Paper I02-2009

Mixed Model Selection

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

ABSTRACT

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. This macro application, ALLMIXED2 will complement the model selection option currently available in the SAS PROC REG for multiple linear regressions and the experimental SAS procedure GLMSELECT that focuses on the standard independently and identically distributed general linear model for univariate responses. 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) and MDL (minimal description length) are used in all possible model selection and summaries of the best model selection are compared graphically. The differences in the degree of penalty factors associated with the model dimension between AICC and MDL are investigated. 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. The ALLMIXED2 SAS macro application is an improved version of the SAS macro application ALLMIXED2 previously reported (Fernandez, 2007). Instructions for downloading and running this user-friendly macro application are included.

INTRODUCTION

Model selection is usually carried out by the automated procedures built into the software including frequently used forward, backward, and stepwise model selection procedures. There is an extensive review and discussion on the theoretical aspects of model selection criteria and procedures (Burnham and Anderson 2002; Hoeting et.al 2006). All possible model selection of fixed effects in general linear model setup using delta AICC , delta SBC and model weights are implemented in the REGDIAG macro (Fernandez, 2002). However, all possible mixed model selection can be tedious, time consuming and complicated due to the presence of additional random terms and optimal variance-covariance structure associated with repeated measures (Littell et.al 2006; Hoeting et.al 2006; Kramer 2004). Ngo and Rand (2002) developed a SAS macro for performing mixed model selection for user-specified models. Keselman et .al (1998) proposed a SAS based method to select the best covariance structure in mixed model repeated measures analysis. However, to apply these SAS macros in model selection, SAS programming experience is a requirement. Kramer (2004) developed an automated model selection application using SAS Mixed and PERL codes for both fixed and random effects. Programming knowledge in PERL is required to use this application. This paper presents a practical and complete solution for automated and efficient mixed model selection and model exploration using a user-friendly SAS macro application named ALLMIXED2. The ALLMIXED2 SAS macro application presented here is an improved version of the SAS macro application ALLMIXED2 previously reported elsewhere (Fernandez, 2007). All new improvements of the previously reported ALLMIXED2 macro applications are emphasized here.

MODEL SELECTION CRITERIA USED IN ALLMIXED2 MACRO

The general form of information criterion (IC)= *-2 log L* + *Penalty factor (pf)*

-2 log L is derived from PROC MIXED method = ML

Δ *-2 log L* = *(-2 log L* $_i$*)- (-2 log L* $_{min}$)

-2 log L $_{min}$ = *The smallest -2 log L* derived from PROC MIXED method ML of all models compared.

AIC = -2 log L + 2(p+k+1) (Hoeting et. al 2006)

AICC= -2 log L + [2(p+k+1) (n/(n-p-k-2))] (Hoeting et. al 2006)

Where

p = number of fixed effect terms
k = number of random effect terms
n = total sample size for random effect model and **number of subjects in case of repeated measures**

* In large sample AIC and AICC are nearly equivalent

$\Delta AICC = AICC_i - AICC_{min}$ *Best candidate models = (ΔAICC <=2)*

$AICC_{sas}$ *= AICC reported by SAS PROC Mixed using ML*

$AICC_{REML}$ *= AICC reported by SAS PROC Mixed using REML*

MDL =1/2 {-2 log L + [log(n) (p+k+1)]} (Hoeting et. al 2006)

Δ *MDL* $= MDL_i - MDL_{min}$ *Best candidate models = (Δ MDL <=1)*
In the ALLMIXED2 macro, the best candidate models selection criterion based on MDL is changed from 1/2(log(n)) (Fernandez 2006) to <=1. This new criterion is comparable to the criterion used for AICC (<=2)

BIC = -2 log L + [log(n) (p+k+1)] *(SAS Institute 2006)*

Penalty factor % = *(pf* $_i$ */ -2 log L* $_{ref}$*) *100*

-2 log L $_{ref}$ = *-2 log L* derived from PROC MIXED method ML that contain optional random and repeated measure covariance parameter and user specified "**Must-Have**" fixed effects.

AICC weights =*Exp(-0.5*Delta AICC* $_i$*) / Sum of (Exp(-0.5*Delta AICC* $_i$*)) all best candidate model (Buckland et al. 1997).*

MDL weights =*Exp(-0.5*Delta MDL* $_i$*) / Sum of (Exp(-0.5*Delta MDL* $_i$*)) all best candidate model*

AICC weight ratio = AICC weight / Max (AICC weight)

MDL weight ratio = MDL weight / Max (MDL weight)

ALL POSSIBLE MODEL SELECTION STEPS

The recommended selection steps for performing the model selection in MIXED model is illustrated in Figure1. Although the recommended sequence of the steps are identified in the figure, it is not a requirement to follow the same sequence. Users are free to choose to run any model selection steps in any order they desire. However, before running these model selection steps the data format must be suitable for running the SAS PROC Mixed (Littell et.al 2006) procedure. The following types of PC data formats can be used with the ALLMIXED2 macro: SAS temporary and permanent data files, Microsoft excel or ACCESS data tables, COMMA or TAB delimited text file. Refer the ALLMIXED2 macro help file available from the author's website for more information regarding inputting data file name and its location in the macro-call window.

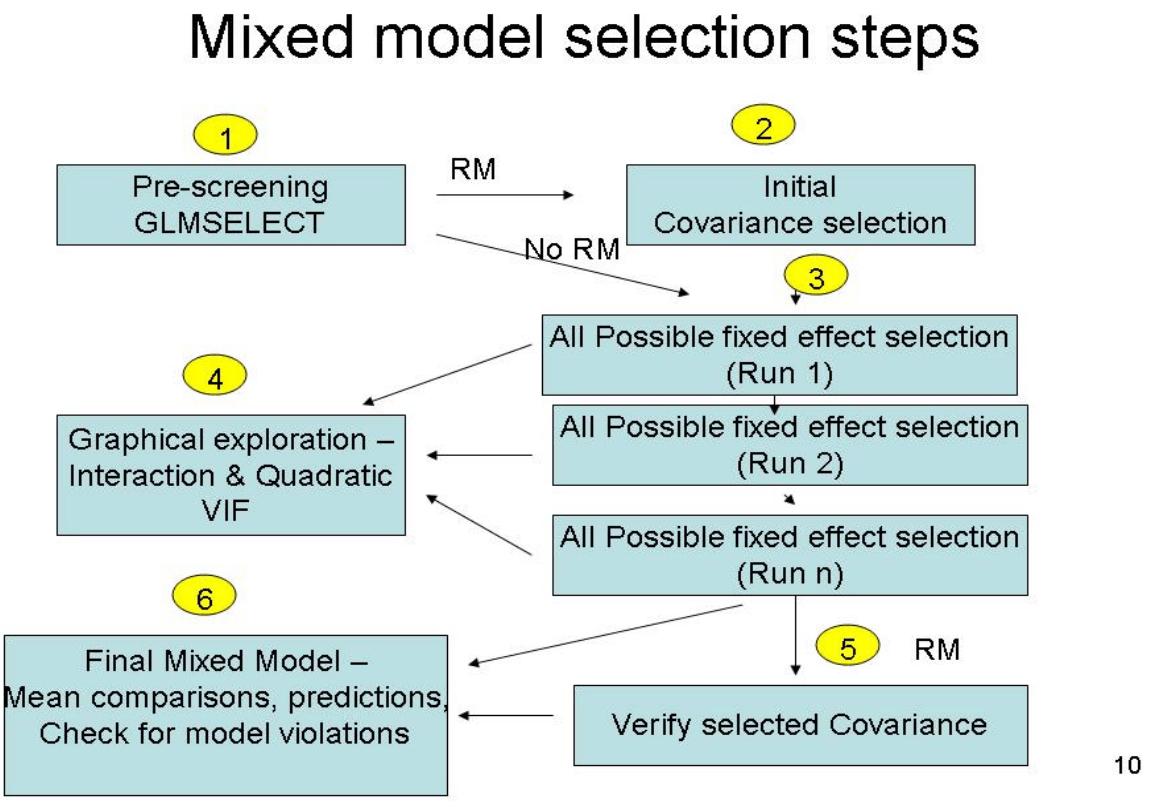


Figure 1 Recommended all possible model selection steps and their sequence

STEP1: PRE-SCREENING

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 drop least contributing model terms. In pre-screening step, the repeated measures covariance structure is ignored and the random effects are treated as fixed. To drop the least contributing model terms and to select the user-specified number of effects, the LASSO (Tibshirani 1996) method implemented in the GLMSELECT (Cohen 2006) the experimental SAS procedure in SAS version 9.13 is used. For more information of the theory and selection features refer SAS Institute (2006). The LASSO model selection options - CHOOSE=NONE and

SELECT=SBC are used in this macro. 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 drop 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 six 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.

