Paper S02-2007

RANDOM and REPEATED statements - How to Use Them to Model the Covariance Structure in Proc Mixed

Charlie Liu, Dachuang Cao, Peiqi Chen, Tony Zagar

Eli Lilly & Company, Indianapolis, IN

ABSTRACT

Proc Mixed, a SAS procedure based on mixed model methodology, has been widely used for longitudinal data analyses since its release in 1992. It provides for convenient modeling of the covariance structure using RANDOM and REPEATED statements, with the RANDOM statement often used to model between-subject variation and the REPEATED statement often used to model within-subject variation. However, the proper use of the RANDOM and REPEATED statements depends on the covariance structure and might not be fully understood by the users.

This paper studies proper use of the RANDOM and REPEATED statements in Proc Mixed to model three commonly used covariance structures - unstructured (UN), compound symmetry (CS), and auto-regressive (AR(1)). Based on mathematical formula and simulation study results, using only the REPEATED statement is recommended with UN and CS. For AR, if the variance of the random between subject effects is significant, using both the RANDOM and REPEATED statements is recommended; otherwise, only the REPEATED statement should be used for proper modeling. However, simulation results showed using only the REPEATED and using both the RANDOM and REPEATED statements for AR structure had similar Type I and II error rates.

INTRODUCTION

In most clinical trials, repeated measurements are taken on the same experimental subject over time. These repeated measurements are correlated. To estimate the treatment effect, it is important to adequately model the covariance structure of the repeated measurements. The SAS procedure PROC MIXED provides such flexibility and thus has been widely used to analyze clinical trial longitudinal data since its release in 1992.

The overall variation in the data can be attributed to between subject variation (at the same time point) and within subject variation (among different time points). PROC MIXED uses the RANDOM statement to model between subject variation and the REPEATED statement to model within subject variation. In practice, we have seen cases where only one of the two statements is needed and either statement yields the same results as the other. When both RANDOM and REPEATED statements are used in a model, sometimes a note (a warning in the earlier version of SAS) would appear in the Log window stating that the "final Hessian is not positive definite" or the "estimated G matrix is not positive definite**",** depending on the covariance structure used. Furthermore, there have been numerous publications or presentations on various SUG meetings addressing how to choose the right covariance structures for the data (Wolfinger, 1993, 1996; Kincaid, 2005; Littell et el. 2006) or select the appropriate models using Proc Mixed (Ngo, 1997), but few on how to properly use the statements once the appropriate covariance structure is selected. This research investigates the appropriate use of either the REPEATED statement or the use of both the RANDOM and REPEATED statements in PROC MIXED with three covariance structures commonly used in clinical trials: unstructured (UN), compound symmetry (CS), and auto-regressive (1) (AR(1)).

The notation for the mixed model and the three commonly used covariance structures are reviewed, then recommendations are given for the proper use of the RANDOM and REPEATED statements based on mixed model theory for the three structures. Simulations are conducted to verify the recommendations for use of the RANDOM and REPEATED statements for the three structures and additional simulations are conducted for the AR(1) structure to investigate the consequences on Type I and II errors of using the REPEATED statement and both RANDOM and REPEATED statements. Finally the proper use of the statements, based on the theory and simulation results, is discussed.

MIXED MODEL NOTATION AND COVARIANCE STRUCTURES

GENERAL MIXED MODEL NOTATION

The typical linear mixed model notation is:

$$Y = X\beta + ZU + \varepsilon,$$

where β denotes fixed effects with design matrix X, U random effects with design matrix Z, and ε random error. U and ε are assumed to be Gaussian random variables that are uncorrelated and have expectations **0** and variances ***G*** and ***R***, respectively, that is U ~ N(0, G) and ε ~ N(0, R). Thus, the variance of ***y*** is Var(Y) = V = ZGZ' + R. Note that, when **R** = σ^2I and **Z** = **0**, the mixed model reduces to the standard linear model, i.e. Y = Xβ + ε.

