

## The Case of the Missing Data

Maribeth Johnson and Mark Litaker  
Medical College of Georgia, Augusta, GA

### ABSTRACT

The MIXED procedure of the SAS® system enables examination of correlational structures and variability changes between repeated measurements 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 try to answer these types of questions. This paper shows how to simulate sets of data where assumptions have been made as to the variance-covariance (V-C) structure of the repeated measurements. Then observations are systematically deleted in a pattern determined from the data in question. The results of the MIXED analyses on the balanced simulated data sets and those from the analysis of the unbalanced sets can be compared to see if and when the underlying assumed V-C structure could be determined. An application of this technique to a longitudinal study of ambulatory blood pressures in children with a family history of hypertension will be shown.

### 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. Not all children in the study consented to wear an ambulatory BP monitor and those that consented to wear the monitor did not do so every year of the study. This resulted in a small and sparse data set when we looked at four consecutive years of data. 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.

We used simulation 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.

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  $n = 92$  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  $\pm$  SD for SBP in each of 4 years would be  $110 \pm 10$  mmHg. The correlation between measurements separated by 1 year would be .70, 2 years would be .60 and 3 years would be .48. This should follow a TOEP covariance structure. In this example, the samples to be generated are of the form

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \sim N \left( \begin{bmatrix} 110 \\ 110 \\ 110 \\ 110 \end{bmatrix}, \begin{bmatrix} 100 & 70 & 60 & 48 \\ 70 & 100 & 70 & 60 \\ 60 & 70 & 100 & 70 \\ 48 & 60 & 70 & 100 \end{bmatrix} \right)$$

For the vector  $\underline{x}$ ,

$$\underline{x} \sim N[\underline{\mu}_x, \Sigma_x],$$

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

where  $B$  and  $\underline{b}$  are constants, is distributed as

$$N[(B\underline{\mu}_x + \underline{b}), B\Sigma_x B'].$$

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

$$B \Sigma_x B' = B I B' = BB'.$$

Thus,  $B$  is the matrix resulting from a Cholesky decomposition of the desired variance-covariance matrix  $BB'$ . In this example,

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

Therefore,

$$\underline{y} = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 7 & \sqrt{51} & 0 & 0 \\ 6 & \frac{28}{51}\sqrt{51} & \frac{4}{51}\sqrt{7905} & 0 \\ \frac{24}{5} & \frac{44}{85}\sqrt{51} & \frac{227}{5270}\sqrt{7905} & \frac{1}{310}\sqrt{4673095} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 110 \\ 110 \\ 110 \\ 110 \end{bmatrix}$$

The SAS code for the simulation follows.

```
* one replication of the simulation *;
data sbp;
* generate 4 independent standard normal
random variables for each subject;
do i = 1 to 92; * where 92 is the desired
sample size;
x1 = rannor(-1);
x2 = rannor(-1);
x3 = rannor(-1);
x4 = rannor(-1);

* 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;
```

```
data sbp;
set sbp;
keep i sbp1 sbp2 sbp3 sbp4;
run;
```

## ANALYSIS

The results from 10 simulations were analyzed using PROC MIXED and three different V-C matrices: CS, TOEP, and UN. The balanced data sets were analyzed as well as the data where some records were deleted to match the data structure of the longitudinal study in question.

The MAKE statements were used to output the model fitting information to SAS data sets. The global macro variable `_PRINT_` was turned off to suppress the printing of the PROC MIXED output.

```
libname sim 'c:\simu';

%global _PRINT_;
%let _PRINT_ = OFF;

%macro simmix;
%do j=1 %to 10;

data all; set sim.sbp&j;
/* *****
/* Start systematic deletion of records */
/* *****
```

```

if i ge 3 and i le 15 then do;
  sbp4=.;
end;
if i ge 16 and i le 21 then do;
  sbp3=.;
end;
if i ge 22 and i le 50 then do;
  sbp3=.;
  sbp4=.;
end;
if i ge 51 and i le 54 then do;
  sbp2=.;
end;
if i ge 55 and i le 78 then do;
  sbp2=.;
  sbp4=.;
end;
if i ge 79 then do;
  sbp2=.;
  sbp3=.;
end;
/* ***** */
/* 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=fit_un&j;
run; quit;

/* Toeplitz V-C Matrix */
proc mixed data=allt; class _name_;
  model col1=_name_;
  repeated /type=toep subject=i;
  make 'fitting' out=fit_tp&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=fit_cs&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 fit_cs&j(rename=(value=val_cs))
        fit_tp&j(rename=(value=val_toep))
        fit_un&j(rename=(value=val_un));
  simu="Sim &j "; output;

%end;
%mend simmix;

%simmix;
run;