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 not contributing 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.

STEP2: REPEATED MEASURES - INITIAL COVARIANCE TYPE SELECTION.

In a repeated measures modeling , the best covariance structure describing the correlation among the repeated measures should be identified first. The best covariance structure can be identified from different user-specified covariance structures by comparing the AICC statistic computed in PROC MIXED using REML method and select the covariance type which gives the smallest AICC value. Refer the ALLMIXED2 macro help file available from the authors website for more information regarding inputting appropriate parameters.

Step3: All possible model selection steps

All combination of models associated with the user-specified fixed effects subset range (start:2 and stop:6) are generated by the ALLMIXED2 macro and their information criteria statistics, AICC and MDL are compared in this step. Users can optionally specify certain fixed effects as "MUST HAVE" and other fixed effects as "SELECTABLE" in the all possible model selection. All combination of mixed model using the fixed effects listed in "SELECTABLE" category are generated in this step and the following statistics are estimated.
Variance inflation statistics (VIF) for each continuous predictor variables in the model.
PRESS statistics generated in GLM proc- To monitor the impact of influential observations the differences between PRESS and SSE are evaluated in each model selection step.
Information criteria estimates based on REML: $AICC_{reml}$
Information criteria estimates based on ML: AIC, AICC, $AICC_{sas}$, MDL, and BIC.

In the ALLMIXED2 macro, the following changes are made in the computation of the IC statistics. When computing the IC statistics for a repeated measures data during the pre-screening and the all possible subset model selection step, the sample size (n=250) is substituted by the number of subjects (50). Thus, the results reported in this paper regarding the performance of AICC and MDL in model selection is contradicting with the previously reported results (Fernandez, 2006) for the same data where a larger sample number N=250 was used.

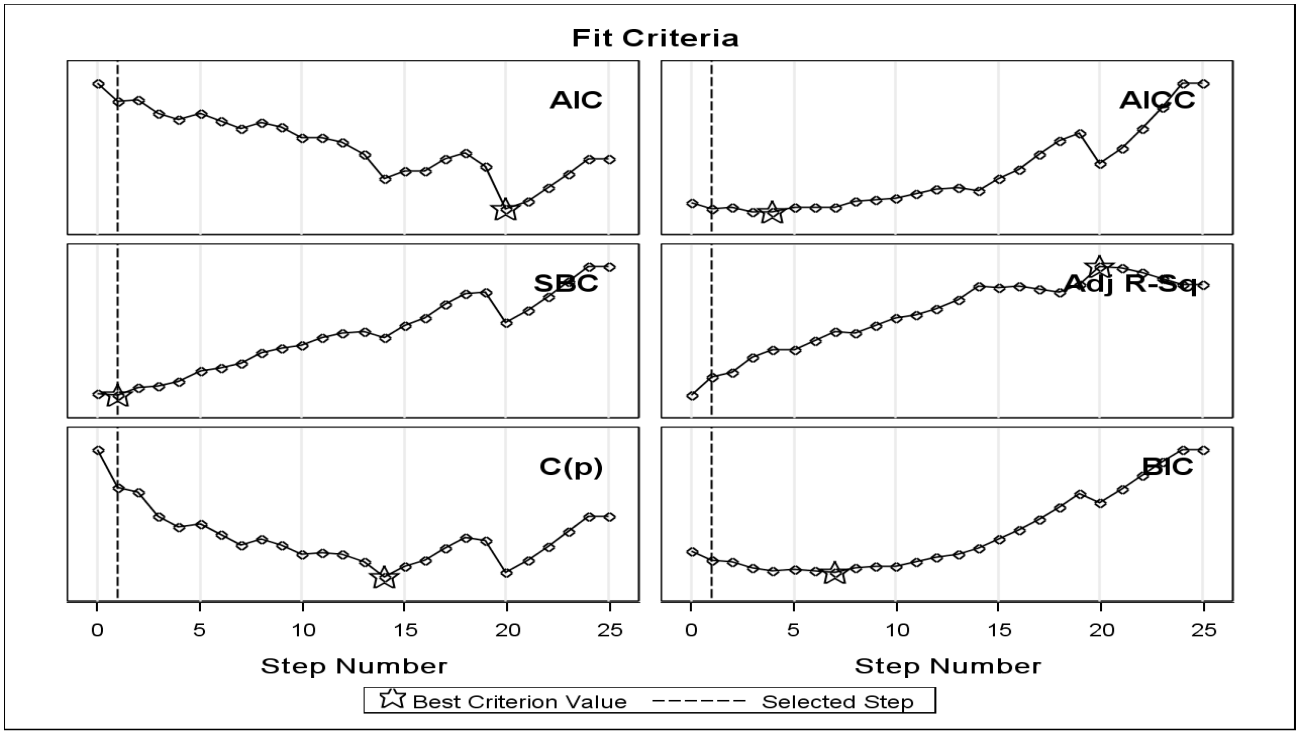


Figure 2 Information criteria estimates computed in the LASSO method of model selection available in SAS procedure GLMSELECT. The model parameters included are two group effects (trt and time) and 20 covariates (x1-x20)

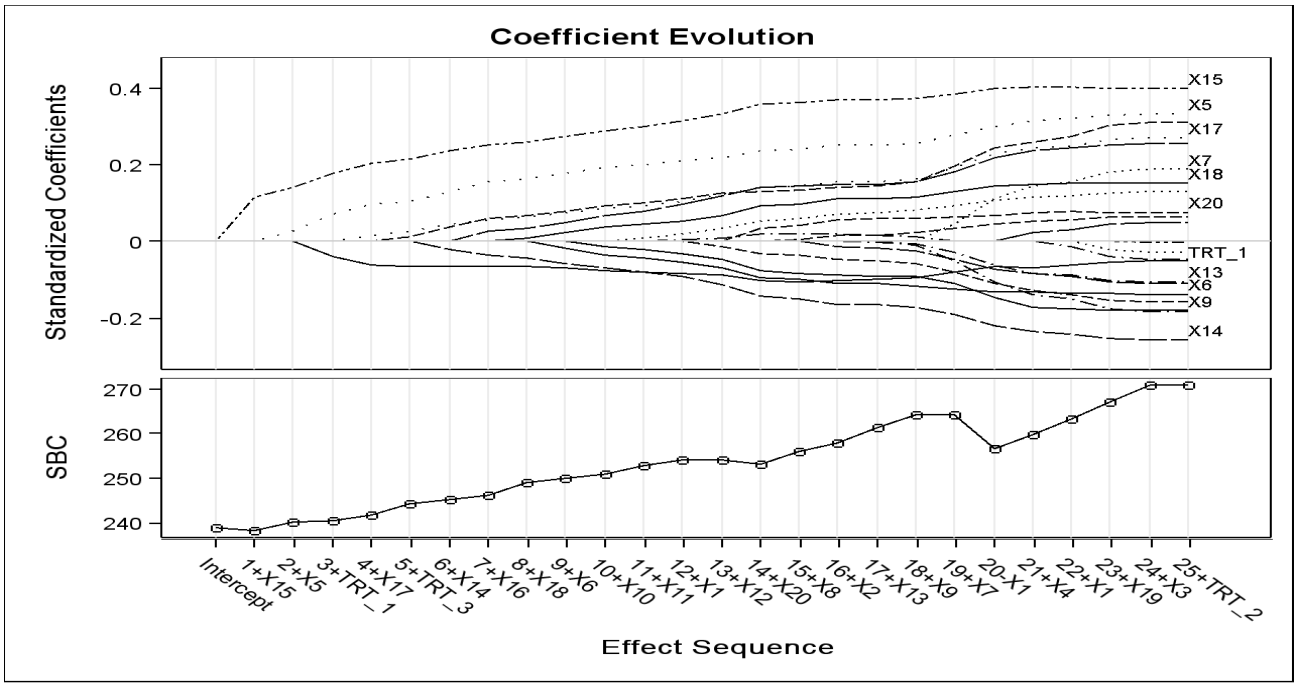


Figure 3 Standardized regression coefficient estimates and SBC computed at each model selection sequence during the LASSO method of model selection available in SAS procedure GLMSELECT. The model parameters included are two group effects (trt and time) and 20 covariates (x1-x20)

