Paper 198-30

# Guidelines for Selecting the Covariance Structure in Mixed Model Analysis

Chuck Kincaid, COMSYS Information Technology Services, Inc., Portage, MI

## INTRODUCTION

Mixed Models is rapidly becoming a very useful tool for statisticians. As a general paradigm it can be used to handle almost every situation, especially if you extend the Linear Mixed Model to the Generalized Linear Mixed Model case or the Nonlinear Mixed Model case. It's also an area in which a lot of research is being done, because the questions are far from being answered. Advanced computing power is giving us the capability to answer those questions. One important question which, unfortunately, still has no good answer is how to select the covariance structure. This paper is an attempt to survey the information available for answering the question.

## SITUATIONS

Mixed Models, i.e. models with both fixed and random effects arise in a variety of research situations. Split plots, strip plots, repeated measures, multi-site clinical trials, hierarchical linear models, random coefficients, analysis of covariance are all special cases of the mixed model. The question of selecting the covariance structure changes with each case, as it does when you throw in missing values or missing treatment combinations. For this paper we will stick to the repeated measures situation with no missing values. For example, suppose we are testing the efficacy of a new drug. We have two groups, treatment and control, and we are taking multiple measurements on each person in the two groups.

## BASICS OF MIXED MODEL

### NOTATION

The Linear Mixed Model (LMM) is a generalization of the Linear Model (LM) and is represented in its most general fashion as

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i$$

where $\vec{X}_i$ and $\vec{Z}_i$ are the fixed and random design matrices, respectively, $\vec{\beta}$ is a vector of unknown fixed effects, $\vec{\gamma}_i$ is a vector of unknown random effects and $\vec{\varepsilon}_i$ is the unknown random error. $\vec{\beta}$ represents parameters that are the same for all subjects; $\vec{\gamma}$ represents parameters that are allowed to vary over subjects. Milliken [] is an excellent source for learning how to determine the appropriate terms for the given design and treatment structures.

We assume that the random effects are normally distributed with

$$E\begin{bmatrix}\vec{\gamma}\\ \vec{\varepsilon}\end{bmatrix} = \begin{bmatrix}\vec{0}\\ \vec{0}\end{bmatrix} \qquad Var\begin{bmatrix}\vec{\gamma}\\ \vec{\varepsilon}\end{bmatrix} = \begin{bmatrix}\vec{G} & \vec{0}\\ \vec{0} & \vec{R}\end{bmatrix}$$

Given these assumptions, the variance of $\vec{y}$, which is the reason we're all here, is $\vec{V} = \vec{Z}\vec{G}\vec{Z}^t + \vec{R}$. We fit the random portion of the model by specifying the terms that define the random design matrix $\vec{Z}$ and specifying the structures of covariance matrices $\vec{G}$ and $\vec{R}$.

### MEANING

The random effects, as stated above, are allowed to vary over subjects. Another way to think of them is as subject-specific regression coefficients that reflect the natural heterogeneity in the population. Suppose site is a random effect. Then the effect of a particular site on the response, $\gamma_i$, is different for each site. The relationship among the effects of all of the sites is, we assume, described by a Normal distribution with mean 0 and variance, say, $\sigma_S^2$.

The repeated measurements could be repeated either as multiple measurements taken on the same experimental unit at the same time or a single measurement taken on the same experimental unit at multiple times or a combination of the two. Moser provides an excellent overview of fitting both types of measurements.

In this case suppose $\varepsilon_i = \{\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \mathrm{K}, \varepsilon_{im}\}$ is a vector of measurements taken at *m* equally spaced time points. The measurements each come from a normal distribution with covariance matrix

$$R_i = Var\left(\underset{\rightarrow}{\varepsilon_i}\right) = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \Lambda & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \Lambda & \sigma_{2m} \\ M & M & O & M \\ \sigma_{m1} & \sigma_{m2} & \Lambda & \sigma_{mm} \end{bmatrix}$$

Since they are all taken on the same experimental unit the measurements are correlated with each other.

We note that, because the variance of $\underset{\rightarrow}{y}$ is made up of two components, $\underset{\rightarrow}{Z}\underset{\rightarrow}{G}\underset{\rightarrow}{Z}'$ and $\underset{\rightarrow}{R}$, we could model the structure in $\underset{\rightarrow}{G}$ or $\underset{\rightarrow}{R}$ or both.

