108-2010

PROC MIXED: Underlying Ideas with Examples

David A. Dickey, NC State University, Raleigh, NC

ABSTRACT

PROC MIXED PROVIDES A SINGLE TOOL FOR ANALYZING A LARGE ARRAY OF MODELS USED IN STATISTICS, ESPECIALLY EXPERIMENTAL DESIGN, THROUGH THE USE OF REML ESTIMATION. A STRATEGY FOR IDENTIFYING MIXED MODELS IS FOLLOWED BY A DESCRIPTION OF REML ESTIMATION ALONG WITH A SIMPLE EXAMPLE THAT ILLUSTRATES ITS ADVANTAGES. A COMPARISON OF SOME OF THE AVAILABLE TESTS FOR VARIANCE COMPONENTS IS GIVEN ALONG WITH SEVERAL EXAMPLES, BOTH REAL AND ARTIFICIAL, THAT ILLUSTRATE THE VARIETY OF MODELS HANDLED BY PROC MIXED.

INTRODUCTION

Mixed models include a wide array of useful statistical approaches, some new and some quite old. SAS® PROC MIXED uses an estimation method similar to maximum likelihood called REML estimation. This is a relatively new method and with it comes some new looking output similar to the traditional analysis of variance table but with some added features that give useful information related to both traditional models and more interesting cases such as random coefficient models, panel data in economics, repeated measures (closely related to panel data) and spatial data. This paper attempts to provide the user with a better understanding of the ideas behind mixed models.

The first section of the paper explains the difference between random and fixed effects and gives a checklist for deciding which effects you have. Mixed models, as the name implies, can have some of each. The next section uses a simple experimental design, the randomized complete block, to investigate the differences between treating block effects as fixed and treating them as random, both in the presence of fixed treatment effects. It includes the definition and computation of so-called "BLUPs "and the intraclass correlation coefficient. The next section formalizes the general mixed model and reviews the concept of REML versus maximum likelihood (ML) estimation. ML (maximum likelihood) and REML are compared in the context of the randomized complete block design. Following this a discussion of several suggested tests for the presence of random effects is given along with a small Monte Carlo study comparing these in the context of a randomized complete block design. In the next section an unbalanced data set with random and fixed effects is shown and analyzed in both PROC GLM and PROC MIXED for comparison purposes. The paper ends with a random coefficient model using a study on activity levels in bears.

RANDOM OR FIXED?

Imagine a clinical trial involving doctors within hospitals. Each doctor has 3 patients with a certain disease and assigns then to drugs O (old) N (new) and C (control) at random, one patient for each drug. Now if there is a fourth drug, the researcher surely could not say anything about the performance of this new untested drug based on the results for the other three. On the other hand, the readers would be quite disappointed if the researcher found the new drug better than the old but then stated that this only holds for the 20 doctors (from 4 clinics) used in the study. Unless one of them is my doctor I have no interest in such a result. Nevertheless there may be a doctor effect so that the researcher needs to include doctor as a source of variation. One possibility is to imagine the doctors (and clinics) used as a random sample from a population of doctors (clinics) whose effects are normally and independently distributed with some doctors having effects less than average and some more than average. It would then be only the variation in the doctor effects that would be of interest and the reader would, as with any sample, assume that inference was for the population of doctors from which the researcher sampled. In this example drugs are fixed effects while doctors and clinics are random effects.

I present, below, a table that I use as a checklist to distinguish fixed versus random effects. Going through this checklist one thinks of doctors as being a random sample form a larger population, though often true randomization and sampling from a complete list of doctors is not actually done. The same applies to clinics. In contrast, the drugs were no doubt selected because they were the only items of interest.

	RANDOM	FIXED
Levels	selected at random from conceptually infinite collection of possibilities	finite number of possibilities
Another Experiment	would use different levels from same population	would use same levels of the factor
Goal	estimate variance components	estimate means
Inference	for all levels of the factor (i.e., for population from which levels are selected)*	only for levels actually used in the experiment*

(* an exception is when Y has a polynomial relationship to some X - usually X would be considered fixed even though the polynomial can predict at any X)

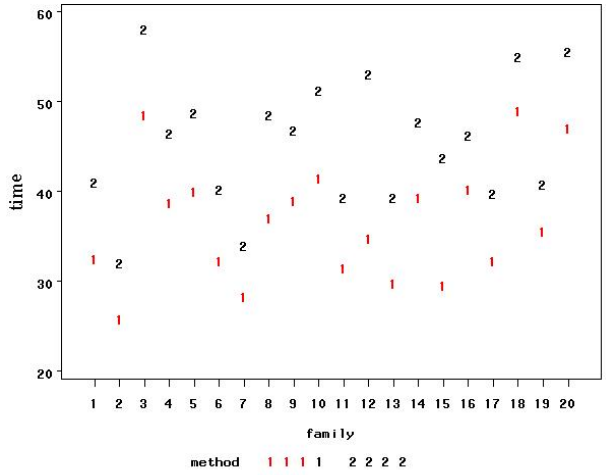
In row 2 of the table, suppose a researcher from another state saw the results and wanted to replicate the experiment. This new experiment would surely use the same drugs if it were a true check on the first experiment, but would likely use a different sample of clinics and doctors (from the same large population).

While the supervisor of the clinic may be interested in specific doctor means and an insurance company might be interested in clinic means, the nature of the experiment as described does not focus on means but rather simply admits that there are variance components for the doctor and clinic factors and estimates these variance components. On the other hand, direct comparison of drug means is surely of interest here. This illustrates the line of the above checklist labeled “Goal”.

Finally there is the issue of the scope of inference. As stated above there would be no thought that this experiment would inform the reader about an untested fourth drug, but surely the scope of inference for doctors and clinics should extend beyond just the 4 clinics and 5 doctors per clinic used.

Example 1: Using some made up data for illustration, here is a run with PROC MIXED. Here we look at twins from 20 families. We train one twin in SAS programming using method A and the other with method B. At the end of the training we give a programming task and record the time it takes to come up with a correctly running solution, this being our response variable TIME. This experiment can be thought of as a randomized complete block design with families serving as blocks, or equivalently, a paired t test with twins paired up by families. The treatment is the training method. Figure 1 is a plot with labels indicating training method, family number on the horizontal axis and programming time on the vertical axis.