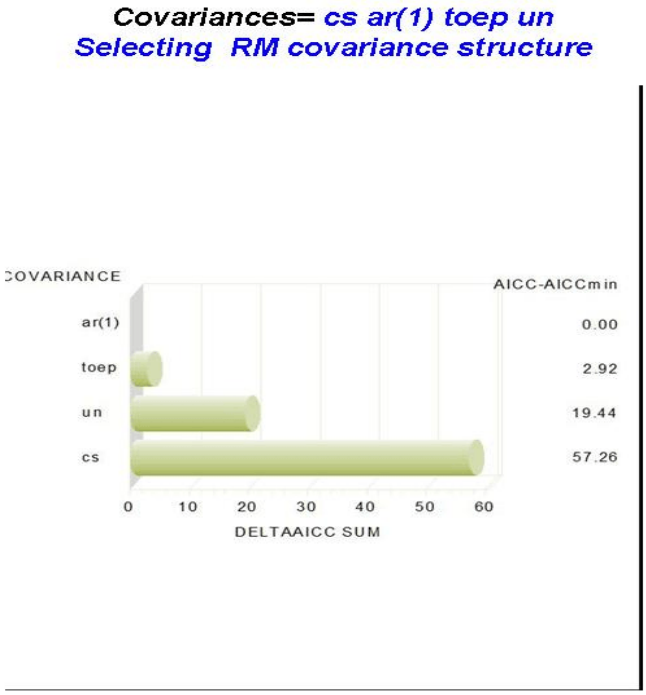


Figure 4 Repeated Measure analysis covariance type selection based on smallest AICC.

The relationships between AIC, AICC, $AICC_{sas}$, $AICC_{reml}$, MDL, and BIC are investigated by the rank correlations using the SAS PROC CORR and scatter plot matrix (Figure 5). Perfect rank correlations (1) are commonly observed between (AICC and $AICC_{sas}$) and (MDL and BIC) indicating that these two sets of IC behave identically in the model selection. Furthermore AIC and AICC didn't behave similarly and the degree of penalty was not the same when the number of fixed effects is relatively larger (*p=19*) compared with total number of observations (*n=number of subjects= 50*) in this simulated repeated measures data. The rank correlations between $AICC_{reml}$ and AIC statistics computed by ML method (AIC, AIC_C, $AICC_{sas}$) were not perfectly correlated and $AICC_{reml}$ behave differently from ML based IC. Big differences were observed in the model selection performance of AIC based (AIC, AIC_C,) and BIC based (MDL) information criteria as evident by the rank correlation.

All IC statistics reported here are made out of two components: Log likelihood estimate (*-2 log L*) and penalty factor (*pf*). For a given model, *-2log L* value is constant and is influenced by degree of model fit, variable included and not included in the model, presence of influential outliers, and model specification errors. The penalty factor is made out of number of fixed (*p*) and random effects (*k)* and the sample size (*n*). When all possible model selection involving only the fixed effects are carried out, the sample size and the number of random effects become constant. Therefore, only the number of fixed factor becomes the determining component of the penalty factor. The relationship between penalty factor and the number of fixed effects between AIC_C, $AICC_{reml}$, and MDL are shown in Figure 6. The penalty factor for the $AICC_{reml}$ becomes constant because this penalty factor does not include any fixed effects and only the number of random effects (which is a constant) is included. The penalty factors for AIC_C and MDL shows a positive linear effect associated with the increase in the number of fixed effects. The MDL penalty factor % increased from 6.5% (Reference model) to about 8.5% where as the AIC_C penalty factor % increased from 2.5% (Reference model) to about 3.2% when the number of fixed effects increased by 5 terms. Thus, the degree of penalty is about 3 times stronger for MDL than the AIC_C and this clearly evident in the ratio

between MDL penalty % and AIC_C penalty and the relationship is clearly shown in Figure 6. These findings clearly

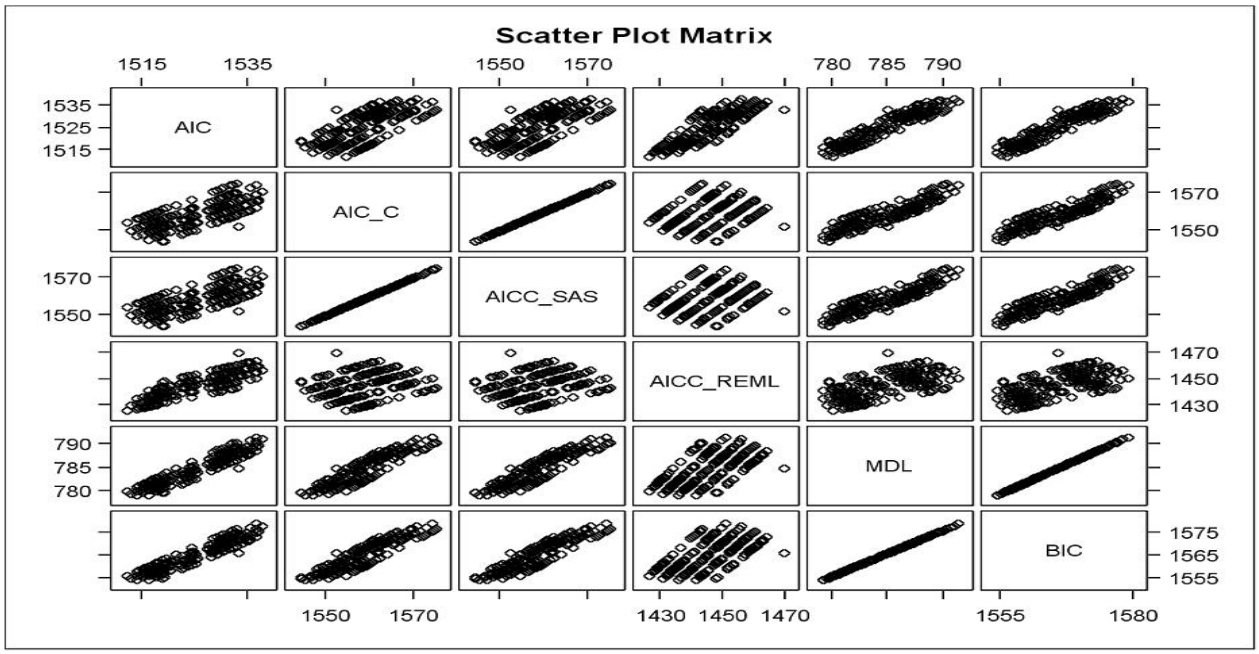


Figure 5 Scatter plot matrix showing rank correlation among 4 AIC and 2 BC based Information criteria statistics in all possible mixed model selection

showed the the relationships between AIC, AICC, $AICC_{sas}$, $AICC_{reml}$, MDL, and BIC are investigated by the rank correlations using the SAS PROC CORR and scatter plot matrix (Figure 5). Perfect rank correlations (1) were observed between (AICC and $AICC_{sas}$) and (MDL and BIC) indicating that these two sets of IC behave identically in the model selection. Furthermore AIC and AIC_C behaved very similarly and the degree of penalty was similar when the number of fixed effects is relatively small (*p=19*) compared with total number of observations (*n=250*) in this simulated data. The rank correlations between $AICC_{reml}$ and AIC statistics computed by ML method (AIC, AIC_C, $AICC_{sas}$) were not perfectly correlated and $AICC_{reml}$ behave differently from ML based IC. Big differences were observed in the model selection performance of AIC based (AIC, AIC_C,) and BIC based (MDL) information criteria as evident by the rank correlation.

The components of AICC and MDL (*-2 log l* and the penalty factor) are graphically compared in Figure 7. For a given model, *-2log L* value is constant when estimating AICC and MDL and it decreases linearly with an increase in the number of fixed terms. But, within a subset (two, three, four variable subset), the *-2 log L* value varies a lot whereas all the models within a subset have the same penalty factor for both AICC and MDL (Figure 7). Also AICC statistic favors parsimonious model (2 and 3 subsets) whereas MDL statistic favors models with large number of model terms (3,4,5 subsets) especially in a small data set (50 subjects in repeated measures data) (Figure 7).This contradicts with the earlier report where MDL favored more parsimonious models when n was considered large (50 subjects x 5 repeated measures =total sample size 250) Fernandez (2006)

Graphical display of best models within each subset based on smallest *ΔAICC* and ΔMDL within each subset are shown in Figure 8. Graphical display of overall best candidate models based on *ΔAICC* <= 2 and ΔMDL <=1 are shown in Figure 9. Refer the ALLMIXED2 macro help file available from the authors website for more information regarding downloading this user-friendly macro file and associated macro help file.

inputting appropriate parameters in the macro-call window.

