Paper 3504-2015

The %IC_MIXED: A SAS® Macro to Produce Sorted Information Criteria List for PROC MIXED for Model Selection

Qinlei Huang, St. Jude Children's Research Hospital

Liang Zhu, St. Jude Children's Research Hospital

ABSTRACT

PROC MIXED is one of the most popular SAS procedures to perform longitudinal analysis or multilevel models in Epidemiology. Model selection is one of the fundamental questions in model building. One of the most popular and widely used strategies is model selection based on information criteria, such as Akaike Information Criterion (AIC) and Sawa Bayesian Information Criterion (BIC). It considers both fit and complexity, and enables multiple models to be compared simultaneously. However, there is no existing SAS procedure to perform model selection automatically based on Information Criteria for PROC MIXED, given a set of covariates. This paper provides a SAS macro `%IC_MIXED` to select a final model with the smallest value of AIC/BIC. Specifically, `%IC_MIXED` will 1) produce a complete list of all possible model specifications given a set of covariates; 2) use do loop to read in one model specification every time and save it in a macro variable; 3) execute PROC MIXED and use SAS/ODS to output AICs and BICs; 4) append all outputs and use SAS/DATA to create a sorted list of information criteria with model specifications; and 5) run PROC REPORT to produce the final summary table. Based on the sorted list of information criteria, researchers can easily identify the best model. This paper includes the macro programming language, as well as examples of the macro calls and outputs.

Keywords: Model Selection, Information Criterion, PROC MIXED, SAS/ODS, SAS Macro

INTRODUCTION

This paper presents an accessible, flexible, and modifiable SAS macro `%IC_MIXED` to fit linear mixed-effects models on correlated data and perform model selection based on information criteria. Introduction part briefly presents the statistical background of linear mixed-effects model and illustrative commands of PROC MIXED. Part I explains the framework and details of the macro program itself. It is intended for advanced users who wish to understand and/or modify the `%IC_MIXED` code. Part II describes the usage of the macro program and provides hands-on examples for different data profiles. Using this macro requires basic knowledge of SAS and thus makes automatic model selection based on information criteria for PROC MIXED available to any PC SAS user.

LINEAR MIXED-EFFECTS MODEL

Linear mixed-effects model (LMM) is an extension of general linear model (LM). Both of them work with continuous response variables and model the linear relationships between responses and explanatory variables. The general linear model assumes independent and identically distributed normal random errors. It is written as

$$y = X\beta + \varepsilon$$

Where *y* denotes the vector of *observed* outcomes, *X* is the *known* matrix of covariates; $\boldsymbol{\beta}$ is the *unknown* vector of fixed-effects parameters; and $\boldsymbol{\varepsilon}$ is the *unobserved* vector of random errors. The general linear model is a useful method for most cross-sectional data with no systematic hierarchies. However, for correlated data (i.e., longitudinal data or hierarchical data), its assumption about $\boldsymbol{\varepsilon}$ is too restrictive.

The linear mixed-effects model includes additional random effect parameters and allows for a more flexible covariance matrix of the random errors. It accommodates both correlated error terms and error terms with heterogeneous variances. The mixed model is written as

$$y = X\beta + Z\gamma + \varepsilon$$

Where everything is the same as in the general linear model except for the addition of the known design matrix *Z* and the unknown vector of random-effects parameters *γ*. The matrix Z can contain continuous or

dummy variables, just like the matrix X. γ represents parameters that are allowed to vary over subjects. γ and ε are normally distributed with

$$E\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \quad var\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}G & 0\\ 0 & R\end{bmatrix}$$

The variance of y is, therefore, V = ZGZ' + R. It is determined by the design matrix of random effects, Z, the covariance matrix of random-effects parameters, G, and the covariance matrix of random errors, R. The covariance structure for the G matrix models the error that represents the natural heterogeneity between subjects (i.e., between-subject sources of variability). The covariance structure in the R matrix models the serial correlations (i.e., within-subject sources of variability), which is directly related to the spacing of measurements.

PROC MIXED is one of the most popular SAS procedures to fit linear mixed-effects model on continuous responses for longitudinal and clustered data. There are three sources of random variations for longitudinal data: variability between subjects, serial correlations within subject, and measurement error. The first two are represented by the random effects, ZGZ', and covariance matrix R in the linear mixed-effects model. PROC MIXED addresses the between-subject variability and intra-subject correlations by 1) specifying covariance matrix R for random errors using REPEATED statement, 2) adding random effects Z and defining covariance matrix G for random effects using RANDOM statement, and 3) adding random effects and specifying covariance matrix using both RANDOM and REPEATED statements. The first model with R covariance matrix specification and with no random effect is often called covariance pattern model. The second model is usually designated as mixed model with random effects. And the third one is denoted as hybrid mixed model. They are all linear mixed-effect models.

Covariance Pattern Model

When measurement error and within-subject serial correlation account for the most variance in the error terms, covariance pattern model should be fit. Since no random effect is specified, the model is written as

$$y = X\beta + \varepsilon$$

$$E[\varepsilon] = 0 \quad var[\varepsilon] = R$$

The covariance pattern model enables correlation among observations and possible non-constant variances through the specification of the R matrix using the REPEATED statement of PROC MIXED.