```

### DETERMINE THE PREFERRED MODEL

The model fitting information is used to determine the preferred covariance structure from each simulation. The results using Akaike's Information Criterion (AIC), Schwarz's Bayesian Criterion (SBC) and Likelihood Ratio Tests (LRT) for CS vs TOEP (LRTC\_T) and for TOEP vs UN (LRTT\_U) are all computed. Models having the largest AIC or SBC are preferred. An LRT for the significance of a more general 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.

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

```

/* ***** */
/* 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 fit1 fit2 fit3 fit4 fit5
              fit6 fit7 fit8 fit9 fit10;

/* Determine the preferred model using */
/* Akaike's Information Criterion */

if descr="Akaike's Information Criterion"
  and (val_cs > val_toep)
  and (val_cs > val_un)
  then aic='CS ';
else
  if descr="Akaike's Information Criterion"
    and (val_toep > val_cs)

```

```

        and (val_toep > val_un)
        then aic='TOEP';
    else
    if descr='Akaike's Information Criterion'
        and (val_un > val_cs)
        and (val_un > val_toep)
        then aic='UN ';
/* Determine the preferred model using */
/* Schwarz's Bayesian Criterion */

    if descr='Schwarz's Bayesian Criterion'
        and (val_cs > val_toep)
        and (val_cs > val_un)
        then sbc='CS ';
    else
    if descr='Schwarz's Bayesian Criterion'
        and (val_toep > val_cs)
        and (val_toep > val_un)
        then sbc='TOEP';
    else
    if descr='Schwarz's Bayesian Criterion'
        and (val_un > val_cs)
        and (val_un > val_toep)
        then sbc='UN ';

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

/* Compound Symmetric vs Toeplitz */

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

/* Toeplitz vs Unstructured */

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

title1 '10 simulations--Unbalanced data';
run;
title2 'Model fit information and tests of
preferred models';
run;

proc print data=pref; id descr;
    where descr='Akaike's Information Criterion'
    or descr='Schwarz's Bayesian Criterion'

```

```

    or descr='-2 REML Log Likelihood';
run;

```

```

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

```

## RESULTS

The results from the comparison of models from the analysis of the balanced simulated data are shown in Table 2.

Table 2. 10 simulations--Balanced data  
Tests of preferred models

| AIC  | Frequency | Percent |
|------|-----------|---------|
| TOEP | 9         | 90.0    |
| UN   | 1         | 10.0    |

| SBC  | Frequency | Percent |
|------|-----------|---------|
| CS   | 3         | 30.0    |
| TOEP | 7         | 70.0    |

| LRTC_T | LRTT_U | Frequency | Percent |
|--------|--------|-----------|---------|
| TOEP   | TOEP   | 9         | 90.0    |
| TOEP   | UN     | 1         | 10.0    |

The conclusion when using both the AIC and the LRT is that TOEP is the preferred covariance structure in 9 out of the 10 simulations, with the more general UN structure preferred in one. The TOEP structure is also preferred in 7 of 10 models when using the SBC as the model comparing criteria, with the more restrictive CS preferred in three. SBC penalizes models with more covariance parameters more than does AIC.

The results from the analysis of the data where records were deleted in a specific pattern are shown in Table 3.

Convergence was attained for all models in only 7 of the 10 simulations. In each case it was the UN model that did not converge. The results for the remaining 7 simulations are mixed. Based on AIC, the CS model was preferred in 3 of 7 simulations with TOEP preferred in the other 4. The results from the LRT show that CS was the preferred model in 5 of 7 simulations and TOEP in the other 2. CS was the preferred model in all 7 simulations when using the SBC as the criteria.

Table 3. 10 simulations--Unbalanced data  
Tests of preferred models

| AIC  | Frequency | Percent |
|------|-----------|---------|
| CS   | 3         | 42.9    |
| TOEP | 4         | 57.1    |

| SBC | Frequency | Percent |
|-----|-----------|---------|
| CS  | 7         | 100.0   |

| LRTC_T | LRTT_U | Frequency | Percent |
|--------|--------|-----------|---------|
| CS     | TOEP   | 5         | 71.4    |
| TOEP   | TOEP   | 2         | 28.6    |

The entire results are shown in tabular form in the Appendix (Tables 4 and 5).

