Effects of Multicollinearity in All Possible Mixed Model Selection

George C J Fernandez, University of Nevada - Reno, Reno NV 89557

ABSTRACT

The effects of multicollinearity in all possible model selection of fixed effects including quadratic and cross products in the presence of random and repeated measures effects are presented here. The user-friendly SAS macro application ALLMIXED2 complements the model selection option currently available in the SAS macro applications ‘REGDIAG’ and ‘LOGISTIC’ for multiple linear and logistic regressions respectively. Options are also available in this macro to select the best covariance structure associated with the user-specified fully saturated repeated measures model; to graphically explore and to detect statistical significance of user specified linear, quadratic, interaction terms for fixed effects; and to diagnose multicollinearity, via the VIF statistic for each continuous predictors involved in each model selection step. The effects multicollinearity and sample size in pre-screening the variables in GLMSELECT using the LASSO selection method and in all possible subset selection within user-specified subset range are investigated using simulated repeated measures data with time independent auto-correlated error structure. A combination of two sample sizes (25 and 100 subjects) and two levels of multicollinarity (r < 0.2 and r > 0.8) among selected covariates were used in simulating 4 repeated measures data sets. Two model selection criteria, AICC (corrected Akaike Information Criterion) and MDL (minimal description length) were used in all possible model selection and summaries of the best model selection are compared graphically. The outcome of the model selection based on information criteria (AICC or MDL) was not influenced by the degree of multicollinarity. However, the sample size had a major impact on the accuracy of best candidate model selection. Instructions for downloading and running this user-friendly macro application, ALLMIXED2 are included.

INTRODUCTION

A user-friendly SAS macro application to perform all possible model selection of fixed effects including quadratic and cross products within a user-specified subset range in the presence of random and repeated measures effects using SAS PROC MIXED is available now (Fernandez, 2007). This macro application, ALLMIXED2 will complement the model selection option currently available in the SAS macro applications ‘REGDIAG’ and ‘LOGISTIC’ (Fernandez, 2002) for multiple linear and logistic regressions respectively. Options are also included in this macro to select the best covariance structure associated with the user-specified fully saturated repeated measures model; to graphically explore and to detect statistical significance of user specified linear, quadratic, interaction terms for fixed effects; and to diagnose multicollinearity, via the VIF statistic for each continuous predictors involved in each model selection step. Two model selection criteria, AICC (corrected Akaike Information Criterion) (Burnham and Anderson, 2002) and MDL (minimal description length) (Hoeting, et.al 2006) are used in all possible model selection and summaries of the best model selection are compared graphically. Complete mixed model analysis of final model including data exploration, influential diagnostics, and checking for model violations using the experimental ODS GRAPHICS option available in Version 9.13 is also implemented within this macro application. The ALLMIXED2 SAS macro application is an improved version of the SAS macro application ALLMIXED previously reported (Fernandez, 2006). Instructions for downloading and running this user-friendly macro application are included. Fernandez (2007) presented the features of this
updated ALLMIXED2 SAS macro application with enhanced features such as searching for the best
candidate models with user-specified subset ranges (e.g. between variable subsets 5 and 7 etc.). The
best subset range can be identified by enhanced model selection method such as the LASSO in the SAS
GLMSELECT procedure. However, it is not very clear how a high degree of multicollinearity can influence
the best candidate model selection. Therefore, the objective of this study is to evaluate the performance
of pre screening of variables using GLMSELECT and the best candidate mode selection based on all-
possible mixed model selection in the presence of severe multicollinearity. The following criteria were used
in evaluating the success of the selection performance:

1) The success of including all the significant linear fixed effects in the pre-screening stage using the
   GLMSELECT LASSO method.
2) The success of selecting the significant linear effects among the best candidate models in the first
   phase of all possible mixed model selection.
3) The success of selecting the significant quadratic and the interaction effects among the best candidate
   models in the second phase of all possible mixed model selection.
4) The success of not including the false effects once the true model has been identified in the validation
   phase of all possible mixed model selection.