## DIFFERENT COVARIANCE STRUCTURES

The table below lists the simpler covariance structures that can be modeled in SAS via PROC MIXED. Each of these can be described in a fairly intuitive manner, though as we'll see they can be very similar to one another.

| Structure | Description | # of Parameters | {i,j}th element |
|---|---|---|---|
| AR(1) | Autoregressive(1) | 2 | $\sigma_{ij} = \sigma^2 \rho^{\lvert i-j \rvert}$ |
| CS | Compound Symmetry | 2 | $\sigma_{ij} = \sigma_1 + \sigma^2 1(i = j)$ |
| UN | Unstructured | t(t+1)/2 | $\sigma_{ij} = \sigma_{ij}$ |
| TOEP | Toeplitz | t | $\sigma_{ij} = \sigma_{\lvert i-j \rvert +1}$ |
| VC | Variance Components | q | $\sigma_{ij} = \sigma_k^2 1(i = j)$ and $i$ corresponds to the *k*th effect |
| ARH(1) | Heterogeneous AR(1) | t+1 | $\sigma_{ij} = \sigma_i \sigma_j \rho^{\lvert i-j \rvert}$ |
| CSH | Heterogeneous CS | t+1 | $\sigma_{ij} = \sigma_i \sigma_j \left[\rho 1(i \neq j) + 1(i = j)\right]$ |
| TOEPH | Heterogeneous TOEP | 2t-1 | $\sigma_{ij} = \sigma_i \sigma_j \rho_{\lvert i-j \rvert}$ |

### VARIANCE COMPONENTS

The VC structure is the standard variance components and is the default.

$$\begin{bmatrix} \sigma_A^2 & 0 & 0 & 0 \\ 0 & \sigma_B^2 & 0 & 0 \\ 0 & 0 & \sigma_{AB}^2 & 0 \\ 0 & 0 & 0 & \sigma_{AB}^2 \end{bmatrix}$$

### AUTOREGRESSIVE(1)

The AR(1) structure has homogeneous variances and correlations that decline exponentially with distance. In our case this means that the variability in a measurement, say white blood cell count, is constant regardless of when you measure it. It also means that two measurements that are right next to each other in time are going to be pretty correlated (depending on the value of $\rho$), but that as measurements get farther and farther apart they are less correlated.

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

### COMPOUND SYMMETRY

The CS structure is the well-known compound symmetry structure required for split-plot designs "in the old days". As can be seen in the table, the variances are homogeneous. There is a correlation between two separate measurements, but it is assumed that the correlation is constant regardless of how far apart the measurements are.

$$\begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{bmatrix}$$

### UNSTRUCTURED

The UN structure is the most “liberal” of all allowing every term to be different. It requires fitting the most parameters of any structure, t(t+1)/2.

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

### TOEPLITZ

The TOEP structure is similar to the AR(1) in that all measurements next to each other have the same correlation, measurements two apart have the same correlation different from the first, measurements three apart have the same correlation different from the first two, etc. However, the correlations do not necessarily have the same pattern as in the AR(1). Technically, the AR(1) is a special case of the Toeplitz.

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

### HETEROGENEOUS VERSIONS OF THE ABOVE

The heterogeneous versions of the covariance structures above are a simple extension. That is the variances, along the diagonal of the matrix, do not have to be the same. Note that this adds more parameters to be estimated, one for every measurement.

## SPECIFYING THE COVARIANCE STRUCTURE

### PROC MIXED NOTATION

A lot of the notation for MIXED is similar to what is in GLM, but often the meaning is different. There are two ways to specify a covariance structure in PROC MIXED, the RANDOM statement and the REPEATED statement. The former specifies the structure for the $\underset{\rightarrow}{G}$ matrix and the latter for the $\underset{\rightarrow}{R}$ matrix. The RANDOM statement imposes a particular covariance structure on the random effect terms. These could be the larger experimental units. You can have multiple RANDOM statements in one model. Effects in the same RANDOM statement might be correlated, but independent in different RANDOM statements.

The SAS 9 documentation explains that the REPEATED statement is used to specify covariance structures for repeated measurements on subjects or, another way, is that the REPEATED statement controls the covariance structure of the residuals. Similar syntax is used for both.