In proc mixed, the RANDOM statement models random effects (including the random between subject variation) by setting up the **Z** and **G** matrices**,** and the REPEATED statement models the within subject variation by setting up the **R** matrix, which is the covariance structure for repeated measurements on subjects (SAS Online Doc 9.1.3). If no REPEATED statement is specified, **R** is assumed to be equal to σ^2**I**, and thus the correlation between measurements over time is constant.

AN EXAMPLE OF MIXED MODEL WITH REPEATED MEASUREMENTS

To illustrate the notation for the mixed model with repeated measurements over time, consider a clinical trial having two treatment groups (i=1, 2), 50 subjects (j=1 to 50) per treatment, and repeated measurements in 5 visits (k=1 to 5). The trial does not have random factors in design. The mixed model for the study is:

$$Y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + u_{ij} + e_{ijk},$$

where Y_{ijk} is response at time k (k=1 to 5) for the j^{th} subject (j=1 to 50) in the i^{th} group (i = 1 to 2); μ, α_i, γ_k and $(\alpha\gamma)_{ik}$ are fixed effects ; u_{ij} is the random effect corresponding to the j^{th} subject in the i^{th} group; and e_{ijk} is random error.

The variance of Y_{ijk} and covariance between Y_{ijk} and Y_{lmn} are as follows:

$$Var(Y_{ijk}) = Var(u_{ij} + e_{ijk}) = \sigma_u^2 + Var(e_{ijk})$$

$$\begin{aligned} Cov(Y_{ijk}, Y_{lmn}) &= Cov(u_{ij}, u_{lm}) + Cov(u_{ij}, e_{lmn}) + Cov(u_{lm}, e_{ijk}) + Cov(e_{ijk}, e_{lmn}) \\ &= Cov(u_{ij}, u_{ij}) + Cov(e_{ijk}, e_{ijn}) \\ &= \sigma_u^2 + Cov(e_{ijk}, e_{ijn}) \end{aligned}$$

From the above formula, we see the variance and covariance are determined by both the random subject effect (σ_u^2) and the correlation between different measurements of the same subject in the same group ($Cov(e_{ijk}, e_{ijn})$).

VARIANCES AND COVARIANCE OF THREE COMMONLY USED COVARIANCE STRUCTURES

Three commonly used covariance structures (CS, UN and AR(1)), including the parameters and the formula for the $(I,j)^{th}$ element are shown in Table 1 (SAS Online Doc 9.1.3).

Table 1: The Three Most Commonly Used Covariance Structures

Structure	Description	Parameters	$(i.j)^{th}$ element
AR(1)	Autoregressive(1)	2	$\sigma^2 \rho^{\lvert i-j \rvert}$
CS	Compound Symmetry	2	$\sigma_1 + \sigma^2 1(i=j)$
UN	Unstructured	t(t+1)/2	σ_{ij}

The variances and co-variances of the three structures are shown below:

Compound Symmetry (CS):

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

The variances are homogeneous in the CS structure and the correlation between two separate measurements is constant no matter how far apart in time the measurements are.

Unstructured (UN):

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

The variances and covariance in the UN structure are allowed to differ at and between different measurements. Among all the possible covariance structures, UN requires the most parameters to be fitted (t(t+1) /2).

Autoregressive (1):

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

The variances in AR(1) structure are homogenous, and correlations decline exponentially with time. This means the variability in a measurement is constant at different measurements times, and consecutive measurements are more highly correlated than non-consecutive measurements.

HOW TO CHOOSE RANDOM AND REPEATED STATEMENTS TO MODEL COVARIANCE?

– RESULTS FROM MATHEMATICAL FORMULA.

