Paper 262

The Effect of Missing Data on Repeated Measures Models

Maribeth Johnson, Medical College of Georgia, Augusta, GA

ABSTRACT

Researchers involved with longitudinal studies are faced with the problem of trying to get study subjects to return for every follow-up visit. There is always some amount of missing data when looking at these types of studies. The MIXED procedure of the SAS® enables examination of correlational structures and variability changes between repeated measurements on experimental units across time. While PROC MIXED has the capacity to handle unbalanced data when the data are missing at random, a question arises as to when the degree of sparseness jeopardizes inference. Simulation is a tool that can be used to answer these types of questions. This paper shows the application of simulation to determine inference problems in a data set with a specific pattern of missing data. This technique is also applied to the topic of initial sample size determination.

INTRODUCTION

Researchers at the Medical College of Georgia have been collecting data on and studying children from families with a history of hypertension for a number of years. A measurement of interest is the systolic blood pressure (SBP) measurement obtained from a monitor that the child wears for 24 hours. SBP measurements are obtained every 20 minutes from 6am to 10pm and every 30 minutes during the night. Daytime and nighttime means are calculated and used in analysis. Because of the nature of these measurements not all children in the study consent to wear an ambulatory BP monitor and those that consent to wear the monitor do not do so every year of the study. When they do wear the monitor, there may be technical problems which result in an insufficient number of readings for analysis.

This resulted in a small and sparse data set when four consecutive years of data were looked at. Table 1 shows the frequency and percent of the 92 children who had at least two of four measurements and the years in which the measurements were obtained. The data set is only 57% complete.

When these data were analyzed using PROC MIXED the preferred V-C structure was determined to be compound symmetry (CS). Toeplitz (TOEP) or autoregressive (AR(1)) would seem to more intuitive given the nature of a year separation between measurements. Correlations between measurements separated by more time are expected to be lower than those between measurements obtained closer together.

A question arises as to the sufficiency of the size of the resulting sample in determining the relationship between these measures taken a year apart. There may not be sufficient power to determine the appropriate V-C structure. Simulation was used to investigate the possibility that the inferences obtained from the MIXED analyses might be due to the small sample size and/or sparseness of the data. It was also used to investigate the sample sizes needed to make correct determinations of the assumed underlying structure when problems do exist.

Table 1. Unbalanced data structure

Y_1	Y_2	Y_3	Y_4	Frequency	Percent
1	.	.	4	14	15.2
1	.	3	.	24	26.1
1	.	3	4	4	4.3
1	2	.	.	29	31.5
1	2	.	4	6	6.5
1	2	3	.	13	14.1
1	2	3	4	2	2.2

SIMULATION

The simulation problem was to generate samples of various sizes from a 4-variable multivariate normal distribution with specified mean vector and variance-covariance matrix. In a separate study of SBP in children it was determined that the mean±SD for SBP in each of 4 years was 110±10 mmHg. The correlation between measurements separated by 1 year was .70, 2 years was .60 and 3 years was .48,. following a TOEP covariance structure. Thus, the samples to be generated are of the form

$$\underline{y} = \begin{vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{vmatrix} \sim N\left[\begin{pmatrix} 110 \\ 110 \\ 110 \\ 110 \end{pmatrix}, \begin{pmatrix} 100 & 70 & 60 & 48 \\ 70 & 100 & 70 & 60 \\ 60 & 70 & 100 & 70 \\ 48 & 60 & 70 & 100 \end{pmatrix}\right].$$

For vectors $\underline{x}$ and $\underline{y}$ such that,

$$\underline{y} = B\underline{x} + \underline{b},$$

$$\underline{x} \sim N\left|\underline{\mu}_x, \sum_x\right|,$$

where **B** and $\underline{b}$ are constants, then

$$\underline{y} \sim N\left[(B\underline{\mu}_x + \underline{b}), B\sum_x B'\right].$$

