GLMMs as a Tool for Simultaneous Study of the Structure of the Covariance Matrix and Experimental Design Effects

K.E. Muller, J.D. Hosking, and R.W. Helms

University of North Carolina, Chapel Hill

Conventional Residual Analysis

When considering the general linear model, the phrase "error covariance matrix" calls to mind a somewhat distasteful and unwieldy collection of mistakes. Even "residual covariance matrix" would seem to be something left over. Be that as it may, this paper has been based on the assumption that in many cases, residuals from a linear model, and particularly the residual covariance matrix, may be of much more interest as, or more interest than, the model itself. Both the analysis of residuals and their covariances will be discussed, first in order to motivate the consideration of the possibilities, second to indicate the many convenient ways of doing so with GLMMs, and finally to sketch the capabilities of existing modules in GLMM5 designed to analyze residuals.

First consider the analysis of residuals about a general linear multivariate model. We shall use the notation defined in the companion paper (Helms, Hosking, and Christiansen, 1979). The model expression is

$$Y = X\beta + \epsilon,$$

where $Y$ is a $N$ x $1$ vector of responses, $X$ is a $N$ x $q$ matrix of independent variables, $\beta$ is a $q$ x $1$ vector of regression coefficients, and $\epsilon$ is a $N$ x $1$ vector of random errors. Here $\epsilon$ is an unobservable matrix of deviations about the model. After fitting the model the matrix of residuals is $R = Y - \hat{X}\hat{\beta}$, where, of course, $\hat{E}$ is the least squares estimate of $\beta$. The most natural application of residual analysis arises when one questions whether the assumptions of the linear model analysis have been met. Least squares estimates are notoriously sensitive to certain types of failures of the model's assumptions. Scattergrams of residuals by dependent variable scores, univariate outlier analysis via PROC MEANS, PROC DATACHK, PROC PLOT, OR PROC SPLOTO can often be profitably employed. All of these procedures are based on well-known statistical methods and any number of good things can fall out. Data entry errors may be trapped, which can occasionally turn an analysis upside down. Transformations may be indicated by systematic heteroscedasticity (e.g., perhaps logarithms of the dependent variable should be analyzed). With time series data, tests of stationarity may be appropriate. Of course, all of this rather unexciting work simply indicates that you have a spiffy model and clean data approximately analyzed by the least squares and normal theory methods in GLMMs.

An Example

In order to introduce some more complex possibilities for analyzing residuals, consider the following more complicated hypothetical situation, a human engineering task. A company wants to evaluate effects of using different soldering irons and different techniques on speed of assembly of circuit boards. Say that two types of soldering iron and three techniques were studied in a factorial design, with roughly 100 to 200 subjects in each of the six cells (soldering iron by technique combinations). Each subject was asked to assemble one of each of the 27 different circuit board types that IBM Company produces in the course of their work. Roughly 70% of the subjects were females, and 30% males. The Itty Bitty personnel people have a list of questions: 1) which soldering iron is best? 2) which technique is best? 3) are the soldering irons easier to use with certain techniques? Their statistician translates that into a multivariate analysis of variance, with 27 dependent variables and a cell mean model. Previous experience indicates that logarithms of the response times should be analyzed. The main effect and secondary parameters will be tested using GLMMs, naturally. These are standard GLMM modeling issues addressed in the companion papers.

Additional questions exist. How do circuit boards vary? The company assumes the type of circuit board affects variability, but has no reasonable classification scheme. We think you will agree that we have now reached the arts and crafts stage of data analysis. We would like to suggest one reasonable approach, and raise the possibilities of others. Consider principal components analysis of the residual covariance matrix, $\Sigma$ which is an estimate of the population covariance matrix of each row of $X$. Recall that the first principal component of $\Sigma$ is that linear combination of variables which has the maximum variance, subject to a normalization constraint. See, for example, Morrison (1967) for a description of component analysis methods. By studying the component weights, which indicate each circuit board's contribution, and/or the correlations of the circuit boards with the components, a useful and interpretable description of the structure of the between and within circuit board variability may be derived. Please note that whatever type of analysis one chooses to do here should almost surely begin by removing the effects of the experimental treatments -- before looking at the structure of the residual covariance matrix. This suggestion contrasts with classical principal components analysis in which the estimate of the covariance matrix is usually adjusted for an overall mean, not for the known experimental design. The former is a special case of the latter. This situation is just one example with one technique. In general we say that the structure of a linear model, although perhaps the original focus of a study, has traditionally obstructed the analyst's attempts to consider covariance structure and to answer other questions addressed to the data.
Residual Analysis with GLMMS