According to the underlying variance and covariance structure of CS, UN and AR(1), we can infer the following formula for variance and covariance:

Autoregressive (1) (AR(1)):

$Var(Y_{ijk}) = \sigma_u^2 + Var(e_{ijk}) = \sigma_u^2 + \sigma^2$

$Cov(Y_{ijk}, Y_{ijn}) = \sigma_u^2 + Cov(e_{ijk},e_{ijn}) = \sigma_u^2 + \sigma^2\rho^{|k-n|}$

There is no redundancy in the formulation, the between-subject variance σ_u^2 must be specified with the RANDOM statement and the within-subject covariance $\sigma^2\rho^{|k-n|}$ must be specified with the REPEATED statement. However, if σ_u^2 is 0 then using only the REPEATED statement is appropriate.

Compound Symmetry (CS):

$Var(Y_{ijk}) = \sigma_u^2 + Var(e_{ijk}) = \sigma_u^2 + \sigma_1 + \sigma^2$

$Cov(Y_{ijk}, Y_{ijn}) = \sigma_u^2 + Cov(e_{ijk},e_{ijn}) = \sigma_u^2 + \sigma_1$

There is a redundancy in the formulation, because $\sigma_u^2 + \sigma_1$ appear only as the sum $\sigma_u^2 + \sigma_1$, and either σ_u^2 or σ_1 must be set to zero in order to be able to estimate the other. This implies there is no need to use both RANDOM and REPEATED statements, only one of them should be adequate.

Unstructured (UN):

$Var(Y_{ijk}) = \sigma_u^2 + Var(e_{ijk}) = \sigma_u^2 + \sigma_k^2$

$Cov(Y_{ijk}, Y_{ijn}) = \sigma_u^2 + Cov(e_{ijk},e_{ijn}) = \sigma_u^2 + \sigma_{kn}$

The same as with CS, there is redundancy in the formulation. Because σ_u^2 always appears in the sum with a σ_{kn} parameter, the either σ_u^2 or σ_{kn} must be set to 0 in order to estimate the other. However, assuming $\sigma_{kn} = 0$ implies that the measurements over time are independent, which violates the nature of longitudinal measurements and the UN structure.

For the CS and UN structures, since there are redundancies in their variance and covariance formulae, we can not estimate uniquely and simultaneously the between subject random variation (σ_u^2) and the within subject variation (σ_1^2, σ_{kn}). Thus, we should not use both the RANDOM and REPEATED statements for CS and UN structures; only one is

needed. For the CS structure, using only one of the statements would produce the same results if the correlations between different measurements in time are positive, as the RANDOM specification constrains the correlation to be positive whereas the REPEATED specification leaves the correlation unconstrained (SAS OnlineDoc 9.1.3). For UN, using only the RANDOM statement, i.e. assuming the covariance components - $\sigma_{kn} = 0$, will imply the measurements over time are independent, which violate the nature of longitudinal measurement and the UN structure. Thus, only the REPEATED statement should be used.

For the AR(1) structure, there is no redundancy, and both the RANDOM and REPEATED statements can be used, but if we know that $\sigma_u^2 = 0$ then using only the REPEATED statement is appropriate. These suggestions for choosing the right statements according to the covariance structures are well described by Littell et al. (1998).

SIMULATION STUDY

In order to test the suggestions based on the mathematical formula presented above and also to answer the questions "What would happen if we fit the CS and UN models with both RANDOM and REPEATED statements?" and "what would happen if we fit the AR with only REPEATED statement?", a simulation study was conducted to compare the performance of using only the REPEATED statement and that of using both the RANDOM and the REPEATED statements for the three selected covariance structures.

SIMULATION METHODS

Data were simulated for 2 treatment groups (treated and untreated), 50 subjects per treatment group, and at 5 visits (see the detailed model described in the example shown in the earlier section). For the untreated group, the mean of the response variable was set to 0 at each of the five visits. For the treated group, the means of the response variable at the five visits were set to 0, 1, 2, 3, and 4. According to the variance-covariance values shown below, 1000 datasets were simulated for the CS structure and the AR structure, and 100 datasets for the UN structure. The small number of UN datasets was due to the computation time estimate the large number of parameters.