The following statements invoke PROC MIXED to fit a model with y as the numeric response variable and treatment group (trt), time (time), and their interaction (trt*time) as the fixed effects. Treatment group is a classification variable and time is a numeric variable.

The REPEATED statement defines a repeated time effect (t). The unstructured covariance R matrix (type = un) is defined within each patient (subject = id) varying by group (group = group). Maximum likelihood (method = ml) is the estimation method for the covariance parameters. Kenward Roger (ddfm = kr) is the method for computing the denominator degrees of freedom for the tests of fixed effects.

```
proc mixed data = data method = ml ;
   class id trt t group ;
   model y = trt time trt*time / ddfm = kr ;
   repeated t / type = un subject = id group = group ;
run ;
```

Mixed Model with Random Effects

When measurement error and between-subject variability account for the most variance in the error terms, mixed model with random effects should be fit. Since no REPEATED statement is specified, R is assumed to be equal to $\sigma^2 I$ and the model is written as

$$y = X\beta + Z\gamma + \varepsilon$$

$$E\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \quad var\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}G & 0\\ 0 & \sigma 2\mathrm{In}\end{bmatrix}$$

Linear mixed model with random effects assumes that the vector of correlated measurements on each subject follows a linear regression model. Some of the regression parameters are population-specific (fixed-effects), but other parameters are subject-specific (random-effects). The random effect represents the deviation of a subject from the population-specific effect.

The following statements invoke PROC MIXED to fit a model with y as the numeric response variable and treatment group (trt), time (time), and their interaction (trt*time) as the fixed effects. Treatment group is a classification variable and time is a numeric variable.

The RANDOM statement defines a random intercept (int) and slope (time) model. The unstructured covariance G matrix (type = un) is defined within each patient (subject = id). Unstructured covariance matrix allows different variances for the intercept and slope and a covariance between them. Inter-independence is assumed across subjects.

```
proc mixed data = data method = ml ;
  class id trt ;
  model y = trt time trt*time / ddfm = kr ;
  random intercept time / type = un subject = id ;
run ;
```

Hybrid Mixed Model

When the variance in the error terms is a hybrid of within-subject serial correlation and between-subject variability, hybrid mixed model (mixed model with random effects and covariance pattern) should be fit. The model is written as

$$y = X\beta + Z\gamma + \varepsilon$$

$$E\begin{bmatrix}\gamma \\ \varepsilon\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix} \quad var\begin{bmatrix}\gamma \\ \varepsilon\end{bmatrix} = \begin{bmatrix}G & 0 \\ 0 & R\end{bmatrix}$$

To fit hybrid models that include both random effects and correlated errors, it is necessary to include both the RANDOM statement and the REPEATED statement.

The following commands define a linear mixed-effects model with randomly varying intercepts and slopes (random intercept time) with unstructured error covariance (type = un), within-subject errors with a first-order autoregressive covariance (type = ar(1)) varying by group, and independent measurement errors using PROC MIXED.

```
proc mixed data = data method = ml ;
  class id trt group t ;
  model y = trt time trt*time /solution ddfm = kr ;
  repeated t / type = ar(1) subject = id group = group ;
  random intercept time / type = un subject = id ;
run ;
```

In sum, linear mixed-effects model assumes independent normal random effects and errors and linear relationship between response and predictors. It allows for correlated and heterogeneous random errors. The name mixed model comes from the fact that the model contains both fixed-effects parameters, β, and random-effects parameters, γ. For its flexibility, linear mixed-effects model can be applied to longitudinal and hierarchical data. Based on the sources of random errors, covariance pattern model, mixed model with random effects, and hybrid model can be easily fitted using PROC MIXED.

Linear Mixed-effects Model	Sources of Error	PROC MIXED Statement	Z	G	R
Covariance Pattern Model	within-, measurement error	RANDOM	×	×	✓
Mixed model with Random Effects	between-, measurement error	REPEATED	✓	✓	$\sigma^2 I$
Hybrid Mixed model	within-, between-, measurement error	REPEATED RANDOM	✓	✓	✓

Table 1. Linear Mixed-effects Model, Sources of Error, and PROC MIXED

INFORMATION CRITERION

Model selection based on information criteria, such as Akaike Information Criterion (AIC) and Sawa Bayesian Information Criterion (BIC), is a standard way and often recommended by reviewers. It considers both fit and complexity, and enables multiple models to be compared simultaneously.

The Akaike Information Criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. AIC deals with the trade-off between the goodness of fit and the complexity of the model. For any statistical model, the AIC value is

$AIC=2k-2\ln(L)$,

where k is the number of parameters in the model, and L is the maximized value of the likelihood function for the model.

The Bayesian Information Criterion (BIC) or Schwarz Criterion (SC) is a criterion for model selection among a finite set of models. It is closely related to the Akaike information criterion (AIC) and introduces a larger penalty term for the number of parameters in the model to solve the problem of over-fitting.