If the elements of $\underline{x}$ are independent standard multivariate normal, then the variance-covariance matrix of $\underline{y}$ is

$$B\sum_x B' = B\,I\,B' = BB'.$$

Thus, **B** is the matrix resulting from a Cholesky decomposition of the desired variance-covariance matrix **BB**′. In this example, if we generate a random standard normal vector $\underline{x}$ and set

$$B = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 7 & \sqrt{51} & 0 & 0 \\ 6 & \frac{28}{51}\sqrt{51} & \frac{4}{51}\sqrt{7905} & 0 \\ \frac{24}{5} & \frac{44}{85}\sqrt{51} & \frac{227}{5270}\sqrt{7905} & \frac{1}{310}\sqrt{4673095} \end{bmatrix},$$

then,

$$\underline{y} = B \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 110 \\ 110 \\ 110 \\ 110 \end{bmatrix}$$

represents a random sample from the distribution of blood pressures.

Each run of the SAS macro simulates 1000 groups of some number of subjects (n). The random number seed is reproducible but changes for each subject within each simulation.

ANALYSIS

The results from the 1000 simulations are analyzed using PROC MIXED and three different V-C matrices: CS, TOEP, and UN. The MAKE statements were used to output the model fitting information to SAS data sets.

```
/*Transpose data into the format */
/*   needed by PROC MIXED        */

 proc transpose data=all out=allt;
    by i; var sbp1 sbp2 sbp3 sbp4;
 run;

 /* Unstructured V-C Matrix */
 proc mixed data=allt; class _name_;
   model col1=_name_;
   repeated /type=un subject=i;
   make 'fitting' out=ftun&j;
 run; quit;

/* Toeplitz V-C Matrix */
 proc mixed data=allt; class _name_;
   model col1=_name_;
   repeated /type=toep subject=i;
   make 'fitting' out=fttp&j;
 run; quit;

/* Compound Symmetric V-C Matrix */
 proc mixed data=allt; class _name_;
   model col1=_name_;
   repeated /type=cs subject=i;
   make 'fitting' out=ftcs&j;
 run; quit;
```

All of the model fit information is merged together in order to make the model comparisons. This is then appended to all previous simulations so that in the end there is one file containing the results from all 1000 simulations.

DETERMINE THE PREFERRED MODEL

Likelihood Ratio Test (LRT)

The model fitting information is used to determine the preferred covariance structure from each simulation. The results using LRTs for CS vs TOEP (LRTC_T) and for TOEP vs UN (LRTT_U) are computed. An LRT for the significance of a more general model can be constructed if one covariance model is a submodel of another by computing -2 times the difference between their log likelihoods. Then this statistic is compared to the chi-square distribution with degrees of freedom equal to the difference in the number of parameters for the two models.

If CS is preferred to TOEP and TOEP is preferred to UN then CS is the preferred model for that sample. If TOEP is preferred to both CS and UN then it is the preferred model. If TOEP is preferred to CS and UN is preferred to TOEP then UN is the preferred model.

Since Type I error for the LRTs is set at 0.05 we are looking for the sample size where TOEP is the preferred V-C structure in approximately 95% of the 1000 samples. A one-sided 95% confidence interval yields a lower limit of 93.9%. When determining initial study sample size the sample sizes were increased until the percent of samples where TOEP is preferred exceeded this lower limit.

Akaike's Information Criterion (AIC)

The model that has the largest value for AIC is the preferred model.

Schwarz's Bayesian Criterion (SBC)

The model that has the largest value for SBC is also the preferred model but SBC penalizes models with more covariance parameters more than AIC. Therefore the two criteria may not agree when assessing model fit.

In these simulations, when CS is preferred then there is an over specification of the V-C structure since the underlying structure is the more general TOEP. When UN is preferred then there is an under specification.

CREATE MISSING OBSERVATIONS

Since the subjects and observations within subjects are randomly simulated, a systematic deletion of observations still produces a random sample.

For instances where a specific pattern of missing data is desired simply write the code that will delete the appropriate values.