- CS: $\sigma^2=0.5$, $\sigma_1=1.0$
- UN
    - $\sigma_1^2=3.0$, $\sigma_{12}=2.0$, $\sigma_{13}=1.0$, $\sigma_{14}=1.0$, $\sigma_{15}=0.5$
    - $\sigma_2^2=4.0$, $\sigma_{23}=1.5$, $\sigma_{24}=0.7$, $\sigma_{25}=0.3$
    - $\sigma_3^2=3.0$, $\sigma_{34}=0.5$, $\sigma_{35}=0.2$
    - $\sigma_4^2=2.0$, $\sigma_{45}=0.6$
    - $\sigma_5^2=4.0$
- AR(1): $\sigma_u^2=6.0$, $\sigma^2=2.0$, and $\rho=0.8$

Each of the datasets was analyzed with Proc Mixed using two models: (a) only the REPEATED statement and (b) both the RANDOM and REPEATED statements. Model fit was assessed by examining the convergence information and AIC values. The following code was use.

```
proc mixed data=&dataset NOCLPRINT NOINFO COVTEST;
        class patient therapy VISIT;
        model Y&n=therapy VISIT therapy*VISIT /SOLUTION DDFM=KR;
        repeated visit / subject=patient(THERAPY) type=&datastr.;
run;
proc mixed data=&dataset COVTEST NOCLPRINT NOINFO;
        class patient therapy VISIT;
        model Y&n=therapy VISIT therapy*VISIT /SOLUTION DDFM=KR;
        random PATIENT(THERAPY)/G V;
        repeated visit / subject=patient(THERAPY) type=&datastr R;
run;
```

SIMULATION RESULTS

Compound Symmetry (CS): Using only the REPEATED statement always resulted in a better fit (smaller AIC) than using both RANDOM and REPEATED statements (Table 2). In addition, with both the RANDOM and REPEATED statements, a note stating “Converge criteria met but final Hessian is not positive definite” occurred in the Log window in more than 96% of the analyses.

Unstructured (UN): Using only the REPEATED statement resulted in better fit (smaller AIC) than using both RANDOM and REPEATED statements (Table 2). In addition, with both the RANDOM and REPEATED statements, the note “Converge criteria met but final Hessian is not positive definite” occurred in 91% of the analyses, and the model stopped running with a warning message “Warning: Stopped because of infinite likelihood” in 6% of the analyses.

For the CS and UN structures, the simulation results confirmed that using only the REPEATED statement is appropriate. For the majority of the analyses (>90%), using both the RANDOM and REPEATED statements resulted in a non-positive definite Hessian matrix. Furthermore, for the UN structure, using both the RANDOM and REPEATED statements requires estimation of a large number of variance and covariance parameters and computational problems, especially with unbalanced data.

Table 2. Simulation results for CS and UN structures.

Structure	N	Model Convergence using BOTH	Model with Smaller AIC	AIC n	Model	Mean ± S.D.	Min ,Max.
CS	962	Converge criteria met but Final Hessian is not positive definite	ONLY	962	ONLY	1328 ± 32	1232, 1411
					BOTH	1330 ± 32	1234, 1413
	38	Converge criteria met	ONLY	38	ONLY	1323 ± 29	1252, 1373
					BOTH	1325 ± 29	1254, 1375
UN	91	Converge criteria met but Final Hessian is not positive definite	ONLY	91	ONLY	1812 ± 32	1736, 1890
					BOTH	1814 ± 32	1738, 1892
	3	Converge criteria met	ONLY	3	ONLY	1790 ± 19	1769, 1807
					BOTH	1792 ± 19	1771, 1809
	6	Warning: Stopped because of Infinite likelihood					

Note: BOTH = both RANDOM and REPEATED statements, ONLY = only REPEATED statement.

