TESTS FOR GAUSSIAN REPEATED MEASURES WITH MISSING DATA IN SMALL SAMPLES

Diane J. Catellier
and
Keith E. Muller

Department of Biostatistics CB#7400
University of North Carolina
Chapel Hill, North Carolina 27599-7400

D. Catellier: telephone 919-962-1895, email DCatelli@BIOS.UNC.EDU

FAX 919-962-3264

Key words: test size, multivariate linear models, MANOVA

Acknowledgements. An earlier version of this paper was one portion of the dissertation research for the first author's Dr. Ph. degree program. The authors thank Dr.'s J. D. Hosking, G. G. Koch, P. W. Stewart, and D. J. Weber for helpful comments. Catellier's work supported in part by USFHHS grant 90CA1467 and IPRIC grant 5-326-12. Muller's work supported in part by NIH grants PO1-CA47982-04, RO1-CA67183-01A1, RO1-CA72875, RO1-CA60193-04, MO1-RR000-46-33, NO1-ES-35356, and MH33127.

SUMMARY

For small samples of Gaussian repeated measures with missing data, Barton and Cramer (1989) recommended using the EM algorithm for estimation and reducing the degrees of freedom for an analog of Rao's F approximation to Wilks' test. Computer simulations led to the conclusion that the modified test was slightly conservative for total sample size of $N = 40$. Here we consider additional methods and smaller sample sizes, $N \in \{12, 24\}$.

We describe analogs of the Pillai-Bartlett trace, Hotelling-Lawley trace and Geisser-Greenhouse corrected univariate tests which allow for missing data. Eleven sample size adjustments were examined which replace N by some function of the numbers of non-missing pairs of responses in computing error degrees of freedom.

Overall, simulation results allowed concluding that an adjusted test can always control test size at or below the nominal rate, even with as few as 12 observations and up to 10% missing data. The choice of method varies with the test statistic. Replacing N by the mean number of non-missing responses per variable works best for the Geisser-Greenhouse test. The Pillai-Bartlett test requires the stronger adjustment of replacing N by the harmonic mean number of non-missing pairs of responses. For Wilks' and Hotelling-Lawley, an even more aggressive adjustment based on the minimum number of non-missing pairs must be used.

1. Introduction

1.1 *Motivation*

Repeated measurements can be distinct variables, or a single variable measured at several points in time, with the spacing consistent across subjects. For ease of presentation, call the experimental unit a “subject” and the metameter on which the measurements are indexed “time”. Traditional linear models are particularly useful if the responses are at least approximately Gaussian and can be explained by some linear function of predictors. When each subject is observed at the same p times and no missing observations, closed-form maximum likelihood (ML) estimates of the model parameters are often available. Often, especially in clinical trials, the response is not observed at all time points for every subject.

A large amount of research has been directed at estimation for linear repeated measures models with missing data. These appear to work well in both large and small samples. In contrast, much less effort has been directed towards methods for inference. Various asymptotic test statistics work well in large samples. However, in small samples the available methods for inference may work very poorly. In particular, the methods produce inflated type I error rates in small samples (Barton and Cramer, 1989; Schluchter and Elashoff, 1990).

We seek to develop hypothesis tests for Gaussian repeated measures with missing data, accurate in small samples. In doing so, we restrict attention to a particular range of studies. We assume that the missing data elements are "missing at random" (MAR; Rubin, 1976).

1.2 *General Strategies for the Analysis of Repeated Measures Designs*

Models in which the expected value of the response vector equals a linear function of the parameters have traditionally been described as (general) linear models. Most often, one of three strategies is used for linear models with repeated measures: the multivariate analysis of variance (MANOVA) approach, the univariate approach to repeated measures, or mixed model analysis. All three models account for the dependencies among the repeated measures, but differ in the special form assumed for the covariance matrix within subjects, $\boldsymbol{\Sigma}$.

See Muller, LaVange, Ramey and Ramey (1992) for an overview of the multivariate and univariate approaches to repeated measures ANOVA. They described the assumptions behind the methods as well as the most widely used tests for both approaches. For the sake of brevity, the information in that article will be assumed throughout.

The mixed model has long been used for the analysis of continuous data, especially for missing and mistimed data. By virtue of modeling the subject as a random component, the mixed model can encompass a broad range of covariance structures. In this paper, we will compare existing inference methods for all three approaches to new ones.

Using the terminology of Rubin (1976), missing responses are said to be missing at random (MAR) if missingness of a particular response does not depend on its unobserved value, but can depend on the covariates or any of the observed responses. Likelihood-based estimation methods assume that the data are MAR.