SAMPLE SIMULATED DATA with known level of multicollinearity

Four repeated measures data with the following characteristics were generated and used in this
investigation. The attributes of the 4 simulated data sets are compared in the following table.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Data1: IND25</th>
<th>Data2: CORR25</th>
<th>Data3: IND100</th>
<th>Data4: CORR100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>25</td>
<td>25</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Correlated covariates</td>
<td>none</td>
<td>x1 x5 x10</td>
<td>none</td>
<td>x1 x5 x10</td>
</tr>
<tr>
<td>Independent covariates</td>
<td>x1-x15</td>
<td>x2 x3 x4 x6 x7 x8 x9 x11 x12 x13 x14</td>
<td>X1-x15</td>
<td>x2 x3 x4 x6 x7 x8 x9 x11 x12 x13 x14</td>
</tr>
<tr>
<td>Significant repeated Measure effect -Time</td>
<td>5 levels</td>
<td>5 levels</td>
<td>5 levels</td>
<td>5 levels</td>
</tr>
<tr>
<td>Significant Treatment effect</td>
<td>4 levels</td>
<td>4 levels</td>
<td>4 levels</td>
<td>4 levels</td>
</tr>
<tr>
<td>Repeated measures correlation</td>
<td>AR(1) &gt;0.8</td>
<td>AR(1) &gt;0.8</td>
<td>AR(1) &gt;0.8</td>
<td>AR(1) &gt;0.8</td>
</tr>
</tbody>
</table>
The degree of correlation among the 15 covariates in one of the data set, IND100 is illustrated in the scatter-plot matrix (Figure 1). This high degree of correlation between X1, X5 and X10 covariates are highlighted in plot. The experimental ODS graphics features available in the SAS PROC CORR was used to generate this scatter plot matrix.

### RESULTS

1) **The success of including all the significant linear fixed effects in the pre-screening stage using the GLMSELECT LASSO method.**

If the number of fixed effects exceeds 10, running all possible models will take very long time to complete. Therefore, under these circumstances, pre-screening is recommended to discard least contributing model terms and select the potential sub-set range. In pre-screening, the repeated measures covariance structure is ignored and all the random effects are treated as fixed. To discard the least important model terms and to select the user-specified subset model range, the LASSO (Tibshirani 1996) method implemented in the GLMSELECT (Cohen 2006) the experimental SAS procedure available in SAS version 9.13 was used. For more information of the theory and selection features refer SAS Institute (2006). The LASSO model selection options - CHOOSE=NONE and SELECT=SBC were used in this investigation to evaluate the multicollinarity effects on pre-screening. The ‘FIT CRITERIA’ and the ‘COEFFICIENT EVALUATION’ plots generated by the SAS ODS GRAPHICS features were utilized in the pre-screening evaluation to identify the potential subset ranges and to discard potential insignificant covariate and to select less than or equal to 10 potentially significant covariates.

The LASSO selection method add or drop an effect and compute several information criteria (IC) statistics in each step. The FIT CRITERIA plots display the trend of 6 IC statistics in each step and the best subset is identified by a STAR symbol (Figure 2). Because, the SELECT=SBC option was used, the FIT CRITERIA plot highlight the best subset based on the SBC criterion. In the large independent sample (IND100), the SBC criterion selected the parsimonious subset and identified the 5th step as the best while as all other Ics selected subset 8. However, in case of the large correlated data (CORR100) all six IC criteria identified step 7 as the best subset. In the small independent sample (IND25), the SBC criterion selected the 6th step as the best where as AICC and BIC identified the step 4 as the best. Thus, in case of small samples AICC and BIC favors more parsimonious models where as in large samples SBC favors parsimonious models. However, in case of the small correlated data (CORR25) the 4 IC criteria (AIC, AICC, SBC, and BIC) identified step 4 as the best subset. Because all six IC failed to select the same step as the best subset, it is not advisable to rely only one IC statistic when performing model selection.

The ‘COEFFICIENT EVALUATION’ plots displayed in Figure 3, shows the magnitude of the standardized
regression coefficients of the selected model effects in each step along with the SBC criterion. This plot could help us to discard the insignificant effects and to select less than or equal 10 covarites which can be used in the all possible mixed model selection in the next step. In case of large samples (IND100 and CORR100) all the true effects (TRT, TIME, X5 and X15) were identified as the potential candidates within the selected range of subset 8 based on the FIT CRITERIA plot. The presence of high degree of multicollinarity did not interfere in the selection of the highly correlated covariate X5 as the potential covariate. However, in case of both small samples one of the true covariate X15 was not selected as the potential effect within the best subset range 8 based on the COEFFICIENT EVALUATION plot. Even though, the highly correlated X5 had a significantly larger standardized regression coefficient (> 2) in the small correlated data (CORR25), X5 was selected as one of the potential covariate within best subset range 8.

