Paper 201-27

A Macro for Examining the Consequences of Error Structure Misspecifications

John Ferron, University of South Florida, Tampa, FL
Kristine Y. Hogarty, University of South Florida, Tampa, FL
Melinda Hess, University of South Florida, Tampa, FL
Jeanine Romano, University of Tampa, Tampa, FL
Jeffrey D. Kromrey, University of South Florida, Tampa, FL
John D. Niles, University of South Florida, Tampa, FL
George R. Dawkins, University of South Florida, Tampa, FL
Christina Sentovich, University of South Florida, Tampa, FL

ABSTRACT

Researchers using mixed linear models often use fit criteria to select among possible covariance structures for their data. Unfortunately, fit criteria do not always lead to the correct specification of the covariance structure, and misspecification can have negative consequences for estimation and inference. A program is presented that allows researchers to explore the sensitivity of Akaike's Information Criterion (AIC) and Schwartz's Bayesian Criterion (SBC) to possible misspecifications. For these potential misspecifications, the program then allows the researcher to examine the bias in the estimation of the variance parameters and the fixed effects, and the Type I error rates for tests of fixed effects. Monte Carlo methods are employed using a SAS® macro in which data are generated using SAS/IML® and analyzed using PROC MIXED of SAS/STAT®. An illustrative example based on a longitudinal study of student achievement at one of the National Science Foundation's Urban Systemic Initiative sites is provided.

INTRODUCTION

Researchers in many fields are frequently confronted with challenges regarding the analysis of data that are inherently hierarchical in nature. This is often the case in educational research, as students are commonly nested in classes that share similar experiences; these classes are nested in schools, and the schools in school districts. In addition, repeated observations are nested within units. Such a structure requires that researchers address the issue of independence among observations and thus implement an analysis of a hierarchical nature in order to obtain reasonable estimates of the standard errors for effects. If the hierarchical structure of the data is ignored, the errors in the statistical model are not likely to be independent. Such non-independence biases estimates of standard errors and invalidates tests for statistical significance (e.g., Hopkins, 1982; Kromrey & Dickenson, 1996).

Hierarchical linear modeling (Bryk & Raudenbush, 1992; Singer, 1998), also referred to as multi-level modeling, mixed linear modeling, or growth curve modeling when one considers longitudinal data, provides an alternative way for dealing with hierarchical data structures. By modeling the hierarchical structure of the data, this approach allows for better estimation of standard errors. In addition, a hierarchical analysis facilitates the estimation of across unit variability, and allows for the exploration of factors that may explain observed differences among units.

A hierarchical model can be represented as a mixed linear model, which has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is the vector of outcome data, $\boldsymbol{\beta}$ is the vector of fixed effects, $\mathbf{X}$ and $\mathbf{Z}$ are known model matrices, $\boldsymbol{\nu}$ is a vector of random effects, and $\boldsymbol{\varepsilon}$ is a vector of errors (Henderson, 1975).

Analysts using mixed linear models must select among a variety of choices regarding the covariance structure for the data. For growth curve models, the analyst needs to decide whether individual growth curve parameters should be allowed to vary, which can be specified through the $\mathbf{G}$ matrix (the covariance matrix of $\boldsymbol{\nu}$); and what structure is most appropriate for the errors of the individual growth trajectories, which can be specified through the $\mathbf{R}$ matrix (the covariance matrix of $\boldsymbol{\varepsilon}$). Options available in PROC MIXED for the $\mathbf{R}$ matrix include diagonal, first-order autoregressive, banded, unstructured, toeplitz, banded toeplitz, compound symmetry, and first-order autoregressive plus a diagonal.

