1}^{J} d_j^2 \\
&= \sum_{j=1}^{J} m_j (\overline{y}_{\mathrm{obs},j} - \hat{\mu}_{\mathrm{obs},j}) \hat{\Sigma}^{-1}_{\mathrm{obs},j} (\overline{y}_{\mathrm{obs},j} - \hat{\mu}_{\mathrm{obs},j})^T \qquad (1)
\end{aligned}
$$

(Little's equation (5)) in the case Σ is unknown. Asymptotically, this has a χ^2 distribution with

$$\sum_{j=1}^{J} p_j - p$$

degrees of freedom under fairly general assumptions.

If this test rejects H_0 it may be possible to gain insight as to why the test was rejected by looking at the individual terms to see which were particularly large. However straightforward comparisons of the sizes of terms are misleading. We describe here a graphical presentation of Little's test which is intended to avoid misleading.

There are two themes which inform our proposed representation. First, and most important, is that the individual terms d_j^2 in (1) should be compared to χ^2 distributions with *different* degrees of freedom. Second, the expected values of the individual terms are equal to certain factors times the nominal degrees of freedom for the terms.

The test statistic (1) is similar to

$$\sum_{j=1}^{J} m_j (\overline{y}_{\text{obs},j} - \mu_{\text{obs},j}) \Sigma_{\text{obs},j}^{-1} (\overline{y}_{\text{obs},j} - \mu_{\text{obs},j})^T. \tag{2}$$

If MCAR holds and Σ is positive definite, and if the underlying distribution is normal (or asymptotically otherwise), then each of the terms in the summation has a χ^2 distribution with p_j degrees of freedom. This continues to hold, asymptotically, if Σ is replaced by $\hat{\Sigma}$, or by any other consistent estimate.

2 Rectangles

We propose to represent the terms d_j^2 graphically by rectangles, one for each pattern j, with width equal to the degrees of freedom p_j, and height d_j^2/p_j. The area of the jth rectangle is d_j^2, and the total area is d^2. The heights of the individual rectangles encode the contribution of that pattern to d^2, normalized by the degrees of freedom. A horizontal line at height 1 (the expected value of a χ^2 variable divided by its degrees of freedom) is added for reference. These rectangles are shown in Figure 1. In this case Little's test does not reject H_0, and none of the rectangles are much taller than 1. The rectangle for missingness pattern 2 is the tallest, indicating that the means for that pattern were the most different from the overall means.

3 Height Factors

If $\mu_{\text{obs},j}$ in (2) is replaced by $\hat{\mu}_{\text{obs},j}$ (as in (1)) the test loses p degrees of freedom and the individual terms tend to be correspondingly smaller, in general by different factors. We propose to calculate these factors and include them in the graphical representation, to help avoid misleading comparisons.

The expected heights (the factors) are

$$E(d_j^2)/p_j \doteq c_j = 1 - \text{tr}((\sum_{\ell} \mathbf{A}_\ell)^{-1} \mathbf{A}_j)/p_j, \tag{3}$$

where $\mathbf{A_k}$ are weight matrices such that $\hat{\mu} = (\sum_\ell \mathbf{A}_\ell)^{-1} \sum_j \mathbf{A}_j \overline{y}_j$. We defer the derivation to the Appendix. Three points are worth noting here. First, $0 \leq c_j \leq 1$. Second, $\mathbf{A}_j$ is