Figure 1: Programming Times



Our data set needs variables FAMILY, TWIN, METHOD, and the response variable TIME. Here is the SAS program and part of the output:

```
PROC MIXED DATA=TWINS;
CLASS FAMILY METHOD;
MODEL TIME = METHOD;
RANDOM FAMILY;
```

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
family	21.2184
Residual	40.8338

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
method	1	19	9.60	0.0059

Our main goal was to compare the training methods. You see strong evidence that they differ based on the Type 3 F tests. The family variance component is 40.8 while the twin to twin (within family) variance component is 21.2. One might ask how much of such ability to learn is inherited. The variance components give a way to estimate this using the so-called intraclass correlation coefficient. A regular correlation coefficient can be computed from two columns of numbers, but in the twins case, which twin goes in column 1 and which in column 2? One can't use the training method for this decision as it is a question of native ability, not of training method. Were the experiment to be rerandomized, a different twin from some pairs would get assigned to method 1 as compared to the twin being assigned to treatment 1 now. So it is unclear how to construct the columns and each different construction will give a different correlation.

Calculation 1: The intraclass correlation coefficient.

To resolve this problem, consider the model to be $Y_{ijk} = \mu + M_i + F_j + e_{ijk}$ where F_j represents a family j effect (variance estimated at 21.2) and e_{ijk} represents the effect of individual twin k. Now the difference between two twins, one from each family, involves a difference of random effects, namely $F_j+e_{ijk} – F_j - e_{i'j'k'}$ and the variance of this is $2(\sigma^2_F+\sigma^2)$ where σ^2_F and σ^2 are the family and individual variance components respectively. The estimate of this is 2(21.2+40.8) = 2(62). The difference between siblings involves the same family so the F parts cancel out and we have $e_{ijk} – e_{i'jk'}$ with variance $2\sigma^2$ and its estimate 2(40.8). Now if the ability to program (i.e. the required intelligence and logical ability) has a genetic component we expect the difference in programming times between siblings to vary less than that between unrelated people. The ratio 40.8/62, about 2/3, is the relevant ratio. About 2/3 of the variation we see comes from individual characteristics and 1/3 from family effects. As the within pairs variation decreases, the genetic component appears stronger and this ratio gets close to 0. Subtracting the ratio from 1 gives a correlation-like statistic that would be close to 1 when the genetic component is strong and near 0 when siblings differ about as much as a randomly selected pair of individuals. This intraclass correlation 1/3 is a very simple and unsophisticated way to measure heritability. The intraclass correlation then is an estimate of $\sigma^2_F /(\sigma^2_F+\sigma^2)$.

Calculation 2: Best Linear Unbiased Predictor (BLUP)

Suppose, for some reason, I want to measure the effect of family j. Without careful thought, one might think that a simple difference between the family j mean and the overall mean would work, but let's think again. The two times that are averaged for family j represent the family effect F_j plus the average of two individual

effects. Even if there were no family effects, there would be a largest family mean which would differ from the overall mean only because of individual differences in people. It would not be fair for the family with the highest mean to boast of its great genes. It could be some individual effects, like a good night's sleep, or even just lucky guesswork that day that caused these two siblings to do well. So what would be a good estimate of the family effect F_1? The deviation D_j of the family j sample mean from the overall mean $\overline{Y}$ would be clear of training method effects and of the overall mean. This difference D_j would be F_j plus the mean of two e's so its variance $\sigma^2_F+\sigma^2/2$ would be approximated as 21.2 + 40.8/2 = 41.6. Suppose I wanted to find the best multiplier for the difference D_j (between the family j mean and the overall mean) as an estimate of F_j. If that multiplier is 1 then I just use the difference D_j between the family mean and the overall mean as my estimate.

Let the multiplier mentioned above be b. To find the best multiplier, meaning that which minimizes $E\{ F_i + bD_j)^2 \} = E\{F_i + b(F_j + (e_{ij1} + e_{ij2})/2)^2 \} = (1-b)^2\sigma^2_F + b^2\sigma^2/2$. This is minimized when $-2(1-b)\sigma^2_F + 2b\sigma^2/2 = 0$, that is, when $b = \sigma^2_F/(\sigma^2_F+\sigma^2/2)$. Notice that this b is less than 1 so the difference D_j between the family j mean and the overall mean is shrunken toward 0 and the modified value bD_j is called the BLUP (standing for Best Linear Unbiased Predictor) of the family effect F_j. Adding the overall mean $\overline{Y}$ gives $\overline{Y}$ $+bD_j$ as the BLUP of the family j mean. As another example, if teachers in a school system with highly varying student quality are evaluated based on student test performance, much of the rating will be the "luck of the draw" in terms of which students a teacher gets.

While PROC GLM also has a random statement, its models are estimated as though all effects are fixed. Thus the variable FAMILY would be treated as fixed and the difference D_j between the family j mean and the overall mean would be the GLM estimate of the family j effect F_j. No BLUP calculation would be done. In PROC GLM, the LSMEAN statement delivers the estimate of the overall mean plus the family j effect. In this balanced data the overall mean of the 40 observations is 39.981 and the mean of the two family 1 observations is 35.625 so D_1 is -4.356. In contrast to this, the BLUP is $\overline{Y}$ + $0.510D_1$ = 39.981 + 0.510(-4.356) = 37.76 where the estimated b based on the PROC MIXED variance component estimates is 21.2/(21.2 + 40.8/2) = 0.510. The LSMEAN statement in PROC MIXED is used for fixed effects and to get the BLUPS, a series of ESTIMATE statements can be used.

To complete the BLUP discussion, here is some code and partial output comparing the LSMEANS form PROC GLM with the BLUPs from PROC MIXED. The numbers above should appear for family 1 within roundoff error.

```
ODS OUTPUT ESTIMATES=BLUPS;

PROC MIXED DATA=TWINS;
CLASS FAMILY METHOD;
MODEL TIME = METHOD;
RANDOM FAMILY;
ESTIMATE "1 " intercept 1 | family  1;
ESTIMATE "2 " intercept 1 | family  0 1;
```

(etc.)

```
ODS OUTPUT LSMEANS=GLMMEANS;
PROC GLM DATA=TWINS; CLASS FAMILY METHOD;
MODEL TIME = FAMILY METHOD;
LSMEANS FAMILY;
```

The outputs can be compared to the calculations above for family 1.