## CONCLUSIONS

When a TOEP type structure is simulated, the model comparisons obtained from the PROC MIXED analyses of the balanced data arrive at that conclusion 9 out of 10 times. The results from the model comparisons from the analyses of the unbalanced data found that not only did the models have convergence problems but that the more restrictive CS model is preferred at least half the time over the TOEP model.

Simulation can be used as a tool to help determine if the pattern of missing data may be clouding the picture of the underlying covariance structure of repeated measurements.

## APPENDIX

Table 4. 10 simulations--Balanced data  
Model fit information and tests of preferred models

| DESCR                          | VAL_CS   | VAL_TOEP | VAL_UN   | SIMU  | AIC  | SBC  | LRTC_T | LRTT_U |
|--------------------------------|----------|----------|----------|-------|------|------|--------|--------|
| Akaike's Information Criterion | -1275.27 | -1266.10 | -1270.00 | Sim 1 | TOEP |      |        |        |
| Schwarz's Bayesian Criterion   | -1279.17 | -1273.89 | -1289.49 | Sim 1 |      | TOEP |        |        |
| -2 REML Log Likelihood         | 2546.545 | 2524.197 | 2520.001 | Sim 1 |      |      | TOEP   | TOEP   |
| Akaike's Information Criterion | -1262.25 | -1259.27 | -1263.96 | Sim 2 | TOEP |      |        |        |
| Schwarz's Bayesian Criterion   | -1266.15 | -1267.06 | -1283.45 | Sim 2 |      | CS   |        |        |
| -2 REML Log Likelihood         | 2520.497 | 2510.532 | 2507.927 | Sim 2 |      |      | TOEP   | TOEP   |
| Akaike's Information Criterion | -1298.02 | -1291.65 | -1294.93 | Sim 3 | TOEP |      |        |        |
| Schwarz's Bayesian Criterion   | -1301.92 | -1299.44 | -1314.42 | Sim 3 |      | TOEP |        |        |
| -2 REML Log Likelihood         | 2592.037 | 2575.290 | 2569.867 | Sim 3 |      |      | TOEP   | TOEP   |

## 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

|                                |          |          |          |        |      |      |      |      |
|--------------------------------|----------|----------|----------|--------|------|------|------|------|
| Akaike's Information Criterion | -1287.09 | -1281.94 | -1283.41 | Sim 4  | TOEP |      |      |      |
| Schwarz's Bayesian Criterion   | -1290.99 | -1289.74 | -1302.90 | Sim 4  |      | TOEP |      |      |
| -2 REML Log Likelihood         | 2570.180 | 2555.885 | 2546.824 | Sim 4  |      |      | TOEP | TOEP |
| Akaike's Information Criterion | -1286.39 | -1278.24 | -1279.57 | Sim 5  | TOEP |      |      |      |
| Schwarz's Bayesian Criterion   | -1290.29 | -1286.03 | -1299.05 | Sim 5  |      | TOEP |      |      |
| -2 REML Log Likelihood         | 2568.788 | 2548.477 | 2539.133 | Sim 5  |      |      | TOEP | TOEP |
| Akaike's Information Criterion | -1274.81 | -1271.70 | -1273.62 | Sim 6  | TOEP |      |      |      |
| Schwarz's Bayesian Criterion   | -1278.71 | -1279.49 | -1293.11 | Sim 6  |      | CS   |      |      |
| -2 REML Log Likelihood         | 2545.618 | 2535.398 | 2527.245 | Sim 6  |      |      | TOEP | TOEP |
| Akaike's Information Criterion | -1271.80 | -1266.26 | -1264.27 | Sim 7  | UN   |      |      |      |
| Schwarz's Bayesian Criterion   | -1275.70 | -1274.06 | -1283.76 | Sim 7  |      | TOEP |      |      |
| -2 REML Log Likelihood         | 2539.604 | 2524.528 | 2508.540 | Sim 7  |      |      | TOEP | UN   |
| Akaike's Information Criterion | -1265.11 | -1257.71 | -1260.11 | Sim 8  | TOEP |      |      |      |
| Schwarz's Bayesian Criterion   | -1269.01 | -1265.51 | -1279.59 | Sim 8  |      | TOEP |      |      |
| -2 REML Log Likelihood         | 2526.221 | 2507.426 | 2500.212 | Sim 8  |      |      | TOEP | TOEP |
| Akaike's Information Criterion | -1270.08 | -1266.48 | -1270.99 | Sim 9  | TOEP |      |      |      |
| Schwarz's Bayesian Criterion   | -1273.98 | -1274.27 | -1290.48 | Sim 9  |      | CS   |      |      |
| -2 REML Log Likelihood         | 2536.165 | 2524.951 | 2521.990 | Sim 9  |      |      | TOEP | TOEP |
| Akaike's Information Criterion | -1274.29 | -1264.41 | -1266.79 | Sim 10 | TOEP |      |      |      |
| Schwarz's Bayesian Criterion   | -1278.19 | -1272.21 | -1286.28 | Sim 10 |      | TOEP |      |      |
| -2 REML Log Likelihood         | 2544.585 | 2520.829 | 2513.583 | Sim 10 |      |      | TOEP | TOEP |

