SAS® GLOBAL FORUM 2018

USERS PROGRAM

Model Selection with Higher Order Interactions in SAS® PROC MIXED and GLIMMIX

A user-friendly SAS macro, INTERACTION_SELECT

April 8 – 11 | Denver, CO

#SASGF

Model Selection with Higher Order Interactions in SAS® PROC MIXED and GLIMMIX

Yin Zhang • Nanhua Zhang

Cincinnati Children’s Hospital Medical Center

ABSTRACT

It is common to model a longitudinal outcome using a linear mixed effect model or generalized linear mixed effect model. For example, the effect of traumatic brain injury on behavioral outcomes over time may be moderated by the genetics and family environment, resulting in a four-way interaction of TBI (vs. no TBI), gene, family environment and time since injury. It is tedious to do variable selection involving high-order interactions due to the number of terms and the hierarchical structure of the terms in the model, especially when we have multiple outcomes to consider in the analysis. A user-friendly SAS macro, INTERACTION_SELECT, to perform backward model selection of fixed effects including higher order interactions with a user-specified random and repeated effects using SAS 9.4 PROC MIXED and GLIMMIX is provided. This macro supports user-specified initial model structure including response variable, subject ID, continuous, categorical, user-forced predictors and their two-way or higher interactions. At each step, type III tests of fixed effects that are not involved in higher order terms will be used as a criteria to eliminate predictors. After model selection, significant (e.g., p < 0.05) predictors that are not elements of any higher order interactions and all their lower order predictors will be included in the optimal model.

METHODS

- **Data Preparation**

The macro generates a full model including all user-specified terms with indicators of forced variable (vs. no forced), variable type, and variable order for further merging purpose. In order to avoid hitting the maximum length of name of variable in SAS, all main effects and categorical variables will be assigned an artificial variable name, e.g. V1, V2, etc..

- **Model Specification**

The type of SAS procedures, PROC MIXED or PROC GLIMMIX, were chosen by users. Options are also included in this macro to define random and repeated statements including covariance structures; to suspend/include intercept in models and to define distribution, link function, and/or event category for PROC GLIMMIX.

- **Backward Elimination**

Predictors are separated into *eliminable* (type III tests p-value > 0.05, not a user forced variable, and not be included in higher order interactions) and *non-eliminable* groups for each step of backward selection. Among those eliminable predictors, the term with largest p-value will be eliminated from the model. The backward model selection finished when all p-values of eliminable terms are less than 0.05 or there is only one predictor left in model. After model selection, significant (e.g., p < 0.05) predictors and all their lower order terms will be included in the optimal model.

- **Model Diagnosis and Comparison**

P-values of main effects and user-forced predictors are graphically summarized. Model selection fit statistics, AIC, BIC, -2 Res Log Pseudo-Likelihood, or adjusted Generalized Chi-square are plotted using ODS GRAPHICS options.

Macro Specification

Macro Variables	Description
PATH_DATA	The directory of input dataset.
PATH_OUT	The directory of results.
INDATA	Name of input dataset.
INDATATYPE	Type of input dataset, e.g. SAS, SPSS (DAT), EXCEL (XLSX), and CSV
NOTE	Indicator variable: 0 = suspend notes in SAS log page 1 = display notes in SAS log page
MODEL	Type of models (GLIMMIX or MIXED)
RANDOM_STATE	Random statement in models, e.g. random intercept / subject=subject
REPEATED_STATE	Repeated statement in models, e.g. repeated / subject=subject
EVENT	Test event level
LINK	Type of link function
DIST	Type of distribution
INTERCEPT	Indicator variable: 0 = no intercept in model 1 = Include intercept in model
RESPONSE	Name of response variable
SUBJECT	Variable name of subject ID
MAX_ORDER	Maximum order of interaction terms e.g. $Y = X1 + X1 \cdot X2 + X1 \cdot X2 \cdot X3$ then MAX_ORDER=3
CATEGORICAL	List of categorical covariates. Separate variables by blank space ' '. e.g. X1 X2 X3
FORCE	List of forced covariates. Separate variables by blank space ' '.
MAIN_INTER	List of main effects and their interactions. Please do *NOT* specify categorical variables here. Separate variables/terms by plus sign ‘+’ e.g. X1 + X2 + X3 + X1*X2 + X1*X3

Model Selection with Higher Order Interactions in SAS® PROC MIXED and GLIMMIX

Yin Zhang • Nanhua Zhang

Cincinnati Children’s Hospital Medical Center

RESULTS

- **Data Simulation**

Normally distributed repeated measured response for 500 subjects and 4 repeated (Time) measures are simulated.