Alternatively, one can use the quasi-likelihood approach of Liang and Zeger (1986) by solving the generalized estimating equations (GEE) to obtain estimates of the regression parameters. Park (1993) compared the GEE approach to the ML approach for multivariate normal outcomes. He showed that with no missing data and an unstructured covariance matrix that the GEE and ML score equations are equivalent and lead to the same estimates of expected value and covariance parameters. With missing observations, however, the equivalence fails. For data missing completely at random (MCAR, Rubin, 1976), the GEE solution produces consistent estimators. Three weaknesses, however, make GEE less desirable than ML estimation in the missing data setting. First, the GEE estimate of the working covariance matrix may not always be positive definite, while the ML estimate from the EM algorithm (Dempster, Laird and Rubin, 1977) is guaranteed to be positive definite (Laird, Lange and Stram, 1987). Second, simulation studies by Stiger, Kosinski, Barnhart and Kleinbaum (1997), Qu, Piedmonte, and Williams (1994), Park (1993), and Emrich and Piedmont (1992) allow concluding that ML estimators perform better in small samples than GEE estimators. The former tend to have less bias, smaller mean squared errors and more accurate test size. Third, under misspecification of the covariance, the GEE estimators will only be consistent provided that the missing observations are MCAR. ML procedures give unbiased estimates under the weaker MAR assumption. The limitations of the GEE procedure in small samples, combined with the focus on Gaussian data, makes ML estimation more attractive and therefore the focus of this paper.

2. Known Methods For Estimation And Inference

2.1 *Complete Data*

The repeated measures model can be represented as a special case of the general linear multivariate model (GLMM). See Muller *et al*. (1992), Davidson (1972), and O'Brien and Kaiser (1985) for further discussion. Suppose that the responses for subject i ($i \in \{1, \ldots, N\}$) are measured at p times (the same for all subjects). Specify the GLMM as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \mathbb{E}, \tag{2.1}$$

with $\boldsymbol{Y}$ the observed matrix of random observations ($N \times p$), $\boldsymbol{X}$ the design matrix ($N \times q$, fixed and known, conditional upon having chosen the subjects) and $\boldsymbol{B}$ the fixed and unknown parameters ($q \times p$). Indicate the i^{th} row of $\boldsymbol{X}$ as $\boldsymbol{X}_i = \text{row}_i(\boldsymbol{X})$. Assuming that $\text{row}_i(\boldsymbol{E})$ is $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, we can test the general linear hypothesis (GLH)

$$H_0\text{: } \boldsymbol{\Theta} = \boldsymbol{C}\boldsymbol{B}\boldsymbol{U} = \boldsymbol{\Theta}_{\text{o}} \text{ versus } H_{\text{A}}\text{: } \boldsymbol{\Theta} \neq \boldsymbol{\Theta}_{\text{o}}. \tag{2.2}$$

Each row of $\boldsymbol{C}$ $(a \times q)$ defines a between-subject contrast and each column of $\boldsymbol{U}$ $(p \times b)$ defines a within-subjects contrast. Define $\nu_E = N - \text{rank}(\boldsymbol{X})$. All tests of H_0 are based on

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}, \tag{2.3}$$

$$\widehat{\boldsymbol{\Theta}} = \boldsymbol{C}\widehat{\boldsymbol{B}}\boldsymbol{U}, \tag{2.4}$$

$$\widehat{\boldsymbol{H}} = (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}']^{-1}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0), \tag{2.5}$$

and

$$\widehat{\boldsymbol{E}} = \boldsymbol{U}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{U}. \tag{2.6}$$

$\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{E}}$ have Wishart distributions $W(a, \boldsymbol{\Sigma}_*, \boldsymbol{\Omega})$ and $W(\nu_E, \boldsymbol{\Sigma}_*)$ respectively, with

$$\boldsymbol{\Omega} = (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}']^{-1}(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)\,\boldsymbol{\Sigma}_*^{-1}. \tag{2.7}$$

Let $s = \min(a, b)$ indicate the rank of $\widehat{\boldsymbol{H}}$ and hence $\widehat{\boldsymbol{H}}\widehat{\boldsymbol{E}}^{-1}$.

The common multivariate tests may be constructed using the eigenvalues $(l_1, \ldots, l_s)$ of $\widehat{\boldsymbol{H}}\widehat{\boldsymbol{E}}^{-1}$. Specifically, for Wilks' Lambda $W = \prod_{i=1}^{s}(1+l_i)^{-1}$, for Pillai-Bartlett trace $V = \sum_{i=1}^{s} l_i(1+l_i)^{-1}$, for Hotelling-Lawley trace $U = \sum_{i=1}^{s} l_i$, and for Roy's largest root $R = \max(l_i)$. When $s > 1$ (or $s > 2$ for Wilks'), closed form expressions for the distributions of these test statistics are not available and approximations are used (§2.2, Muller, *et al*. 1992). In general, no uniformly most powerful test exists. Hence the optimal choice depends on the alternative hypothesis. See Olson (1974, 1976, 1979), Anderson (1984, pp. 330-333), and Muller, *et al.* (1992) for detailed discussions. Concerns about robustness and power led to not considering Roy's test any further.