It is often not possible to know the underlying structure in advance, so researchers will often examine multiple

structures and rely on fit indices to select among possible covariance structures (Singer, 1998; Wolfinger, 1993). Among the indices commonly used are Akaike's Information Criteria (Akaike, 1974) and Schwartz's Bayesian Criterion (Schwartz, 1978).

$$AIC = \log(L) - q$$

where q is the number of covariance parameters.

$$SBC = \log(L) - (q\log(N - p))/2$$

Both AIC and SBC start with the log likelihood value and then penalize for the number of covariance parameters estimated, with SBC employing a stiffer penalty. For each of these indices values closer to zero represent better fit, so typically the model with the value closest to zero is selected.

This approach, however, does not always lead to identification of the correct covariance model, especially when data are somewhat limited. For example, with repeated measures data, it is difficult to correctly select the covariance structure when the series length is short (Ferron, Dailey, & Yi, in press; Keselman, Algina, Kowalchuk, & Wolfinger, 1998). Furthermore, misspecification can affect estimation and inference (Ferron, Dailey, & Yi, in press; Lange & Laird, 1989).

Consequently, careful researchers may wish to explore the chances of misspecification given the type of data they are analyzing, and the ramifications of plausible misspecifications. A MACRO was developed to help researchers examine and consider these issues in the context of growth curve models for longitudinal data.

EXAMPLE

Over the past few years, educational researchers have been examining the effects of systemic reform in education. Such efforts include the Statewide, Urban, and Rural Systemic Initiatives (SIs) established by the National Science Foundation (NSF). To gauge the effects of reforms like these, achievement data are collected over time and analyzed. To study the effects of NSF's reforms on one particular urban site, researchers examined science achievement data for 80 schools over a six-year period.

Growth curves containing linear and quadratic components were found to adequately reflect these data. The parameters of the growth trajectories appeared to vary across schools, and variability in these parameters was modeled as a function of years of participation in reform activities. The MIXMOD macro was used to get a sense of the probability of error structure misspecification and to explore the resulting consequences.

MACRO MIXMOD

The macro MIXMOD uses PROC IML to generate growth curve data based on user specifications. These data are then analyzed twice using PROC MIXED. For the first analysis the covariance structure is misspecified, whereas in the second analysis the covariance structure is specified correctly.

The data generation and analysis procedures are embedded in a loop so that performance can be monitored across a large numbers of samples. In particular, AIC and SBC are compared for the two models, to estimate the proportion of times these indices lead to the correct identification of the covariance structure. Also the results from the misspecified model are aggregated across samples to estimate bias in the estimates of the variance parameters and fixed effects, as well as to estimate the Type I error rates for the tests of fixed effects.

The macro MIXMOD was developed to allow flexibility in the nature of the predictor variables and the outcome variable. Code is included to generate a normally distributed predictor (x1), a dichotomous predictor (x2), and a discrete ordinal predictor (x3). Similarly, outcome variables are produced that reflect linear growth (y1) and quadratic growth (y2). The arguments supplied to the macro include the number of observational units, the number of observations for each unit, the population autocorrelation used for data generation, and the number of replications to be simulated.

However, the macro as illustrated is set up to generate data that mirror those collected in the longitudinal achievement study. In particular, the number of observational units (nobs) was set to 80 since data were gathered on 80 schools, and the series length (t) was set to 6, since achievement data were gathered for 6 years. The variable x3, which was used as a predictor in the models, was generated as an ordered categorical variable with a distribution that matched the distribution of the years of participation variable in the longitudinal study. The variable y2, which is based on quadratic growth trajectories, was used as the outcome variable since the science achievement trajectories appeared curvilinear. The growth trajectories underlying y2 were generated with random variability in the intercept, linear, and quadratic components, since the schools' growth trajectories appeared to vary in each of these ways. Finally, the errors associated with the growth trajectories were generated to follow a first-order autoregressive structure.