PROC MIXED reports AIC and BIC in the model fit statistics.

Fit Statistics

-2 Log Likelihood	2395.1
AIC (smaller is better)	2411.1
AICC (smaller is better)	2411.3
BIC (smaller is better)	2434.3

Output 1. Model fit statistics from PROC MIXED

MODEL SELECTION

There is no existing SAS procedure to perform model selection automatically based on information criteria for PROC MIXED given a set of covariates and / or their interactions. A SAS macro `%IC_MIXED` is written to address the issue.

The number of models to be run grows at an exponential rate with the increase of covariates. For example, for four covariates, there will be 16 (2^4) possible model specifications; seven covariates require 128 (2^7) model specifications; ten covariates need 1,024 (2^{10}) model specifications; and fifteen covariates demand 32,768 (2^{15}) model specifications.

# of covariates	# of models	# of covariates	# of models
1	1	**11**	2048
2	4	**12**	4096
3	8	**13**	8192
4	16	**14**	16 384
5	32	**15**	32 768
6	64	**16**	65 536
7	128	**17**	131 072
8	256	**18**	262 144
9	512	**19**	524 288
10	1024	**20**	1 048 576

Table 2. Number of model specifications for specific number of covariates

It is very time consuming to copy paste programs and change model specifications manually, even with an average size of covariates. Also, it may cause errors easily by typos, and thus, not reliable. The SAS macro `%IC_MIXED` offers an efficient, automatic, and reliable tool to solve the problem. It is adopted to choose a subset of covariates based on the smallest AIC or BIC.

DEVELOPMENT OF THE SAS MACRO %IC_MIXED

The macro introduced here, %IC_MIXED, is actually a group of smaller macros (%MODELCOMB, %MODELREADIN, %MIXED_REPEAT, %MIXED_RANDOM, %MIXED, %DATAAPPEND, %DATAFINAL, %REPORT_IC, and %IC_MIXED). All macros are saved in a central directory. Following conventional use, each macro is defined in the SAS program file of the same name (e.g., %MODELCOMB is defined in modelcomb.sas). Some of these macros call other macros. Here, we introduce the workflow of %IC_MIXED and main ideas behind each macro program.

WORKFLOW

Framework

%IC_MIXED produces the sorted list of information criteria via five steps:

1) Execute %MODELCOMB to produce a complete list of all possible model specifications given a set of covariates and / or their interactions;

2) Execute %MODELREADIN to read in one model specification each time and save it in a macro variable;

3) Execute one of the three macros on your selection (%MIXED/%MIXED_REPEATED/%MIXED_RANDOM) to run PROC MIXED and use SAS/ODS to output AICs and BICs;

4) Execute %DATAAPPEND and %DATAFINAL to create a sorted list of information criteria; and

5) Execute %REPORT_IC to PROC REPORT the final summary table.