From PROC MIXED:

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
1	37.7612	3.3024	19	11.43	<.0001
2	34.5760	3.3024	19	10.47	<.0001
3	45.5330	3.3024	19	13.79	<.0001
		(etc.)			
20	42.6663	3.3024	19	12.92	<.0001

From PROC GLM:

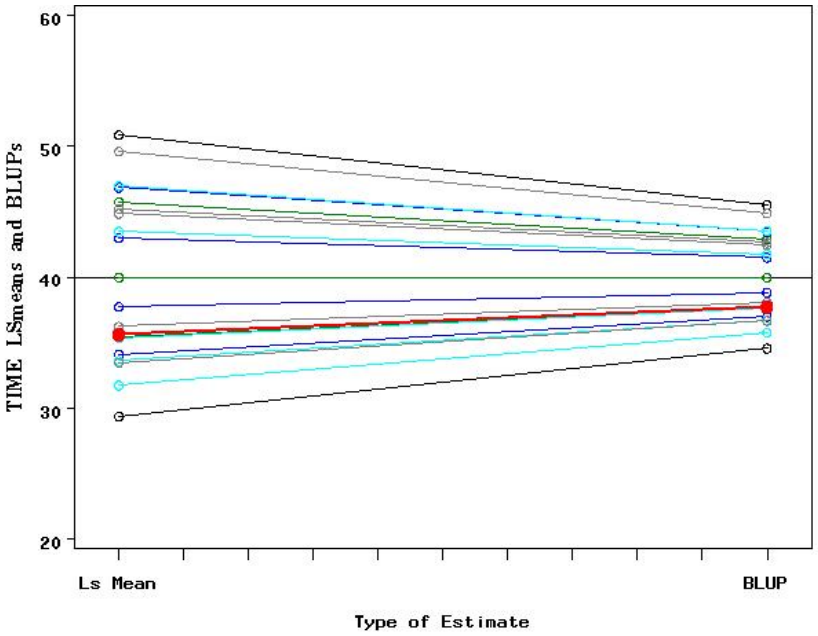
The GLM Procedure
Least Squares Means

family	time LSMEAN	Standard Error	Pr > |t|
1	35.6250000	4.5185064	<.0001
2	29.3750000	4.5185064	<.0001
3	50.8750000	4.5185064	<.0001
	(etc.)		
20	45.2500000	4.5185064	<.0001

Not only are the LSMEANS further from the overall mean 39.981 but in addition their standard errors are different than those of the BLUPs. The estimated family and individual variance components are the same in both cases, however PROC GLM, by considering blocks fixed, uses $\sqrt{40.8338/2}$ as a standard error whereas the BLUP, being an optimal combination of the overall mean and the family 1 mean has a smaller standard error.

To illustrate what is happening here, the datasets GLMMEANS and BLUPS are merged and transposed by family to give a dataset amenable to plotting. The resulting plot in Figure 2 has a horizontal reference line at the overall mean and larger red dots for family 1 to illustrate the hand calculations above.

Figure 2. LSmeans (left) and BLUPS



THE GENERAL MIXED MODEL AND REML ESTIMATION:
The model for the first 6 observations in the twins data can be expressed as

$$\begin{bmatrix} 38.25 \\ 33.00 \\ 28.75 \\ 30.00 \\ 46.50 \\ 55.25 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ M_1 \\ M_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

which has the matrix form $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma\square\square\square\mathbf{e}$. Had we used PROC GLM, our design matrix would join **X** and **Z** side by side to get a 6x6 matrix with the coefficient vector consisting of β stacked on top of γ. The variance matrix for **e** would be $\mathbf{I}\sigma^2$. In the mixed model, **X** contains only the fixed effects and its nature is determined by the MODEL statement. The **Z** matrix contains the random effects, family in this example, which are assumed to have some distribution with variance covariance matrix **G** and the error vector e is assumed to have some variance matrix **R** where neither of these could be, but does not have to be, a diagonal matrix. The structure of **G** is specified by the RANDOM statement and that of **R** by the REPEATED statement. For these first 3 families these matrices would be

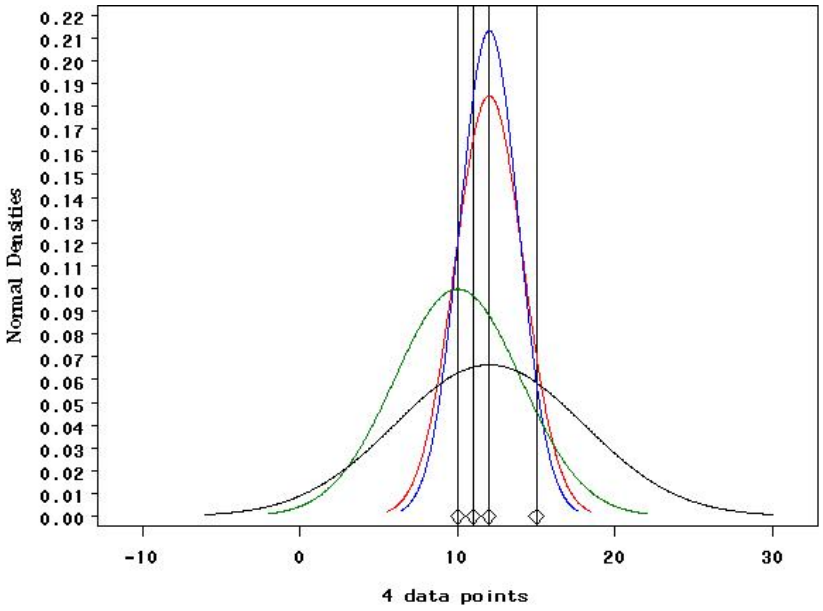
$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \sigma_F^2 \quad \text{and } \mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \sigma^2$$

For all 20 families **G** would be $\mathbf{I}\,\sigma_F^2$ where now **I** is a 20x20 rather than 3x3 matrix and likewise **R** would be a 40x40 matrix for these 40 observations.

There are two common kinds of estimation, ML = maximum likelihood and REML = residual (or restricted) maximum likelihood. To illustrate and compare these, an extremely simple example will be used.

Suppose a sample of 4 observations, 10, 11, 12, and 15, is available to estimate the mean and variance of a normal population. These observations along with some normal distributions are shown in figure 3.

Figure 3: A sample of 4 points.



The green curve there (leftmost peak) has mean 10 which seems too far to the left. The others have mean 12. The green and black (lowest peak) curves appear to display more variation than the four points, shown as diamonds near the bottom, suggest. The blue (highest peak) and red (next highest) curves look reasonable for the 4 points shown. The mean of 10, 11, 12 and 15 is 12. The deviations are -2, -1, 0, and 3 with sum of squares SSq = 4+1+0+9=14. The blue curve has variance SSq/n = 14/4 and the red one SSq/(n-1) = 14/3 . For any given distribution, a set of points has a "likelihood" defined as the product of the given probability density function values at each of the points. For each curve shown we take the product of the heights of the 4 vertical lines where they cross the curve. Thus for each curve there is a "likelihood" that the point came form that distribution. Looking at the blue curve, the four heights seem about the same or larger than for any of the other 3 curves so their product would exceed any of the other products, indicating that, if these were the only choices, we would pick blue. Finding the maximum likelihood over all possible normal curves gives us the maximum likelihood estimators of the mean and variance and as it happens, the blue curve is exactly that.