Primary deletion scenario

For subsequent optimal sample size determinations the 10, 20, and 25% deletions of observations were evenly distributed across the last three years and no subjects had more than one missing observation. No observations were deleted from the year 1 since all recruited subjects are assumed to have an observation in the first year.

For example, there are 600 total observations for a sample size of 150 subjects with measurements taken over four years.

For 10% deletion, 60 observations are deleted, 20 each from year 2, 3, and 4. Therefore, for the first 20 subjects the observation in year 2 is missing, for the next 20 the year 3 observation is missing and the next 20 are missing the year 4 observation. The remaining 90 subjects have no missing data.

For 20% deletion of sample size 150, 120 observations are deleted, 40 each from year 2, 3, and 4. For 25% deletion, 150 observations are deleted, 50 each from year 2, 3, and 4, i.e. each subject is missing one observation.

To determine the effect of clustered missing data, two different patterns were simulated at the optimal sample size determined above.

Crop failure scenario (CF2, CF3, CF4)

CF occurs when all the data are missing in the same year.

Using the same example of 600 observations for 150 subjects, for 10% deletion all 60 records are deleted in year 2, or year3, or year 4, i.e. three separate simulations are run.

For 20% deletion, 120 of the 150 observations are deleted in year2, then year3, and again for year 4. 25% deletion is not possible for CF since an entire year of data is lost and it would simply become a 3 year study.

Lost to follow-up scenario (LFU)

This scheme simulates the type of data most often seen when dealing with repeated measurements on human subjects. Once a person misses an appointment it is very difficult to get them to return.

For these simulations all subjects must have the first 2 years of data but years 3 and 4 can be missing. The missing records are evenly distributed between year 3 and both year 4 possibilities. Table 2 shows the data distribution with the number of subjects for the 3 levels of deletion.

For a sample size of 150 subjects and 10% deletion, 20 subjects are missing year 3 and year 4 records and 20 are missing only year 4. 110 subjects have complete records. For 20% deletion, 40 subjects are missing both year 3 and 4 and 40 are missing year 4 only. For 25% deletion, 50 year 3 records and 100 year 4 records are missing

Table 2. Data structure for LFU
Example for 600 observations of 150 subjects

Y_1	Y_2	Y_3	Y_4	10%	20%	25%
1	2	.	.	20	40	50
1	2	3	.	20	40	50
1	2	3	4	110	70	50

For all levels of deletion, two thirds of the missing data are concentrated in year 4.

RESULTS

Simulation of a specific pattern of missing data

The results from the comparison of models from the analyses of the data simulated for a sample of 92 subjects are shown in Table 3.

Using the LRT, the TOEP V-C structure was correctly determined in 922 of the 1000 simulations when there are no missing data for the 92 subjects. This rate is less than the lower error limit for the LRT. At this sample size even when the data are not missing there are inference problems.

The results are similar when using the AIC to determine the preferred model but the SBC shows that the TOEP model is preferred in only 81.6% of the simulations. This is not surprising and will be seen throughout the results section since there are 4 parameters to be estimated in the TOEP model and only 2 parameters for CS. SBC penalizes models with more covariance parameters.

Table 3. 1000 simulations – n=92
Tests of preferred models (%)

	Data Structure	
	Balanced	Specified deletions
LRT		
CS	3.6	62.3
TOEP	92.2	34.5
UN	4.2	3.3
AIC		
CS	1.1	43.0
TOEP	93.0	49.6
UN	5.9	7.4
SBC		
CS	18.4	86.7
TOEP	81.6	13.3
UN	0.0	0.0

When the data are deleted in the pattern specified by the data in question convergence of all models was attained for only 676 of the 1000 simulations. The TOEP structure is correctly determined in only 233 of these 676 simulations when using an LRT. The incorrect determination of CS as the preferred model is made the majority of the time. The TOEP structure is

correctly determined only 50% of the time when using AIC for model comparison. There are major problems with inference at this sample size when there is this much missing data.