Binary response is generated by separating the simulated continuous data by its 50% quantile.

Two binary predictors with different probabilities (x1, x2)

One ordinal predictors (x3)

Five continuous predictors (x4, x5, x6, x7, x8)

Subject ID (Subject) are generated for this example.

Significant main effects and interactions: x1, x2, time, x5*x6, x2*x5*x6, and x2*x5*x6*x7

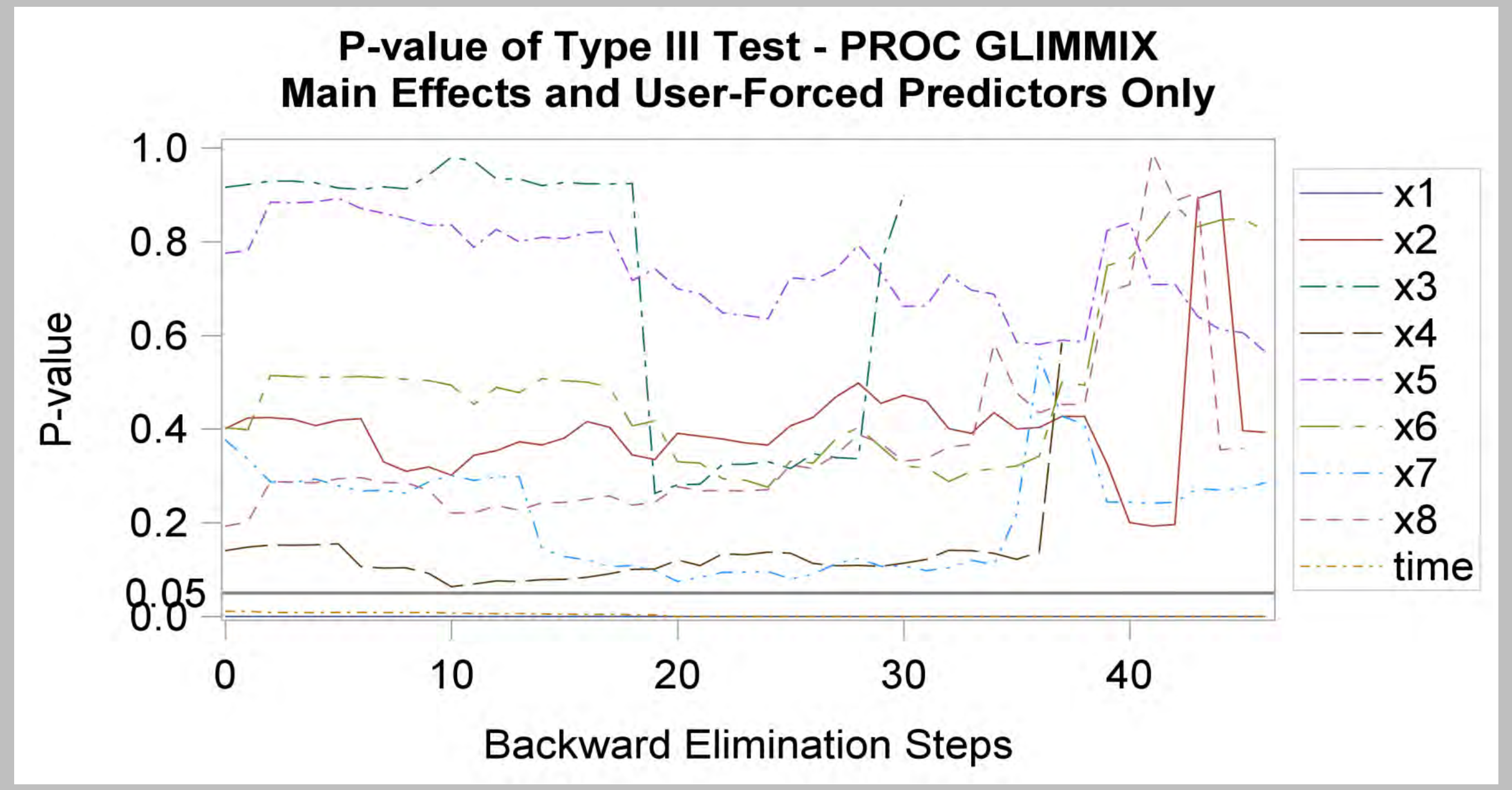
Covariance structure:

Spearman Correlation	Baseline	First Week	Second Week	Third Week
Baseline		0.160	0.105	0.102
First Week	0.160		0.125	0.095
Second Week	0.105	0.125		0.102
Third Week	0.102	0.095	0.102	

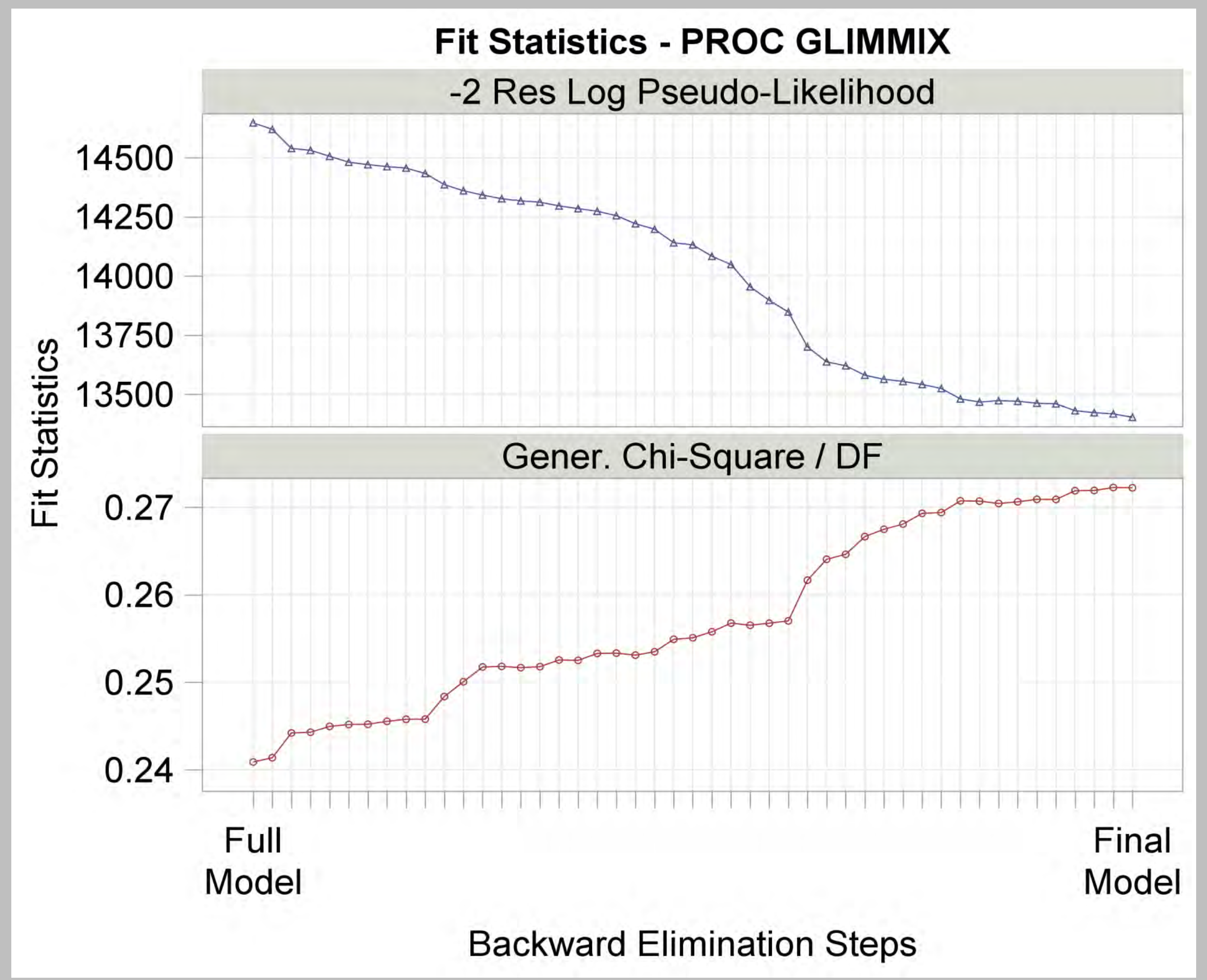
- **Backward Elimination**

Significant main effects, x1 and time are consistently significant through backward steps.