Autoregressive (1) (AR(1)): Among the 1000 simulated datasets, the between subject variation was significant in 85% of the datasets, non-significant in 11% of the datasets, and 0 or missing in 3% of the datasets (Table 3). Using both the RANDOM and REPEATED statements, all analyses converged to a solution but the analyses for the 34 datasets with zero or missing between subject variation resulted in a note stating that the "Estimated G Matrix is not positive definite", which is the result when one or more variance components are equal to zero (SAS Online Doc 9.1.3).

Table 3. Simulation results for AR(1) structure

N	Between Subject Variation	Model Convergence using BOTH	Model with Smaller AIC	AIC n	Statement	Mean ± S.D	Min., Max.
852	> 0 and Significant	Converge criteria met	BOTH	643	ONLY	1540 ± 31	1448, 1624
					BOTH	1537 ± 31	1448, 1619
			ONLY	209	ONLY	1539 ± 32	1436, 1615
					BOTH	1540 ± 32	1438, 1615
114	> 0 and Non-significant	Converge criteria met	ONLY	114	ONLY	1528 ± 31	1441, 1602
					BOTH	1530 ± 31	1443, 1604
34	0 or does not exist	Converge criteria met Estimated G Matrix is not positive definite.	EQUAL	34	ONLY	1529 ± 31	1449, 1584
					BOTH	1529 ± 31	1449, 1584

Note: BOTH = both RANDOM and REPEATED statements, EQUAL = either only REPEATED or both RANDOM and REPEATED statements, ONLY = only REPEATED statement.

Among the datasets with significant between subjects variation, 75% had a better fit with both the RANDOM and REPEATED statements than with the REPEATED alone. For all datasets with non-significant between subjects variation, using the REPEATED statement alone resulted in a better fit. For the 34 datasets with 0 or missing between subject variation, using the REPEATED statement alone and using both the RANDOM and REPEATED statements resulted in the same AIC values, however, the Log note "Estimated G Matrix is not positive definite" indicated that the

RANDOM statement was not needed.

Based on the mathematical formula, using both the RANDOM and REPEATED statements should be more appropriate for the AR structure, especially when the random effect has a non-zero variance, which will be the case with significant between subject variation. The simulation study showed that if the between subject variation is significant, using both the RANDOM and REPEATED statements resulted in a better fit 75% of the time. Thus if large between subject variation is expected in a clinical trial, a test for the significance of the between subject variation should be performed and the use of both the RANDOM and REPEATED statements is recommended if the between subject variation is significant.

SIMULATION ON TYPE I & II ERRORS FOR AR(1) STRUCTURE

For the AR structure, simulations were also conducted to examine the impact of using only the REPEATED statement and using both the RANDOM and REPEATED statements on Type I and II errors. This was examined under 2 levels of between subject variation (significant and non-significant).

Simulation for Type I Error:

- Simulate 1000 AR structure datasets with $\mu 0 = \mu 1 = (0, 0, 0, 0, 0)$, $\sigma_u^2=6.0$, $\sigma^2=2.0$, and $\rho=0.8$.
- 1st run: Fit the 1000 datasets with Proc Mixed using both the RANDOM and REPEATED statements.
- 2nd run: Fit the same 1000 datasets with Proc Mixed using only the REPEATED statements.
- 3rd run: Fit the same 1000 datasets with Proc Mixed using both the RANDOM and REPEATED statements for those with significant between subject variations; and using only the REPEATED statement for those with no significant between subject variations.
- Compare the Type I error between the three runs: the simulations resulted in significant difference between treatment groups would be the cases failed to accept H_0 when H_0 is true and thus making Type I error.