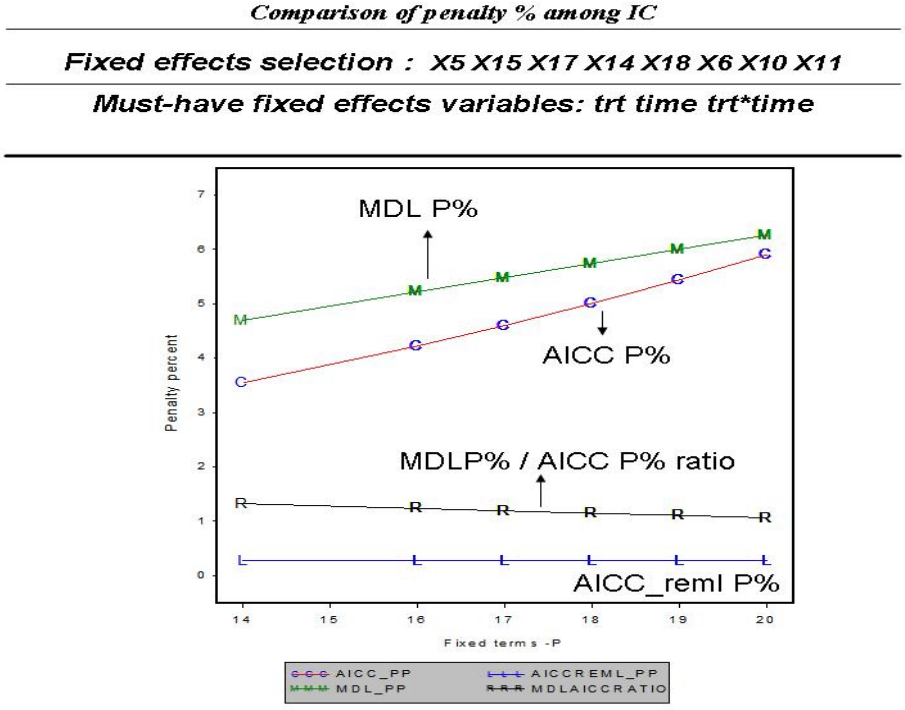


Figure 6 Comparison of penalty factor among AICC, AICC $_{reml}$ and MDL versus number of fixed effect terms.

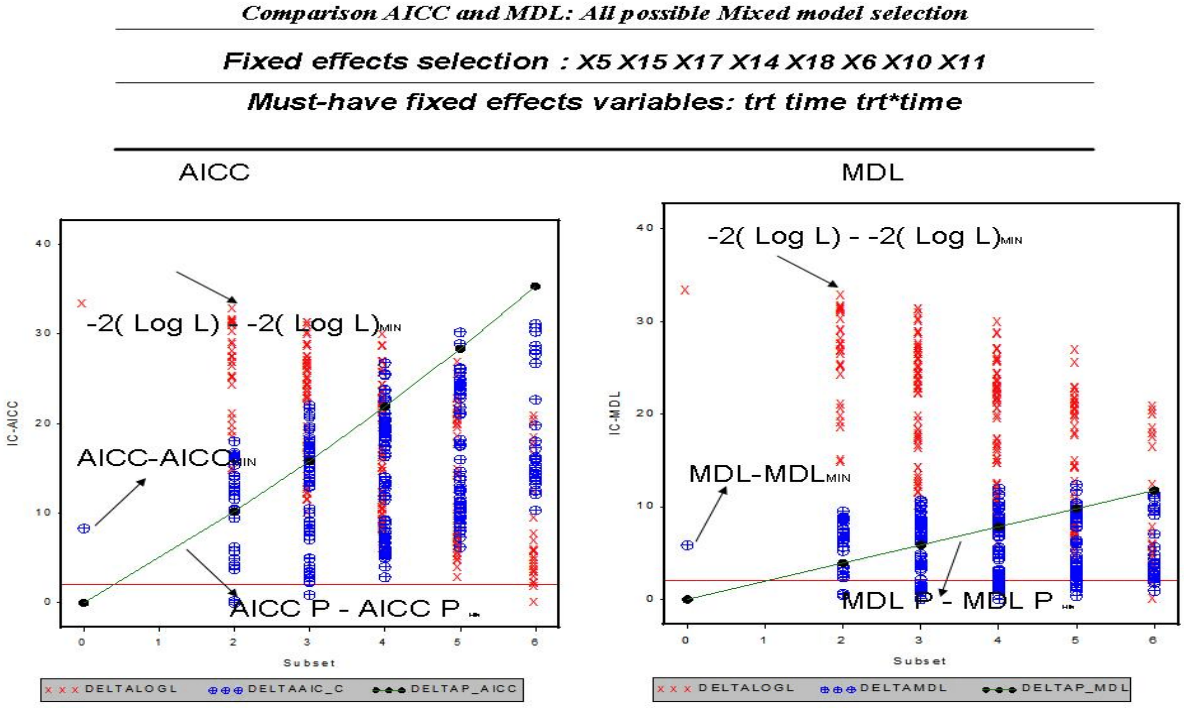


Figure 7 Comparison of the association between AICC and MDL components and the number of fixed effect terms

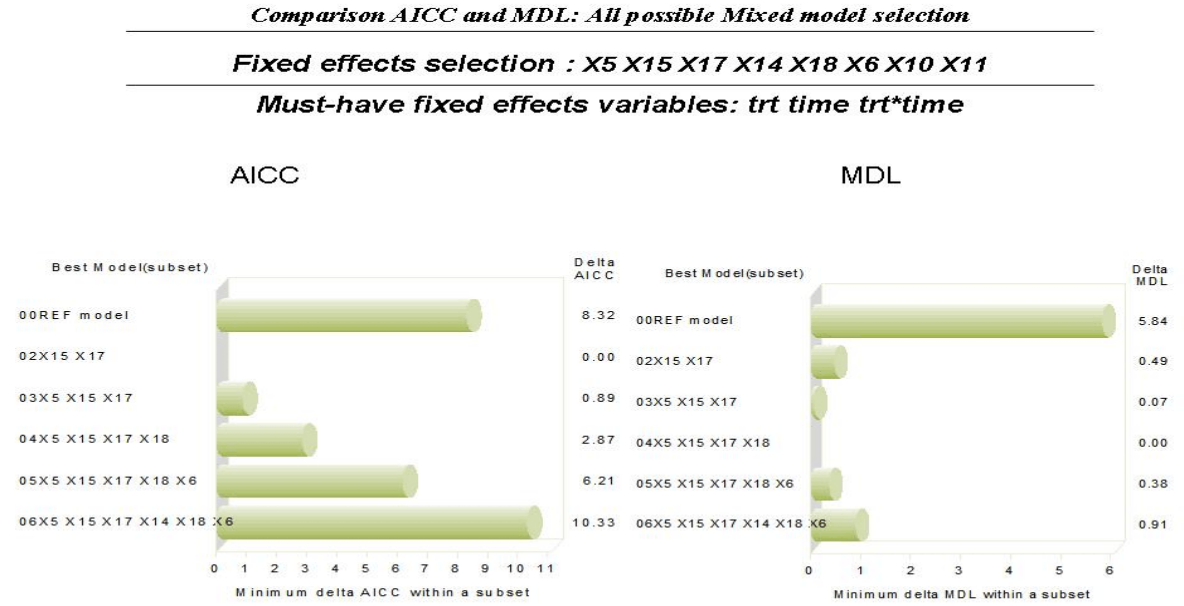


Figure 8 Graphical display of best model within each subset identified by the AICC