In the context of GLMMS a number of alternate paths exist for accessing the residuals. The most natural, of course, is to compute $R = Y - X\beta$ and store the result in a SAS dataset. Then any procedure in SAS and in turn, in any other packages can itself be used to conveniently do many analyses not available otherwise. Unfortunately, this approach requires being able to store the $X$ and $Y$ matrices in core which is not always possible. Figure 1 outlines available strategies with residuals with GLMMS. We just mentioned the first alternative. A simple variation would be to bring the matrices in core in pieces rather than all at once. DATA steps preceding the GLMMS step and following would be required. The second alternative would be compute residuals through the use of compute statements in a DATA step. The third alternative would be to use PROC SCORE to compute the residuals. These three approaches apply in many more situations than just GLMMS. The last two methods listed invoke modules which are part of GLMMS. The remainder of this paper will describe these features.

Before jumping into the nitty-gritty of modules, however, note that they are not intended to replace SAS procedures or programs in other statistical packages. As stated earlier, we recommend the use of GLMMS when its power is required or its generality leads to greater rather than lesser convenience.

The PUTOUT Module

We bestowed the name "PUTOUT" on the module of most general use. PUTOUT has only three options: 1) produce matrices needed to create a TYPE=COR dataset, 2) produce matrices needed to create a TYPE=COV dataset, and 3) produce the matrices needed to create a TYPE=SSCP dataset. Each matrix is based on residuals from a model fit using GLMMS. The user receives matrices from PUTOUT within PROC MATRIX and creates the dataset with two OUTPUT statements and one subsequent simple DATA step. The matrices contain all observations needed for the type of dataset specified. The resulting datasets conform to SAS standards as we understand them and therefore allow convenient access to many SAS procedures such as PROC FACTOR. The existence of PUTOUT saves the user who needs such a dataset a good bit of nuts and bolts programming. In a companion paper, Hosking and Helms (1979) discuss the tremendous potential for core savings with such an approach. PUTOUT is quite easy to use since only seven extra SAS statements are required to invoke it and to create the dataset.

The PRINCOV Module

Figure 2 outlines the characteristics of module PRINCOV. The module computes a principal components analysis based on the residual covariance matrix computed by the LINMOD modules. Note that PUTOUT combined with PROC FACTOR may be used to perform components analysis based on the correlation matrix, and that, presently, PROC FACTOR does not work on covariance matrices. PRINCOV could very neatly be used in the Itty Bitty Machine example. A components analysis finds that ordered orthogonal set of linear combinations each of which has maximum variance, given the preceding ones. A study of which circuit boards define each factor could have substantial help in interpreting its results. Figure 3 lists outputs available. The eigenvalues are the variances for each linear combination, eigenvectors are the weights applied to the variables to form the combinations, and compute the covariances of each variable with each component (the last are the factors of the matrix). Naturally, all of these standard outputs are provided. The module also provides a significance test for the number of factors and also provides confidence bounds on each eigenvalue. The correlations of variables and small analysis of the goodness of fit of the factorization are also available. As with all modules in GLMMS, all of this output, if asked for, remains available as matrices in PROC MATRIX and hence can be manipulated by the user. In addition, a matrix is produced which can be plotted by PROC PLOT to display the component pattern. Please do not try to rig up a TYPE=FACTOR dataset from PRINCOV output, since the result is nonsense. The user should be aware of the important differences between factor analysis and components analysis. For example, one may not wish to do any rotation of components, in distinct contrast to factors. As with residuals, MATRICES, FACTOR statements, and so forth on DATA statements may be used to compute component scores.