Now the mean of the population from which these data came would not likely be the same as the sample mean and the sample mean minimizes the sum of squared deviations, that is, any other mean than 12 would make that sum of squares larger. This means that the average squared deviation from the sample mean is larger than the average squared deviation from the population mean. This means that we have a biased low estimate of the true population variance by using maximum likelihood, and this is why we usually divide by the degrees of freedom, n-1 (3 in our example) rather than n (4 for our example). That gives an unbiased estimate of the variance.

The problem above stems from having to estimate the mean. Were the mean known, we would compute differences from that known mean and compute a better sum of squares. There is a way to get 3 observations from this distribution that have known mean 0. We can do it by contrasts. Three contrasts with their means and variances are shown below:

$Z_1 = (Y_1 + Y_2 - Y_3 - Y_4)/2$, Mean $\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$, Variance $4\sigma^2/4$
$Z_2 = (Y_1 - Y_2 - Y_3 + Y_4)/2$, Mean 0, Variance σ^2
$Z_3 = (Y_1 - Y_2 + Y_3 - Y_4)/2$, Mean 0, Variance σ^2

These contrasts are orthogonal to each other. A fourth orthogonal linear combination, namely the sum of the 4 Y values, can be computed, but it involves the unkown mean and thus will not be used. We now have n-1 = 3 observations, for our data they are

(10+11-12-15)/2 = -3
(10-11-12+15)/2 = 1
(10-11+12-15)/2 = -2

The sum of squares here is 9+1+4 = 14, the same as when we used the estimated mean. The average squared deviation here is 14/3 because we had to eliminate one contrast for estimating the mean. Thus 14/3 is the REML estimate of the variance which can be recognized as the usual unbiased estimate of the true error variance. Likewise with 5 treatment groups and 20 observations in a standard ANOVA, we would eliminate 5 of 20 orthogonal contrasts leaving 20-5 or 15 orthogonal contrasts to compute the REML or Residual Maximum Likelihood Estimate. The red curve in Figure 3 uses this REML estimate for the variance. Clearly it does not maximize the usual likelihood but it does maximize the residual likelihood. The normal density formula is

$$f(y) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left(-0.5\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

and for each of the 4 densities one can insert the mean and variance into f(y) then evaluate this at the 4 Y data points or the 3 Z contrasts then multiply these 4 (or 3) numbers together to compute the likelihood. Here are the results for the 4 curves (labeled by the curve color) for each of our two methods:

Likelihoods for 4 data points:
red=0.0002595289 blue=0.000279844 green=0.0000638847 black=0.0000160913

Likelihoods for 3 REML contrasts:
red=0.0014053317 blue=0.0013123199 green=0.000640541 black=0.0002420086

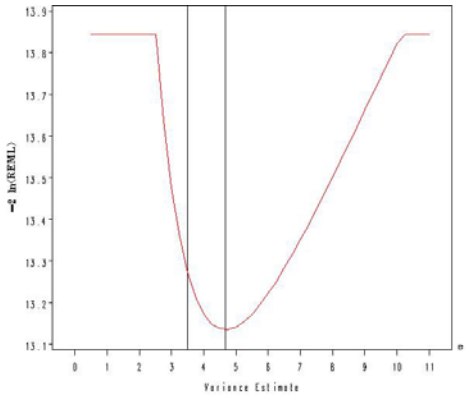
As expected the red curve (variance 14/3) maximizes the REML likelihood while the blue curve (variance 14/4) maximizes the likelihood.

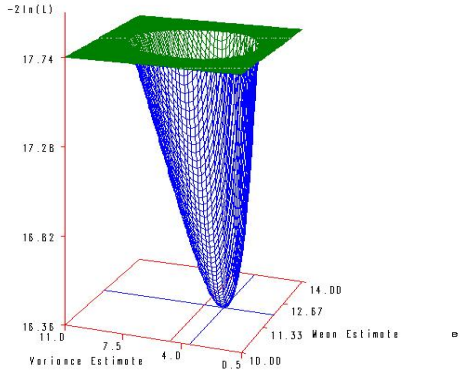
Once the 4 observations or the 3 contrast values are inserted into the likelihood (residual likelihood) the result is a function of μ and σ^2 (or just σ^2) that can be plotted against various values of μ and/or σ^2 to find a maximum. Instead one often plots -2ln(L) where L is one of the two forms of likelihood and looks for a minimum. The reason for this is that in large samples, the difference between -2ln(L) for a full model and -2ln(L) for a reduced model is approximately a Chi-square under the null hypothesis corresponding to the restriction provided the restriction sets the parameter to a value strictly inside the interior of the parameter space. The degrees of freedom is the number of restricted parameters. On both likelihood plots, the 80th percentile of a Chi-square with 2 (ML) or 1(REML) degrees of freedom is computed and added to the minimum -2ln(L) for that plot. When -2ln(L) exceeds this number it is reset to that number, providing an

upper “ceiling” to both plots. The set of points for which -2ln(L) is not truncated, then, forms a type of 95% confidence region for the model’s parameters in that this is the set of parameters that would not be rejected with a 20% level hypothesis test. This kind of test that compares likelihoods is called a “likelihood ratio test” and is very common in statistics.

On each plot (figures 4 A and B) the σ^2 estimates for both likelihoods are shown. Recall that these are 4.667 for REML and 3.5 for ML. It should be obvious which one minimizes the plot shown and the other serves to illustrate how far apart are the two estimates for this very small sample.

Figure 4 A and B: REML likelihood (left, A) and ML (right, B)





RANDOM EFFECTS: VARIANCE COMPONENT TESTS

Usually the focus of interest in mixed models is on the fixed effects. However it is sometimes of interest to test a random effect. Notice again that the likelihood ratio test is justified for restrictions that set a parameter to a value strictly inside its parameter space. Since variances are positive, 0 is on the boundary of the parameter space so a test that a variance is 0 is not justified by likelihood ratio theory. It is also fairly common to do a test that, roughly speaking, is an approximation to the likelihood ratio test. Again speaking loosely, this test assumes the log likelihood is approximately quadratic in a region near the maximum. It is referred to as a “Wald test” after its inventor and it still has the interior of the parameter space restriction. The plots we have seen are clearly not exactly quadratic, but still have a somewhat parabolic (quadratic) shape. The Wald test is available in PROC MIXED with the COVTEST option. For such variance component tests, the Wald test is *very* approximate and should only be used when many levels of the random effects have been observed. Further, one can fit the full and reduced model and compute the likelihood ratio test from the -2 log Likelihoods that appear in both outputs. For REML variance component tests, the same fixed effects should appear in both the full and reduced models.

The following SAS programs and partial outputs illustrate these tests for a hypothetical data set on times to death of tumor affected rats. Nine rats from each of 10 families (genetic lines) are randomly assigned to 9 dose levels of a drug and times to death reported.