- **Figure 1** Main Effect P-value



- **Figure 2** Fit Statistics



CONCLUSIONS

- **Final Optimal Model**

After the backward elimination, significant (e.g., p < 0.05) predictors and all their lower order predictors will be included in the final optimal model. Significant predictors are X1, X2*time, time, X5*X6, X2*X5*X6, and X2*X5*X6*X7 which capture all the significant interaction terms in the true model (**Table 2**). Also, the final model has the lowest -2 Res Log Pseudo-Likelihood (**Figure 2**) than the full model and all interim models.

- **Conclusions and Discussion**

This user-friendly SAS macro provide an effective way to detect potential higher order interactions and contribute to significant cost and time savings for investigators who are interested in observational studies.

Currently, the maximum order of interactions is six for considering computational expenses. For dataset with 2000 observations and 9 main effects, the total computational time is about 10 minutes. The time also varies by different initial model structures.

- **Table 2** Final optimal model

Variables	User-forced Variable	Eliminable	P-value Type III Test
x1	No	Yes	<.0001
time	No	No	<.0001
x5*x6	No	No	<.0001
x2*time	No	Yes	0.0046
x2*x5*x6	No	No	0.0117
x2*x5*x6*x7	No	Yes	0.0117
x5*x6*x7	No	No	0.0607
x2*x6	No	No	0.1958
x2*x5	No	No	0.2183
x6*x7	No	No	0.2363
x7	No	No	0.2855
x5*x7	No	No	0.3382
x2	No	No	0.3927
x5	No	No	0.5643
x2*x6*x7	No	No	0.5910
x2*x7	No	No	0.6403
x2*x5*x7	No	No	0.7225
x6	No	No	0.8215

REFERENCES

George Fernandez (2007). Model Selection in PROC MIXED – A User-friendly SAS® Macro Application. SAS Global Forum 2007

Fan Pan, Jin Liu (2016). SAS® Macro for Automated Model Selection Involving PROC GLIMMIX and PROC MIXED. SESUG 2016

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF

Paper 2676-2018

Model Selection with Higher-Order Interactions in SASÂ® MIXED and GLIMMIX Procedures

Yin Zhang, Cincinnati Children’s Hospital Medical Center, Cincinnati OH 45069

Nanhua Zhang, Cincinnati Children’s Hospital Medical Center, Cincinnati OH 45069

ABSTRACT

It is common to model a longitudinal outcome using a linear mixed effect model or generalized linear mixed effect model. For example, the effect of traumatic brain injury on behavioral outcomes over time may be moderated by the genetics and family environment, resulting in a four-way interaction of TBI (vs. no TBI), gene, family environment and time since injury; the model would involve even higher interaction if we test whether gender or parental education moderate the effect of traumatic brain injury on behavioral outcomes. It is tedious to do variable selection involving high-order interactions due to the number of terms and the hierarchical structure of the terms in the model, especially when we have multiple outcomes to consider in the analysis. A user-friendly SAS macro, INTERACTION_SELECT, to perform backward model selection of fixed effects including higher order interactions with a user-specified random and repeated effects using SAS 9.4 PROC MIXED and GLIMMIX is provided. This macro supports user-specified initial model structure including response variable, subject ID, continuous, categorical, user-forced predictors and their two-way or higher interactions. Options are also included in this macro to define random and repeated statements including covariance structures. At each step, type III tests of fixed effects that are not involved in higher order terms will be used as a criteria to eliminate predictors. After model selection, significant (e.g., p < 0.05) predictors that are not elements of any higher order interactions and all their lower order predictors will be included in the optimal model. Model selection fit statistics, AIC, AICC, BIC for PROC MIXED or -2 Res Log Pseudo-Likelihood, Generalized Chi-square, and adjusted Generalized Chi-square for PROC GLIMMIX are summarized graphically.

KEY WORDS: backward selection, PROC MIXED, PROC GLIMMIX, high-order interactions, type III tests, model selection fit statistics

INTRODUCTION

It is common to model a longitudinal outcome using a linear mixed effect model or generalized linear mixed effect model. For example, the effect of traumatic brain injury on behavioral outcomes over time may be moderated by the genetics and family environment, resulting in a four-way interaction of TBI (vs. no TBI), gene, family environment and time since injury; the model would involve even higher interaction if we test whether gender or parental education moderate the effect of traumatic brain injury on behavioral outcomes. It is tedious to do variable selection involving high-order interactions due to the number of terms and the hierarchical structure of the terms in the model, especially when we have multiple outcomes to consider in the analysis.