Table 5. 10 simulations--Unbalanced data  
Model fit information and tests of preferred models

| DESCR                          | VAL_CS   | VAL_TOEP | VAL_UN   | SIMU   | AIC  | SBC | LRTC_T | LRTT_U |
|--------------------------------|----------|----------|----------|--------|------|-----|--------|--------|
| Akaike's Information Criterion | -747.439 | -746.712 | -747.138 | Sim 1  | TOEP |     |        |        |
| Schwarz's Bayesian Criterion   | -750.772 | -753.378 | -763.802 | Sim 1  |      | CS  |        |        |
| -2 REML Log Likelihood         | 1490.878 | 1485.424 | 1474.277 | Sim 1  |      |     | CS     | TOEP   |
| Akaike's Information Criterion | -744.995 | -744.794 | -749.400 | Sim 2  | TOEP |     |        |        |
| Schwarz's Bayesian Criterion   | -748.328 | -751.459 | -766.063 | Sim 2  |      | CS  |        |        |
| -2 REML Log Likelihood         | 1485.991 | 1481.587 | 1478.799 | Sim 2  |      |     | CS     | TOEP   |
| Akaike's Information Criterion | -754.438 | -754.973 | -757.596 | Sim 3  | CS   |     |        |        |
| Schwarz's Bayesian Criterion   | -757.771 | -761.639 | -774.260 | Sim 3  |      | CS  |        |        |
| -2 REML Log Likelihood         | 1504.876 | 1501.947 | 1495.193 | Sim 3  |      |     | CS     | TOEP   |
| Akaike's Information Criterion | -756.741 | -753.679 | -756.125 | Sim 5  | TOEP |     |        |        |
| Schwarz's Bayesian Criterion   | -760.073 | -760.345 | -772.789 | Sim 5  |      | CS  |        |        |
| -2 REML Log Likelihood         | 1509.481 | 1499.358 | 1492.250 | Sim 5  |      |     | TOEP   | TOEP   |
| Akaike's Information Criterion | -744.367 | -746.197 | -746.141 | Sim 7  | CS   |     |        |        |
| Schwarz's Bayesian Criterion   | -747.700 | -752.862 | -762.804 | Sim 7  |      | CS  |        |        |
| -2 REML Log Likelihood         | 1484.735 | 1484.394 | 1472.281 | Sim 7  |      |     | CS     | TOEP   |
| Akaike's Information Criterion | -753.139 | -753.673 | -756.075 | Sim 9  | CS   |     |        |        |
| Schwarz's Bayesian Criterion   | -756.471 | -760.338 | -772.738 | Sim 9  |      | CS  |        |        |
| -2 REML Log Likelihood         | 1502.277 | 1499.346 | 1492.149 | Sim 9  |      |     | CS     | TOEP   |
| Akaike's Information Criterion | -744.407 | -742.810 | -747.129 | Sim 10 | TOEP |     |        |        |
| Schwarz's Bayesian Criterion   | -747.739 | -749.476 | -763.793 | Sim 10 |      | CS  |        |        |
| -2 REML Log Likelihood         | 1484.813 | 1477.621 | 1474.259 | Sim 10 |      |     | TOEP   | TOEP   |