Three things are done. First the full model is fit in REML adding the COVTEST option. Next the reduced (no family effect) model is fit and the two -2 log Likelihood values are used to construct a likelihood ratio test. Despite previous comments lauding PROC MIXED for its handling of random effects as compared to PROC GLM, there is still the advantage in GLM that the F test for random effects are exactly distributed as F for simple models like this with balanced data. Recently a TYPE3 option has been added to PROC MIXED that enables it to reproduce these exact F tests in the cases in which they are justified. Here we will just use PROC GLM to produce the F tests.

Program 1

```
PROC MIXED DATA=MICE COVTEST;
CLASS FAMILY DOSE;
MODEL TIME = DOSE;
RANDOM FAMILY;
```

Program 1 partial output

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z
family	8.3191	7.0243	1.18	0.1181
Residual	57.6759	9.6127	6.00	<.0001

Fit Statistics

-2 Res Log Likelihood	586.5

The Wald test Z=1.18 does not find a family variance component at the 5% level.
The statistic -2 Res Log Likelihood = 586.5 associated with this full model will be compared to that of the reduced model.

Program 2

```
PROC MIXED DATA=MICE;
CLASS FAMILY DOSE;
MODEL TIME = DOSE;
```

Program 2 partial output

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	65.9951

Fit Statistics

-2 Res Log Likelihood	589.9

The likelihood ratio test is 589.9-586.5 = 3.4 and for a Chi-square with 1 degree of freedom we can compute the p-value 0.065196 for this result. While we still do not find a family variance component at the 5% level we come a lot closer to doing so here than with Wald. Results in Self and Liang (1987) suggest that in the case a single variance component is restricted to 0 under the null hypothesis, the boundary effect implies that the p-value can be divided in half, giving p-value 0.0376 and implying that there is a family variance component. In setting several parameters to their boundary values, the complexity of Self and Liang's

results make it extremely difficult to compute the proper adjustment. Finally, we look at the exact F test from PROC GLM, which could be reproduced in PROC MIXED with the proper METHOD=TYPE3 option.

Program 3:

```
PROC GLM DATA=MICE;
CLASS FAMILY DOSE;
MODEL TIME = DOSE FAMILY;
run;
```

Program 3 partial output:

Dependent Variable: time

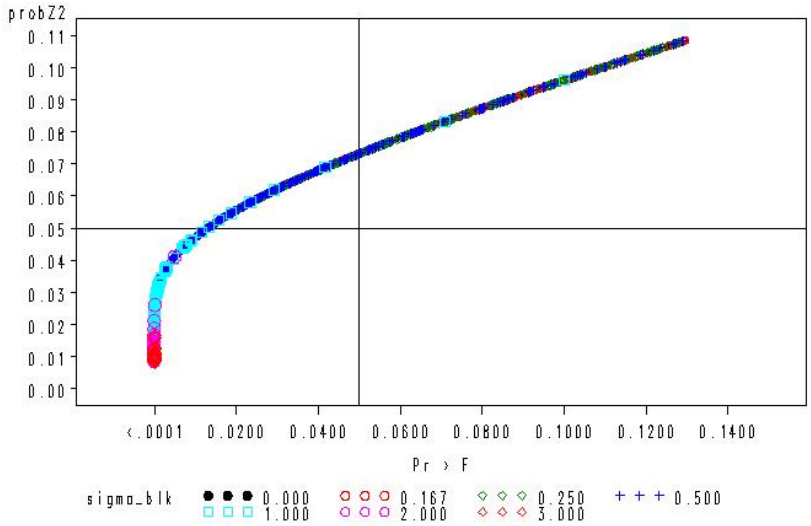
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	5761.155556	338.891503	5.88	<.0001
Error	72	4152.666667	57.675926		
Corrected Total	89	9913.822222			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DOSE	8	4568.222222	571.027778	9.90	<.0001
family	9	1192.933333	132.548148	2.30	0.0249

The F test is not only exact, it also gives the strongest evidence yet of a family effect (p-value 0.0249). Notice that this difference is not because of different error estimates. The MSE here, 57.6759, is exactly the same as the corresponding residual variance component estimate from PROC MIXED. Using expected mean squares, the PROC GLM estimate of the family variance component is (132.548148-57.675926)/9 = 8.31913, the same as PROC MIXED.

The question now arises as to whether this behavior is typical. To investigate, we use 1000 runs of the model above, putting out the COVTEST values as well as the GLM values, for family variance component equaling the error variance times 0, 0.25, 1, 4, 9, and 16.

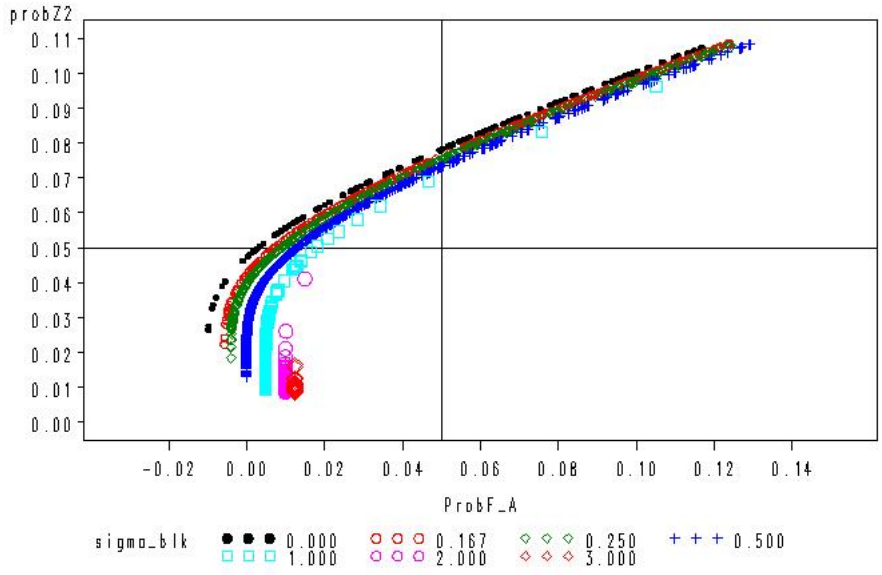
Figure 5: Wald p-values vs. F p-values



In Figure 5 are the P-values for the Wald Z test, divided by 2. These are plotted against the p-values for the F test and a different symbol is used for each block standard deviations 0, 1/6, 1/4, 1/2, 1, 2, and 3. The crosshairs are at 0.05 on each axis. Note that they all seem to lie along a smooth curve.