METHOD

The SAS macro %INTERACTION_SELECT implements backward model selection based on type III tests p-values through PROC MIXED or PROC GLIMMIX. Predictor with largest p-value among those independent terms that are not elements of any higher order interactions is eliminated from model. After model selection, significant (e.g., p < 0.05) predictors that are not elements of any higher order interactions and all their lower order predictors will be included in the optimal model. Model selection fit statistics, AIC, AICC, BIC for PROC MIXED or -2 Res Log Pseudo-Likelihood, Generalized Chi-square, and adjusted Generalized Chi-square for PROC GLIMMIX for all models including full model, optimal model and all intermediate models are summarized graphically. Likelihood ratio test will be performed on the full model vs. the final model.

STEP1: DATA PREPARATION

Based on user-defined response variable, subject ID, continuous, categorical, forced predictors and their interactions, the macro generates a full model including all user-specified terms with indicators of forced variable (vs. no forced), variable type, and variable order for further merging purpose. In order to avoid hitting the maximum length of name of variable in SAS, all main effects and categorical variables will be assigned an artificial variable name, e.g. V1, V2, etc.. All these model information will be saved as MODEL_FULL in a temporary SAS data in WORK library. The input data file can be any of permanent SAS, EXCEL (XLSX), SPSS (DAT), or CSV files. Variable names in the input data will be changed to be matched with names in MODEL_FULL.

STEP2: MODEL SPECIFICATION

The type of SAS procedures, PROC MIXED or PROC GLIMMIX, were chosen by users. Options are also included in this macro to define random and repeated statements including covariance structures; to suspend/include intercept in models and to define distribution, link function, and/or event category for PROC GLIMMIX.

STEP3: BACKWARD ELIMINATION

Predictors including categorical variables are separated into eliminable and non-eliminable groups for each step of backward selection. A term is eliminable if it is not a component of a higher order interaction. For example, suppose a model is like: $Y = X_1 + X_2 + X_3 + X_4 + X_1 \cdot X_2 + X_1 \cdot X_3 + X_1 \cdot X_2 \cdot X_3$ then X_1, X_2, X_3, $X_1 \cdot X_2$, and $X_1 \cdot X_3$ are non-eliminable; and $X_4 \text{ and } X_1 \cdot X_2 \cdot X_3$ are eliminable. Among those eliminable predictors, the term with largest type III tests p-value will be eliminated from the model. User-forced predictors are included in each model without elimination. The backward model selection finished when all p-values of eliminable terms are less than 0.05 or there is only one predictor left in model. After model selection, significant (e.g., $p < 0.05$) predictors that are not elements of any higher order interactions and all their lower order predictors will be included in the optimal model.

STEP4: MODEL DIAGNOSIS AND COMPARISON

P-values of main effects and user-forced predictors are graphically summarized (Figure 1). Model selection fit statistics, AIC, AICC, BIC for PROC MIXED or -2 Res Log Pseudo-Likelihood, Generalized Chi-square, and adjusted Generalized Chi-square for PROC GLIMMIX are plotted using ODS GRAPHICS options (Figure 2). Likelihood ratio test will be performed on the full model vs. the final model. Output Delivery System (ODS) are used to save the figures and tables.

MACRO

Macro Variables	Description
PATH_DATA	The directory of input dataset.
PATH_OUT	The directory of results.
INDATA	Name of input dataset.
INDATATYPE	Type of input dataset, e.g. SAS, SPSS (DAT), EXCEL (XLSX), and CSV
NOTE	Indicator variable: 0 = suspend notes in SAS log page 1 = display notes in SAS log page
MODEL	Type of models (GLIMMIX or MIXED)
RANDOM_STATE	Random statement in models, e.g. random intercept / subject=subject
REPEATED_STATE	Repeated statement in models, e.g. repeated / subject=subject
EVENT	Test event level
LINK	Type of link function
DIST	Type of distribution
INTERCEPT	Indicator variable: 0 = no intercept in model 1 = Include intercept in model
RESPONSE	Name of response variable
SUBJECT	Variable name of subject ID