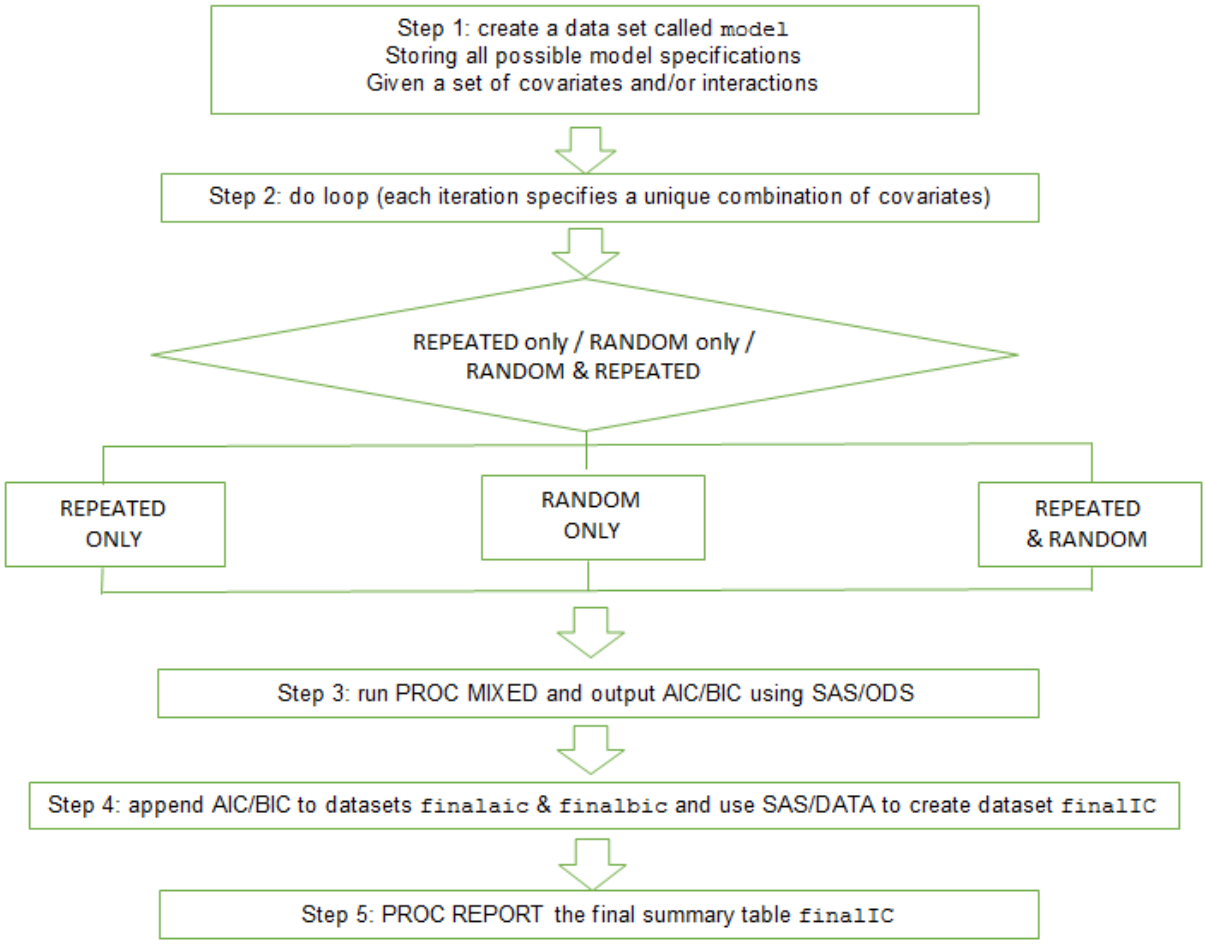

Figure 1. Flowchart for Macro %IC_MIXED

Below is the SAS code of %IC_MIXED. It considers the three typical linear mixed-effects models: covariance pattern model using REPEATED statement only, mixed model with random effects using RANDOM statement only, and hybrid model with random effects and covariance pattern using both REPEATED and RANDOM statements.

```
%macro ic_mixed (data = ,         y = ,
                 xnum = ,        n1 = ,
                 xcat = ,        n2 = ,
                 xint = ,        n3 = ,
               xforce = ,   xforcec = ,
                model = ,    method = ,
                 subj = ,     group = ,
             repeated = ,     type1 = ,
               random = ,     type2 = ) ;

   %modelcomb() ;

   proc sql noprint ;
      select count(distinct id) into: n_model from model ;
   %put total number of model specifications is &n_model ;

     ods select none ;

   %do i = 1 %to &n_model ;

      %modelReadin() ;

          * scenario 1: model with random statement ;

         %if &model = 1 %then %do ;
            %mixed_random() ;
             %end ;

          * scenario 2: model with repeat statement ;

         %if &model = 2 %then %do ;
            %mixed_repeated() ;
             %end ;

          * scenario 3: model with random & repeat statements ;

         %if &model = 3 %then %do ;
            %mixed() ;
             %end ;

      %dataAppend ;

   %end ;

   %dataFinal() ;

     ods select all ;

   %report_ic() ;

%mend ic_mixed ;
```

Step 1: `%MODELCOMB` to create a complete list of possible model specifications

The SAS macro `%modelcomb` is used to produce a full combination sheet of covariates for model selection and store it in the SAS dataset `model`.

`%modelcomb` first uses do loop, ARRAY statements, COMB and LEXCOMB functions to generate two separate sheets of distinct combinations of numeric covariates and categorical covariates. It next uses PROC SQL to create a Cartesian Product table joining the two sheets. That is, all possible combinations of rows from each table is produced.

For example, if there are 4 numeric covariates and 3 categorical covariates, `%modelcomb` will produce a sheet called '`num`' containing the 16 (i.e., 2^4) combinations of numeric covariates and a sheet called '`cat`' including the 8 (i.e., 2^3) combinations of categorical covariates. It then join '`num`' and '`cat`' to create the table '`model`' which contains 128 (i.e., 16x8) rows.

`%modelcomb` then handles interaction terms one by one. It first reads in one interaction each time and creates a Cartesian Product table of all original combinations and the interaction. The first part is the outputs of the 128 original combinations. The second part is the output of the 128 new combinations with the interaction term added in. `%modelcomb` next checks if both of the two covariates constituting the interaction are included in the original combinations. If one of them is not in the original combination, the new combination will be deleted.

`%modelcomb` then uses CATX function to return a concatenated string of selected covariates and interactions with space as delimiters. It stores the concatenated string in a column "modelvar". `%modelcomb` next uses CATX function to return a concatenated string of selected categorical covariates with space as delimiters. It stores the concatenated string in a column "classvar".