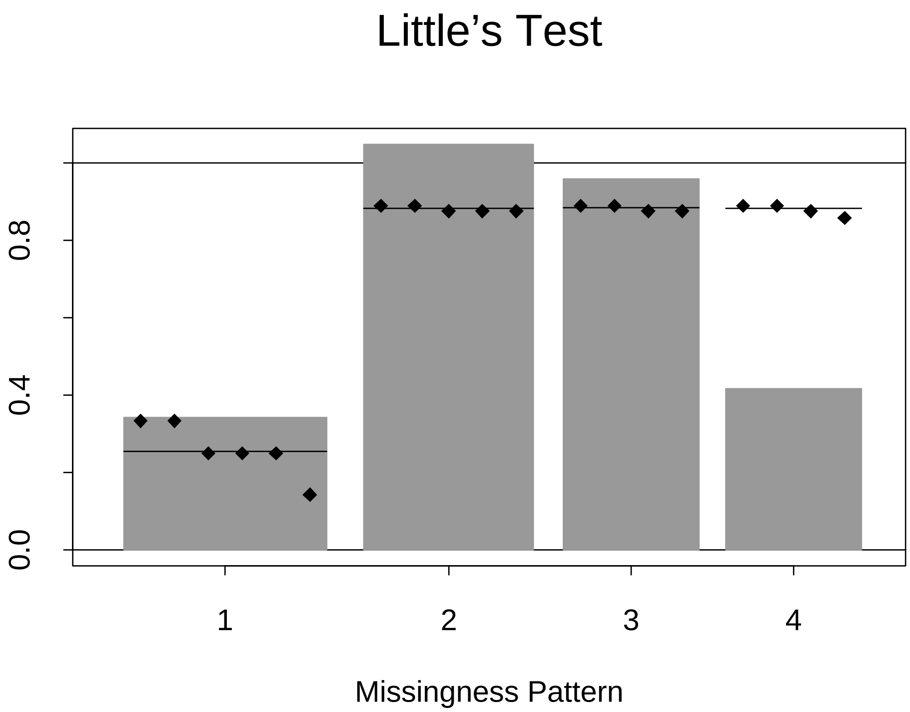

Figure 1: Graphical representation of Little's Test. There is one rectangle for each term in the test, with widths proportional to the nominal degrees of freedom for the term, and area proportional to the term. There is a single horizontal reference line at 1, and individual reference lines for each rectangle at the expected height for that rectangle. Individual variable factors are shown as well; these are optional.

proportional to m_j, the number of observations in pattern j, so that rectangles for "large" patterns (those which many observations match) tend to be shorter. Third, $\sum_j p_j c_j = \sum_j p_j - p$. In other words, adding the products of the width and approximate expected height of each rectangle gives the degrees of freedom of the test statistic.

The expected heights are shown as horizontal black lines in Figure 1. In this case the first three rectangles are slightly taller than their expected values under the null hypothesis, and so make positive contributions to the the overall χ^2 statistic, but the last rectangle is shorter than expected.

An alternate representation would modify the rectangles, multiplying the width and dividing the height of each rectangle by the factor. However then the widths of the rectangles would no longer indicate the number of non-missing variables in the corresponding missingness pattern, which is particularly useful in the next section.

3.1 Individual Variable Factors

Here we derive an alternative way to indicate the expected height of each pattern, in which the single factor c_j for term j is replaced by multiple factors c_{jk}, each of which corresponds to one of the p_j nonmissing variables in pattern j.

The idea is to focus on one variable at a time, say variable k. We estimate μ_k by $\overline{y}_{\mathrm{obs},k}$, the average of the available data for that variable. Then for any missingness pattern j for which variable k is observed,

$$\begin{aligned} c_{jk} &= E(m_j(\overline{y}_{\mathrm{obs},j,k} - \overline{y}_{\mathrm{obs},k})\Sigma_{kk}^{-1}(\overline{y}_{\mathrm{obs},j,k} - \overline{y}_{\mathrm{obs},k})^T) \\ &= (1 - m_j/\sum_{\ell} m_\ell)E(m_j(\overline{y}_{\mathrm{obs},j,k} - \mu_k)\Sigma_{kk}^{-1}(\overline{y}_{\mathrm{obs},j,k} - \mu_k)^T) \\ &= (1 - m_j/\sum_{\ell} m_\ell). \end{aligned} \tag{4}$$

The terms here are the expected values of χ^2_1 variables times factors which depend only on the patterns of missing values; see the Appendix for a proof. The same results hold asymptotically if Σ_{kk} is replaced by a consistent estimate.

In the multivariable situation, each term d_j^2 is a normalized Mahalanobis distance, based on p_j nonmissing variables for that pattern. For each of those variables we may compute a factor $c_{jk} = (1 - m_j/\sum_\ell m_{\ell,k})$. Figure 1 shows dots corresponding to these individual variable factors over the corresponding rectangle, at a height equal to the factor. The average height of the dots is roughly equal to the expected height of the rectangle, under the null hypothesis.

The dots also give an idea of the number of observations in each pattern, because the distance $1 - c_{jk}$ between each dot and the horizontal line at $y = 1$ is equal to the proportion of nonmissing observations for variable k which are found in missingness pattern j. In this case pattern 1, the pattern with no missing variables, had the most observations (6 out of 9 total observations, and a higher proportion of the nonmissing observations for some variables).

These factors also satisfy an equality for the overall degrees of freedom, $\sum_{jk} c_{jk} = \sum p_j - p$.