The test statistics W, V, and U can be accurately approximated by an F random variable. Rao's (1973; also detailed in Muller *et. al.,* 1992) approximation for W works well, even in very small samples. Although widely used in current statistical packages, Pillai's (1954) F approximation for V may be very conservative in small samples. Muller (1998) developed an F approximation for V that provides substantially better accuracy. Hence Muller's approximation will be used, with $\nu_1(V) = Kab$, $\nu_2(V) = K \cdot s(\nu_E + s - b)$, with

$$K = \frac{1}{s(\nu_E + a)}\left[\frac{s(\nu_E + s - b)(\nu_E + a + 2)(\nu_E + a - 1)}{\nu_E(\nu_E + a - b)}\right]. \tag{2.8}$$

McKeon (1974) provided a slightly better F approximation than the more widely used Pillai-Sampson approximation for the Hotelling-Lawley statistic. Write the McKeon approximation

$$F = \frac{(U/h)/(ab)}{1/\nu_2(U)}\,, \tag{2.9}$$

with $\nu_1(U) = ab$, $\nu_2(U) = (4 + ab + 2)g'$,

$$g' = \frac{\nu_E^2 - \nu_E(2b+3) + b(b+3)}{\nu_E(a+b+1) - (a+2b+b^2-1)}\,, \tag{2.10}$$

and

$$h = \frac{\nu_2(U) - 2}{\nu_E - b - 1} \,. \tag{2.11}$$

See Muller and Barton (1989, 1991) or Muller *et al.* (1992) for details of the computation and approximation of the distribution of the various tests for the "univariate" approach to repeated measures. Muller and Barton (1989) suggested that the Geisser-Greenhouse (GG) test provides the best compromise in controlling type I error rate with excellent power. Hence only the GG test statistic will be examined here.

2.2 *Data Missing at Random*

For the GLMM, both ML and REML estimation methods have been extensively investigated for MAR data. For a general review, see Little and Rubin (1986, chapters 7-10). For some special case patterns of missing data, such as monotone missing data, the likelihood equations have a closed form solution (Rubin, 1974). For arbitrary missing data patterns, the solution must be obtained iteratively. The computational efficiency and simplicity of the EM algorithm (Orchard and Woodbury, 1972; Beale and Little, 1975; Dempster, Laird and Rubin, 1977) make it an attractive choice for ML estimation in the setting of interest. Barton and Cramer's experience with the method for the application at hand also strongly supports the choice. The best algorithm for more general models is not as obvious. See for example, Mensah, Elswick and Chinchilli (1993) or Callahan and Harville (1990).

Except in special cases, no known method exists for providing accurate and efficient inference in small multivariate normal samples with missing data. Hypothesis tests constructed from complete observations only, while accurate in small samples, are inefficient. A number of approximate methods have been proposed for the problem of testing equality of means for a bivariate normal sample with data missing on one variable (Morrison, 1973; and Little, 1976 and 1988).

Barton and Cramer (1989) suggested an appealing technique for testing the general linear hypothesis in a GLMM with an arbitrary pattern of missing data. The approach involves using the EM algorithm for ML estimation, and modifying Rao's F approximation to Wilks' test, F_W, with adjusted error degrees of freedom. Let $N_{jj'}$ indicate the number of observations for which both Y_{ij} and $Y_{ij'}$, for $i \in \{1 \ldots N\}$, have non-missing values. Note that N_{jj} equals the number of cases observed for the j^{th} response. All adjustments considered by Barton and Cramer, and in this paper, involve replacing N by N_* in $\nu_E = N - \text{rank}(\boldsymbol{X})$. In all cases N_* equals a function only of $\{N_{jj'}\}$. For samples of size 40 and up to 20% missing data, test statistics with degrees of freedom based on the naive choice $N_* = N$ produced inflated type I error rates ranging from 0.10 to 0.23 assuming a nominal rate of 0.05. In contrast, choosing N_* as the number of complete cases gave very conservative rates (0.004-0.014). The best choice was the average number of non-missing pairs of responses. An analog of Wilks' test based on this adjustment produced acceptable test sizes across all simulated conditions.