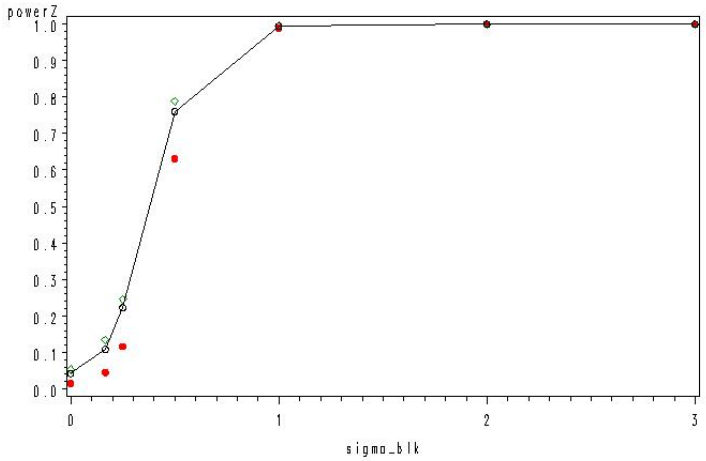
In the upper left quadrant are points that would reject the 0 block effects hypothesis with F but not with the Wald Z, even after dividing by 2. The lower right quadrant is empty indicating that Z never found block effects when F did not. Also notice that all the points seem to lie along the same curve of rejection probabilities. The different symbols suggest that as the actual block variance grows, the p-values move left along this curve so that more and more fall into the lower left quadrant indicating rejection by both F and Z (that is, increasing power). To see this effect more clearly, in Figure 6 the sets of points for the different block standard deviations were shifted by moving the sets of points with smaller standard deviations to the left and with larger to the right. The horizontal axis no longer retains much meaning but with the collections of points thus spread, it is clear how the larger standard deviations (lower right sets of points in magenta circles and orange diamonds) are moving both tests into their rejection regions (left and below the crosshairs) while the smaller block standard deviations (small black dots toward the upper left) have only a few points falling into the lower left quadrant where both tests reject the hypothesis of no block variability. The sizes of the plot symbols also increase with increasing block standard deviation.

Figure 6: Wald p-values vs. Shifted F p-values



The proportion of points falling left of the vertical crosshair (F) or below the horizontal (Z) gives the empirical power of the test. Graphing this power for the F test, the Z test with the dividing by 2 rule, as well as the likelihood ratio test with the dividing by 2 rule against the block standard deviation produces the power plot, figure 7 .

Figure 7: Powers of F, Wald, and Likelihood Ratio Tests



The connected black dots are the F test results while the green diamonds, which are about the same as the dots, are the likelihood ratio results. The red circles are the Wald Z tests and we see that the Z test has somewhat less power. As the number of blocks drops, the difference between Z and the other two becomes much more dramatic. My recommendation is to use F when available (Type III method in PROC MIXED) as it has an exact F distribution under the null hypothesis.

UNBALANCED DATA

A nice feature of MIXED is that you use the same code whether the data are balanced or not. To illustrate, I use another artificial dataset. Suppose I have 3 types of earplugs to test on a noisy factory floor. My concern is with temporary hearing loss after an 8 hour shift in the factory. I can test each ear separately and want to do a randomized complete block study using workers in a shift as blocks, but unless I can find workers with 3 ears, I cannot get *complete* blocks. I enlist 7 workers (blocks, random effects) and assign each to 2 different ear plugs. Here are the data, including hearing loss and which ear was used.
Each row has 2 nonmissing values (2 ears per worker)

Table of Temporary Hearing Losses

Plugs → ---------- Workers	I	II	III
A	25 L	-	22 R
B	19 L	8 R	-
C	-	7 L	7 R
D	29 R	23 L	-
E	-	16 R	14 L
F	16 R	-	12 L
G	25 L	24 R	-

PROC GLM treats WORKER as fixed for the estimation part, even if you have issued a random statement. In the analysis of variance table, then, the Type III F test for PLUG would be adjusted for WORKER, that is, it is a completely within worker comparison. For example, using only workers with plugs I and II we compare (19+29+25)/3 to (8+23+24)/3, a difference of (73-55)/3 = 6 and this estimate does not involve worker effects (but it is also not the best estimate delivered by PROC GLM where blocks are treated as fixed). There is information in comparing the sum of Worker A's numbers (Plug I + Plug III), 47 to that of worker E (Plug II+Plug III), 30. The difference, 47-30=17, estimates the Plug I versus Plug II difference but the error therein

involves the difference of two worker effects plus the difference of errors for the two workers whereas in PROC GLM, the comparison has a standard error not involving worker effects because they have been mathematically eliminated. PROC MIXED automatically combines the information between and within blocks in an optimal way. In this particular example this is just enough to drop the p-value for Plugs from just above 0.05 (GLM) to just below (MIXED). Here is the code and partial output:

```
proc glm;  class plug worker;
model loss = worker plug;  Random Worker;
Estimate "I vs III - GLM" Plug -1 0 1;  run;
proc mixed;  class plug worker;
model Loss=Plug;  Random Worker;
Estimate "I vs III - Mixed" Plug -1 0 1; run;
```

Relevant GLM output:

Dependent Variable: Loss

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	666.3705357	83.2963170	13.50	0.0054
Error	5	30.8437500	6.1687500		
Corrected Total	13	697.2142857			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
worker	6	451.9062500	75.3177083	12.21	0.0074
plug	2	62.6562500	31.3281250	5.08	0.0625

Parameter	Estimate	Standard Error	t Value	Pr > |t|
I vs III - GLM	-4.81250000	1.96353476	-2.45	0.0579

The Type I and III sums of squares are not the same. Type III is the one to use here. It just fails to show Plug effects at the 0.05 level. The RANDOM statement has no effect on any of this. An estimate of the difference in hearing loss between plugs I and III also shows no significance.

Relevant MIXED output:

Covariance Parameter Estimates

Cov Parm	Estimate
worker	37.5785
Residual	6.1674

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
plug	2	5	5.79	0.0499

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
I vs III - Mixed	-5.2448	1.9347	5	-2.71	0.0422

The F test and I vs. III estimate are now significant. The estimate itself is different as it combines within and between block information. The standard error is also slightly different. The residual variance estimates, 6.1688 and 6.1674, differ slightly as well. In unbalanced data this is typical and unlike what we saw for the balanced data. The estimate of the worker variance component from GLM using expected mean square information (not shown) is (75.3177 – 6.1688)/1.8333 = 37.7177, slightly differing from the 37.5785 given by PROC MIXED.