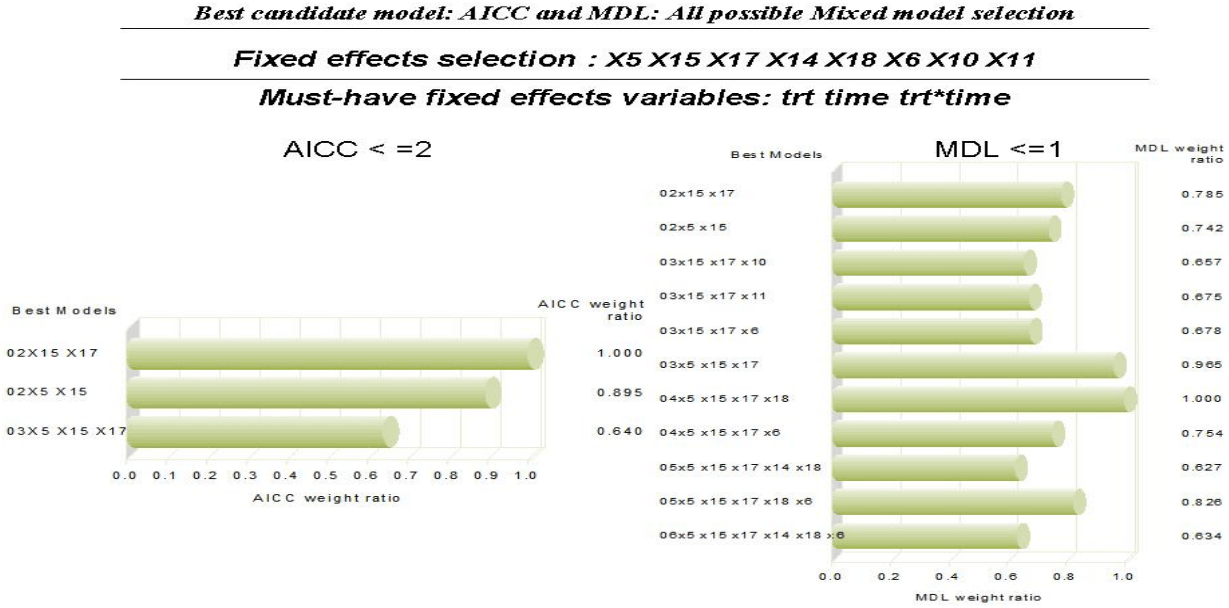
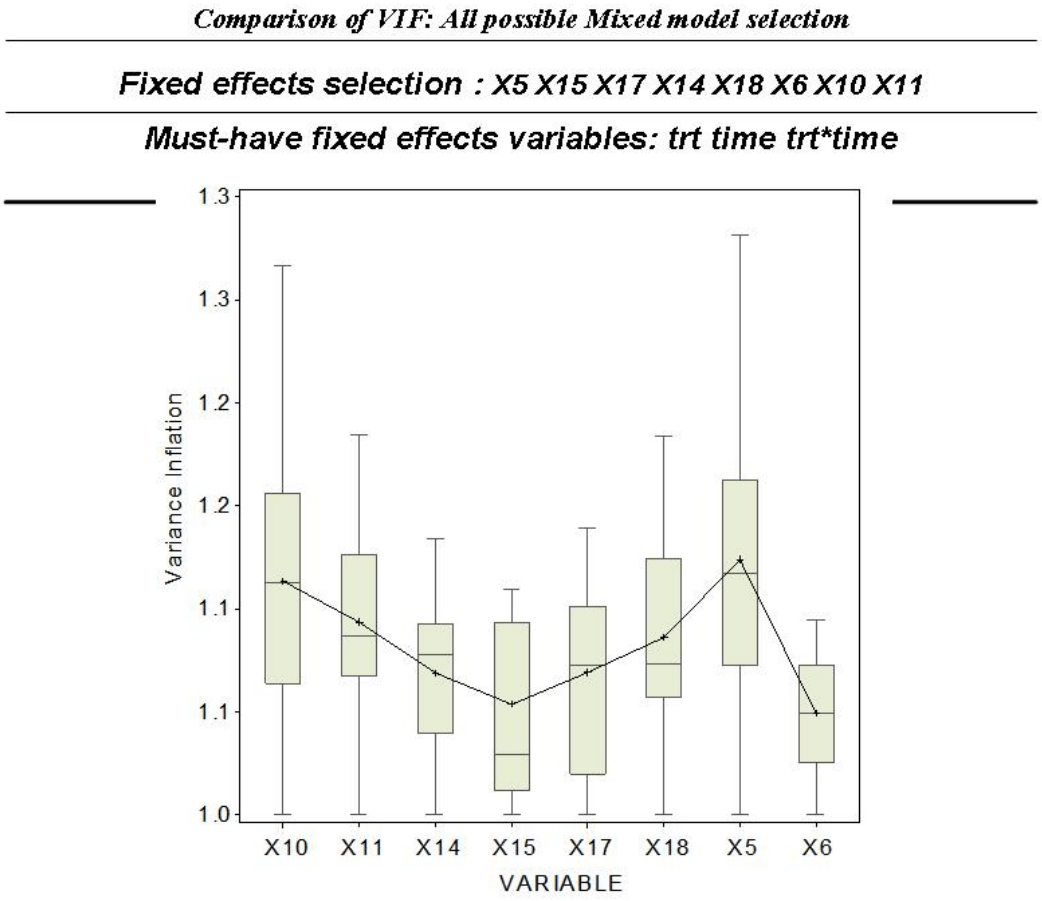


Figure 9 Graphical display of best candidate models identified by AICC (<=2) and the MDL (<=1).

STEP4: GRAPHICAL EXPLORATION FOR MULTICOLLINARITY AND MODEL SPECIFICATION ERROR

Severe multicollinarity (Variance inflation factor > 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. Therefore, presence of high degree of multicollinarity can impact fixed effect selection. Therefore, assessing the degree of multicollinarity for each of the continuous fixed effects in all possible model selection can help to select the best model from the set of best candidate models. Variable(s) not contributing multicollinarity could be preferred over the variables significantly contributing to multicolliniarity. Figure 10 shows the box-plot display of VIF distribution for all the continuous predictors included in model selection. Because the data used in the study are simulated from known properties multicollinarity should not exists and it is clearly shown in Figure10 where VIF values were less than 2 for all the predictor variables. Also, to diagnose multicollinearity in each model selection step (when VIF value > 10) the VIF statistic for each continuous predictors involved in multicollinarity is sent to an output table for further exploration.



Model selection success can also be influenced by model specification error when significant higher order model terms (quadratic and cross-product) omitted from the mixed model. The need for an quadratic term or an interaction between any two predictor variables could be evaluated in the 'quadratic' or 'interaction detection plot respectively. To detect the need for a significant quadratic term, first fit the full model including the quadratic term for the given predictor variable (X_i)and examine the Type III P-value for statistical significance and output the predicted values ($YHAT_{full}$) for the full model. Then drop both the linear and the quadratic terms for this given predictor from the model and estimate the predicted values for this reduced model ($YHAT_{red}$). Then a graphical display between the Δyhat ($YHAT_{full}$ -$YHAT_{red}$) and X_i can reveal the nature and the strength of quadratic effects (Figure 11).

Similarly, to detect the need for a significant interaction term between two predictors, first fit the full model including the cross-product term for the two predictor variable (X_1 and X_2) and examine the Type III P-value for statistical significance of the interaction term and output the predicted values ($YHAT_{full}$) for the full model. Then drop the cross product from the model and estimate the predicted values for this reduced model ($YHAT_{red}$). Then a 3-D graphical display between the Δyhat ($YHAT_{full}$ -$YHAT_{red}$) and X_1 and X_2 can reveal the nature and the strength of interaction effects (Figure 12). Refer the ALLMIXED2 macro help file available from the authors website) for more information regarding inputting appropriate parameters in the macro-call window.

Step5: Final covariance type selection.(Optional step- for repeated measures model)

After several runs of all possible model selection steps, many data exploration, and multicollinarity checks, we can select the final fixed effect model. But, before finalizing the final mixed effect model it is important to verify whether the covariance type used in the model selection step is still the best type for the selected model. Again user-specified covariance types, can be compared and the final covariance type selection can be made based on $\Delta AICC_j$ ($AICC_j$ - $AICC_{min}$) using PROC MIXED REML method. Refer the ALLMIXED2 macro help file available from the authors website) for more information regarding inputting appropriate parameters in the macro-call window.

Step6: Complete mixed model analysis

After selecting the final repeated measures mixed model dimensions, complete mixed model analysis can be performed including the data exploration by box plots (Figure 13), mixed model analysis, LSMEAN comparisons with alphabet mean separation (Figure 14-15) suggested by Saxton (2002), model predictions, checking for normality of studentized conditional residuals (Figure 16), and performing influential diagnostics (Figure 17) in one step. Refer the ALLMIXED2 macro help file available from the authors website for more information regarding inputting appropriate parameters in the macro-call window.