MAX_ORDER	Maximum order of interaction terms e.g. $Y = X1 + X1 \cdot X2 + X1 \cdot X2 \cdot X3$ then MAX_ORDER=3
CATEGORICAL	List of categorical covariates. Separate variables by blank space ' '. e.g. X1 X2 X3
FORCE	List of forced covariates. Separate variables by blank space ' '.
MAIN_INTER	List of main effects and their interactions. Do NOT specify categorical variables here. Separate variables/terms by plus sign '+', e.g. X1 + X2 + X3 + X1*X2 + X1*X3

EXAMPLE

DATA SIMULATION

Normally distributed repeated measured response for 500 subjects and 4 repeated (Time) measures are simulated. Binary response is generated by separating the simulated continuous data by its 50% quantile. Two binary predictors with different probabilities (x1 and x2), one ordinal predictors (x3), five continuous predictors (x4 – x8), and subject ID (Subject) are generated for this example.
Significant predictors: x1, x2, time, x5*x6, x2*x5*x6, and x2*x5*x6*x7

MACRO SPECIFICATION

```
%Interaction_select(
   path_data= Directory of input data,
   path_out=Directory of output,
   indata=sim,
   indatatype=sas,
   note=1,
   intercept=0,
   model=glimmix,
   dist=binary,
   event=1,
   link=logit,
   response=Y,
   subject=subject,
   random_state=random intercept / subject=subject,
   repeated_state=,
   categorical=x1 x2 x3 time,
   force=,
   max_order=4,
   main_inter=x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + time + x2*x3 + x2*x4
   + x2*x5 + x2*x6 + x2*x7 + x2*x8 + x2*time + x3*x4 + x3*x5 + x3*x6 +
   x3*x7 + x3*x8 + x3*time + x4*x5 + x4*x6 + x4*x7 + x4*x8 + x4*time +
   x5*x6 + x5*x7 + x5*x8 + x5*time + x6*x7 + x6*x8 + x6*time + x7*x8 +
   x7*time + x8*time + x2*x3*x4 + x2*x3*x5 + x2*x3*x6 + x2*x3*x7 +
   x2*x3*x8 + x2*x3*time + x2*x4*x5 + x2*x4*x6 + x2*x4*x7 + x2*x4*x8 +
   x2*x4*time + x2*x5*x6 + x2*x5*x7 + x2*x5*x8 + x2*x5*time + x2*x6*x7 +
   x2*x6*x8 + x2*x6*time + x2*x7*x8 + x5*x6*x7 + x5*x6*x8 + x5*x7*x8 +
   x6*x7*x8 + x2*x5*x6*x7 + x2*x5*x6*x8 + x2*x5*x7*x8 + x2*x6*x7*x8
);
```

FINAL OPTIMAL MODEL

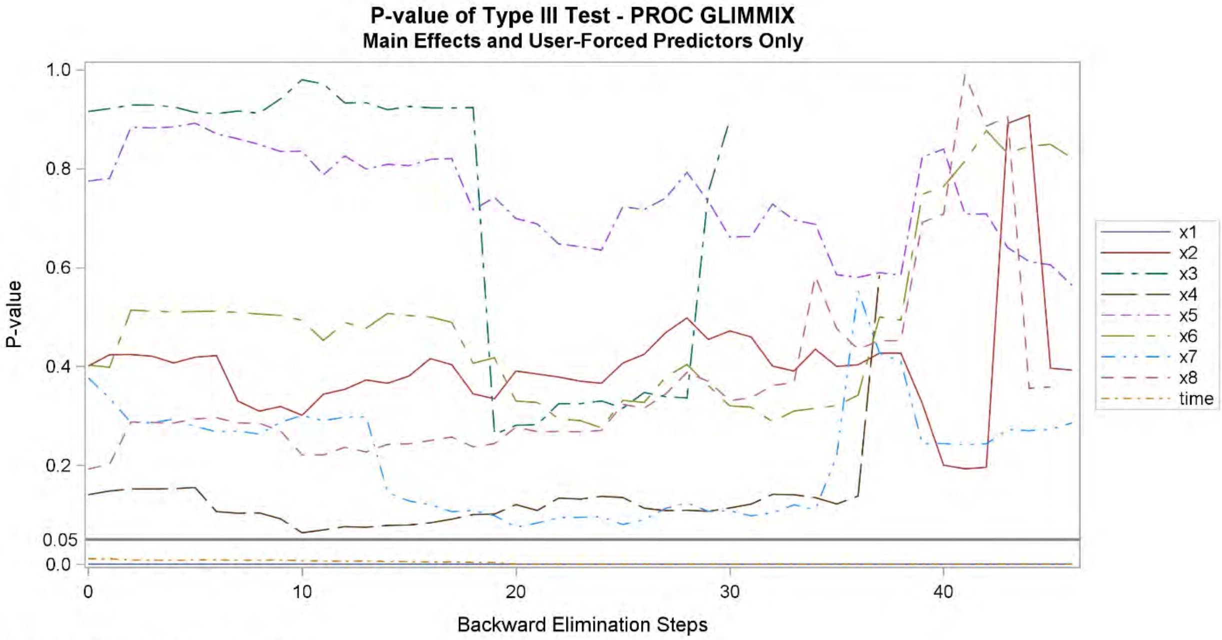
After the backward elimination, significant (e.g., $p < 0.05$) predictors and all their lower order predictors will be included in the final optimal model. Significant predictors are X1, X2*time, time, X5*X6, X2*X5*X6, and X2*X5*X6*X7 which capture all the significant interaction terms in the true model. Also, the final model has the lowest -2 Res Log Pseudo-Likelihood than the full model and all interim models.