From the SAS Help Files we have

```
        RANDOM random-effects < / options > ;
        REPEATED < repeated-effect >< / options > ;
```

The random-effects can be continuous or classification variables and multiple RANDOM statements can be used at the same time. A lot of times you do not have to specify a repeated effect, however, when you do only one REPEATED statement may be used at a time.

Some of the primary options for specifying the structure of the covariance matrix are below. The other options have mostly to do with tests or displaying matrices and the like.

| | |
|---|---|
| TYPE=covariance-structure | specifies the covariance structure of $\underset{\rightarrow}{G}$ or $\underset{\rightarrow}{R}$. TYPE=VC (variance components) is the default and it models a different variance component for each random effect or repeated effect. |
| SUBJECT=effect | identifies the subjects in your mixed model. Complete independence is assumed across subjects; thus, for the RANDOM (REPEATED) statement, |

| | |
|---|---|
| | the SUBJECT= option produces a block-diagonal structure in $\underrightarrow{G}$ ($\underrightarrow{R}$) with identical blocks.<br>With the RANDOM statement the $\underrightarrow{Z}$ matrix is modified to accommodate the block-diagonality specified by the SUBJECT option. In fact, specifying a subject effect is equivalent to nesting all other effects in the RANDOM statement within the subject effect. |
| GROUP=effect | Allows the user to change the covariance parameters from one group to another. This can greatly increase the number of covariance parameters needing to be estimated. |

Our goal is to select the covariance structure of the random effects. The first step, then, is to know what random effects we are modeling. Milliken or Milliken and Johnson can be very helpful in determining the treatment structure and the design structure. These structures will identify the different sizes of experimental units which typically correspond to the random design effects.

You will use the RANDOM statement for random effects that are not at the lowest level of experimental unit. The REPEATED statement will be used to specify the correlation among the experimental errors.

### EXAMPLES

Suppose a researcher wants to study the effect of three levels of drug A and four levels of drug B on bacteria growth in his laboratory. He has three batches of blood serum each of which is partitioned into 12 parts. Each treatment combination is assigned at random to one part in each batch. The treatment structure is a two-way factorial arrangement with 12 treatment combinations. The design structure is a randomized complete block design. Thus BATCH is a random effect we want to include in our model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + B_k + e_{ijk}$$

where

$y_{ijk}$ is the observation

$\mu$ is the overall mean

$\alpha_i$ is the drug A effect

$\beta_j$ is the drug B effect

$B_k$ is the batch effect

$e_{ijk}$ is the random error

We assume $B_k \sim N\left(0, \sigma_B^2\right)$ and $e_{ijk} \sim N\left(0, \sigma^2\right)$ and are independent of each other. The SAS code would be

```
proc mixed data=lab;
     class batch A B;
     model growth = A B;
     random batch;
run;
```

This code uses the default variance component (VC) structure give us an estimate of $\sigma_B^2$. Because of our assumption regarding the distribution of the errors we do not need to specify a REPEATED statement.

For the second example, consider a multicenter trial to compare 4 treatments. [Littell, Milliken, Stroup, Wolfinger] The treatments were observed at each of 9 centers and at each center a RCB design with 3 blocks was used. The model we'll use is

$$y_{ijk} = \mu + \tau_i + L_j + R(L)_{jk} + (\tau L)_{ij} + e_{ijk}$$

where

$y_{ijk}$ is the observation

$\mu$ is the overall mean

$\tau_i$ is the treatment effect

$L_j$ is the random Location effect, $\sim N\left(0, \sigma_L^2\right)$

$R(L)_{jk}$ is the effect of block within location, $\sim N\left(0, \sigma_R^2\right)$

$\left(\tau L\right)_{ij}$ is the treatment by location effect, $\sim N\left(0, \sigma_T^2\right)$

$e_{ijk}$ is the random error, $\sim N\left(0, \sigma^2\right)$

The treatment structure is a one-way arrangement. The design structure is a randomized block design. Both the location and the block within location are block effects. The SAS code we'll use to fit the data is the following.

```
        proc mixed;
                class loc block trt;
                model resp = trt / ddfm=satterth;
                random loc block(loc) loc*trt;
        run;
```

Again, we are using the default variance components structure for the covariance and it's not necessary to use a REPEATED statement.