Following is the SAS code to call `%modelcomb`.

```
%modelcomb(xnum = ,n1 = ,xcat = ,n2 = ,xint = ,n3 = ,xforce = ,xforcec =);
```

- xnum: the list of all numeric covariates of interest for model selection, each enclosed inside a pair of single quotation marks and separated by space (e.g., xnum = 'age' 'time')
- n1: the total number of numeric covariates of interest for model selection
- xcat: the list of all categorical covariates of interest for model selection, each enclosed inside a pair of single quotation marks and separated by space (e.g., xcat = 'group' 'sex')
- n2: the total number of categorical covariates of interest for model selection
- xint: the list of all interaction terms of interest for model selection, each enclosed inside a pair of single quotation marks and separated by space (e.g., xint = 'group*time' 'group*sex')
- n3: the total number of interaction terms of interest for model selection
- xforce: the list of all force-in covariates and/or interactions, separated by space (e.g., xforce = group time group*time)
- xforcec: the list of categorical force-in covariates, separated by space (e.g., xforcec = group)

The below four examples are provided to give you a better idea of how `%modelcomb` works. They consider situations with or without force-in covariates and scenarios with or without interaction terms.

	Model Selection Covariates			**Force in Covariates**	
	Numeric	**Categorical**	**Interaction**	**All**	**Categorical**
Example 1	Time Age	Treatment	-	-	-
Example 2	Time	Treatment	Treatment * Time	-	-
Example 3	Time Age	Sex	-	Treatment	Treatment
Example 4	Time	Sex	Treatment * Time	Treatment	Treatment

Table 3. List of four examples for `%modelcomb`

The first example has 2 numeric covariates and 1 categorical covariate for model selection. There is no interaction and force-in covariates. 8 possible combinations are produced and stored in dataset `model` by `%modelcomb`.

```
%modelcomb(
           xnum = %str('time' 'age'),
             n1 = 2,
           xcat = %str('trt'),
             n2 = 1,
           xint = %str(),
             n3 = 0,
         xforce = %str(),
        xforcec = %str() );
```

forcevar	modelvar	classvar	id
			1
	age		2
	time		3
	age time		4
	trt	trt	5
	age trt	trt	6
	time trt	trt	7
	age time trt	trt	8

Table 4. List of model specifications for example 1

The second example has 1 numeric covariates, 1 categorical covariate, and 1 interaction for model selection. There is no force-in covariates. 5 possible combinations are produced and stored in dataset `model` by `%modelcomb`. Please note that 3 combinations 'trt*time', 'trt trt*time', and 'time trt*time' are deleted from the 8 possible combinations for incomplete main effects.

```
%modelcomb(
           xnum = %str('time'),
             n1 = 1,
           xcat = %str('trt'),
             n2 = 1,
           xint = %str('trt*time'),
             n3 = 1,
         xforce = %str(),
        xforcec = %str() );
```

forcevar	modelvar	classvar	id
			1
	trt	trt	2
	time		3
	time trt	trt	4
	time trt trt*time	trt	5

Table 5. Number of model specifications for specific number of covariates

The third example has 1 force-in covariate and 2 numeric and 1 categorical covariates for model selection. There is no interaction term. 8 possible combinations are produced and stored in dataset `model` by `%modelcomb`. Please note that the classvar include both selected and force-in categorical covariates.

```
%modelcomb (
           xnum = %str('time' 'age'),
             n1 = 2,
           xcat = %str('sex'),
             n2 = 1,
           xint = %str(),
             n3 = 0,
         xforce = %str(trt),
        xforcec = %str(trt)  ) ;
```

forcevar	modelvar	classvar	id
trt		trt	1
trt	age	trt	2
trt	time	trt	3
trt	age time	trt	4
trt	sex	sex trt	5
trt	age sex	sex trt	6
trt	time sex	sex trt	7
trt	age time sex	sex trt	8

Table 6. Number of model specifications for specific number of covariates

The fourth example has 1 force-in covariate. It has 1 numeric covariate, 1 categorical covariate, and 1 interaction for model selection. 6 possible combinations are produced and stored in dataset `model` by `%modelcomb`. You may notice that it is 6 instead of 5 possible combinations as in the second example. It is because one of the two covariates constituting the interaction term 'trt*time' is a force-in variable.

```
%modelcomb (
           xnum = %str('time'),
             n1 = 1,
           xcat = %str('sex'),
             n2 = 1,
           xint = %str('trt*time'),
             n3 = 1,
         xforce = %str(trt),
        xforcec = %str(trt)  );
```

forcevar	modelvar	classvar	id
trt		trt	1
trt	sex	sex trt	2
trt	time	trt	3
trt	time sex	sex trt	4
trt	time trt*time	trt	5
trt	time sex trt*time	sex trt	6

Table 7. Number of model specifications for specific number of covariates

Step 2: `%MODELREADIN` to read in model specification

`%modelreadin` is used to work with the `%do` loop to read in one model specification each time. It uses the CALL SYMPUT routine to assign the values in "modelvar" and "classvar" that were produced in step 1 to the macro variables `&modelvar` and `&classvar`. `&modelvar` is to be used in the PROC MIXED MODEL statement. `&classvar` is to be used in the PROC MIXED CLASS statement.