Table 1

Summary of Backward Model Selection - PROC GLIMMIX
Final Optimal Model

Variables	User-forced Variable	Eliminable	P-value Type III Test
x1	No	Yes	<.0001
time	No	No	<.0001
x5*x6	No	No	<.0001
x2*time	No	Yes	0.0046
x2*x5*x6	No	No	0.0117
x2*x5*x6*x7	No	Yes	0.0117
x5*x6*x7	No	No	0.0607
x2*x6	No	No	0.1958
x2*x5	No	No	0.2183
x6*x7	No	No	0.2363
x7	No	No	0.2855
x5*x7	No	No	0.3382
x2	No	No	0.3927
x5	No	No	0.5643
x2*x6*x7	No	No	0.5910
x2*x7	No	No	0.6403
x2*x5*x7	No	No	0.7225
x6	No	No	0.8215

Table 1 Final optimal model with type III test p-values.



No varibale was forced in model.

Figure 1 Main effects' and user-forced predictors' p-value trends through backward model selection.

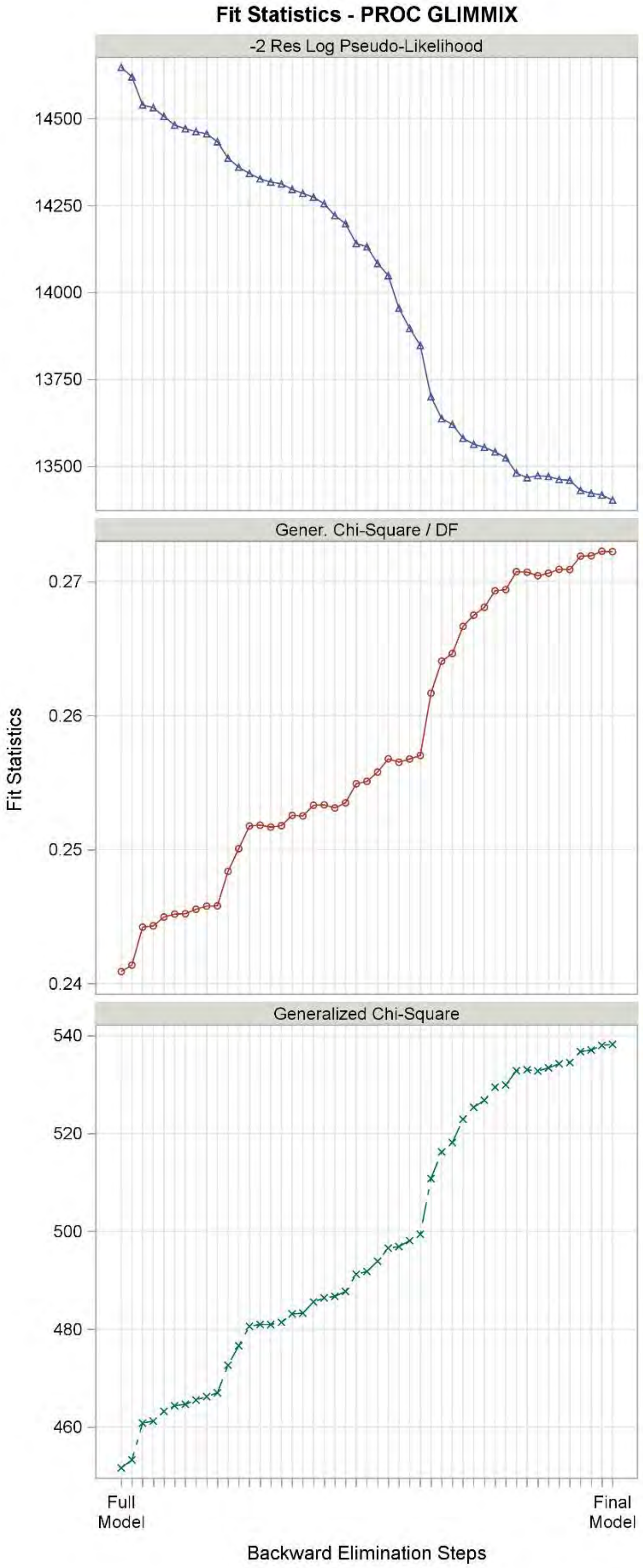


Figure 2 Model fit criteria -2 Res Log Pseudo-Likelihood and Generalized Chi-square by series plot.

Summary of Backward Model Selection - PROC GLIMMIX
Eliminated Variables at Each Step

Variables	Backward Steps	Eliminable	P-value Type III Test
x2*x3*x7	0	Yes	0.9649
x2*x3*time	1	Yes	0.9496
x2*x4*x5	2	Yes	0.9959
x2*x3*x5	3	Yes	0.9673
x2*x3*x4	4	Yes	0.8935
x4*x5	5	Yes	0.8680
x2*x4*x8	6	Yes	0.8126
x2*x5*x7*x8	7	Yes	0.7745
x3*x7	8	Yes	0.7548
x3*x4	9	Yes	0.6938
x2*x3*x6	10	Yes	0.6554
x3*x6	11	Yes	0.8315
x2*x6*x7*x8	12	Yes	0.6599
x6*x7*x8	13	Yes	0.7241
x2*x7*x8	14	Yes	0.6983
x2*x4*x6	15	Yes	0.6147
x2*x4*x7	16	Yes	0.5659
x2*x3*x8	17	Yes	0.5358
x3*x8	18	Yes	0.9739
x8*time	19	Yes	0.5185
x4*x7	20	Yes	0.5799
x2*x4*time	21	Yes	0.4989
x2*x4	22	Yes	0.8373
x4*time	23	Yes	0.7938