For the second example, we consider the following repeated measures data taken from Littell, Milliken, Stroup, and Wolfinger. Subjects in an exercise therapy study were assigned to one of three weightlifting programs. Strengths of the subjects were measured every other day for two weeks following the beginning of the study.

$$y_{ijk} = \mu + \alpha_i + d_{ij} + \tau_k + (\alpha\tau)_{ik} + e_{ijk}$$

where

$y_{ijk}$ is the observation

$\mu$ is the overall mean

$\alpha_i$ is the program effect, $\sim N\left(0, \sigma_L^2\right)$

$\tau_k$ is the time effect

$\left(\alpha\tau\right)_{ik}$ is the treatment by location effect

$d_k$ is the random subject effect, $\sim MVN\left(\underset{\rightarrow}{0}, \sigma_S^2 \underset{\rightarrow}{I}\right)$

$e_{ijk}$ is the random error, $\sim MVN\left(\underset{\rightarrow}{0}, \sigma_T^2 \underset{\rightarrow}{I}\right)$

This model places the compound symmetric structure on the covariance matrix of $\underset{\rightarrow}{Y}$. There are actually three ways in SAS of specifying this same model. They are

```
   /*  Using variance components. Covariance structure combination from both the
   random effects G and the error R .  */

       proc mixed data=weight2;
          class program subj time;
          model strength = program time program*time;
          random subj(program);
       run;

   /*  Same as above, but explicitly states Compound Symmetry. Has no real
   value, because fits duplicate parameters.   */

       proc mixed data=weight2;
          class program subj time;
          model strength = program time program*time;
          random subj(program) / type=cs;
       run;
```

```
/*  Covariance structure entirely in $\underset{\rightarrow}{R}$ .    */

    proc mixed data=weight2;
       class program subj time;
       model strength = program time program*time;
       repeated / type=cs sub=subj(program) r rcorr;
    run;
```

The REPEATED statement here allows for more flexibility in fitting the data. We could, as Littell, et al do, fit an autoregressive or unstructured covariance to the data. By selecting SUBJ(PROGRAM) as the subject, we are saying that the compound symmetric structure pertains to the subject.

## SELECTING THE COVARIANCE STRUCTURE

### STRATEGIES

Once you have the random effects determined, then you can move on to selecting the covariance structure. There are a variety of considerations when selecting the covariance structure. They include the number of parameters, the interpretation of the structure, diagnostic results, and effects on fixed effects.

### BY PARSIMONY

If the data suffices, one could always fit the unstructured covariance structure and go with it. However, just as in traditional regression we want to have as few parameters in the model as possible. The more data you have the more parameters you can fit, but they do not always add to our knowledge and often take away. The more complex the model the more specific to the data it will be and the less generalizable. Our belief that Nature follows simple, elegant rules leads us to look for the simpler model when possible. From the fixed effects perspective, selecting a structure that is too simple increases the fixed effects Type I error rate, and selecting a structure that is too complex sacrifices power and efficiency. [Hallahan]

A small side note. There are times when one could fit duplicate parameters. This *might be* indicated by getting the error message “Convergence criteria met but final Hessian is not positive definite.” This can happen regardless of the number parameters in the model. Thinking through the effects of the statements you’ve specified will help you to understand the model that you are fitting better and avoid errors like above. Note that the error could occur for other reasons.

### BY MEANING

There are many more covariance structures than those listed in this paper. It is not recommended to fit all possible structures and take the one that seems to fit best. Using your understanding of the design and treatment structures and the meaning of the covariance structures will usually give you a few candidate structures to work with. Therefore, even though choosing the candidate structures before the data is collected, i.e. as part of the SAP, may be required in a clinical trial, it is definitely the best method, since it avoids allowing chance to drive the covariance estimates.

### BY IC

Specifying the IC option on the PROC statement gives three default Information Criteria, AIC = Akaike’s Information Criteria, AICC = AIC Corrected, and BIC = Bayesian Information Criteria. These statistics are functions of the log likelihood and can be compared across models, if you keep the fixed effects part of the model constant. SAS provides them in the smaller is better format. One could fit models with competing covariance structures and compare the IC. They should give you comparable results and you can use the democratic rule to determine the best model. If they give conflicting results, then the simpler model is probably better. Unfortunately, though they seem to help, the information criteria are not guaranteed to lead you to the correct structure.