RANDOM COEFFICIENT MODELS

PROC MIXED also allows the estimation of an overall relationship, a line for example, relating Y (let's say blood pressure) to X (dose of a drug) over all patients. In many cases, each patient has their own line and it might make sense to think of the slopes and intercepts from these lines as including random deviations from the overall slope and intercept. In the case of lines, for data with Y and X positive, one might expect patients with higher than average slopes to have lower than average intercepts. Thus the 2x2 covariance matrix between the slope and intercept deviations would be expected to show a negative covariance off the diagonal. There would be no reason to expect the slopes and intercepts to have the same variance either so a 2x2 ***unstructured*** matrix would usually be fitted to the model, using the REPEATED statement in PROC MIXED.

As a last example, here are some real data and a random coefficients model with a small twist: the fitted function has to be periodic to make sense.

Example: Black Bears

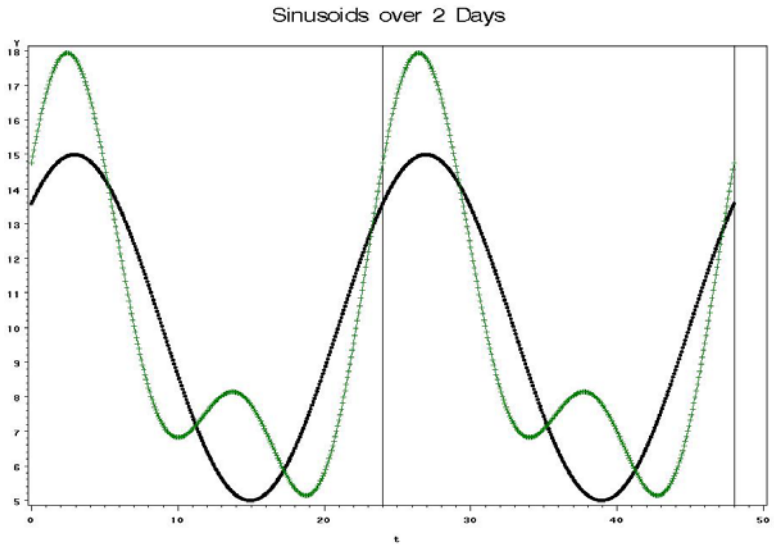
A study of activity levels in black bears was done by Francesca Antonelli, a visiting researcher in zoology at North Carolina State University in 2001 who kindly shared the data used here. Black bears were anesthetized and fitted with transmitter collars. These collars allowed the researcher to monitor the bear activity in terms of movement. Movement above a certain level was categorized as VA – Very Active and an assessment of the activity category was made every 5 minutes for each of 12 bears for a lilt over 10 days (2884 observations per bear).

In modeling these data some interesting features must be accounted for. Interest lies in describing the within day pattern of activity for the bears. First, the time of day would not be useful as a linear predictor. As time goes from 0 to 24 hours in 5 minute steps, either a linear increase or decrease will result in an unreasonable discontinuity as we pass the midnight point going into the next day. We need a periodic function of period 24 hours if the model is to be believable. One possibility is to use $\sin(2\pi jt/24)$. When j is 1 this goes through one cycle per day, j=2 gives 2 per day etc. The problem with just using sines is that for a given amplitude, the predicted response is then forced to be half way between its high and low value exactly at midnight. This is unreasonable for many situations. To fit the data well, it may be necessary to move the sine wave right or left in addition to adjusting its amplitude to fit the data. This gives something called a sinusoid – it is a shifted sine wave with the shift being called a "phase" shift.

A fact from trigonometry is that sin(A+B) = cos(B)sin(A) + sin(B)cos(A) so if the rotation angle is A = $2\pi jt/24$ and the phase shift B (we would also have an amplitude C) then we see that $C\sin(2\pi jt/24 + B) =$

[Ccos(B)]sin($2\pi jt/24$) + [Csin(B)]cos($2\pi jt/24$) or $\beta_1 \sin(2\pi jt/24) + \beta_2 \cos(2\pi jt/24)$. For example, Figure 8 shows, using black dots, a sine wave of period 1 day.

FIGURE 8: PURE SINUSOIDS



It is easy to create variables S= sin($2\pi jt/24$) and C= cos($2\pi jt/24$) in the data set then regress Y on S and C. The results are:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7212.84270	3606.42135	Infty	<.0001
Error	574	0	0		
Corrected Total	576	7212.84270			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.00000	0	Infty	<.0001
S	1	3.48353	0	Infty	<.0001
C	1	3.58678	0	Infty	<.0001

The fit is perfect (0 error sum of squares) and both the sine and cosine are needed. Since 3.5868/3.4835 is the tangent of the phase shift angle and is just over 1, the phase shift must be a little over 45 degrees. The figure also shows, in green, the sum of the black sinusoid with a second that goes through 2 cycles per year. The result is another periodic function with period 24 hours but now having a little more interesting pattern. The multiple period per day waves, like the one added to the black sinusoid, are called harmonics of the fundamental frequency. With enough harmonics (2 per day, 3 per day, 4 per day …) we can get arbitrarily close to any reasonably smooth periodic function.

With this background in mind, we fit a model with a fundamental sine wave and a few harmonics, this being thought of as a fixed effect diurnal pattern across all bears, and then sinusoidal deviations for each individual bear. We will assume these random coefficients, be they sines or cosines, high or low frequency, come from the same normal distribution. Different variances for different frequencies could also be tried and would

be a sensible approach but with a large number of effects to be estimated. Convergence problems can occur as models grow more complex. These can be data problems or the result of accidentally introducing an exact dependency in the model. For illustration, we stick with the simple model. Using a single variance component for all the trigonometric components, we:

(1) Use the repeated statement interacting all of our sines and cosines (continuous variables) with the class variable BEAR

(2) Force a structure $\mathbf{I}\sigma_{\text{Trig}}^2$ on these deviations by specifying TYPE=TOEP(1).

We use σ_{trig}^2 to denote the common variance of all the trigonometric function random coefficients. A Toeplitz matrix has a "striped" appearance like this:

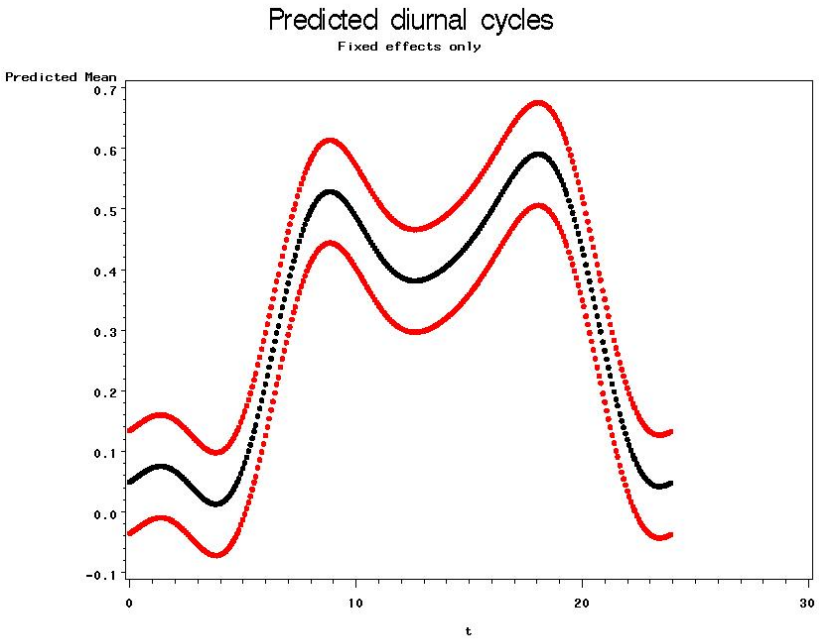
$$\begin{pmatrix} a & b & c \\ b & a & b \\ c & b & a \end{pmatrix}$$ and TOEP(1) in SAS means that only the first letter (a) is nonzero.