In conclusion, the pre-screening step based on LASSO method identified all the true fixed effects in case of large samples (IND100 and CORR100) even in the presence of multicollinarity. However, incase of small samples (IND25 and CORR25) LASSO method failed to identify one of the true fixed effects within

---

**Figure 1** Scatter plot matrix showing the degree of correlation between the 15 covariates (x1-x15) in the simulated data CORR100. The highlighted correlations are greater than 0.8.
Figure 2: Comparison of the information criteria estimates computed in the LASSO method of model selection available in SAS procedure GLMSELECT. The model parameters included are two group effects (trt and time) and 15 covariates (x1-x15)
Figure 3  Comparison of standardized regression coefficient estimates and SBC computed at each model selection sequence during the LASSO method of model selection available in SAS procedure GLMSELECT. The model parameters included are two group effects (trt and time) and 15 covariates (x1-x15).
the selected subset range. Thus, small sample size has a greater impact on model selection than the multicollinearity effects.

2) The success of selecting the significant linear effects among the best candidate models in the first phase of all possible mixed model selection.

All combination of models associated with the fixed effects within the user-specified range of subsets are generated by the ALLMIXED2 macro and their information criteria statistics, AICC and MDL are compared in this step. Severe multicollinearity (Variance inflation factor (VIF) > 10) among predictor variables in mixed model analysis can result in unstable parameter estimates with inflated standard errors. When a fixed effect predictor involved in a collinear relationship is dropped from the model, the sign and size of the remaining predictor variable estimates can change dramatically. If any of the covariates are highly correlated with other predictors and involved with severe multicollinearity, then their VIF statistics can range above 10. Variance inflation statistics for each continuous predictor variables included in the model selection step were computed and compared by the box-plot display (Figure 4). If the covariates included in the model selection are independent and not involved with severe multicollinearity the range VIF statistics for these covariates should fall below 2. The VIF statistics of correlated variables X1, X5, and X10 were relatively larger than 10 in the correlated data (CORR100 and CORR25) (Figure 4).

In the all possible mixed model selection, users can optionally differentiate certain fixed effects as “MUST HAVE” and other fixed effects as “SELECTABLE” group. All combination of mixed model effects within the range of user specified subsets using the fixed effects listed in “SELECTABLE” category are generated in this step and the AICC and MDL statistics are estimated. The best candidate model in all possible subset model selection were identified by the $\Delta$ AICC (AICC$_i$ - AICC$_{\text{min}}$) $\leq 2$ or $\Delta$ MDL (MDL$_i$ - MDL$_{\text{min}}$) $\leq 1$ criteria.

The details of first phase of all possible mixed model selection include:

MUST HAVE effects included: TRT TIME TRT*TIME

SELECTABLE effects included: X5 X15 X1 X10 X4 X6 X8 X14

REPEATED STATEMENT: Repeated Time /sub=id type=AR(1)

Among the eight selectable covariates included, the model selection routine should select only the two variable model X5 and X15 as the best model. Variables X1, X5 and X10 were highly correlated and severely involved with multicollinearity in CORR100 and CORR25 data sets. Variables X1, X4, X6, X8, X10, and X14 were the non-significant covariates.

In case of large independent data set (IND100), $\Delta$ AICC failed to select the true two variable model as one of the best candidates where as $\Delta$ MDL selected the true two variable model as one of the best candidates out 3 models (Figure 5). In case of CORR100, both $\Delta$ AICC and $\Delta$ MDL selected the true two variable model as one of the best candidates. However, $\Delta$ MDL identified two best candidate models whereas $\Delta$ AICC identified only three as the best candidates.

In case of small independent (IND25) and correlated (CORR25) data set, $\Delta$ AICC failed to select the true two variable model as one of the best candidates. In small dataset, $\Delta$ AICC become very conservative and thus failed to detect the true fixed effects. Also, in small dataset $\Delta$ MDL failed to detect the true two variable model as one of the best candidates in case of small independent (IND25) and correlated (CORR25) data sets (Figure 6).
Figure 4 Comparison of the variance inflation (VIF) estimates of all the continuous predictors used in the all possible mixed model selection step within the user specified subset range.
All possible Mixed best subset selection - delta AICC

Best candidate models (delta AICC <= 2)

Fixed effects selection: x1 x5 x10 x15 x4 x6 x8 x14
Must-have fixed effects variables: trt time trt*time
repeated time /sub=id type=ar(1)

Data: Ind100  Corr100

All possible Mixed best subset selection - delta MDL

Best candidate models (delta MDL <= 1)

Fixed effects selection: x1 x5 x10 x15 x4 x6 x8 x14
Must-have fixed effects variables: trt time trt*time
repeated time /sub=id type=ar(1)

Data: Ind100  Corr100

Figure 5 Best candidate linear effects selected using delta AICC and MDL criteria in all possible mixed model selection within the user specified subset range - Large independent and correlated data (100 subjects)
Figure 6 Best candidate linear effects selected based on delta AICC and MDL criteria in all possible mixed model selection within user specified subset range - Small independent and correlated data with 25 subjects.
Furthermore, one of the best candidate model (X1 x5 x10) selected by the MDL was severely involved with high degree multicollinarity. Thus this finding clearly indicated that best candidate model selection based on AICC or MDL were not influenced by the presence of multicollinearity. Only the sample size had a greater impact on the outcome of the model selection.