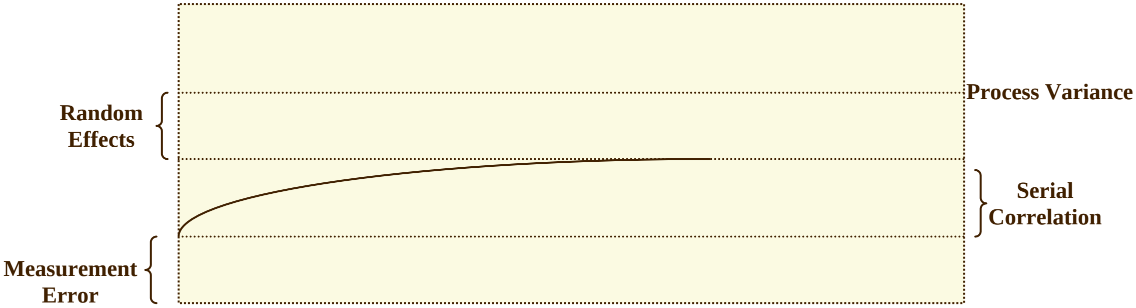
Keselman, Algina, Kowalchuk, and Wolfinger 1998 investigated using the AIC and BIC for various conditions (e.g., covariance heterogeneity, nonnormality, unequal group sizes) and found that they performed rather poorly. The AIC selected the correct structure only 47 percent of the time on average and the BIC only 35 percent of the time. An examination of using these two criteria to fit a growth curve model, under simpler conditions, was made by Ferron, et al. They found the success rate for the AIC to be 70 percent and 45 percent for the BIC.

### GRAPHICALLY

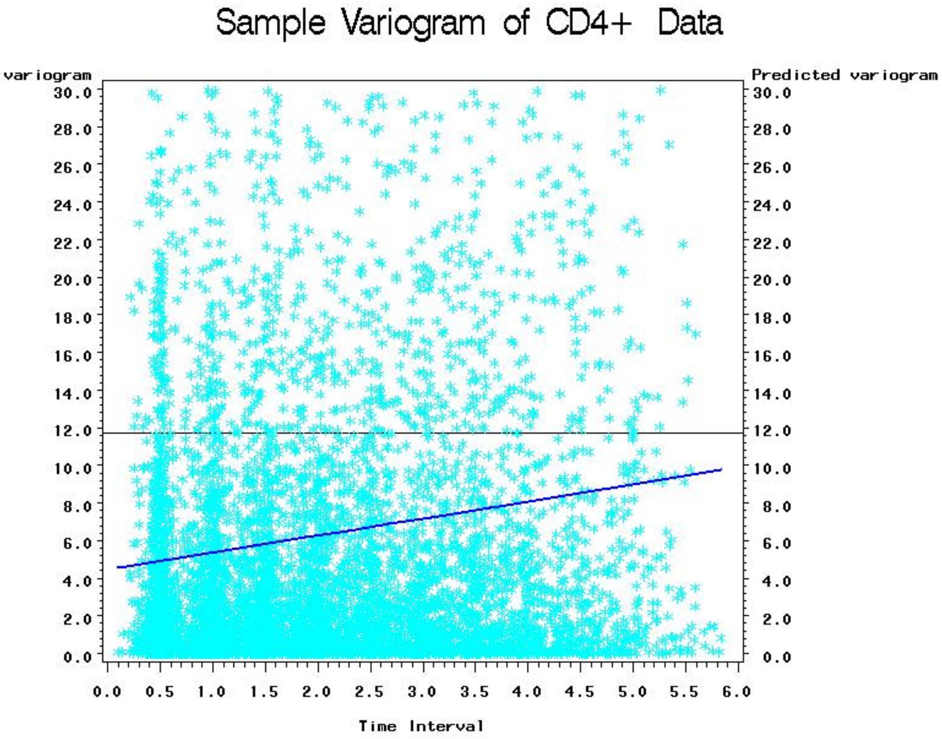
For us visual learners, there are a few graphical techniques that have been developed. Littell recommends to fit the data with an unstructured covariance matrix asking for the residual correlations or covariances. Plotting the covariances separately for each starting time can provide diagnostic information. That is, plot lag 1 covariance, lag 2 covariance, lag 3 covariance, etc. for errors starting at time 0. Do the same for errors starting at time 1 and so on. Then, if one sees linearly declining covariances with increasing lags one might fit an AR1 structure. And if the lines overlay each other, then a constant variance would be appropriate, otherwise the “H” or heterogeneous version of the structure would be best.