AVAILABILITY OF THE ALLMIXED MACRO:

Users can download the ALLMIXED2 .SAS macro-call file from the authors website at http://www.ag.unr.edu/gf and 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 18) (You need to have access to INTERNET to download and execute the macro while running this ALLMIXED2 macro in your system). Input all the required macro input parameters and submit the macro to perform the all possible mixed model selection. Please refer the required SAS modules listed below for running this macro successfully.

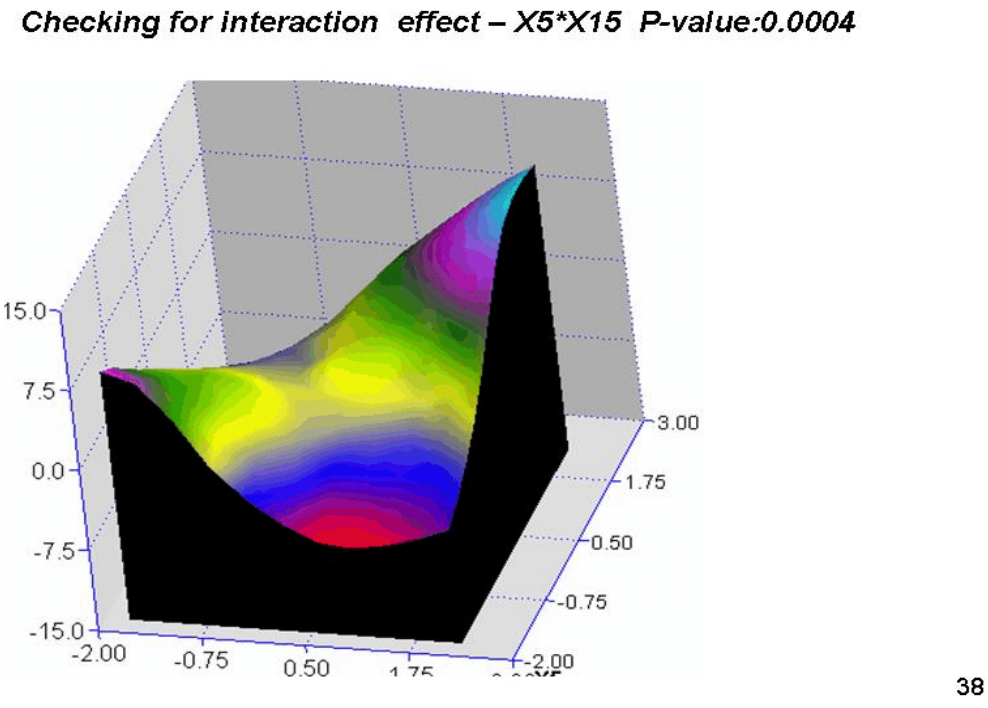


Figure 12 Graphical exploration and the statistical significance of the user-specified cross product

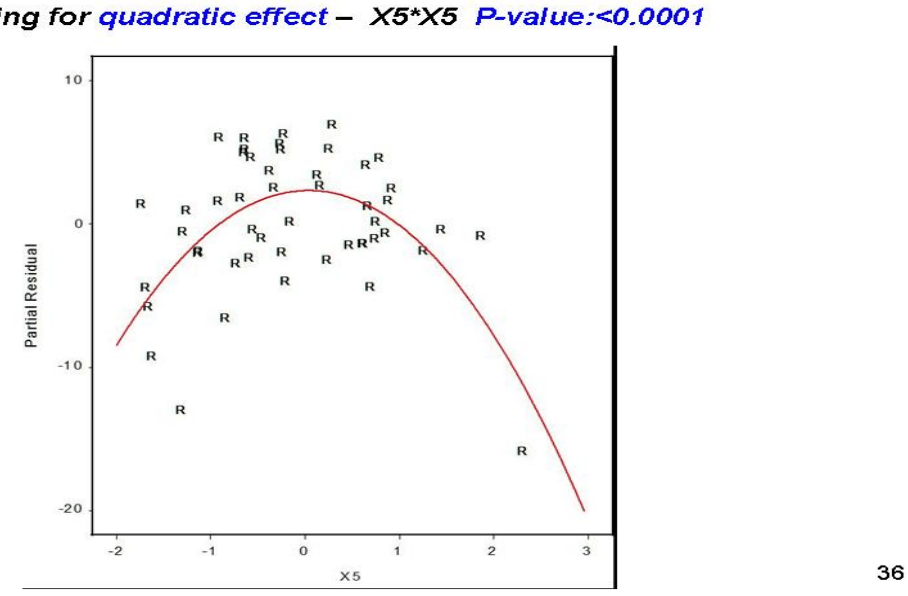


Figure 11 Graphical exploration and the statistical significance of the user-specified quadratic effect.

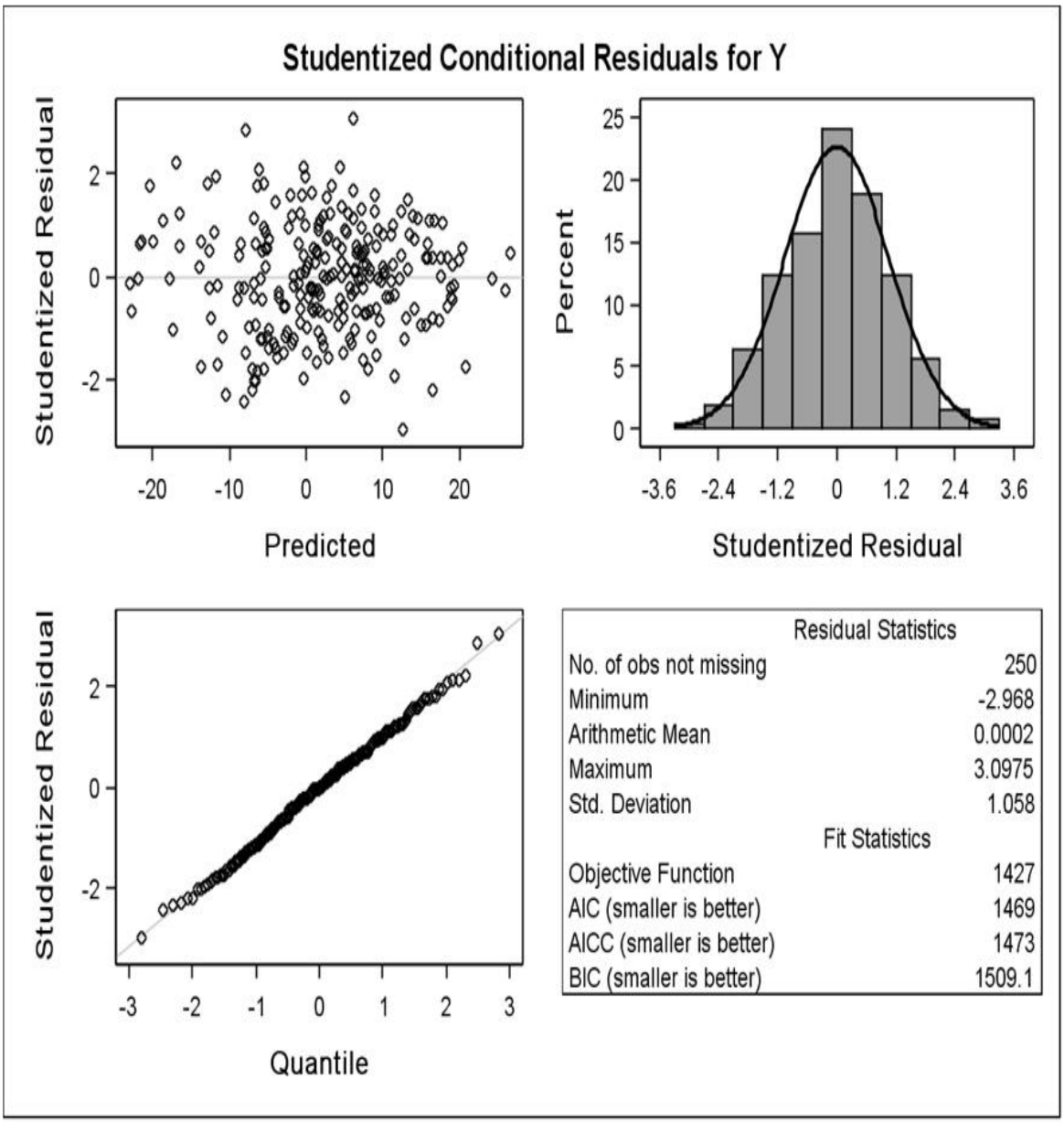
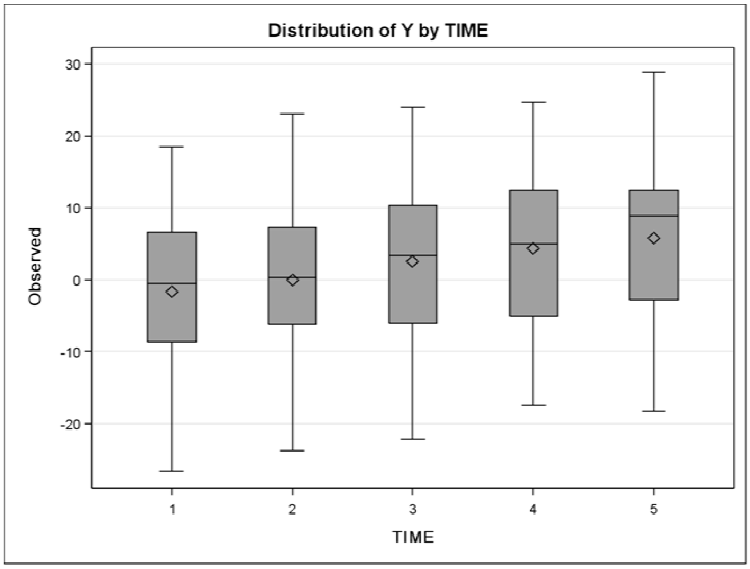