3) The success of selecting the significant quadratic and the interaction effects among the best candidate models in the second phase of all possible mixed model selection.

The details of second phase of all possible mixed model selection include:

MUST HAVE effects included: TRT TIME TRT*TIME X5 X15
SELECTABLE effects included: X5*X5 X5*X15 X1 X10 X4 X6 X8 X14
REPEATED STATEMENT: Repeated Time /sub=id type=AR(1)

Among the eight selectable higher order terms and covariates included, the model selection routine should select only the two higher order terms X5*X5 X5*X15 as the best model. Variables X1, X5 and X10 were highly correlated and severely involved with multicollinearity in CORR100 and CORR25 data sets. Variables X1, X4, X6, X8, X10, and X14 were the non-significant covariates.

In large independent and correlated data sets (IND100 and CORR100), both ∆AICC and ∆MDL selected the true higher order terms (X5*X5 X5*X15) as one of the best candidates (Figure 7). However, ∆MDL identified 2 best candidate models whereas ∆AICC identified 3-4 as the best candidates.

In small independent (IND25) and correlated (CORR25) data set, ∆AICC failed to select the true two higher order terms as one of the best candidates. In small dataset, ∆AICC become very conservative and thus failed to detect the true higher order fixed effects. However, in small independent dataset (IND25) ∆MDL estimate failed to select models with only the true two higher order terms as one of the best candidates. However, incase of correlated small data (CORR25) ∆MDL estimate failed to detect the true two variable higher order fixed effects (Figure 8) Thus, the sample size had a greater impact on the outcome of the model selection.

4) The success of not including the false effects once the true model has been identified in the validation phase of all possible mixed model selection.

The details of second phase of all possible mixed model selection include:

MUST HAVE effects included: Full model: TRT TIME TRT*TIME X5 X15 X5*X5 X5*X15
SELECTABLE effects included: False effects: X1 X10 X4 X6 X8 X14
REPEATED STATEMENT: Repeated Time /sub=id type=AR(1)

In this evaluation, the must have effects included all the true effects. Therefore, among the six selectable false effects included, the model selection routine should NOT select any false covarites as the best model.

In case of large independent and correlated data sets (IND100 and CORR100), both ∆AICC and ∆MDL selected the base reference model as one of the best candidates (Figure 9). However, ∆AICC falsely detected models containing 2-3 terms as best candidate models whereas ∆MDL identified only the.
Figure 7 Best candidates linear, quadratic and interaction effects selected based on Delta AICC and MDL in all possible mixed model selection with user-specified subset range - Large independent and correlated data with 100 subjects
**Figure 8** Best candidates linear, quadratic and interaction effects selected based on Delta AICC and MDL in all possible mixed model selection with user-specified subset range - small independent and correlated data with 25 subjects
Figure 9 False effects detected based on Delta AICC and MDL after the true model specified in all possible mixed model selection with user-specified subset range - Large independent and correlated data with 100 subjects
Figure 10 False effects detected based on Delta AICC and MDL after the true model specified in all possible mixed model selection with user-specified subset range - Small independent and correlated data with 25 subjects
reference model as the best candidates. In case of small independent (IND25) and correlated (CORR25) data set, similar to other previous evaluation $\Delta$ AICC selected only the reference model as the best candidate (Figure 10). This is because in evaluating small sample $\Delta$ AICC behave very conservative and thus it correctly picked the reference model only. However, in small independent dataset (IND25) $\Delta$ MDL estimate becomes more liberal and falsely selected 1 model with false effects in addition to selecting the reference model. However, incase of correlated small data (CORR25) $\Delta$ MDL estimate failed to detect the correct reference model but selected the 7 models with false effects (Figure 10). Thus, the sample size had a greater impact on the outcome of the model selection.