The CANCORR Module

The final module to be discussed, CANCORR, performs a canonical correlation analysis on the residual covariance/correlation matrix. Canonical correlation attempts to assess the strength of relation between two sets of variables, in contrast to more common multiple regression which assesses strength of relation between one variable and a set. The latter is a special case of the former. SAS72 has a CANCORR procedure, and SAS76 will soon have one. The GLMMS CANCORR module exists for two reasons. Most importantly, it allows extremely convenient analysis of linear model residuals. Second, it provides extensive output needed, in the authors' opinion, to sort out such an analysis. Figure 4 briefly summarizes the outputs. Traditional output consists of canonical and co-variance matrices, significance tests and both standardized and non-standardized canonical
weights. The other outputs are less common. The canonical factor structure is simply the matrix of correlations between the variables and the canonical variates. The redundancy analysis is a collection of summary statistics devised by Stewart and Love (1967) specifically designed to aid interpretation of canonical analysis. Despite the fact that some controversy surrounds it, we expect it to slowly seep into wider use as more information about its properties is gathered. The module will also provide a cornucopia of multiple R's: each variable with all others in its set and each variable with all variables in the other set. Finally, the regression weights for predicting each set from the other are provided. All output is retained as matrices in PROC MATRIX for the user to treat further, if desired. Canonical scores may be computed either within PROC MATRIX, PROC SCORE, or DATA steps. Canonical correlation is frequently difficult to interpret, but when needed it can be very helpful.

Summary

We hope to leave you with two ideas from this paper. First we hope you have considered the possibilities in analyzing linear model residuals and residual covariance matrices. Second, we hope you have grasped the powerful facilities GLMM5 has available for helping you do so. Remember, simply because a covariance is a second order moment, it need not be treated as second rate, at least with GLMM in town.

Figure 1

PATHS TO ANALYZING RESIDUALS WITH GLMM

1. Direct computation of residuals in PROC MATRIX
   a. All at once, \( R = Y - X \hat{\beta} \)
   b. Chunk-wise, several rows of \( R \) in each chunk

2. Output Beta matrix, use DATA step to compute
   \[ z_{i,j} = x_{i,j} - \sum (x_{i,j} \hat{\beta}_{k,j}) \]

3. Output Beta matrix, use PROC SCORE

4. Use PUTOUT model
   a. Produce TYPE=COR, residual correlation matrix
   b. Produce TYPE=COV, residual correlation matrix
   c. Produce TYPE=SSCP, residual sums of squares

5. Use analysis modules
   a. PRINCOV, principal components on residual covariance
   b. CANCORN, canonical correlation analysis

Figure 2

PUTOUT MODULE CHARACTERISTICS

1. Allows convenient saving of residual matrix in compact form for later analysis

2. Options/uses
   a. Produce TYPE=COR matrices
   b. Produce TYPE=COV matrices
   c. Produce TYPE=SSCP matrices

3. Works on any subset of variables

4. Reduces detailed user programming in PROC MATRIX

5. Can save tremendous amounts of storage

6. Gives access to large range of procedures

Figure 3

MODULE PRINCOV CHARACTERISTICS

1. Performs principal components analysis of residual covariance matrix. Components are orthogonal set of linear combinations of residuals, each with maximum variance

2. Features
   a. Test last \( m \) roots equal (number of factors test)
   b. Provides confidence bounds on eigenvalues
   c. Provides eigenvalues and eigenvectors
   d. Provides component matrix
   e. Provides correlations of variables with components
   f. Provides matrix suitable for plotting loading with PROC PLOT
   g. Provides evaluation of discrepancy \( (E - EE') \)

3. Cautions
   a. Be certain components on covariance appropriate
   b. Don't try to force output into PROC FACTOR
   c. Recognize limitations in applications of components
FIGURE 4

MODULE CANCORR CHARACTERISTICS

1. Performs canonical correlation based on residual covariance/correlation matrix

2. Features
   a. Easy to use in PROC MATRIX/GLMM setting
   b. Convenient for analyzing linear model residuals
   c. Provides correlation and covariance matrices
   d. Provides significance tests for number of variates
   e. Provides canonical weights
   f. Provides canonical factor structure
   g. Provides redundancy analysis
   h. Provides multiple R's within and between sets
   i. Provides between set regression weights