The obvious question to ask is how large the sample size must be to overcome these problems.

Initial study sample size determination

The results from the comparison of models for the 1000 samples from the simulation and analysis of the data when no observations are missing are shown in Table 4.

A sample size of 150 subjects is required to correctly determine the underlying TOEP V-C structure within the limits of error of the LRT when no data are missing. At smaller sample sizes, the less general CS structure is incorrectly determined at a higher rate. A lack of power to distinguish the actual V-C pattern, which leads to an over specified model, is a problem when sample sizes are too small.

Table 4. 1000 simulations – No deletions
Tests of preferred models (%)

	Sample size		
	n=100	n=125	n=150
LRT			
CS	2.3	0.7	0.0
TOEP	92.6	93.9	94.6
UN	5.1	5.4	5.4
AIC			
CS	0.8	0.2	0.1
TOEP	92.7	93.1	92.2
UN	6.5	6.7	7.7
SBC			
CS	14.9	6.4	3.4
TOEP	85.1	93.6	96.6
UN	0.0	0.0	0.0

Table 5 shows the results when 10% of the observations are deleted. A sample size of 150 is no longer sufficient to determine the underlying TOEP structure within error limits of the LRT. A small increase in sample size to 185 subjects is needed to correct this problem.

Table 5. 1000 simulations – 10% deletion
Tests of preferred models (%)

	Sample size	
	n=150	n=185
LRT		
CS	0.8	0.1
TOEP	93.9	94.4
UN	5.3	5.5
AIC		
CS	0.0	0.0
TOEP	93.5	93.7
UN	6.5	6.3
SBC		
CS	8.7	4.0
TOEP	91.3	96.0
UN	0.0	0.0

The results when 20% of observations are deleted are shown in Table 6.

Only 90.4% of samples when the number of subjects is 150 determine that the preferred model is TOEP. A sample size of 185 subjects is also no longer sufficient to determine the TOEP structure within error limits of the LRT. An increase to a sample size of 225 is required.

Table 6. 1000 simulations – 20% deletion
Tests of preferred models (%)

	Sample size		
	n=150	n=185	n=225
LRT			
CS	3.3	1.2	0.2
TOEP	90.4	93.3	94.6
UN	6.3	5.5	5.3
AIC			
CS	1.1	0.3	0.1
TOEP	91.4	92.5	93.6
UN	7.5	7.2	6.3
SBC			
CS	21.7	12.8	4.7
TOEP	78.3	87.2	95.3
UN	0.0	0.0	0.0

Table 7 shows the results when 25%, or one observation per subject, is deleted.

Table 7. 1000 simulations – 25% deletion
Tests of preferred models (%)

	Sample size		
	n=150	n=225	n=250
LRT			
CS	5.2	0.9	0.5
TOEP	89.9	93.7	94.7
UN	4.9	5.4	4.8
AIC			
CS	2.2	0.1	0.1
TOEP	91.0	92.9	93.7
UN	6.8	7.0	6.2
SBC			
CS	27.4	10.5	6.3
TOEP	72.6	89.5	93.7
UN	0.0	0.0	0.0

More serious inference errors occur at the original sample size of 150, the CS structure is preferred in 5.2% of the samples. The sample size of 225 subjects is also no longer adequate in detecting the underlying V-C structure. It is determined that an increase to a sample size of 250 subjects is required to determine the underlying TOEP structure within error of the LRT.

Clustered missing data effect

In order to determine whether missing data in certain years is more harmful to inferences than evenly spaced missing data, two clustered missing data scenarios were simulated at the optimal sample size determined in the prior section.

The results of these simulations at the optimal sample size of 185 that was determined for 10% deletion are shown in Table 8.