Figure 13 Mixed model violation detection using studentized conditional residuals

Mixed model exploration – ODS GRAPHICS



37

Figure 14 Exploration of repeated measures-Time effect by box plot

Mixed model analysis data=simdata1 response=y
Mean separation based on Saxton, A.M(1998) 23rd SUGI pp1243-1246

Effect=TRT **Method=Tukey-Kramer(P<.05)** **Table=1**

TRT	TIME	LSmeans	Standard Error	Alpha	Lower CI	Upper CI	Alphabet Group
3	_	8.6556	1.3078	0.05	6.0344	11.2769	A
2	_	2.3088	1.4381	0.05	-0.5736	5.1913	B
1	_	-2.2621	1.0871	0.05	-4.4410	-0.08309	C

Effect=TIME **Method=Tukey-Kramer(P<.05)** **Table=2**

TRT	TIME	LSmeans	Standard Error	Alpha	Lower CI	Upper CI	Alphabet Group
_	5	6.6457	0.8735	0.05	4.9126	8.3789	A
_	4	5.1903	0.8735	0.05	3.4572	6.9235	A
_	3	3.4085	0.8735	0.05	1.6754	5.1416	B
_	2	0.6109	0.8735	0.05	-1.1222	2.3441	C
_	1	-1.3515	0.8735	0.05	-3.0846	0.3817	D

41

Figure 15 Main effect LSMEAN Comparison - using alphabet notation

Mixed model analysis data=simdata1 response=y
Mean separation based on Saxton, A.M(1998) 23rd SUGI pp1243-1246

Effect=TRT*TIME **Method=Tukey-Kramer(P<.05)** **Table=3**

TRT	TIME	LSmeans	Standard Error	Alpha	Lower CI	Upper CI	Alphabet Group
3	5	13.2799	1.5724	0.05	10.1595	16.4003	A
3	4	11.6835	1.5724	0.05	8.5631	14.8039	AB
3	3	10.0406	1.5724	0.05	6.9202	13.1610	ABC
3	2	6.6650	1.5724	0.05	3.5446	9.7854	CDE
2	5	5.6627	1.7168	0.05	2.2548	9.0706	ABCDF
2	4	5.0490	1.7168	0.05	1.6411	8.4569	BCDF
2	3	3.2151	1.7168	0.05	-0.1929	6.6230	CDEFGH
3	1	1.6093	1.5724	0.05	-1.5111	4.7297	FGHI
1	5	0.9946	1.3044	0.05	-1.5941	3.5834	DEFG
2	2	-0.6784	1.7168	0.05	-4.0863	2.7296	EGHI
1	4	-1.1615	1.3044	0.05	-3.7502	1.4273	FGHI
2	1	-1.7042	1.7168	0.05	-5.1121	1.7037	GH
1	3	-3.0302	1.3044	0.05	-5.6190	-0.4414	H
1	1	-3.9595	1.3044	0.05	-6.5482	-1.3707	GH
1	2	-4.1538	1.3044	0.05	-6.7426	-1.5651	H

42

Figure 16 LSMEAN Comparison - Interaction means with alphabet notations

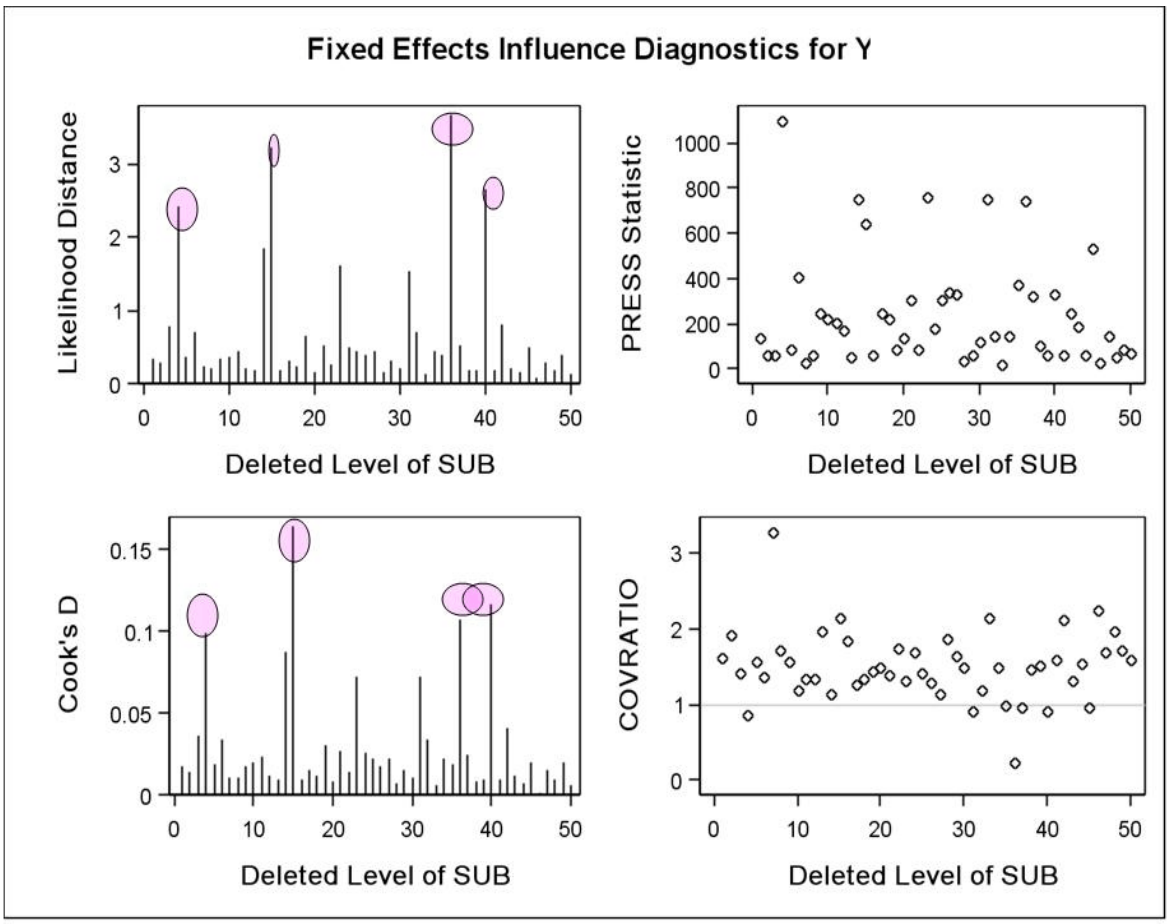
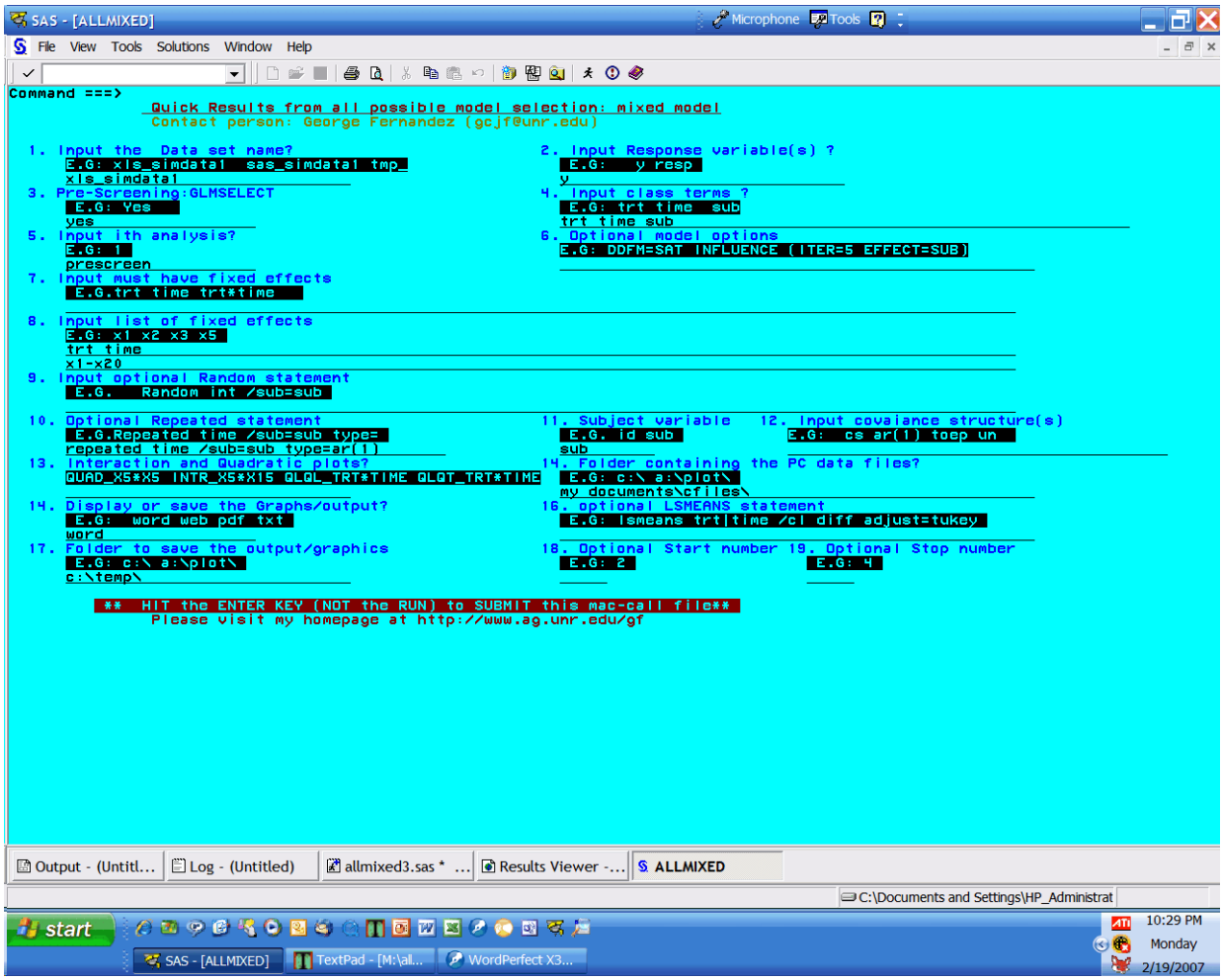


