

USING PROC MIXED TO ANALYZE A REPEATED MEASURES DESIGN

Keith Dunnigan
Trilogy Consulting Company

Introduction:

The purpose of this presentation is to provide an introductory primer to PROC MIXED regarding its basic use and syntax. After providing a quick overview of the procedure and motivating its use, the syntax for a basic design will be presented, followed by the slightly more complicated details involved in a repeated measures design.

Overview:

The General Linear Model fit by PROC GLM is:

$$y = X\beta + \varepsilon, \quad \text{where} \quad E(\varepsilon) = 0 \quad \text{and} \quad \text{Var}[\varepsilon] = \sigma^2$$

This model assumes the components of ε are independent and normally distributed with equal variance. The model is fit by the method of least squares with the formulas for the coefficient vector and sums of squares given below, where L is computed differently for each effect in the model depending on whether it is type 1, 2, 3, or 4.

$$b = (X'X)^{-1} X'y \quad \text{and} \quad SS(Lb = 0) = (Lb)'(L(X'X)^{-1}L')^{-1}(Lb)$$

The Mixed Model fit by PROC MIXED is a generalization of that fit by GLM. It is:

$$y = X\beta + Zv + \varepsilon, \quad \text{where} \quad E\begin{bmatrix} v \\ \varepsilon \end{bmatrix} = 0 \quad \text{and} \quad E\begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} = 0$$

This model assumes the components of ε and v are normally distributed, the independence and constant variance assumptions of the general linear model are not required. The General Linear Model is a special case of the Mixed Model with $Z = 0$ and $R = \sigma^2$. Estimation is much harder in the mixed model due to the extra parameters. A likelihood type approach is used in which G and R are first estimated using either maximum likelihood (ML) or Restricted Maximum Likelihood (REML) and the estimates put into a Generalized Least Squares (GLS) equation and the final estimators derived. The resulting equations for the coefficient estimator vectors b and v are:

$$b = (X'V^{-1}X)^{-1} X'V^{-1}y, \quad \text{and} \quad v = GZ'V^{-1}(y - Xb)$$

The sums of squares are not calculated directly by `proc mixed`, but the F tests are estimated directly, with the formula being:

$$F = \frac{1}{\text{rank}(L)} \left[\frac{b}{v} \right]' L' (L' \hat{C} L)^{-1} \left[\frac{b}{v} \right]$$

Where $\hat{C}$ is the estimated covariance matrix of b and v . For further details, see the references cited later.

Properties of the Estimators

Since Estimates were used for G and R in the derivations, b is not the Best Linear Unbiased Estimator (BLUE) of β and v is not the Best Linear Unbiased Predictor (BLUP) of v , nor do the usual T and F statistics have exact T and F distributions. Rather the estimators are described as *empirical* and are described as EBLUE and EBLUP, and the distributions as *approximate* T and F distributions.

Practical Applications

The above suggests several ways in which mixed models can be very useful. For example a simple design such as a 2 way factorial may be analyzed much more simply and accurately using `proc mixed` if it has unequal variances in each cell, unequal numbers of observations per cell, random components, or is incomplete. More complicated designs (such as the repeated measures to follow later) which have intercorrelated or clustered observations may be dealt with also.

The syntax in PROC MIXED in general is similar to PROC GLM. One difference is that effects included in the random statement are *not* included in the model statement. For example the following shows the syntax for GLM and MIXED for a 2 way mixed factorial model.

```
proc glm data=twoway
  class method variety;
  model yield=method variety method*variety/ss3;
  random variety method*variety/q;
run;

proc mixed data=twoway
  class method variety;
  model yield=method;
  random variety method*variety;
run;
```

Another difference in syntax is the use of the repeated statement, which is completely different than in GLM and is explained in the next section. Finally the printed output produced for Proc Mixed does not include Sums of Squares calculations in its tests for fixed effects. As noted earlier, the F statistics are calculated directly in PROC MIXED as opposed to PROC GLM which uses the sums of squares to construct them.

Repeated Measures Designs

Earlier some motivation for using mixed models to analyze simple designs was given. There may be other situations where mixed models (and PROC MIXED) may be more appropriate. One such case is in a repeated measures design. In many repeated measures experiments, strange intercorrelations may exist between successive measurements of a variable on the same subject. One has few options under PROC GLM to model or deal with these correlations. The great advantage that PROC MIXED has is the ability to model the covariance matrix for a subject in many different ways and incorporate that into the analysis of the design.

When using PROC MIXED on a repeated measures design then, the analysis is broken into 2 parts: A) Modeling the Covariance and B) Analyzing the Design. The harder and more time consuming is the first part, with the actual analysis being very easy. To illustrate the syntax, consider the following example. A group of heart patients consisting of depressed and non depressed patients are measured for a certain blood chemical at resting, standing (2 minutes), standing (5 minutes), and standing (ten minutes). The question is whether there is a significant effect corresponding to depression or not. The baseline (resting) measurement is subtracted from the other 3 measurements. The syntax for analyzing the design is shown below, which corresponds to the second part of the analysis and assumes the correct covariance matrix found in the first part. To illustrate the use of the random statement once more, a second program is shown for the case when the baseline is not subtracted off, but controlled for as a random effect (subject within depression).

```
proc mixed data=plasma;
  class depress subj time;
  model plasma=depress time depress*time;
  repeated / type=un subject=subj(depress) group=depress;
run;
```

If the differences are not subtracted:

```
proc mixed data=plasma;
  class depress subj time;
  model plasma=depress time depress*time;
  random subj(depress);
  repeated / type=un subject=subj(depress) group=depress;
run;
```

The use of the repeated statement in PROC MIXED is to set up the model for the covariance matrix. The type option selects the covariance structure to be used (in this case, unstructured). The subject option defines the blocks (which have the particular covariance structure). In this case the blocks are the patients (the repeated measurements on a subject make up a block). The group option defines a category which may be used to specify heterogeneity in the covariance structure of R, that is each level of the group variable produces a new set of covariance parameters having the same structure as the original group.