The mixed model is often used for multivariate data with some missing observations. Let $\boldsymbol{y}_i$ be an $(N_i \times 1)$ vector of measurements for the i^{th} subject, and $N_+ = \sum_{i=1}^{N} N_i$. In the mixed model, $\boldsymbol{y}_+ = [\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N']'$ is modeled as

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}_+, \tag{2.12}$$

with $\boldsymbol{X}$ and $\boldsymbol{Z}$ the known design matrices for the fixed and random effects respectively, and the $\boldsymbol{b}$ the vector of unknown random effects. The key assumptions for inference are that $\boldsymbol{b}$ and $\boldsymbol{e}$ are independent and multivariate Gaussian. Define vec$(\boldsymbol{M})$ as the vector created by stacking the columns of $\boldsymbol{M}$. Also let $\boldsymbol{A} \otimes \boldsymbol{B} = \{a_{ij}\boldsymbol{B}\}$ indicate the (left) Kronecker product. For the cases of interest, the mixed model may be written so that $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B}')$, with $\boldsymbol{B}$ from the GLMM, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \mathbb{E}$. For complete data $\boldsymbol{X}_+ = \boldsymbol{X} \otimes \boldsymbol{I}_p$. For missing data merely delete each row corresponding to a missing response. Let $\boldsymbol{\Sigma}_i$, of dimension p_i, be the sub-matrix of $\boldsymbol{\Sigma}$ with rows and columns corresponding to data observed for subject i. Let $\boldsymbol{\Lambda} = \mathrm{Dg}(\boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_N)$ indicate the block-diagonal matrix created by placing $\boldsymbol{\Sigma}_1$ in the upper left diagonal, *etc*. Then $\boldsymbol{e}_+ = \mathcal{N}_{N_+}(\mathbf{0}, \boldsymbol{\Lambda})$. Likelihood-based estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ typically requires iterative methods. The software used here (SAS®, PROC MIXED) employs a method of Lindstrom and Bates (1988).

Exact methods are not available to test hypotheses about $\boldsymbol{\beta}$. The approximate, large-sample test statistics can be unreliable in small samples. Schluchter and Elashoff (1990, §6) examined the test size of various ML Wald-type statistics. Their small sample simulation results led them to suggest approximating a modified Wald statistic by an F distribution with denominator degrees of freedom based on the number of complete cases. A simple χ^2 approximation to the LR statistic (Hocking, 1985, §8.3.1) proves unreliable in small samples (Woolson, Leeper and Clarke, 1978; Woolson and Leeper, 1980; and Leeper and Woolson, 1982). The version of PROC MIXED studied here used (SAS, 1997, p644):

$$F = \frac{\widehat{\boldsymbol{\theta}}'\left(\boldsymbol{X}_+'\widehat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}_+\right)^{-}\widehat{\boldsymbol{\theta}}}{\mathrm{rank}(\boldsymbol{C})}, \tag{2.13}$$

with numerator degrees of freedom equal to $\mathrm{rank}(\boldsymbol{C})$. Although several approximations are available for the denominator degrees of freedom (see SAS, 1997, page 607), only the Satterthwaite approximation was considered in this paper.

3. New Tests For Data Missing At Random

The success of the Barton and Cramer strategy encouraged us to consider a number of generalizations. First, the approach will be applied to other test statistics. Second, some additional functions of the sample sizes merit consideration. Third, even smaller sample sizes will be studied. In all cases the EM algorithm will be used for estimation.

In addition to W, the U, V and GG tests may be applied to missing data. First, use the EM algorithm to compute maximum likelihood estimates of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$. Second, use the estimates to compute an analog of $\widehat{\boldsymbol{H}}\widehat{\boldsymbol{E}}^{-1}$, as well as analogs of U, V and GG statistics. Third, compute F approximations, changing only the error degrees of freedom by replacing N with some form of N_*.

Overall, eleven forms for N_* will be examined for each of four test statistics. They are listed in Table 1 in rank order from smallest to largest, with the exception that N_{*3} can be either less than or greater than N_{*6} (and hence N_{*4} and N_{*5},). In all cases $N/N_* \to 1$. Consequently, in large samples (as $N \to \infty$, with fixed p, q, and proportion missing) the choice of N_* has less and less effect. The form of the results of Rothenberg (which assume a sequence of local alternatives), as cited in Anderson (1984, §8.6.5) support this position.

4. Numerical Evaluations

4.1 *Methods*

All simulations involved a small range of research designs. In all cases, the designs included 1) one within-subject factor with $p = 3$ or 6 levels, 2) one between-subject factor with $q = 4$ levels, 3) $N \in \{12, 24\}$, and 4) 0%, 5% or 10% of the data missing. No subject's data were allowed to be completely missing. The procedure for producing missing data generated data that are MCAR. Other factors considered are the relative error variance of response variables (equal, unequal), and the error correlation structure (medium, high). See Tables I-II in Barton and Cramer (1989) for details. In addition, a third level was added to the correlation structure factor, which allowed assessing the effect of very low correlation between the responses. The structure specified equal correlation ($\rho = 0.1$) for each pair of responses. The overall test for the presence of a trend (linear, quadratic or cubic) with respect to the between-subject factor for each of the response measures was of primary interest. Under the null, of course, $\boldsymbol{B} = \mathbf{0}$. For 5,000 replications and assuming a true type I error rate of 0.05, the 95% confidence bounds around the type I error rate estimates are approximately $\pm$ 0.006.