Figure 17 Repeated measurers mixed model influential diagnostics - at the subject level


SAS - [ALLMIXED]
File View Tools Solutions Window Help
Command ===>
Quick Results from all possible model selection: mixed model
Contact person: George Fernandez (gcjf@unr.edu)
1. Input the Data set name?
E.G: xls_simdata1 sas_simdata1 tmp_
xls_simdata1
2. Input Response variable(s) ?
E.G: y resp
y
3. Pre-Screening:GLMSELECT
E.G: Yes
yes
4. Input class terms ?
E.G: trt time sub
trt time sub
5. Input ith analysis?
E.G: 1
prescreen
6. Optional model options
E.G: DDFM=SAT INFLUENCE (ITER=5 EFFECT=SUB)
7. Input must have fixed effects
E.G.trt time trt*time
8. Input list of fixed effects
E.G: x1 x2 x3 x5
trt time
x1-x20
9. Input optional Random statement
E.G. Random int /sub=sub
10. Optional Repeated statement
E.G.Repeated time /sub=sub type=
repeated time /sub=sub type=ar(1)
11. Subject variable
E.G. id sub
sub
12. Input covaiance structure(s)
E.G: cs ar(1) toep un
13. Interaction and Quadratic plots?
QUAD_X5*X5 INTR_X5*X15 QLQL_TRT*TIME QLQT_TRT*TIME
14. Folder containing the PC data files?
E.G: c:\ a:\plot\
my documents\cfiles\
14. Display or save the Graphs/output?
E.G: word web pdf txt
word
16. optional LSMEANS statement
E.G: lsmeans trt|time /cl diff adjust=tukey
17. Folder to save the output/graphics
E.G: c:\ a:\plot\
c:\temp\
18. Optional Start number
E.G: 2
19. Optional Stop number
E.G: 4
** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
Please visit my homepage at http://www.ag.unr.edu/gf
Output - (Untitl...
Log - (Untitled)
allmixed3.sas * ...
Results Viewer -...
ALLMIXED
C:\Documents and Settings\HP_Administrat
start
SAS - [ALLMIXED]
TextPad - [M:\all...
WordPerfect X3...
10:29 PM
Monday
2/19/2007

Required SAS Modules for Running the All mixed 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 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.
- 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 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.
- Best candidates models can be selected by the delta AICC and delta MDL based model weight statistics
- 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 boxplot display of VIF estimates by all the continuous fixed effects are generated for the overall assessment of multicollinarity 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 avilable in Version 9.1.
- 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.

REFERENCES

1. Buckland, S. T., K. P. Burnham, and N. H. Augustine. 1997. Model selection: an integral part of inference. Biometrics 53:603-618.
2. Burnham, K. P, and Anderson D. R. (2002) Model selection and inference: a practical information theoretic approach. (Second edition) Springer-Verlag, New York, New York, USA.
3. Cohen R. A (2006) Introducing the GLMSELECT PROCEDURE for Model Selection SUGI31 proceedings http://www2.sas.com/proceedings/sugi31/207-31.pdf
4. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression (with discussion),Annals of Statistics, 32, 407–499.
5. Fernandez, G (2002) Data Mining using SAS applications Book 384 p CRC-Chapman Hall FL USA
6. Hoeting, J.A, Davis R.A Merton A.A and Thompson S. E (2006) Model selection for Geostatistical Models Ecological Applications, 16(1), pp. 87–98
7. Keselman, H. J., Algina, J., Kowalchuk, R. K., and Wolfinger, R. D. (1998) A comparison of two approaches for selecting covariance structures in the analysis of repeated measurement , 27, Communications in Statistics, Simulation & Computation 591-604.
8. Littell, R.C, Milliken, G A., Stroup, WW, and Wolfinger, R D. (2006). SAS System for Mixed Models (second edition) , Cary, NC: SAS Institute Inc.
9. Kramer M (2004) Automatic model selection in the mixed model framework KSU applied statistics conference proceedings.
10. Ngo, L and Rand, R. (2002). Model Selection in Linear Mixed Effects Models Using SAS® Proc Mixed. SUGI 22 http://www2.sas.com/proceedings/sugi22/STATS/PAPER284.PDf
11. SAS Institute (2006) The GLMSELECT procedure (Experimental) Cary, NC http://support.sas.com/rnd/app/papers/glmselect.pdf
12. Saxton, A.M A Macro for Converting Mean Separation Output to Letter Groupings in PROC MIXED SUGI 23 http://www2.sas.com/proceedings/sugi23/Stats/p230.pdf
13. Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso Journal of the Royal Statistical Society Series B, 58, 267–288.

CONTACT INFORMATION

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

Name: George C. Fernandez, PhD
Enterprise: Director / CRDA University of Nevada - Reno
Address: CRDA/088 Reno, NV 89557
Work phone: (775)-784-4206
Email: gcjf@unr.edu
Web: Http://www.ag.unr.edu/gf