The first model estimated using PROC MIXED uses quadratic growth trajectories to model y2 (science achievement), and x3 (years of participation) to model variability in the growth trajectories. The modeling fails, however, to recognize that **R** is first-order autoregressive. In the second model estimated by PROC MIXED, everything is the same, except **R** is correctly specified.

```
*++++++++++++++++++++++++++++++++++++++++++++
   Inputs to the macro:
      n: number of replications to generate
      nobs: number of observational units in each sample
      t: number of time points or observations per individual
      phi: the autocorrelation in the first-level errors
++++++++++++++++++++++++++++++++++++++++++++++;

%macro mixmod (n, nobs, t, phi);

%do i=1 %to &n;
```

```
proc iml;

*++++++++++++++++++++++++++++++++++++++++++++
   This part of the program creates the initial data set,
   which contains the following variables:
        person:  individual observed
        time:  time period of observation
        time2: time squared
        x1:  normally distributed predictor variable
        x2:  dichotomous predictor variable
        x3:  ordered categorical predictor
        y1:  dependent variable, linear trajectory output
        y2:  dependent variable, quadratic trajectory output
        wave: time period of observation
++++++++++++++++++++++++++++++++++++++++++++;

 create j1 var{person time time2 x1 x2 x3 y1 y2 wave};

 *++++++++++++++++++++++++++++++++++++++++++++
      The variable time is created to range from 0 to t - 1.
      In the current example t=6, so the time vector for
      each school is {0,1,2,3,4,5}.
++++++++++++++++++++++++++++++++++++++++++++;

 do per=1 to &nobs;
   time=j(&t,1);
   do tim=1 to &t;
      time[tim,1]=tim-1;
   end;

 time2=time##2;
 wave=time;
 length=nrow(time);

*++++++++++++++++++++++++++++++++++++++++++++
      person is a variable containing a unique ID for each
      participant in the study.
++++++++++++++++++++++++++++++++++++++++++++;

 person=repeat(per,length);

*++++++++++++++++++++++++++++++++++++++++++++
      x1 is a normally distributed predictor variable
      (not used in the current modeling example)
 ++++++++++++++++++++++++++++++++++++++++++++;

 x1=repeat(rannor (0), length);

group = ranuni(0);
 if group < .095 then group = 0;
 else group = 1;

*++++++++++++++++++++++++++++++++++++++++++++
      x2 is a dichotomous predictor variable
      (not used in the current modeling example)
 ++++++++++++++++++++++++++++++++++++++++++++;

 x2 = repeat(group,length);

*++++++++++++++++++++++++++++++++++++++++++++
      x3 is an ordinal predictor variable, that is currently
      set up to have a distribution that matches the
      distribution of the of years of participation in reform
      activities, a predictor variable considered in the
      current modeling example.
++++++++++++++++++++++++++++++++++++++++++++;

 year={5};
 cohort=ranuni(0);
   if cohort<.7920 then year=4;
   if cohort<.6501 then year=3;
   if cohort<.4681 then year=2;
   if cohort<.4610 then year=0;
 x3=repeat(year, length);

 *++++++++++++++++++++++++++++++++++++++++++++
       y1 is an outcome variable based on linear growth
       trajectories where there is random variability in the
       intercepts and linear components.  The errors of the
       individual growth trajectories follow a first-order
       autoregressive model.  y1 is not used in the current
       modeling example.

       y2 is an outcome variable based on curvilinear growth
       trajectories where there is random variability in the
       intercept, linear, and quadratic components.  The
       errors of the individual growth trajectories follow a
       first-order autoregressive model.  y1 is used in the
       current modeling example for science achievement.