Simulation for Type II Error:

- Simulate 1000 AR structure datasets with $\mu_0 = (0, 0, 0, 0, 0)$, $\mu_1 = (0, 0.80, 1.6, 2.4, 3.2)$, $\sigma_u^2=6.0$, $\sigma^2=2.0$, and $\rho=0.8$.
- 1st run: Fit the 1000 simulated datasets with Proc Mixed using both the RANDOM and REPEATED statements.
- 2nd run: Fit the same 1000 datasets with Proc Mixed using only the REPEATED statements.
- 3rd run: Fit the same 1000 datasets with Proc Mixed using both RANDOM and REPEATED statements for those with significant between subject variations; and using only the REPEATED statement for those with no significant between subject variations.
- Check and compare the Type-II error and power between the three runs: the simulations resulted in no

significant difference between treatment groups would be the cases failed to reject H_0 when H_a is true and thus making Type II error, and power = 1 – Type II error.

The type I error rate was 5.7% when using both the RANDOM and REPEATED statements, 5.8% when using the REPEATED statement alone, and 5.5% with the choice of the model depending on between subject variation. The type II error rate was 18.4% (power = 81.6%) when using the REPEATED statement alone, 18.1% (power = 81.9%) with both the RANDOM and REPEATED statements, and 18.3% (power = 81.7%) with the choice of the model depending on between subject variation. The three methods produced similar results for Type I and II error rates. thus the impact on Type I and II errors of using REPEATED statement only or both the RANDOM and REPEATED statements is minimal for AR(1) structure.

DISCUSSION AND CONCLUSION

The results from the simulation study for CS and UN structured datasets were similar to those predicted from the mathematical formula. Using only the REPEATED statement produced better model fit based on the AIC criteria and using both RANDOM and REPEATED statements would result in over-modeling and a note stating that the “Hessian is not positive definite”..

For the AR(1) structure, the mathematical theory justifies using both RANDOM and REPEATED statements if the between subject variation is significant. The simulation study also showed that 75% of the time using both RANDOM and REPEATED statement was appropriate. However, the impact on Type I and II errors of choosing only the REPEATED or both the RANDOM and REPEATED statements is minimal.

Combining the results from the mathematical formula and simulation studies, we recommend the following:

- For CS and UN structures: use only the REPEATED statement.
- For AR(1) structure: If great variation among patients is expected, test the between subject variation and if it is significant, use both the RANDOM and REPEATED statements ; otherwise use only the REPEATED statement.

REFERENCES

Kincaid C. 2005. Guidelines for selecting the covariance structure in mixed model analysis. SUGI 30 Proceedings: Paper 198-30. Maintained at: http://www2.sas.com/proceedings/sugi30/198-30.pdf

Little, R.C. et el. 1998. Statistical analysis of repeated measures data using SAS procedures. J. Anim. Sci. 76:1216-1231.

Little, R.C. et. e.l. 2006. SAS for Mixed Models, Second Ediction. Cary, NC: SAS Institute Inc.

Ngo L. 1997. Model selection in linear mixed effects models using SAS PROC MIXED. SUGI 22 Proceedings. Maintained at: http://www2.sas.com/proceedings/sugi22/STATS/PAPER284.PDF

SAS OnlineDoc 9.1.3. http://support.sas.com/onlinedoc/913/docMainpage.jsp

Wolfinger, R.D. (1993). Covariance structure selection in general mixed models. Communications in Statistics, Simulation and Computation 22: 1079-1106.

Wolfinger, R.D. (1996). Heterogeneous variance covariance structures for repeated measures. Journal of Agricultural, Biological, and Environmental Statistics 1: 205-230.

ACKNOWLEDGMENTS

The authors would like to express thanks to Yuqin Li who helped to review and write the SAS simulation programs, and to Yongming Qu and Cindy Lee who reviewed the manuscript and provided valuable feedback.

CONTACT INFORMATION

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

Charlie Liu
Eli Lilly & Company
Lilly Corporate Center
Indianapolis, IN 46285
Work Phone: 317.655.1817
Email: Liuch@Lilly.com

Dachuang Cao
Eli Lilly & Company
Lilly Corporate Center
Indianapolis, IN 46285
Work Phone: 317. 433. 3425
Email: Cao_Dachuang@Lilly.com