AVAILABILITY OF THE ALLMIXED2 MACRO:

The main features of the user-friendly SAS macro application, ALLMIXED2 are summarized below:

- The users can input, temporary and permanent SAS data files, Microsoft Excel and Access and comma and TAB delimited text files as input data set.
- Users can input multiple response variable and perform all the model selection steps simultaneously.
- Users can optionally pre-screen the fixed effects and drop obvious non-significant fixed effects if the number of fixed effects exceed 10 using the SAS 9.1 experimental GLMSELECT procedure implemented within the macro. The new model selection method, LASSO is used in this macro to pre-screen the many fixed effects covariates.
- In case of repeated measures mixed model analysis, the best covariance structure selection from the user specified covariance structures are implemented by comparing the AICC value estimated in the PROC Mixed using REML method and then best covariance structures is graphically identified by searching for the covariance structure with the smallest AICC value.
- Options for performing all possible fixed effect model selection within the ranges of subsets identified by the user in the pre-screening step with and without repeated and random effects and selecting the best candidate models using AICC and MDL estimates using PROC MIXED method ML. In this step, users can differentiate the “must- keep” effects and “selectable” effects. The all possible model selection will be performed using the fixed effects identified in the “Selectable” list of terms.
- Options are also available for graphical exploration and statistical significance of user specified linear, quadratic, interaction terms for fixed effects. Also, to diagnose multicollinearity (when VIF value > 10) the VIF statistic for each continuous predictors involved in each model selection step are sent to an output table. Also, a box-plot display of VIF estimates by all the continuous fixed effects are generated for the overall assessment of multicollinearity in the model selection process.
- Options are also available for performing complete mixed model analysis of final model including data exploration, influential diagnostics, and checking for model violations using the experimental ODS GRAPHICS option available in Version 9.13.
- Users can save all SAS output and graphics in Word, HTML, or PDF formats. In addition, full details all model selection diagnostic statistics are automatically sent to MS excel data tables. SAS log messages are automatically saved to external text log files and only the ERROR and WARNING messages are extracted and displayed as HTML output for easy error checks.
- Download instructions are given above to download this macro-call file and to perform all possible model selection.

Users can download the ALLMIXED2.zip file containing the ALLMIXED2.SAS macro-call file, excel data sets used in the simulation, and the ALLMIXED2 help files from the authors website at http://www.ag.unr.edu/gf by clicking the “Running puppy dog” clip art. Save the ALLMIXED2.SAS macro-call file in your PC first and open it in SAS.
display manager and submit to view the blue macro-call window (Figure 11) (You need to have access to
INTERNET while running this SAS macro in your system. Input all the required macro input parameters
and submit the macro to perform any one of the steps in the all possible mixed model selection steps. Please refer
the required SAS modules listed below for running this macro successfully.

Required SAS Modules for Running the Allmixed2 SAS Macro in Version 9.13:

- SAS /STAT : PROCs MIXED, CORR, REG and GLMSELECT
- SAS/GRAph : PROC GCHART, PROC GPLOT, PROC G3D
- SAS/BASE SAS ODS (RTF, HTML, PDF)
- SAS/ACCESS: PC FILES – PROC IMPORT and EXPORT

SUMMARY

The effects of multicollinearity in pre-screening phase and in all possible mixed model selection were evaluated
using 4 simulated repeated measures data with different sample sizes and different degree of multicollinearity. The
user-friendly SAS macro application ALLMIXED2 was used to perform the pre-screening and the all possible
subset mixed model selection. Pre-screening step based on LASSO method identified all the true fixed effects in
large samples (number of subjects >100) even in the presence of multicollinearity. In small samples (number of

Figure 11 Screen shot of the ALLMIXED2 SAS macro application window
LASSO method failed to identify true fixed effects. High degree of multicollinearity didn’t influence model selection based on information criteria in the pre-screening step of the model selection. However, small sample size had a greater adverse effect on model selection. In all possible mixed model selection, the best candidate model selection based on AICC or MDL were not influenced by the presence of multicollinearity. Only the sample size had a greater impact on the outcome of the model selection. When the sample size used is very small (total number of subjects <= 25) the AICC becomes very conservative and failed to select the true covariates whereas the MDL becomes liberal and selected few non-significant covariates as the best candidate models. Instructions for downloading and running this user-friendly macro application ALLMIXED2 are included.

REFERENCES


CONTACT INFORMATION

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

Name: George C. Fernandez PhD, Professor of Applied Statistics
Enterprise: University of Nevada - Reno
Address: CABNR/204 Reno, NV 89557
Work phone: (775)-784-4206
Email: gcjf@unr.edu
Web: http://www.ag.unr.edu/gf

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.