 ++++++++++++++++++++++++++++++++++++++++++++;

 x=round(1000000*ranuni(0));
 resid=armasim({1,&phi},0,0,1,length,x);
 intercep=repeat(normal(0),length);
 slope=rannor(0);
 y1=intercep+slope*time+resid;
 slope2=rannor(0);
 y2=intercep + slope2*time2 +slope*time + resid;

append;
end;
close j1;

*++++++++++++++++++++++++++++++++++++++++++++
      The following set of commands uses curvilinear
      growth trajectories to model y2, and x3 to model
      variability in the intercept, linear, and quadratic
      components of the growth trajectories.  The G matrix
      is specified as unstructured, allowing for random
      variation in the intercept, linear, and quadratic
      components.  The R matrix is specified as simple, a
      specification inconsistent with the data generation,
      which used a first-order autoregressive model.

      Proc Mixed is used to generate three more datasets,
      j2a, j2b and j2c, which contain the estimates of the
      variance parameters, the estimates of the fixed
      effects, and fit information, respectively.   These data
      sets are then rearranged so that the resulting data
      sets have one row, with each column containing a
      value of interest.
++++++++++++++++++++++++++++++++++++++++++++;

proc mixed data=j1;
  class person;
  model y2=time time2 x3 time*x3 time2*x3
     /s ddfm=bw;
  random int time time2
     /sub=person type=un;
  ods output CovParms=j2a
```

```
   (keep=Estimate);
  ods output SolutionF=j2b
   (keep=Estimate Probt);
  ods output FitStatistics=j2c
   (keep=Value);
  ods listing close;

proc transpose data=j2a out=j2aa
  (rename=(col1=G11 col2=G21 col3=G22
  col4=G31 col5=G32 col6=G33 col7=R11));

proc transpose data=j2c out=j2cc
  (keep=col2 col4
   rename=(col2=AIC_not
   col4=SBC_not));

data j2bb;
  set j2b;
  w=Estimate; output;
  w=Probt; output;
  drop Estimate Probt;


proc transpose data=j2bb
  out=j2bbb
  (rename=(col1=int col2=p_int
          col3=time col4=p_time
          col5=time2 col6=p_time2
          col7=x3 col8=p_x3
          col9=time_x3 col10=p_time_x3
          col11=time2_x3 col12=p_time2_x3));

*++++++++++++++++++++++++++++++++++++++++++++
      The following set of commands differs in two
      ways from the preceding set.  First, the R matrix
      is specified correctly for the model as first-order
      autoregressive.  Second, only the fit information
      is output into a data set.
 ++++++++++++++++++++++++++++++++++++++++++++;

proc mixed data=j1;
  class person wave;
  model y2=time time2 x3 time*x3 time2*x3
    /s ddfm=bw;
  random int time time2
    /sub=person type=un;
  repeated wave
    /sub=person type=ar(1);
  ods output FitStatistics=j2d
   (keep=value);
  ods listing close;

proc transpose data=j2d out=j2dd
  (keep=col2 col4
  rename=(col2=AIC_OK
  col4=SBC_OK));

*++++++++++++++++++++++++++++++++++++++++++++
      The following statements merge the output data
      sets resulting with one row of data containing the
      AIC for the incorrect model, the SBC for the
      incorrect model, the variance estimates for the
      incorrect model, the fixed effect estimates for the
      incorrect model, the p-values for the tests of the
      fixed effects in the incorrect model, the AIC for
      the correct model, and the SBC for the correct
      model.  Two new variables are then created
      which indicate whether the AIC correctly
      identified the model, and whether the SBC
      correctly identified the model.
++++++++++++++++++++++++++++++++++++++++++++;

data j3;
  merge j2aa j2bbb j2cc j2dd;
   AIC_id=0; SBC_id=0;
   if AIC_OK<AIC_not then AIC_ID=1;
   if SBC_OK<SBC_not then SBC_ID=1;
   if (AIC_OK=. or AIC_not=.)
     then AIC_ID=.;
   if (SBC_OK=. or SBC_not=.)
     then SBC_ID=.;

*++++++++++++++++++++++++++++++++++++++++++++
      This set of commands appends the data set as