Hallahan suggests analyzing the OLS residuals with the sample variogram to determine the sources of variability.

Three error sources, Random Effects, Serial Correlation, and Measurement Error, can be broken out with this graphical tool. His description of using the tool is spelled out in the reference along with macros to compute and display the sample variogram. A brief (!) summary is given here. The 'theoretical' variogram can be drawn as



His example of fitting CD4+ data gives the following sample variogram.



The graph shows that all three sources of error can be found in this data. Since the smoothed line (from PROC LOESS) does not approach zero there is measurement error. Since the line is sloped upwards serial correlation exists. And since the line does not approach the process variance, random effects are present.

## EFFECT OF DIFFERENT CHOICES

### BIAS IN FIXED EFFECTS

Keselman, et al found that the F-tests were prone to inflated Type I Error rates in the conditions they investigated. Ferron, et al, investigating the simple case of balanced growth curves did not find bias in the fixed effects results. The macro they provide can be modified to investigate other scenarios.

One part of specifying the covariance structure that can affect the computational requirements is the subject. It's better to specify the subject than to not specify the subject. It's also better to sort the observations by the subject. If you sort a continuous subject variable then you don't have to put the variable in the CLASS statement. This saves memory and time.

## CONCLUSION

Working with mixed models is an exciting area and the computational power we have available to us today is allowing statisticians to do some great work. However, there is still a lot of work to do in the area. There is still not a definitive method for determining the best covariance structure, nor even a set of definitive methods given certain conditions. This paper tried to describe the framework of the random portion of the mixed model emphasizing the important statistical concept of understanding the experiment. Starting from the concepts, you do the best you can with the statistical and graphical techniques available. In the standard conditions you should be fine.

## REFERENCES

- Milliken, George A., How to be Successful in Implementing PROC MIXED, ASA Traveling Course
- Littell, et al. (2002) SAS for Linear Models maintained at http://faculty.ucr.edu/~hanneman/linear_models/index.html
- McCulloch and Searle, Generalized, Linear, and Mixed Models, Wiley, 2001
- Goldstein, Multilevel Statistical Models, Arnold, 1995
- Littell, Milliken, Stroup and Wolfinger, SAS System for Mixed Models, SAS Institute, 1996
- Verbeke and Molenberghs, Linear Models for Longitudinal Data, Springer, 2000
- Keselman, Algina, Kowalchuk, and Wolfinger, A Comparison of Two Approaches For Selecting Covariance Structures in The Analysis of Repeated Measurements, 1998, http://home.cc.umanitoba.ca/~kesel/cis1998.pdf
- Keselman, Algina, and Kowalchuk, Graphical Procedures, SAS' PROC MIXED, and Tests of Repeated Measures Effects, 2000, http://home.cc.umanitoba.ca/~kesel/as2000.pdf
- The MIXED Procedure (SAS/STAT Software) http://support.sas.com/techsup/faq/stat_proc/mixeproc.html
- McLerran, Re: REPATED/TYPE=covariance-structure in Proc Mixed®, SAS-L Discussion, March 2000, week 3 (#309).
- Milliken and Johnson, Analysis of Messy Data Volume I: Designed Experiments, New York, Chapman & Hall, 1989.
- Schabenberger, Mixed Model Tools in SAS/Stat®, ASA Statistical Consulting Section Roundtable Conference Call, September, 2004.
- Bland and Altman, Correlation, Regression and Repeated Data, British Medical Journal, 308, 1994.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.  Contact the author at:

Chuck Kincaid<br>
Director, SAS Center of Excellence<br>
COMSYS Information Technology Services, Inc.<br>
5278 Lovers Lane<br>
Portage MI 49002<br>
Work Phone: (269) 344-4100<br>
Fax: (269) 344-6849<br>
Email: ckincaid@comsys.com<br>
Web: www.comsyssas.com