Step 3: %MIXED to run PROC MIXED and output AIC/BIC using SAS/ODS

%MIXED, %MIXED_RANDOM, and %MIXED_REPEATED are used to run PROC MIXED. %MIXED_REPEATED is written for covariance pattern model using REPEATED statement. %MIXED_RANDOM is written for mixed model with random effects using RANDOM statement. %MIXED is written for mixed model with random effects and covariance pattern using REPEATED and RANDOM statements. ODS OUTPUT is used to output the fit statistics AIC and BIC and stores them in SAS datasets AIC and BIC. All the three macros share the same strategy with minor differences. Below are the illustrative commands of %MIXED.

```
ods output FitStatistics = bic(where=(descry="BIC (smaller is better)"))
           FitStatistics = aic(where=(descr ="AIC (smaller is better)"));

proc mixed data = &data method = &method ;
   class &classvar &subj &group &repeated ;
   model &y = &xforce &modelvar / ddfm = kr ;
   repeated &repeated / type = &type1 sub = &subj group = &group ;
   random   &random   / type = &type2 sub = &subj ;
run ;
```

- data: the name of the input SAS dataset, including the name of the SAS library where the input dataset is located
- y: the name of the response variable
- classvar: the concatenated string of the categorical covariates in a model specification, separated by space
- modelvar: the concatenated string of the covariates in a model specification, separated by space
- xforce: force-in covariates
- method: the estimation method for the covariance parameters
- subj: the variable identifying each subject
- group: the categorical variable identifying heterogeneous groups
- repeated: repeated effect which should be a classification variable
- type1: type of covariance matrix R for random errors
- random: random effect which could be intercept and slopes
- type2: type of covariance matrix G for random effects

Step 4: %DATAAPPEND and %DATAFINAL to create a sorted list of ICs with model specifications

%datafinal is implemented to sort and merge the SAS datasets finalAIC and finalBIC to create the summary SAS dataset finalIC. finalIC contains the complete set of model specifications and corresponding AICs and BICs and is sorted by the values of AIC.

Step 5: %REPORT_IC to PROC REPORT the final summary table

%report_ic is used to PROC REPORT the final summary table finalIC.

```
proc report data = finalic nowindows headskip center split='#' ;
     column id aic bic forcevar modelvar ;
   define id       / display  "No." ;
   define aic      / order    "AIC" format=8.3 ;
   define bic      / analysis "BIC" format=8.3 ;
   define forcevar / display  "Force-in" ;
   define modelvar / display  "Selected" ;
run ;
```

USAGE OF THE SAS MACRO %IC_MIXED

INSTALLATION

The macro introduced here, %IC_MIXED, is actually a group of smaller macros (%MODELCOMB, %MODELREADIN, %MIXED_REPEATED, %MIXED_RANDOM, %MIXED, %DATAAPPEND, %DATAFINAL, %REPORT_IC, and %IC_MIXED). To use it within the SAS program, put these macros (all found in the IC directory) into a central directory (e.g., c:\SAS\ic). Then add the following to the SAS program before calling %IC_MIXED:

```
%let icroot = c:\SAS\IC\IC_MIXED;
options mautosource sasautos=("&icroot",sasautos);
```

This command tells SAS to look at the contents of the &icroot directory to find new macro definitions (in particular, %IC_MIXED and its component macros).

PARAMETERS

In this section, we explain how to use %IC_MIXED. Examples are provided for covariance pattern model, mixed model with random effects, and mixed model with random effects and covariance pattern. Some of the parameters might best be understood via the examples that follow.

```
%ic_mixed( data = ,          y = ,
           xnum = ,         n1 = ,
           xcat = ,         n2 = ,
           xint = ,         n3 = ,
         xforce = ,    xforcec = ,
          model = ,     method = ,
           subj = ,      group = ,
       repeated = ,      type1 = ,
         random = ,      type2 = ) ;
```

- data: the name of the input SAS dataset, including the name of the SAS library where the input dataset is located
- y: the name of the response variable
- xnum: the list of all numeric covariates of interest for model selection, each enclosed inside a pair of single quotation marks and separated by space
- n1: the total number of numeric covariates of interest for model selection
- xcat: the list of all categorical covariates of interest for model selection, each enclosed inside a pair of single quotation marks and separated by space
- n2: the total number of categorical covariates of interest for model selection
- xint: the list of all interaction terms of interest for model selection, each enclosed inside a pair of single quotation marks and separated by space
- n3: the total number of interaction terms of interest for model selection
- xforce: the list of all force-in covariates and/or interactions, separated by space
- xforcec: the list of categorical force-in covariates, separated by space
- model: 1 = covariance pattern model, 2 = mixed model with random effects, 3 = hybrid model
- method: the estimation method for the covariance parameters
- subj: the variable identifying each subject
- group: the categorical variable identifying heterogeneous groups

- repeated: repeated effect
- type1: type of covariance matrix R for random errors
- random: random effect
- type2: type of covariance matrix G for random effects

EXAMPLES

Our examples are based on an artificial dataset with eight variables: id, y, trt, time, timespl, t, age, and sex. It is designed to be a two arm clinical trial longitudinal data. Patients are randomly assigned to treatment or control group, with age and sex checked at baseline and outcome measured at five time points (baseline, 1, 2, 3, and 4 years after baseline). ID identifies subject; y is a continuous response; trt indicates treatment group; time and t represents time points; timespl is a spline variable of time; age is a numeric covariate; and sex is a binary covariate. Table 8 shows the first ten observations of the dataset.