Here is the model statement

```
ods output solutionR=BLUPbear;
proc mixed data=bears covtest;
   class bear;
   model Y = s1--c4/solution outpm=means outp=pbear;
   random bear;
   random s1*bear c1*bear s2*bear c2*bear s3*bear
          c3*bear s4*bear c4*bear
          /type=toep(1) solution; run; quit;
```

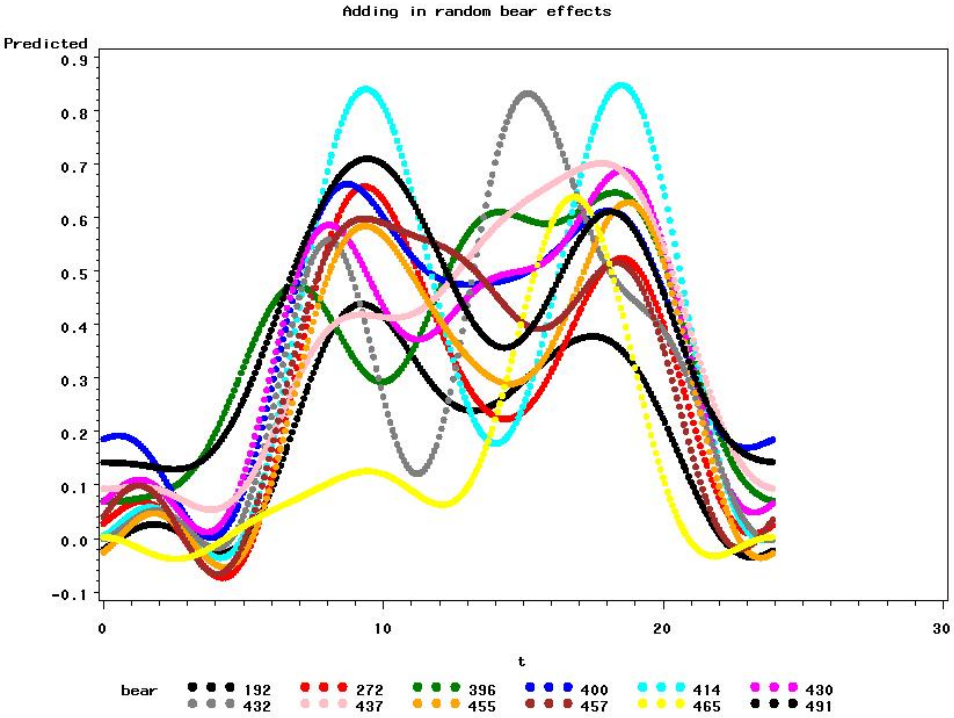
Here s1 and c1 are the fundamental sine and cosine waves with the other s and c variables giving 3 harmonics. The command outpm=means creates a dataset with the fixed sinusoid common to all bears as well as upper and lower confidence limits. That plot is shown in Figure 9

Figure 9: Bear Diurnal Activity Levels – Fixed Part.



The command outp=bear creates a dataset that adds in the random effects, adjustments to the curve, for each bear. These are not the same as would be obtained by fitting sinusoids to each bear as these have had the same kind of BLUP adjustment shown earlier. A plot of these individual bear curves is shown below in Figure 10. Some bears have activity patterns differing somewhat what from the norm. Finally, because the response is 0-1, the pattern indicates a probability of being very active at any given time. Thus a logistic approach, a generalized linear mixed model, is an alternative and perhaps preferred method for modeling these data.

Figure 10: Bear Diurnal Activity Pattern with Individual (BLUP) Effects.



Having called for solutions and using ODS to output them, we can print those with p-values less than 0.0001. These identify bears that have somewhat different patterns than average. We find several of these. There are more than 10,000 data points going into the model and thus we expect good statistical power for our tests. Even small magnitude effects can be statistically significant.

It should also be noted that the 0-1 response variable was not transformed with a logistic function as would be done in logistic regression. Nevertheless, the predictions from the model fall into the 0-1 range nicely.

The following list shows the individuals that are different than the average bear.

Obs	Effect	bear	Estimate	StdErr Pred	DF	tValue	Probt
1	bear	192	-0.1189	0.02872	11E3	-4.14	<.0001
22	s2*bear	396	0.08687	0.02339	11E3	3.71	0.0002
24	s3*bear	396	-0.1047	0.02339	11E3	-4.47	<.0001
40	s2*bear	414	-0.1601	0.03248	11E3	-4.93	<.0001
42	s3*bear	414	0.1308	0.03248	11E3	4.03	<.0001
60	s3*bear	432	-0.1094	0.03247	11E3	-3.37	0.0008
65	s1*bear	437	-0.09941	0.02134	11E3	-4.66	<.0001
91	bear	465	-0.1702	0.03234	11E3	-5.26	<.0001
94	s2*bear	465	0.1461	0.03246	11E3	4.50	<.0001
101	s1*bear	491	0.09137	0.02340	11E3	3.91	<.0001

CONCLUSIONS

PROC MIXED uses REML estimation and thus has some nice properties in terms of getting reasonable estimates of variance components, computing BLUPs, and automatically adjusting fixed effects tests for complex variance structures. Interesting kinds of models that would not be possible to fit without computer intensive methods, can now be fit. A rather broad set of statistical models fit into the MIXED framework.

REFERENCE

Self, S. G. and K. Y. Liang. 1987. “Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions.” *Journal of the American Statistical Association,* 82:605-610.

CONTACT INFORMATION

Name: Professor David A. Dickey
Enterprise: Department of Statistics
Address: Box 8203, North Carolina State University
City, State ZIP: Raleigh, NC 27695-8203
E-mail: dickey@stat.ncsu.edu
Web: http://www.stat.ncsu.edu/~dickey/