4.2 *Results*

On average, higher levels of correlation within subjects were associated with modestly higher type I error rates. Since this pattern was consistent for each of the test statistics, only the results for the low and high correlation conditions will be presented.

The empirical type I error rates for the mixed model F statistic are given in Table 2. The results indicate that the test has poor small sample properties, producing inflated type I error rates, even when none of the data were missing. For $N = 24$, test sizes increased from slightly greater (0.07-0.10) to considerably greater (0.16-0.32) than the nominal $\alpha = 0.05$ level as the number of repeated measures increased from 3 to 6.

For the conditions with no missing data, all four univariate and multivariate test statistics succeeded in controlling the type I error rate at or below the nominal rate (see Table 3). This illustrates that the sample sizes, while small, are large enough that the approximate F tests are essentially unbiased for complete data. Hence any discrepancy from the desired test size may be attributed to the influence of missing responses, and not to any inaccuracy in test approximations for complete data.

Tables 4, 5, 6, and 7 summarize the empirical test sizes for W, U, V, and GG for 5% and 10% missing data. All tables give results for tests based on $N_{*2} = \min\{N_{jj'}\}$ and $N_{*11} = N$ in order to define bounds on test size.

The EM algorithm failed roughly 90% of the time for the condition with $p = 6$, $N = 12$, and even 5% missing data. Estimates are well defined for complete data. The table cells for these conditions were left blank. Not surprisingly, the results indicate that the worst accuracy tends to occur with more repeated measures, fewer subjects, more missing data and more correlation within subjects.

From Tables 4 and 5, it is evident that the adjusted F_{W} and F_U tests based on N_{*11} give inflated test sizes, and those based on N_{*4}, while accurate for $N = 24$ were inflated for $N = 12$. On the other hand, tests based on N_{*2} controlled test size at or below the nominal rate under all simulated conditions, with the exception of the condition with $p = 6$, $N = 12$ and 10% missing data, in which case test size was as high as 0.09.

Table 6 contains test size for modified F_V tests. The test based on N_{*11} provided inflated type I error rates, and the N_{*2}-adjusted test was conservative. Test sizes for N_{*4} were extremely accurate, with the exception of the condition with $p = 6$, $N = 12$ and 10% missing data where the type I error rates were approximately 0.1.

The results given in Table 7 suggest that N_{*9} is a reasonable adjustment function for the F_{GG} test. When $N = 12$, this test is slightly conservative, however this corresponds to the modest conservatism found when no data are missing (Table 1 in Muller and Barton, 1989).

When $s = \min(a, b) = 1$, all of the multivariate test statistics are equivalent. This can occur if the rank of the $\boldsymbol{C}$ contrast matrix is one ($a = 1$) or if the rank of U is one ($b = 1$). The empirical test sizes shown in Table 8 suggest that when $a = 1$, the best adjustment for the degrees of freedom is based on N_{*2}, while N_{*4} appears to work well when $b = 1$.

4.3 *Example*

Table 9 contains data from a study that examined the effects of choline deprivation on plasma choline concentration over 35 days, in healthy male subjects (Zeisel, DaCosta, Franklin, Alexander, Lamont, Sheard and Beiser, 1991). Subjects were given a standard diet which included 500 mg/day of choline for one week, and then were randomly assigned into two diet groups, one that contained choline and one that did not. During the fifth week of study all subjects again consumed a diet containing choline. Blood samples for choline analyses were obtained before the start of the study (day 0) and on days 7, 14, 21, 28, and 35. One subject had data missing for day 35.

A multivariate analysis of covariance model of difference scores allowed testing the effects of diet on the plasma choline concentration over time, while controlling for treatment group differences in baseline choline levels. The null hypothesis of interest is a test of the trends by treatment interaction. This particular design has two treatment groups. With no missing data, all multivariate F tests would coincide. Hence we chose to use a single choice of N_*. The value of the unadjusted (multivariate) F statistic ($N_* = N$) is 2.77, with $\nu_1 = 4$, $\nu_2 = 8$ and p value of 0.10. For the same hypothesis, unadjusted $F_{GG} = 2.67$, $\nu_1 = 2.95$, $\nu_2 = 32.4$ and $p = 0.065$. Using N_{*2} give an adjusted (multivariate) F of 2.65 with $\nu_1 = 4$ and $\nu_2 = 7$, leading to $p = 0.1443$. The results for using the univariate approach to repeated measures are $F_{GG} =$ 2.58, $\nu_1 = 2.95$, $\nu_2 = 31.4$ and $p = 0.072$. Both missing data analysis methods led to a smaller p value than the analysis based on 13 complete cases ($F =$ 2.00, $\nu_1 = 4$, $\nu_2 = 7$ and $p = 0.20$; $F_{GG} = 2.40$, $\nu_1 = 2.8$, $\nu_2 = 28.3$ and $p = 0.092$).

