The Effect of Missing Data on Sample Sizes for Repeated Measures Models

Maribeth Johnson, Medical College of Georgia, Augusta, GA
Pete Davis, University of Georgia, Athens, 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 show how to simulate sets of data where an assumption of a Toeplitz structure has been made for the variance-covariance (V-C) relationship of the repeated measurements. Then observations are systematically deleted at rates of 10%, 20%, and 25% in specific patterns. Comparisons of the suitability of the Toeplitz versus the unstructured or compound symmetric models were made using Likelihood Ratio Tests (LRTs). Sample sizes can be increased in the simulations until the underlying covariance structure is determined 95% of the time (the p-value for the LRT is set at 0.05).

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.

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. We used simulation to investigate the sample sizes needed to make correct determinations of the assumed underlying structure.

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,

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

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

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 distibution of blood pressures.

The SAS code for the simulation and subsequent analyses follows. Each run simulates 1000 groups of some number of subjects (n) (the example shows 150). The random number seed is reproducible but changes for each subject and simulation.

The global macro variable _PRINT_ was turned off to suppress the printing of the PROC MIXED output.

```
%global _PRINT_;
%let _PRINT_ = OFF;

%macro simulate;
  %do j=1 %to 1000;  * j = the simulation
                       index;

* one rep of the simulation starts here*;
data sbp;
* generate 4 independent standard normal
  random variables for each subject;
do i = 1 to 150;  * where 150 is the desired
       sample size, i is the subject index;
  x1 = rannor(647 + i + &j*99);
  x2 = rannor(372 + i + &j*99);
  x3 = rannor(425 + i + &j*99);
  x4 = rannor(162 + i + &j*99);

* transform to correlated variables;
  sbp1 = 10*x1 +110;
  sbp2 = 7*x1 + sqrt(51)*x2 + 110;
  sbp3 = 6*x1 + (28/51)*sqrt(51)*x2 +
        (4/51)*sqrt(7905)*x3 + 110;
  sbp4 = (24/5)*x1 + (44/85)*sqrt(51)*x2 +
          (227/5270)*sqrt(7905)*x3
          + (1/310)*sqrt(4673095)*x4 + 110;
output;
end;
run;
```

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.

This is also where observations are deleted.

```
  data all; set sbp;
  /* ***************************************/
  /*  Start systematic deletion of records */
  /* ***************************************/

  * Insert code here that deletes the proper
       number of observations, an example is
       shown later in this paper;

  /* ************************************ */
  /*  End systematic deletion of records  */
  /* ************************************ */

 /*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;
 /* ************************************** */
 /* Merge together the model fitting       */
 /* information from the different models  */
 /* Note: the merged dataset will be empty */
 /*       if one of the datasets does not  */
 /*       exist, this happens if a model   */
 /*       fails to converge                */
 /* ************************************** */
  data fit&j;
     merge ftcs&j(rename=(value=val_cs))
           fttp&j(rename=(value=val_toep))
           ftun&j(rename=(value=val_un));
    attrib simu length=$8;
    simu="Sim &j "; output;

  proc datasets;
    delete sbp all allt ftcs&j fttp&j ftun&j;
    append base=fit new=fit&j;
  run;

 %end;
%mend simulate;

%simulate;
run;
```

DETERMINE THE PREFERRED MODEL

The model fitting information is used to determine the preferred covariance structure from each simulation. The results using Likelihood Ratio Tests (LRT) 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.

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. 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%. Sample sizes are increased until the percent of samples where TOEP is preferred exceeds this lower limit.

The results are printed out in tabular form as well as compiled into frequency tables.

```
/* ************************************** */
/*  Set all the model fitting information  */
/*  datasets from all simulations          */
/*     Note: some may not exist due to the */
/*           nonconvergence of some models */
/* ************************************** */

  data pref; set fit;

  /* Determine the preferred model using a */
  /*     Likelihood Ratio Test             */

   /*  Compound Symmetric vs Toeplitz */

    if descr="-2 Res Log Likelihood" and
       probchi((val_cs-val_toep),2) > .95
         then lrtc_t='TOEP';
    else
    if descr="-2 Res Log Likelihood" and
       probchi((val_cs-val_toep),2) le .95
         then lrtc_t='CS  ';

   /*  Toeplitz vs Unstructured */

    if descr="-2 Res Log Likelihood" and
       probchi((val_toep-val_un),6) > .95
         then lrtt_u='UN  ';
    else
    if descr="-2 Res Log Likelihood" and
       probchi((val_toep-val_un),6) le .95
         then lrtt_u='TOEP';

title1 '1000 simulations--No deletions';
run;
title2 'Model fit information and tests of
preferred models';
run;

proc print data=pref; id descr;
 where descr="-2 Res Log Likelihood";
run;

proc freq data=pref;
   tables lrtc_t*lrtt_u / list;
run;
```

DELETE OBSERVATIONS

Since the subjects and observations within subjects are randomly simulated, a systematic deletion of observations still produces a random sample. The 10, 20, and 25% deletion 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. Sample code is as follows (remember that i is the subject index) and is inserted into the program immediately before the PROC MIXED analyses,

```
     if i le 20 then sbp2=.;
     if i ge 21 and i le 40 then sbp3=.;
     if i ge 41 and i le 60 then sbp4=.;
```

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.

RESULTS

The results from the comparison of models for the 1000 samples from the simulation and analysis of the data where no observations are missing are shown in Table 1.

Table 1. 1000 simulations--No deletions
Tests of preferred models (%)

Sample size	CS	TOEP	UN
n=100	2.3	91.9	5.1
n=125	0.7	93.7	5.6
n=135	0.4	94.4	5.2
n=140	0.3	93.1	6.6
n=145	0.2	93.7	6.1
n=150	0.0	94.6	5.4

A sample size of 150 subjects is required to correctly determine the underlying TOEP V-C structure within the limits of error 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 2 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. An increase in sample size to 185 subjects is needed to correct this problem.

Table 2. 1000 simulations--10% deletion
Tests of preferred models (%)

Sample size	CS	TOEP	UN
n=150	0.8	93.9	5.3
n=185	0.1	94.4	5.5

The results when 20% of observations are deleted is shown in Table 3.

Table 3. 1000 simulations--20% deletion
Tests of preferred models (%)

Sample size	CS	TOEP	UN
n=150	3.3	90.4	6.3
n=185	1.2	93.3	5.5
n=225	0.2	94.6	5.2

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. An increase to a sample size of 225 is required.

Table 4 shows the results when 25%, or one observation per subject, is deleted.

Table 4. 1000 simulations--25% deletion
Tests of preferred models (%)

Sample size	CS	TOEP	UN
n=150	5.2	89.9	4.9
n=225	0.9	93.7	5.4
n=250	0.5	94.7	4.8

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.

CONCLUSIONS

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

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