      the macro loops through the specified number of
      iterations.  After 10,000 iterations the data set
      would contain 10,000 records, one based on the
      analysis of each generated data set.
 ++++++++++++++++++++++++++++++++++++++++++++;

data j4;
  set j3;
  counter=&i;

%if &i=1 %then %do;
data j5;
 set j4;
%end;

%else %do;
data j5;
 merge j5 j4;
 by counter;
%end;

%end;

*++++++++++++++++++++++++++++++++++++++++++++
      The following set of commands creates a series
      of indicator variables based on the tests of the
      fixed effects.  For each test the p-value is
      compared to .05 to indicate whether the null
      hypothesis would have been rejected.
 ++++++++++++++++++++++++++++++++++++++++++++;

data j7;
  set j5;
  int_05=0; time_05=0; time2_05=0;
   x3_05=0; time_x3_05=0; time2_x3_05=0;
  if p_int<=.05 then int_05=1;
  if p_time<=.05 then time_05=1;
  if p_time2<=.05 then time2_05=1;
  if p_x3<=.05 then x3_05=1;
  if p_time_x3 <=.05 then time_x3_05=1;
  if p_time2_x3<=.05 then time2_x3_05=1;

*++++++++++++++++++++++++++++++++++++++++++++
      Means of various parameters of the macro are
      then calculated, giving estimates of the
      probability of correct identification of the
      covariance structure, estimates of the variance
```

```
      parameters of the G and R matrices, and
      estimates of the fixed effects, as well as their
      Type Ierror rates.
 ++++++++++++++++++++++++++++++++++++++++++++;

proc means noprint data = j7;
  var AIC_ID SBC_ID
  G11 G21 G22 G31 G32 G33 R11
  int time time2 x3 time_x3 time2_x3
  int_05 time_05 time2_05 x3_05 time_x3_05
  time2_x3_05;

   output out=j8
   mean = AIC_ID SBC_ID
   G11 G21 G22 G31 G32 G33 R11
   int time time2 x3 time_x3 time2_x3
   int_05 time_05 time2_05 x3_05 time_x3_05
   time2_x3_05
   n = n_sims;
   ods listing;

*++++++++++++++++++++++++++++++++++
  The last step of the macro prints the results.
 ++++++++++++++++++++++++++++++++++;

data out;
 set j8;
file print;
put @10 'Output from MIXMOD'    /
@1 '_____________________________________________'//
@10 'Number of Iterations =' @39 n_sims 5./
@1 '_____________________________________________'//
@17 'Model Fit' @40 'Success'/
@42 'Rate'//
@3 "Akaike's Information Criteria (AIC)"
@42 AIC_ID 4.2 /
@3 "Schwartz's Bayesian Criterion (SBC)"
@42 SBC_ID 4.2//
@1 '_____________________________________________'//
@15 'Covariance Matrices'//
@9 ' Parameter' @38 'Estimate'/
@9 ' ---------' @38 '--------'//
@3 'G Covariance Matrix'/
@3 '------------------'//
@3 ' Var Intercepts (G11)' @38 G11 8.3/
@3 ' Cov Intercept & Linear (G21)' @38 G21 8.3/
@3 ' Var Linear Components (G22)'  @38 G22 8.3/
@3 ' Cov Intercept & Quadratic (G31)'  @38 G31 8.3/
@3 ' Cov Linear & Quadratic (G32)' @38 G32 8.3/
@3 ' Var Quadratic Components (G33)' @38 G33 8.3///
@3 'R Covariance Matrix'/
@3 '------------------'//
@3 ' Var Residuals (R11)' @38 R11 8.3/
@1 '_____________________________________________'//
@18 'Fixed Effects'//
@3 ' Parameter' @24 'Bias' @35 'Rejection Rate'/
@3 ' ---------' @22 '-------' @34 '---------------'//
@3 ' intercept' @20 int 8.3 @41 int_05  5.3/
@3 ' time'   @20 time 8.3 @41 time_05 5.3/
@3 ' time2' @20 time2 8.3 @41 time2_05 5.3/
@3 ' x3'  @20 x3 8.3 @41 x3_05 5.3/
@3 ' time*x3'  @20 time_x3 8.3 @41 time_x3_05 5.3/
@3 ' time2*x3'  @20 time2_x3 8.3 @41 time2_x3_05 5.3/;
run;

%mend;
```

```
*++++++++++++++++++++++++++++++++++++++++++++
 EXAMPLE:
 Based on the characteristics of the actual data used,
 input for the macro are:
   n = 10000 (number of samples to generate)
   nobs = 80 (number of observations, schools, in study)
   t = 6 (number of time points each school was observed)
   phi = -.3  (moderately low level of + autocorrelation)
++++++++++++++++++++++++++++++++++++++++++++;

%mixmod(10000,80,6,-.3);

run;
```