5. **Conclusions**

Conclusion 1. For all tests considered, accuracy decreases with more repeated measures, fewer subjects, more missing data and more correlation within subjects.

Conclusion 2. The mixed model F statistic used by PROC MIXED in SAS® with Satterthwaite approximated denominator degrees of freedom gives very liberal test size for $N \leq 24$, even with complete data.

Conclusion 3. For 6 responses and 12 subjects, the simple version of the EM algorithm used here failed roughly 90% of the time.

Conclusion 4. A degree of freedom adjustment can always control test size at or below the nominal level, even for conditions as extreme as $N = 12$ and 10% missing data.

Conclusion 5. The choice of adjustment varies with the test.

5.1 N_{*2}, the min$N_{jj'}$, is best for the Wilks' and Hotelling-Lawley tests.

5.2 N_{*4}, the harmonic mean of $N_{jj'}$, is best for the Pillai-Bartlett test.

5.3 N_{*9}, the mean N_{jj}, is best for the Geisser-Greenhouse test.

6. Future Research

The simulated data were generated in such a way as to create data that are MCAR. Hence MAR data deserve attention in future research.

Techniques for power analysis, given that test size can be controlled, would be very useful. The approach taken here is more intuitive than analytical. Nevertheless, we believe we have succeeded in describing an approach to ensure that test size does not exceed the nominal rate in small, missing data samples for the GLMM. A more formal approach must necessarily involve a rather sophisticated attack, due to the complexity of the distributions for the multivariate test statistics, even with complete data.

REFERENCES

Anderson, T. W. (1984). *An Introduction To Multivariate Statistical Analysis*. New York: Wiley.

Barton, C. N. and Cramer, E. C. (1989). Hypothesis testing in multivariate linear models with randomly missing data. *Communications in Statistics - Simulations* **18**, 875-895.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society-B* **37**, 129-145.

Callahan, T. P. and Harville, D. A. (1990). Some new algorithms for computing maximum likelihood estimates of variance components. *Journal of Statistical Computation and Simulation* **38**, 239-259.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin* **77**, 446-452.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society-B* **39**, 1-38.

Emrich, L. J. and Piedmonte, M. R. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* **41**, 19-29.

Hocking, R. R. (1985). *The Analysis of Linear Models*. Brooks/Cole Publishing Co., Monterey, CA.

Laird, N. M., Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association* **82**, 97-105.

Leeper, J. D. and Woolson, R. F. (1982). Testing hypotheses for the growth curve model when the data are incomplete. *Journal of Statistical Computation and Simulation* **15**, 97-107.

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

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association* **83**, 1014-1022.

Little, R. J. A. (1976). Inference about means from incomplete multivariate data. *Biometrika* **63**, 593-604.

Little, R. J. A. (1988). Approximate calibrated small sample inference about means from bivariate normal data with missing values. *Computational Statistics and Data Analysis* **7**, 161-178.

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

McKeon, J. J. (1974). F approximations to the distribution of Hotelling's T_0^2. *Biometrika* **61**, 381-383.

Mensah, R. D., Elswick, R. K. and Chinchilli, V. M. (1993). Consistent estimators of the variance-covariance matrix of the GMANOVA model with missing data. *Communications in Statistics - A* **22**, 1495-1514.

Morrison, D. F. (1973). A test for equality of means of correlated variates with missing data on one response. *Biometrika* **60**, 101-105.

Muller, K. E. (1998). A new F approximation for the Pillai-Bartlett trace under H_0. *Journal of Computational and Graphical Statistics* **7**, 131-137.

Muller, K. E. and Barton, C. N. (1989). Approximate power for repeated measures for repeated measures anova lacking sphericity. *Journal of the American Statistical Association* **84**, 549-555.

Muller, K. E. and Barton, C. N. (1991). *Correction to* Approximate power for repeated measures for repeated measures anova lacking sphericity. *Journal of the American Statistical Association* **86**, 255-256.

Muller, K. E., LaVange, L. M., Ramey, S. L. and Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association* **87**, 1209-1226.

O'Brien, R. G. and Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin* **97**, 316-333.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association* **69**, 894-908.

Olson, C. L. (1976). Choosing a test statistic in multivariate analysis. *Psychological Bulletin* **86**, 579-586.

Olson, C. L. (1979). Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. *Psychological Bulletin* **86**, 1350-1352.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697-715. Berkeley, California: University of California Press.

Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine* **12**, 1723-1732.

Pillai, K. C. S. (1954). On some distribution problems in multivariate analysis. *Institute of Statistics Mimeo Series No. 88*, University of North Carolina, Chapel Hill.

Qu, Y. S., Piedmonte, M. R. and Williams, G. W. (1994). Small sample validity of latent variable models for correlated binary data. *Communications in Statistics - Simulations* **23**, 243-269.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: Wiley.