Appendix

We prove two results in this appendix. We defer till last (3), which requires new notation and is more involved, and begin instead with equation (4). This is equivalent to the simpler one-variable, two-group situation in which $\overline{y}_j$ is the sample average of m_j observations in group j, $j = 1, 2$, with the grand average $\overline{y} = (m_1\overline{y}_1 + m_2\overline{y}_2)/(m_1 + m_2)$. Let μ and σ^2 be the mean and variance of the underlying distribution. Then

$$\begin{aligned} \overline{y}_j - \overline{y} &= \overline{y}_j - (m_1\overline{y}_1 + m_2\overline{y}_2)/(m_1 + m_2) \\ &= (1 - \frac{m_j}{m_1 + m_2})(\overline{y}_j - \overline{y}_{-j}) \end{aligned} \tag{5}$$

where $\overline{y}_{-j}$ is the sample average of the "other" (not j) group, so

$$\begin{aligned} E(m_j\sigma^{-2}(\overline{y}_j - \overline{y})^2) &= m_j\sigma^{-2}(1 - \frac{m_j}{m_1 + m_2})^2(\sigma^2/m_1 + \sigma^2/m_2) \\ &= 1 - m_j/(m_1 + m_2) \end{aligned} \tag{6}$$

Furthermore, $\overline{y}_j - \overline{y}_{-j}$ is normally distributed if the underlying distribution is normal, or asymptotically otherwise, in which case $m_j\sigma^{-2}(\overline{y}_j - \overline{y})^2)$ is (asymptotically) distributed as $1 - m_j/(m_1 + m_2)$ times a χ^2_1 random variable.

Now turn to (3). We begin by giving some notation, then derive a matrix expression for the the maximum likelihood estimate of μ for given Σ, then use that expression to find the desired expected value.

Let $\tilde{j}$ be an index vector of length p corresponding to missingness pattern j, such that $\mu_{\tilde{j}}$ is a vector of length p which has the same elements as μ in those positions corresponding to variables which are present in pattern j, and which has structural zeros in the other positions. Similarly let $\overline{y}_{j,\tilde{j}}$ be the vector of length p containing the observed means for pattern j.

Let $\Sigma_{\tilde{j},\tilde{j}}$ be $p \times p$ covariance matrix which matches Σ except for structural zeros in those rows and columns for which the corresponding variables are missing for pattern j, and let $\Sigma^{-1}_{\tilde{j},\tilde{j}}$ be its generalized inverse with structural zeros in the same positions. Note that the variance (matrix) of $\overline{y}_{j,\tilde{j}}$ is $m_j^{-1}\Sigma_{\tilde{j},\tilde{j}}$. Let $\Sigma_{\cdot,\tilde{j}}$ match Σ except for structural zeros in columns which are missing for pattern j, so that $\Sigma_{\cdot,\tilde{j}} = \Sigma_{\tilde{j},\tilde{j}} + \Sigma_{\tilde{-j},\tilde{j}}$ where $\tilde{-j}$ is the complement of $\tilde{j}$.

Now given an observation from pattern $\tilde{j}$, the expected values for the missing observations are given by the usual multiple linear regression, written here as

$$E(Y_{\tilde{-j}}|Y_{\tilde{j}}) = \mu_{\tilde{-j}} + \Sigma_{\tilde{-j},\tilde{j}}\Sigma^{-1}_{\tilde{j},\tilde{j}}(Y_{\tilde{j}} - \mu_{\tilde{j}}).$$

Rearranging this slightly, and combining it with the known value for $Y_{\tilde{j}}$, we obtain

$$E(Y|Y_{\tilde{j}}) = \Sigma_{\cdot,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}Y_{\tilde{j}} + (\mathbf{I} - \Sigma_{\cdot,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1})\mu.$$

This forms the basis for the EM algorithm for estimating μ if Σ is known, with

$$\begin{aligned}
\hat{\mu}^{(t+1)} &= n^{-1}\sum_j m_j E(Y|Y_{\tilde{j}} = \overline{y}_{j,\tilde{j}}; \hat{\mu}^{(t)}) \\
&= n^{-1}\sum_j (m_j \Sigma_{\cdot,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}\overline{y}_{j,\tilde{j}}) + (m_j(\mathbf{I} - \Sigma_{\cdot,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1})\hat{\mu}^{(t)}) \\
&= \mathbf{C} + \mathbf{M}\hat{\mu}^{(t)} \qquad (7)
\end{aligned}$$

where $\mathbf{C}$ is a matrix that depends on the observed data but not on $\hat{\mu}^{(t)}$ and $\mathbf{M}$ is a matrix depends only on Σ. This iterative equation has the fixed point solution

$$\begin{aligned}
\hat{\mu} &= (\mathbf{I} - \mathbf{M})^{-1}\mathbf{C} \\
&= (\sum_k \mathbf{A}_k)^{-1}\sum_j \mathbf{A}_j \overline{y}_{j,\tilde{j}} \qquad (8)
\end{aligned}$$

where $\mathbf{A}_j = (m_j/n)\Sigma_{\cdot,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}$.