OUTPUT FROM MACRO MIXMOD

Table 1 provides an example of the output produced by the macro MIXMOD.

Table 1

```
         Output from MIXMOD
_____________________________________________

         Number of Iterations =       10000
_____________________________________________

              Model Fit              Success
                                       Rate

  Akaike's Information Criteria (AIC)    0.70
  Schwartz's Bayesian Criterion (SBC)    0.45

_____________________________________________

            Covariance Matrices

         Parameter                   Estimate
         ---------                   --------

  G Covariance Matrix
  -------------------

    Var Intercepts (G11)                1.438
    Cov Intercept & Linear (G21)       -0.191
    Var Linear Components (G22)         1.277
    Cov Intercept & Quadratic (G31)     0.019
    Cov Linear & Quadratic (G32)       -0.046
    Var Quadratic Components (G33)      1.009


  R Covariance Matrix
  -------------------

    Var Residuals (R11)                 0.751
_____________________________________________
```

Fixed Effects

Parameter	Bias	Rejection Rate
intercept	-0.001	0.053
time	0.001	0.050
time2	0.000	0.052
x3	0.001	0.054
time*x3	-0.000	0.052
time2*x3	0.000	0.055

In this example, it can be seen that the AIC correctly identifies the model 70% of the time and SBC correctly identifies the model only 45% of the time. Thus, with these data it seems plausible that an underlying **R** matrix that was first-order autoregressive may be incorrectly identified as simple.

The remaining sections of the output illustrate the consequences of incorrectly specifying the **R** matrix as simple. Bias can be seen in the estimates of the elements of the variance/covariance matrices. In particular, the variance in the intercepts and the variance in the linear components tend to be overestimated, while the variance in the first-level errors, residuals, tends to be underestimated.

If attention is then focused on the fixed effects there does not appear to be any bias, nor does there appear to be any substantive problems with the tests for the fixed effects. As our team of researchers was primarily interested in interpreting the fixed effects (the effect of years of participation on the growth trajectories), it was possible to proceed somewhat more comfortably in interpreting these effects.

Of course, the macro can easily be modified to explore this data context more fully. For example, the researchers could entertain the consequences of modeling **R** with a first-order autoregressive structure when a simple structure underlies the data.

Naturally, this macro can also be modified for other contexts depending on the specifications of the user. For example, one could change the number of observations, the number of time points, the form of the growth curves, or the amount of variability in the growth curves.

REFERENCES

Akaike, H. (1974). A new look at the statistical model of identification. *IEEE Transaction on Automatic Control*, *19*, 716-723.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage Publications.

Ferron, J., Dailey, R., & Yi, Q. (in press). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*, 423-447.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal, 19*, 5-18.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics: Simulation and Computation*, *27*, 591-604.

Kromrey, J. D., & Dickenson, W. B. (1996). Detecting unit of analysis problems in nested designs: Statistical power and Type I error rates of the F test for groups-within-treatments effects. *Educational and Psychological Measurement*, *56,* 215-231.

Lange, L., & Laird, N. M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*, *84*, 241-247.

Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, *6*, 461-464.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*, 323-355.

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

ACKNOWLEDGMENTS

This work was supported, in part, by the University of South Florida and the National Science Foundation, under Grant No. REC-9988080. The opinions expressed are those of the authors and do not reflect the views of the National Science Foundation or the University of South Florida.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact John Ferron at:

John Ferron
University of South Florida
4202 East Fowler Ave. EDU 162
Tampa, FL 33620
Work Phone: 813-974-5361
Fax: 813-974-4495
Email: ferron@tempest.coedu.usf.edu