Rubin, D. B. (1974). Characterization of estimation of parameters in incomplete data problems. *Journal of the American Statistical Association* **69**, 467-474.

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

SAS Institute Inc. (1997), *SAS/STAT Software: Changes and enhancements, Release 6.12*, SAS Institute Inc., Cary, NC.

Schluchter, M. D. and Elashoff, J. D. (1990). Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computing - Simulations* **37**, 69-87.

Stiger, T. R., Kosinski, A. S., Barnhart, H. X. and Kleinbaum, D. G. (1997). ANOVA for repeated ordinal data with small sample size? A comparison of ANOVA, MANOVA, WLS and GEE methods by simulation. JSM Abstract Book: p. 246.

Woolson, R. F., Leeper, J. D. (1980). Hypothesis testing in multivariate linear models with randomly missing data. *Communications in Statistics - Theory and Methods* **A9**, 1491-1513.

Woolson, R. F., Leeper, J. D. and Clarke, W. R. (1978). Analysis of incomplete data from longitudinal and mixed longitudinal studies. *Journal of the Royal Statistical Society-A* **141**, 242-252.

Table 1

Sample Size Adjustments for Error Degrees of Freedom

Name	Function of $\{N_{jj'}\}$
N_{*1}	$=$ number of complete cases
N_{*2}	$= \min(\{N_{jj'}\})$
N_{*3}	$= \min(\{N_{jj}\})$
N_{*4}	$=$ harmonic mean$(\{N_{jj'}\})$
N_{*5}	$=$ geometric mean$(\{N_{jj'}\})$
N_{*6}	$=$ arithmetic mean$(\{N_{jj'}\})$
N_{*7}	$=$ harmonic mean$(\{N_{jj}\})$
N_{*8}	$=$ geometric mean$(\{N_{jj}\})$
N_{*9}	$=$ arithmetic mean$(\{N_{jj}\})$
N_{*10}	$= \max(\{N_{jj}\})$
N_{*11}	$= N$

Table 2

Test Size for Mixed Model F (5000 Replications, $\pm$ 0.006)

N	$\rho_{jj'}$	σ_j^2	% Missing	$p = 3$	6
12	Low	$=$	0	.126	.58
12	Low	$\neq$	0	.134	.60
12	High	$\neq$	0	.125	.59
24	Low	$\neq$	0	.069	.16
24	High	$\neq$	0	.074	.16
12	Low	$=$	5	.182	
12	Low	$\neq$	5	.171	
12	High	$\neq$	5	.172	
24	Low	$\neq$	5	.080	.21
24	High	$\neq$	5	.081	.23
12	Low	$=$	10	.250	
12	Low	$\neq$	10	.244	
12	High	$\neq$	10	.262	
24	Low	$\neq$	10	.086	.303
24	High	$\neq$	10	.095	.322

Table 3

Test Size for GLMM F Tests (0% Missing, 5000 Replications, $\pm$ 0.006)

			F_W		F_U		F_V		F_{GG}	
N	$\rho_{jj'}$	σ_j^2	$p = 3$	6	3	6	3	6	3	6
12	Low	$=$	.050	.046	.048	.051	.044	.046	.027	.013
12	Low	$\neq$	.052	.054	.050	.055	.049	.059	.040	.025
12	High	$\neq$	.053	.053	.048	.053	.042	.045	.054	.058
24	Low	$\neq$	.049	.051	.050	.050	.048	.049	.041	.039
24	High	$\neq$	.053	.049	.051	.048	.051	.049	.049	.053

Table 4
Adjusted Degree of Freedom Test Size for F_W
(5%, 10% Missing, 5000 Replications, $\pm$ 0.006)

				N_{*2}		N_{*4}		N_{*11}	
N	$\rho_{jj'}$	σ_j^2	% Missing	$p=3$	6	3	6	3	6
12	Low	$=$	5	.029		.073		.145	
12	Low	$\neq$	5	.033		.072		.134	
12	High	$\neq$	5	.032		.074		.141	
24	Low	$\neq$	5	.030	.017	.049	.067	.087	.142
24	High	$\neq$	5	.031	.022	.053	.067	.089	.146
12	Low	$=$	10	.047		.148		.345	
12	Low	$\neq$	10	.042		.144		.335	
12	High	$\neq$	10	.051		.152		.354	
24	Low	$\neq$	10	.020	.078	.051	.171	.133	.379
24	High	$\neq$	10	.025	.094	.062	.199	.158	.411

Table 5

Adjusted Degree of Freedom Test Size for F_U

(5%, 10% Missing, 5000 Replications, $\pm$ 0.006)