Table 8. Optimal sample size for 10% deletion (n=185)
1000 simulations
Tests of preferred models (%)

	Missing data scenario			
	CF2	CF3	CF4	LFU
LRT				
CS	0.2	0.2	0.5	0.5
TOEP	94.0	94.7	93.9	94.7
UN	5.8	5.1	5.6	4.8
AIC				
CS	0.1	0.0	0.0	0.2
TOEP	93.1	93.4	93.3	93.6
UN	6.8	6.6	6.7	6.2
SBC				
CS	3.4	4.3	9.4	4.9
TOEP	96.6	95.7	90.6	95.1
UN	0.0	0.0	0.0	0.0

When the missing data are clustered in year 2 (CF2) or 3 (CF3) the sample size of 185 is still sufficient to determine the underlying TOEP structure within error limits of the LRT. This is also true of the LFU scenario where two-thirds of the missing data are concentrated in year 4. If all of the missing data at the 10% level is found in year 4 (CF4) then slight inference problems arise since the model using the TOEP structure is preferred in 93.9% of the simulations. Since this is the lower limit of the 95% CI of the LRT error rate, sample size should be increased until this limit is exceeded.

The results of the clustered missing data simulations at the optimal sample size of 225 determine for 20% deletion are shown in Table 9.

Table 9. Optimal sample size for 20% deletion (n=225)
1000 simulations
Tests of preferred models (%)

	Missing data scenario			
	CF2	CF3	CF4	LFU
LRT				
CS	0.0	0.0	4.6	1.1
TOEP	94.2	95.0	90.2	93.7
UN	5.8	5.0	5.2	5.2
AIC				
CS	0.0	0.0	1.2	0.3
TOEP	92.8	93.5	92.6	93.6
UN	7.2	6.5	6.2	6.1
SBC				
CS	2.6	2.4	27.6	9.8
TOEP	97.4	97.6	72.4	90.2
UN	0.0	0.0	0.0	0.0

Again we see that if 20% of the data is missing and clustered in year 2 (CF2) or 3 (CF3) then TOEP is the preferred model within error limits of the LRT at this sample size. However, if all (CF4) or two-thirds (LFU) of the missing data are clustered in year 4 then inference problems arise and sample sizes should be increased.

The results of 25% missing data under the LFU scenario are shown in Table 10.

Table 10. Optimal sample size for 25% deletion (n=250)
1000 simulations
Tests of preferred models (%)

	Missing data scenario
	LFU
LRT	
CS	1.7
TOEP	93.3
UN	5.0
AIC	
CS	0.6
TOEP	93.1
UN	6.3
SBC	
CS	16.6
TOEP	83.4
UN	0.0

When two-thirds of the missing data are clustered in year 4 then an increase in sample size is needed in order to correctly determine the TOEP structure within error limits of the LRT. At the sample size of 250 the TOEP structure is correctly determined in only 93.3% of the simulations.

When determining initial study sample size, the concentration as well as the amount of potential missing data should be considered.

CONCLUSIONS

Simulation can be a valuable tool for showing the effect that missing data can have on inferences made from repeated measures models.

Simulation can help determine if a specific pattern of missing data may be clouding the picture of the underlying V-C structure of repeated measurements. It can also help determine the sample size that is required to make correct inferences.

Initial study sample sizes are determined by power calculations under the assumption of no missing data. Even relatively small amounts of missing data require substantial increases in sample size to preserve the correct size of the LRT for inferring the correct V-C structure. The distribution and concentration of the missing data can also have an effect on the number of subjects required for correct inferences of certain models.

An understanding of the subjects being studied and the nature of the measurements being taken is needed to estimate the amount of data expected to be missing before study sample size can be determined.

REFERENCES

SAS Institute Inc. *SAS/STAT® Software: Changes and Enhancements through Release 6.11*, Cary, NC: SAS Institute Inc., 1996, 1104 pp.

Author Contact

Maribeth Johnson
Office of Biostatistics, CI-104
Medical College of Georgia
Augusta, GA 30912-4900
Phone: (706) 721-3785
E-mail: maribeth@stat.mcg.edu