Id	y	trt	time	timespl	t	age	sex
1	5.4	0	1	0	1	2.8	2
1	4.9	0	2	0	2	2.8	2
1	5.0	0	3	0	3	2.8	2
1	4.3	0	4	1	4	2.8	2
1	4.7	0	5	2	5	2.8	2
2	5.0	1	1	0	1	5.6	2
2	5.0	1	2	0	2	5.6	2
2	3.8	1	3	0	3	5.6	2
2	4.0	1	4	1	4	5.6	2
2	5.1	1	5	2	5	5.6	2

Table 8. Dataset for model selection using %IC_MIXED (first ten observations)

Example 1: covariance pattern model

The first example fits covariance pattern models on response y. Trt, time, and timespl are forced in. Age, sex, trt*time, and trt*timespl are of interest for model selection. A repeated time effect (repeated = t) is defined with compound symmetry R matrix (type2 = cs) within each patient (subj = id). Maximum Likelihood estimation method is selected (method = ml). Please note that since no group effect is expected, a group variable with single value ‘1’ is created for the group parameter.

```
%ic_mixed(
           data = work.example,
              y = y,
           xnum = %str('age'),
             n1 = 1,
           xcat = %str('sex'),
             n2 = 1,
           xint = %str('trt*time' 'trt*timespl'),
             n3 = 2,
         xforce = %str(trt time timespl),
        xforcec = %str(trt),
          model = 1,
         method = ml,
           subj = id,
          group = group,
       repeated = t,
          type1 = cs,
         random = ,
          type2 = ) ;
```

No.	AIC	BIC	Force-in	Selected
13	2364.033	2393.086	trt time timespl	trt*time trt*timespl
14	2365.750	2397.708	trt time timespl	sex trt*time trt*timespl
15	2365.933	2397.891	trt time timespl	age trt*time trt*timespl
16	2367.685	2402.548	trt time timespl	age sex trt*time trt*timespl
1	2369.631	2392.874	trt time timespl	
9	2370.551	2396.699	trt time timespl	trt*timespl
2	2371.323	2397.471	trt time timespl	sex
5	2371.428	2397.575	trt time timespl	trt*time
3	2371.492	2397.640	trt time timespl	age
10	2372.265	2401.317	trt time timespl	sex trt*timespl
11	2372.419	2401.472	trt time timespl	age trt*timespl
6	2373.111	2402.164	trt time timespl	sex trt*time
4	2373.227	2402.280	trt time timespl	age sex
7	2373.288	2402.341	trt time timespl	age trt*time
12	2374.173	2406.131	trt time timespl	age sex trt*timespl

Table 9. AIC-sorted information criteria for model selection: covariance pattern model

The SAS macro %IC_MIXED produces a summary dataset finalIC including sorted AIC and BIC for 16 model specifications (i.e., 2^4), from a force-in-only model (#1) to a full model (#16) including force-in and all covariates of interest (Table 9). Based on the table, model #13 (the model with interactions trt*time and trt*timespl selected) is the best model if AIC is the preferred criterion because it has the smallest AIC (2364.033). However, if BIC is the preferred criterion, then model #1 (the force-in effects only model) is the best model because it has the smallest BIC (2392.874). Determining which information criteria to use is out of the scope of this paper; readers desiring more information about choosing information criteria are referred to statistical methods articles about model selection for reference.

Example 2: mixed model with random effects

The second example fits mixed model with random effects on response y (model = 2). Trt, time, and timespl are forced in. Age, sex, trt*time, and trt*timespl are of interest for model selection. Random effects intercept and time (random = int time) are defined with independent G matrix (type2 = vc). Maximum Likelihood estimation method for covariance parameters is selected (method = ml).

```
%ic_mixed(
           data = work.example,
              y = y,
           xnum = %str('age'),
             n1 = 1,
           xcat = %str('sex'),
             n2 = 1,
           xint = %str('trt*time' 'trt*timespl'),
             n3 = 2,
         xforce = %str(trt time timespl),
        xforcec = %str(trt),
          model = 2,
         method = ml,
           subj = id,
          group = ,
       repeated = ,
          type1 = ,
         random = int time,
          type2 = vc ) ;
```

No.	AIC	BIC	Force-in	Selected
13	2389.757	2415.905	trt time timespl	trt*time trt*timespl
14	2391.428	2420.481	trt time timespl	sex trt*time trt*timespl
15	2391.434	2420.487	trt time timespl	age trt*time trt*timespl
16	2392.979	2424.937	trt time timespl	age sex trt*time trt*timespl
1	2394.754	2415.091	trt time timespl	
9	2395.343	2418.585	trt time timespl	trt*timespl
2	2396.425	2419.667	trt time timespl	sex
3	2396.475	2419.717	trt time timespl	age
5	2396.647	2419.889	trt time timespl	trt*time
10	2397.025	2423.173	trt time timespl	sex trt*timespl
11	2397.055	2423.203	trt time timespl	age trt*timespl
4	2398.030	2424.177	trt time timespl	age sex
6	2398.315	2424.463	trt time timespl	sex trt*time
7	2398.366	2424.514	trt time timespl	age trt*time
12	2398.621	2427.674	trt time timespl	age sex trt*timespl