x2*x3	24	Yes	0.5214
x7*time	25	Yes	0.5230
x2*x6*time	26	Yes	0.4029
x6*time	27	Yes	0.7967
x3*time	28	Yes	0.3787
x3*x5	29	Yes	0.6448
x3	30	Yes	0.8986
x2*x5*time	31	Yes	0.3505
x5*time	32	Yes	0.3759
x4*x8	33	Yes	0.3207
x5*x7*x8	34	Yes	0.2567
x7*x8	35	Yes	0.2674
x4*x6	36	Yes	0.1639
x4	37	Yes	0.5847
x2*x5*x6*x8	38	Yes	0.0789
x2*x6*x8	39	Yes	0.8519
x5*x6*x8	40	Yes	0.6441
x6*x8	41	Yes	0.8527
x2*x5*x8	42	Yes	0.1459
x5*x8	43	Yes	0.7652
x2*x8	44	Yes	0.5192
x8	45	Yes	0.3587

Table 2 Eliminated variables at each step of backward model selection.

CONCLUSION

This user-friendly SAS macro provide an effective way to detect potential higher order interactions and contribute to significant cost and time savings for investigators who are interested in observational studies.

Currently, the maximum order of interactions is six for considering computational expenses. For dataset with 500 subjects and 4 repeated measures, the total computational time is about 10 minutes. The time also varies by different initial model structures.

REFERENCES

George Fernandez (2007). Model Selection in PROC MIXED – A User-friendly SAS® Macro Application. SAS Global Forum 2007

Fan Pan, Jin Liu (2016). SAS® Macro for Automated Model Selection Involving PROC GLIMMIX and PROC MIXED. SESUG 2016

CONTACT INFORMATION

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

Name: Yin Zhang
Organization: Cincinnati Children’s Hospital Medical Center
Address: 3333 Burnet Ave., Cincinnati, OH 45229
Work Phone: (513)803-0932
Email: Yin.Zhang@cchmc.org