				N_{*2}		N_{*4}		N_{*11}	
N	$\rho_{jj'}$	σ_j^2	% Missing	$p = 3$	6	3	6	3	6
12	Low	$=$	5	.032		.072		.142	
12	Low	$\neq$	5	.033		.067		.130	
12	High	$\neq$	5	.034		.074		.134	
24	Low	$\neq$	5	.028	.021	.048	.069	.084	.144
24	High	$\neq$	5	.031	.025	.052	.068	.088	.148
12	Low	$=$	10	.052		.163		.341	
12	Low	$\neq$	10	.047		.153		.333	
12	High	$\neq$	10	.055		.169		.352	
24	Low	$\neq$	10	.021	.093	.052	.184	.135	.379
24	High	$\neq$	10	.029	.086	.061	.217	.158	.417

Table 6

Adjusted Degree of Freedom Test Size for F_V

(5%, 10% Missing, 5000 Replications, $\pm$ 0.006)

			%	N_{*2}		N_{*4}		N_{*11}	
N	$\rho_{jj'}$	σ_j^2	Missing	$p = 3$	6	3	6	3	6
12	Low	$=$	5	.023		.052		.102	
12	Low	$\neq$	5	.021		.051		.099	
12	High	$\neq$	5	.022		.050		.096	
24	Low	$\neq$	5	.029	.015	.046	.054	.081	.125
24	High	$\neq$	5	.030	.017	.049	.052	.088	.128
12	Low	$=$	10	.011		.057		.206	
12	Low	$\neq$	10	.010		.051		.215	
12	High	$\neq$	10	.013		.058		.214	
24	Low	$\neq$	10	.017	.019	.044	.106	.127	.334
24	High	$\neq$	10	.020	.022	.055	.120	.148	.354

Table 7

Adjusted Degree of Freedom Test Size for F_{GG}

(5%, 10% Missing Data, 5000 Replications, $\pm$ 0.006)

				N_{*2}		N_{*9}		N_{*11}	
N	$\rho_{jj'}$	σ_j^2	% Missing	$p=3$	6	3	6	3	6
12	Low	$=$	5	.010		.035		.053	
12	Low	$\neq$	5	.019		.043		.059	
12	High	$\neq$	5	.026		.050		.066	
24	Low	$\neq$	5	.030	.010	.047	.037	.061	.057
24	High	$\neq$	5	.049	.018	.049	.041	.061	.053
12	Low	$=$	10	.002		.030		.073	
12	Low	$\neq$	10	.008		.038		.078	
12	High	$\neq$	10	.009		.040		.075	
24	Low	$\neq$	10	.018	.004	.051	.044	.091	.095
24	High	$\neq$	10	.017	.009	.048	.043	.076	.070

Table 8

Adjusted Degree of Freedom Test Size for Multivariate F Test
With $s = \min(a, b) = 1$ (5%, 10% Missing, 5000 Replications, $\pm$ 0.006)

				$a = 1, b = 3$			$a = 3, b = 1$		
			%	N_{*2}	N_{*4}	N_{*11}	N_{*2}	N_{*4}	N_{*11}
N	$\rho_{jj'}$	σ_j^2	Missing	$p = 3$	3	3	$p = 3$	3	3
12	Low	$=$	5	.038	.072	.114	.031	.051	.078
12	Low	$\neq$	5	.034	.066	.105	.029	.048	.071
12	High	$\neq$	5	.039	.074	.110	.027	.042	.065
24	Low	$\neq$	5	.036	.052	.077	.036	.046	.067
24	High	$\neq$	5	.037	.050	.078	.030	.043	.064
12	Low	$=$	10	.058	.132	.253	.018	.043	.096
12	Low	$\neq$	10	.054	.131	.251	.016	.040	.010
12	High	$\neq$	10	.056	.137	.258	.014	.036	.088
24	Low	$\neq$	10	.028	.052	.102	.026	.043	.089
24	High	$\neq$	10	.028	.048	.104	.018	.036	.073

Table 9

Choline Measurements Over 5-Week Period in Male Subjects

Treatment	Day					
	0	7	14	21	28	35
Control	9.93	12.29	9.30	9.51	10.84	9.24
	9.77	8.14	11.43	9.44	11.10	10.56
	12.56	10.90	11.19	12.31	9.95	.
	10.15	10.32	8.86	9.23	8.56	12.78
	11.00	9.20	8.78	9.37	7.54	12.39
	10.46	8.72	8.13	8.14	11.76	9.74
Deficient	12.15	9.52	9.05	9.07	6.76	9.39
	12.88	9.66	7.71	7.29	6.37	10.61
	7.94	9.86	7.87	8.89	8.69	12.28
	9.42	12.82	7.17	8.18	8.30	12.61
	9.57	10.95	9.01	8.98	6.56	9.66
	11.54	10.43	8.66	8.60	7.87	9.69
	11.65	10.64	9.81	8.04	7.52	8.76
	8.73	8.08	7.70	6.44	6.42	8.93