Table 10. AIC-sorted information criteria for model selection: mixed model with random effects

The SAS macro %IC_MIXED produces a summary dataset finalIC including sorted AIC and BIC for 16 model specifications (i.e., 2^4), from a force-in-only model (#1) to a full model (#16) including force-in and all covariates of interest (Table 10). Based on the table, model #13 (the model with interactions trt*time and trt*timespl selected) is the best model if AIC is the preferred criterion because it has the smallest AIC (2389.757). However, if BIC is the preferred criterion, then model #1 (the force-in effects only model) is the best model because it has the smallest BIC (2415.091).

Example 3: hybrid model

The third example fits hybrid models with both repeated and random effects on response y (model = 3). Trt, time, and timespl are forced in. Age, sex, trt*time, and trt*timespl are of interest for model selection. A repeated time effect (repeated = t) is defined with compound symmetry R matrix (type2 = cs) within each patient (subj = id). Random effects of intercept and time (random = int time) are defined with independent G matrix (type2 = vc). Maximum Likelihood estimation method is selected (method = ml).

```
%ic_mixed(
           data = work.example,
              y = y,
           xnum = %str('age'),
             n1 = 1,
           xcat = %str('sex'),
             n2 = 1,
           xint = %str('trt*time' 'trt*timespl'),
             n3 = 2,
         xforce = %str(trt time),
        xforcec = %str(trt),
          model = 3,
         method = ml,
           subj = id,
          group = group,
       repeated = t,
          type1 = cs,
         random = int time,
          type2 = vc ) ;
```

No.	AIC	BIC	Force-in	Selected
13	2365.536	2400.399	trt time timespl	trt*time trt*timespl
14	2367.148	2404.917	trt time timespl	sex trt*time trt*timespl
15	2367.488	2405.257	trt time timespl	age trt*time trt*timespl
16	2369.129	2409.803	trt time timespl	age sex trt*time trt*timespl
1	2371.082	2400.135	trt time timespl	
9	2371.834	2403.792	trt time timespl	trt*timespl
2	2372.631	2404.589	trt time timespl	sex
5	2372.919	2404.877	trt time timespl	trt*time
3	2372.997	2404.955	trt time timespl	age
10	2373.411	2408.275	trt time timespl	sex trt*timespl
11	2373.759	2408.623	trt time timespl	age trt*timespl
6	2374.461	2409.324	trt time timespl	sex trt*time
4	2374.586	2409.450	trt time timespl	age sex
7	2374.834	2409.697	trt time timespl	age trt*time
12	2375.373	2413.142	trt time timespl	age sex trt*timespl

Table 11. AIC-sorted information criteria for model selection: hybrid model

The SAS macro %IC_MIXED produces a summary dataset finalIC including sorted AIC and BIC for 16 model specifications (i.e., 2^4), from a force-in-only model (#1) to a full model (#16) including force-in and all covariates and interactions (Table 11). Based on the table, model #13 (the model with interactions trt*time and trt*timespl selected) is the best model if AIC is the preferred criterion because it has the smallest AIC (2365.536). However, if BIC is the preferred criterion, then model #1 (the force-in effects only model) is the best model because it has the smallest BIC (2400.135).

CONCLUSION

This paper introduces the SAS macro `%IC_MIXED` to perform model selection based on information criteria for PROC MIXED in an automatic, efficient, and reliable way. `%IC_MIXED` is actually a group of smaller macros. It fits linear mixed-effects model on longitudinal and clustered data including covariance pattern model, mixed model with random effects, and hybrid model. It can also be easily modified and extended to support users' own needs.

REFERENCES

Bozdogan H. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*. 1987;52(3):345-370.

Fitzmaurice GM, Laird N, and Ware JH. *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons; 2004.

ACKNOWLEDGMENTS

My sincere gratitude goes to Dr. James Boyett and the department of Biostatistics at St. Jude Children's Research Hospital for providing an exceptionally supportive working environment and great learning resources; to Dr. Guolian Kang for her continuous support of my work; and to Dr. Kumar Srivastava and experienced biostatisticians Catherine Billups, Yinmei Zhou, Wei Liu, and Deqing Pei for their valuable comments on my work.

RECOMMENDED READING

- *Longitudinal Data Analysis with Discrete and Continuous Responses Course Notes*
- SAS® Certification Prep Guide: Advanced Programming for SAS® 9

CONTACT INFORMATION

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

Qinlei Huang
St. Jude Children's Research Hospital
Phone: 901.595.2027
E-mail: qinlei.huang@stjude.org
Web: www.stjude.org