As with any statistical modeling process there are two key things to consider when choosing a covariance model, quantitative/statistical considerations and expert/subjective knowledge. The various covariance models available under PROC MIXED can be found in reference 2 below. From our particular experiment several stand out as likely candidates, particularly those in which the covariance decreases proportionately with time. The act of standing up is a shock to the cardiovascular system which speeds up the heart rate and makes changes in the chemical being

studied, with it slowly returning to normal. This suggests that models for which the covariance or correlation between measurements decrease as the time interval between points increase, in addition to the unstructured model. Further, we postulate that the correlation between successive time point measurements may be different for the depressed and nondepressed groups. These models may be compared using various statistical measures, the AIC criterion being used here (lower scores on this criterion fit better). The following code obtains the AIC for each model chosen and ranks them. Variations for depression as the grouping variable are also fitted. Finally, the results are shown.

```
*****;
** Create macro runmixed to compare the AIC criterion;
** obtained for various covariance structures.;
*****;

%macro runmixed (z=, comment=, out=);
  %let _print_=off;
  proc mixed data=plasma;
    class depress subj time;
    model plasma=depress time depress*time;
    &z;
    make 'fitting' out=fitinfo;
  run;
  data fitinfo;
    length z $20;
    set fitinfo;
    z="&comment";
  run;
  proc append base=&out data=fitinfo ;
  run;
  %let _print_=on;
%mend;

** Now fit several models;

* Ustructured;
%runmixed (z=repeated / type=un sub=subj(depress),
  comment=unstr,
  out=allaic);

* Block diagonal Ustructured-different R for each depress state;
%runmixed (z=repeated / type=un sub=subj(depress)
  group=depress,comment=bl_diag_unstr,out=allaic);

* Ustructured correlations;
%runmixed (z=repeated / type=unr sub=subj(depress),comment=unstr_corr,
  out=allaic);
```

```

* Block diagonal Ustructured correlations-different R for each depress
state;
%runmixed (z=repeated / type=unr sub=subj(depress) group=depress,
            comment=bl_diag_unstr_corr,out=allaic);

* first order antedependence;

%runmixed (
    z=repeated/type=ante(1) sub=subj(depress), comment=antedep, out=allaic);

* first order antedependence, diff r for each depress state;

%runmixed (z=repeated/type=ante(1) sub=subj(depress) group=depress,
            comment=blddiag_antedep1_,out=allaic);

* First Order Heterogenous Autoregressive;
%runmixed (z=repeated / type=arh(1) sub=subj(depress), comment=arh1,
            out=allaic);

* Block Diagonal First Order Heterog. Autoregressive-diff. R for each
depress state;
%runmixed (z=repeated/ type=arh(1) sub=subj(depress) group=depress,
            comment=bl_diag_arh1,out=allaic);

* Heterogenous Toeplitz;
%runmixed (z=repeated / type=toeph sub=subj(depress), comment=toeph,
            out=allaic);

* Block Diagonal Toeplitz-diff. R for each depress state;
%runmixed (z=repeated / type=toeph sub=subj(depress) group=depress,
            comment=bl_diag_toeph,out=allaic);

proc sort data=allaic;
    where descr =: 'Akaike';
    by descending value;
run;

proc print data=allaic;
run;

```

OBS	Z	DESCR	VALUE
1	bl_diag_unstr	Akaike's Information Criterion	-2152.42
2	bldiag_antedep1_	Akaike's Information Criterion	-2152.78
3	unstr	Akaike's Information Criterion	-2182.58
4	unstr_corr	Akaike's Information Criterion	-2182.58
5	bl_diag_unstr_corr	Akaike's Information Criterion	-2182.58
6	antedep	Akaike's Information Criterion	-2183.88
7	arh1	Akaike's Information Criterion	-2183.91
8	bl_diag_arh1	Akaike's Information Criterion	-2183.91
9	toeph	Akaike's Information Criterion	-2184.82
10	bl_diag_toeph	Akaike's Information Criterion	-2184.82

Based on these analysis, the block diagonal unstructured model was chosen (the block diagonal structure with depression as the grouping variable). The final code and results are shown below. From this we can see that the effect of depression did not significantly affect the chemical measurement.

```
proc mixed data=plasma;
  class depress subj time;
  model plasma=depress time depress*time;
  repeated / type=un sub=subj(depress) group=depress;
run;
```

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
DEPRESS	1	87	1.20	0.2767
TIME	3	253	73.41	0.0001
DEPRESS*TIME	3	253	0.18	0.9098

References:

SAS Institute Inc., SAS/STAT Software: Changes and Enhancements through Release 6.11, Cary NC: SAS Institute Inc., 1996.

Latour, D., Latour, K., and Wolfinger, R.D. (1994), Getting Started with Proc Mixed, Software Sales and Marketing Department, SAS Institute Inc., Cary, NC.

SAS is a registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

For further information contact:

Keith Dunnigan
e-mail: Harry.Dunnigan@mkg.com