STARTING VALUES FOR PROC MIXED WITH REPEATED MEASURES DATA

J. C. Recknor

W. W. Stroup

STARTING VALUES FOR PROC MIXED WITH REPEATED MEASURES DATA

J.C. Recknor and W.W. Stroup
Department of Biometry, University of Nebraska-Lincoln

1. Abstract

A major advantage of PROC MIXED for repeated measures data is that one could choose from many different correlated error models. However, MIXED uses default starting values that may cause difficulty obtaining REML estimates of the covariance parameters for several of the models available. This can take the form of excessively long run times or even failure to converge. We have written a program to obtain initial covariance parameter estimates that result in greatly improved performance of the REML algorithm. We will use two covariance models frequently of interest in animal health experiments, the first-order ante-dependence model [ANTE(1)] and the Toeplitz model with heterogeneous variances [TOEPH], to illustrate the use of our procedure.

2. Introduction

Repeated measures experiments are common in agricultural research. When analyzing data from such experiments, it is important to account for the correlation among observations at different times on the same subject. Failure to account for correlated errors results in inflated type I error rates, if positively correlated, which is what we would expect in experiments such as our example. Whereas overmodelling correlation, i.e. using a needlessly complex covariance model, reduces power. SAS PROC MIXED has become a widely used tool for analyzing repeated measures data because it allows the data analyst to choose from many different covariance structures to model correlated errors.

Because PROC MIXED allows such a wide choice of covariance models, there is always the possibility of choosing the "wrong" covariance model. Users need to be able to examine plausible covariance models given the biological context and the experiment's protocol. Often, the examination process is impeded when estimates of plausible covariance models cannot be obtained because MIXED's REML algorithm fails to converge. Failure to converge is often taken as evidence that the covariance model fits poorly. However, failure to converge also happens simply as a result of inadequate starting values; better starting values often result in convergence and may also reveal that the model is the best among those under consideration. Thus, there is a need to come up with better starting values to give those models a fair comparison.

In this paper, we present an example of a case where several biologically plausible correlated error models are fit using PROC MIXED. The models that ultimately provided the best fit initially resulted in failure of the REML algorithm to converge. We present a method for obtaining alternative starting values and show how conclusions could have been seriously distorted had these covariance models not been examined. Section 3 presents the example and Section 4 presents the method for obtaining starting values.

3. Example

We consider an experiment that used two groups of six cattle, three of each sex. There were two treatments. Treatment 1 was randomly assigned to one group; the other group of animals received treatment 2. Measurements were taken on each animal at twelve different times. The times spacing

between measurements were not equal. Also, the variances were not constant over time, indicating that covariance structures allowing heterogeneous variances over time should be considered.

Several covariance models were considered. These included heterogeneous compound symmetry (type=CSH in PROC MIXED), unstructured (UN), heterogeneous Toeplitz (TOEPH), heterogeneous auto-regressive [ARH(1)] and first-order ante-dependence [ANTE(1)] models. The Akaike Information Criterion (AIC) was used as the model-fitting criterion. [NOTE: PROC MEXED also compute Schwarz' Bayesian Information Criterion. However, it is computed incorrectly in Release 6.12 of SAS. This has been corrected in Version 7.0 and subsequent releases. This work was done with release 6.12; hence we used AIC]. Other covariance structures with homogeneous variances over time were also tried but they were not competitive in terms of adequate fit. Also, the unstructured covariance model was evaluated with various options to set upper triangle covariances to zero [UN(q)].

The results for the best fitting models are shown in Table 1. Note that ARH(1), ANTE(1), and TOEPH provide similar fit. The two unstructured models provide the next best fit. CSH (not shown) was even worse.

Originally, the researchers analyzed the data using the UN covariance model. They were able to make modest improvement by setting various upper triangle elements to zero, the optimum being UN(5). Because the parameter estimates, as well as the underlying biology and experimental protocol, suggested various heterogeneous variance models, e.g. ANTE(1) and TOEPH, these alternatives were tried but they resulted in failure of the REML algorithm to converge because the starting values yielded an "infinite likelihood" diagnostic. However, when alternative starting values were used (see Section 4 below), ANTE(1) and TOEPH not only yielded REML estimates, but the improvement in fit was dramatic, as was the impact on conclusions about treatment effects. This is important: data analysts might be tempted to take the initial failure to converge as evidence that ANTE(1) and TOEPH are unsuitable models. In doing so, and reporting results based on UN(5), the results of this experiment would be severely misinterpreted, as indicated by the impact of covariance model on the various F-statistics. As mentioned above, heterogeneous compound symmetry (CSH) was also considered. The results are not shown here because it has an even worse fit than UN. Its F-values were severely inflated.

Table 2 given the results for treatment differences and times 24, 36, and 48. The results for UN and UN(5) are virtually identical, as are the results for ARH(1), ATE(1), and TOEPH. Only the UN and ANTE(1) results are shown. Using UN – or UN(5) – one would conclude that there is insufficient evidence of a treatment effect at 36 hours and after. However, using ANTE(1) – or ARH(1) or TOEPH – one would conclude that treatments are still significantly different after 48 hours. Obvious, that could have a huge impact upon one's decision regarding the value or effectiveness of the treatments.