Now the variance of $\hat{\mu}$ is given by

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= (\sum_k \mathbf{A}_k)^{-1}\sum_j \mathbf{A}_j \mathrm{Var}(\overline{y}_{j,\tilde{j}})\mathbf{A}_j'(\sum_k \mathbf{A}_k)^{-1'} \\
&= (\sum_k \mathbf{A}_k)^{-1}\sum_j \mathbf{A}_j m_j^{-1}\Sigma_{\tilde{j},\tilde{j}}\mathbf{A}_j'(\sum_k \mathbf{A}_k)^{-1'} \\
&= (\sum_k \mathbf{A}_k)^{-1}\sum_j \mathbf{A}_j n^{-1}\Sigma_{\tilde{j},\cdot}(\sum_k \mathbf{A}_k)^{-1'} \\
&= n^{-1}(\sum_k \mathbf{A}_k)^{-1}\sum_j \mathbf{A}_j \Sigma(\sum_k \mathbf{A}_k)^{-1'} \\
&= n^{-1}\Sigma(\sum_k \mathbf{A}_k)^{-1'} \\
&= n^{-1}(\sum_k \mathbf{A}_k)^{-1}\Sigma. \qquad (9)
\end{aligned}$$

The last step is by symmetry. The change from $\Sigma_{\tilde{j},\cdot}$ to $\Sigma = \Sigma_{\cdot,\cdot}$ is justified because the $\mathbf{A}_j$ has structural zeros in the appropriate columns.

We next use the variance of $\hat{\mu}$ to find an expected value we need below; assuming that $\mu = 0$ (this will not cause a loss of generality later),

$$\begin{aligned}
E(m_j\hat{\mu}_{\tilde{j}}'\Sigma_{\tilde{j},\tilde{j}}^{-1}\hat{\mu}_{\tilde{j}}) &= E(m_j\hat{\mu}'\Sigma_{\tilde{j},\tilde{j}}^{-1}\hat{\mu}) \\
&= m_j \mathrm{tr}(\mathrm{Var}(\hat{\mu})\Sigma_{\tilde{j},\tilde{j}}^{-1})
\end{aligned}$$

$$
\begin{aligned}
&= (m_j/n)\mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}\Sigma\Sigma_{\tilde{j},\tilde{j}}^{-1}) \\
&= (m_j/n)\mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}\Sigma\cdot,\tilde{j}\Sigma_{\tilde{j},\tilde{j}}^{-1}) \\
&= \mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}A_j). \qquad (10)
\end{aligned}
$$

Two of the steps here are justified by the locations of structural zeros; the first equality involving the trace of the matrix is justified because for any variable Y and matrix M, $E(Y'MY) = \mathrm{tr}(\Sigma_Y M)$.

We are now ready to find the desired expected value,

$$
\begin{aligned}
E(d_j^2) &= E(m_j(\overline{Y}_{j,\tilde{j}} - \hat{\mu}_{\tilde{j}})'\Sigma_{\tilde{j},\tilde{j}}^{-1}(\overline{Y}_{j,\tilde{j}} - \hat{\mu}_{\tilde{j}})) \\
&= E(m_j\overline{Y}'_{j,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}\overline{Y}_{j,\tilde{j}}) - 2E(m_j\overline{Y}'_{j,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}\hat{\mu}_{\tilde{j}}) + E(m_j\hat{\mu}'_{\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}\hat{\mu}_{\tilde{j}}) \\
&= p_j - 2E(m_j\overline{Y}'_{j,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}(\sum_k \mathbf{A}_k)^{-1}(\sum_\ell \mathbf{A}_\ell\overline{Y}_{\ell,\tilde{\ell}}) + \mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}A_j) \\
&= p_j - 2E(m_j\overline{Y}'_{j,\tilde{j}}\Sigma_{\tilde{j},\tilde{j}}^{-1}(\sum_k \mathbf{A}_k)^{-1}\mathbf{A}_j\overline{Y}_{j,\tilde{j}}) + \mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}A_j) \\
&= p_j - 2\mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}\mathbf{A}_j) + \mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}A_j) \\
&= p_j - \mathrm{tr}((\sum_k \mathbf{A}_k)^{-1}A_j). \qquad (11)
\end{aligned}
$$

Of the three terms in the second line above, the third was simplified earlier and the first simplifies because $m_j\Sigma_{\tilde{j},\tilde{j}}^{-1}$ is the generalized inverse of the covariance matrix of $\overline{Y}_{j,\tilde{j}}$. For the second term one step is justified by the independence of $\overline{Y}_{j,\tilde{j}}$ and $\overline{Y}_{\ell,\tilde{\ell}}$ for $\ell \neq j$, and another because for any variable Y and matrix M, $E(Y'MY) = \mathrm{tr}(\Sigma_Y M)$.

References

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.