4. Starting Values

In many cases, starting values drastically affect the fitting of different covariance structures. As mentioned above, an unfortunate choice of starting value may lead one to discard a desirable model.

SAS PROC MIXED uses Restricted Maximum Likelihood (REML) to estimate variance and covariance parameters. Depending on the covariance model. REML can be very sensitive to starting values. Poor starting values can result in failure of convergence. Heterogeneous variance models, e.g. TOEPH and ANTE(1), are especially sensitive to choice of starting value. There are three methods one can use to obtain starting values. The default in PROC MIXED uses MIVQUE0. A commonly used PROC MIXED option is OLS, which sets' starting values of all variances at 1 and all covariances at 0. The third option is to use a PARMS statement to enter one's "best guess". These maybe obtained from previous closely related experiments or some other method. For the example in Section 3,

we obtained starting values by using the Unstructured model (UN). We chose this method since it obtains an estimate of all possible parameters and can be obtained by a number of methods, so estimates are always available.

The procedures for using the estimated UN correlation matrix to obtain starting values for the TOEPH and ANTE(1) covariance models are illustrated in Figure 3. Additional explanation follows.

To get starting values for the TOEPH model, we used the average of the correlation values that are one time period away to estimate to initial ρ_1. Then use the average of the correlation vales that are two time periods away to estimate ρ_2 and etc. We also used the estimate of the variances directly from the UN model.

For the ANTE(1) model, we obtained a starting value for ρ_i by taking the covariance $\sigma_{i,i+1}$ from the UN model and divided it by the square root of $\sigma_i^2 * \sigma_{i+1}^2$. As with TOEPH, estimates of the variances from UN were used as initial estimates of the σ_i^2.

5. Final Comments

Using the methods illustrated to get starting values for these two models can be helpful in many cases buy this method is not guaranteed to work. We did have various cases where this method did "work," i.e. REML estimates were obtained for covariance models that other SAS options yielded only failure to converge. However, we also had other cases were it did not work. Occasionally, the methods shown here "work" but the covariance models for which they worked do not fit the data as well as competing covariance structures. For example, our starting value procedure may allow TOEPH to be estimated, but the resulting model fitting criteria may be inferior to a simple model, e.g. AR(1) or CS. However, given that the choice of covariance model has a clear potential impact upon final conclusions, we conclude that all covariance models that seem reasonable, given the way the study is conducted and the underlying biology, should be examined.

Finally, we stress that convergence or failure to converge using PROC MIXED starting values, default or optional OLS, is not necessarily a good indicator of the possible fit of the model. No procedure for obtaining starting values is suited to all applications. However, the procedures present here give users another option for studying covariance models suggested by the design or biology of a study.

Table 1

Covariance Structure	AIC[1]	Sex*Time F values	Sex*Time P values	T*S*T F values	T*S*T P values
UN	-110.45	.7	.7136	.58	.8025
UN(5)	-88.782	1	.4522	.96	.4888
ANTE(1)[2]	-3.7280	1.86	.0564	2.03	.0351
ARH(1)	-5.8379	2.92	.0026	2.24	.019
TOEPH[2]	-7.6776	3.51	.0004	2.17	.0238

[1] Akaike's Information Criterion, largest value indicates "best fit"
[2] Encountered problems with convergence

Table 2

Cov Structure	Time	Difference	*Std. Er.*	*P Values*
UN[1]	24	-.471666	.15466	0.0158
	36	-.243333	.14214	0.1253
	48	-.130731	.15033	0.4098
ANTE(1)[2]	24	-.471666	.06593	0.0001
	36	-.243333	.02511	0.0001
	48	-.130731	.02676	0.0001

[1] UN(5) had similar results
[2] ARH(1) and TOEPH have very similar results

Figure 3

Derivation of Both Formulas

Unstructured (UN)

Diag↘

$$\begin{bmatrix} 1 & \rho_{21} & \rho_{31} & \rho_{41} \\ \rho_{21} & 1 & \rho_{32} & \rho_{42} \\ \rho_{31} & \rho_{32} & 1 & \rho_{43} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix}$$

$$\sum_{k=1}^{t-i} \rho_{k+1,k} = \sum_{j=1}^{t-i} \rho_i \qquad \sum_{k=1}^{t-i} \rho_{k+1,k} = (t-i)\rho_i \qquad \frac{\sum_{k=1}^{t-i} \rho_{k+1,k}}{t-i} = \rho_i$$

Heterogeneous Toeplitz (TOEPH)

Diag↘

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Unstructured (UN)

⇓

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{bmatrix}$$

$\sigma_{ji} = \sigma_i \sigma_j \rho_i$

$\sigma_{ji} / (\sigma_i^2 \sigma_j^2)^{1/2} = \rho_i$

First -Order Ante-dependence (ANTE(1))

⇓

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 \\ \sigma_1\sigma_2\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \sigma_2\sigma_3\rho_2\rho_3 \\ \sigma_1\sigma_3\rho_1\rho_2 & \sigma_2\sigma_3\rho_2 & \sigma_3^2 & \sigma_3\sigma_4\rho_3 \\ \sigma_1\sigma_4\rho_1\rho_2\rho_3 & \sigma_2\sigma_3\rho_2\rho_3 & \sigma_3\sigma_4\rho_3 & \sigma_4^2 \end{bmatrix}$$