Paper 194-27

Automating the Building of 2nd and 3rd Order Interactions and The Construction of Hierarchical Logistic Regression Models

Mazen Abdellatif, Robert J. Anderson, and Domenic J. Reda
VA Cooperative Studies Program Coordinating Center, Hines, IL

ABSTRACT

We have developed a SAS [1] macro to automate the building of selected second and third order interaction terms of dichotomous variables and the process of the hierarchical backward elimination method in PROC LOGISTIC [2] to build several logistic regression models. The macro creates needed interaction effects, includes them in the data set, and performs the hierarchical backward elimination method until a final logistic regression model is reached. The macro was developed on Release 6.12 TS045 of the SAS® system running on ALPHASERVER Model 1200 5/533 4MB. A system option of LS=132 is required to accurately read and extract needed components from PROC LOGISTIC output.

INTRODUCTION

In the VA Cooperative Study of Sulfasalazine for the Treatment of Seronegative Spondyloarthropathies (CSP #341) [3] [4] [5], we sought to model the binary variable treatment response by examining treatment assignment (Sulfasalazine and Placebo) and various sets of five to twelve dichotomous variables that indicated the presence or absence of selected HLA antigens as independent variables. We used PROC LOGISTIC of SAS to perform variable selection by backward elimination while adhering to the hierarchical model building principal, that is, all effects contained in a significant interaction term must remain in the model, even if these effects are not conditionally significant. For example, if the interaction term ABC is significant, the main effects A, B, and C, as well as the AB, AC, and BC interaction effects must remain in the model.

At the time of this analysis, the hierarchical backward elimination process needed to arrive at a final logistic regression model was very tedious to implement when dealing with 2nd and 3rd order interactions. Unlike PROC GLM, interactions had to be constructed in the SAS dataset. Then, starting with a full model the p-values for the individual effects from the logistic regression output were hierarchically examined in multiple runs to eliminate non-significant effects that are not contained in significant higher order effects, until no more terms could be eliminated. The more independent variables used, the more tedious the process became.

# Main Effects	# Possible 2nd Order Interaction Terms	# Possible 3rd Order Interaction Terms
1	0	0
2	1	0
3	3	1
4	6	4
5	10	10
6	15	20
7	21	35
8	28	56
9	36	84
10	45	120

Figure 1 possible 2nd and 3rd order interactions

Figure 1 illustrates the number of 2nd and 3rd order interactions for 1-10 independent variables. To simplify and speed up this process, we developed the macro described herein.

THE MACRO DESCRIPTION

The first section of the macro defines the input parameters. The code below shows the needed LIBNAME and FILENAME declarations and the descriptions of the needed input parameters followed by macro %LET assignment statements that assign the input values to the input parameters.

```
options ls=132;
/*current directory     */
libname x     '[abdellat.backward]';
/*initial output file   */
filename lst1 '[abdellat.backward]list1.out';
/*subsequent output file*/
filename lst2 '[abdellat.backward]list2.out';
%MACRO backward;
**********************************************;
*                input parameters            *;
* variable/label/input format                *;
* LIB/library of data set/"'[a.a.a.]'"       *;
* DSET/data set/"aaa"                        *;
* DEP/dependent variable/"aaa"               *;
* NMEFF/# main effects/#                     *;
* MEFF/list of main effects/a b c d e f g h  *;
* ORD2/2nd order interaction control/('a','b')*;
* ORD3/3rd order interaction control/('a','b')*;
* PVAL1/main effects p-value/#.##            *;
* PVAL2/2nd order interactions p-value /#.## *;
* PVAL3/3rd order interactions p-value /#.## *;
* WER/Subsetting Where/"           "         *;
* TIT1/Title1/"'aaaaaaaaaaa'"                *;
* TIT2/Title2/"'aaaaaaaaaaa'"                *;
**********************************************;
%GLOBAL LIB DSET DEP INDEP ORD2 ORD3 PVAL;
%LET LIB  ="'[abdellat.backward]'";
%LET DSET ="hla3";
%LET DEP  ="respon";
%LET NMEFF=8;
%LET MEFF =a1 b27 dr1 dr8 dr14 dq2 dq7 v23152;
%LET ORD2 =
('a1','b27','dr1','dr8','dr14','dq2','dq7','v231
52');
%LET ORD3 =('v23152');
%LET PVAL1=0.1500;
%LET PVAL2=0.0500;
%LET PVAL3=0.0500;
%LET WER  ="axper=1";
%LET TIT1 =”’...’”;
%LET TIT2 =”’...’”;
```

The example shows that the dependent variable is treatment response (“respon”) and eight main effects including seven HLA antigens (‘a1’, ‘b27’, ‘dr1’, ‘dr8’, ‘dr14’, ‘dq2’, ‘dq7’) and treatment assignment (‘V23152’). For the second order interactions, all possible unique interactions will be included in the model statement. For the third order interactions, only those that involve treatment assignment (‘V23152’) are included. Main effects will be kept in the model if their p-values are less than or equal to 0.15 while second and third order interactions will be kept in the model if their p-values are less than or equal to 0.05. The (“axper=1”) in the WHERE clause macro variable determines whether to run the model for the axial or peripheral diagnostic group [6].

Next, the macro creates the TERMS SAS data set, shown in Figure 2, of all needed terms from the provided input. Since there are eight main effects given, the macro creates the new eight variables one001, one002, …, one008 and labels them with their corresponding actual main effect names. The number of unique second order interactions for eight main effects is twenty-eight. The macro creates the twenty-eight interactions two001, two002, …, two028 and labels them accordingly. Similarly, it creates and labels the twenty-one unique third order interactions that include treatment assignment.

```
OBS    V                                  ID

  1    one001=a1;                          .
         ...
  8    one008=v23152;                      .
  9    label                               .
 10    one001='a1'                         .
         ...
 17    one008='v23152'                     .
 18    ;                                   .
 19    two001=a1*b27;                      .
        ...
 46    two028=dq7*v23152;                  .
 47    Label                               .
 48    two001='a1*b27'                     .
        ...
 75    two028='dq7*v23152'                 .
 76    ;                                   .
 77    three001=a1*b27*v23152;             .
        ...
 97    three021=dq2*dq7*v23152;            .
 98    Label                               .
 99    three001='a1*b27*v23152'            .
        ...
119    three021='dq2*dq7*v23152'           .
120    ;                                   .
121    ***                                 .
122    one001                           1001
        ...
129    one008                           1008
130    two001                           2001
        ...
157    two028                           2028
158    three001                         3001
        ...
178    three021                         3021
```

Figure 2 the TERMS SAS dataset

Observations above the “***” observation are used to create new variables in the analysis data set, which the macro creates and calls TEMP1.SAS. Observations below the “***” observation are used as independent variables in the PROC LOGISTIC Model statement of TEMP2.SAS, which the macro also creates.

After the macro creates the TERMS SAS data set, it constructs and runs the TEMP1.SAS program shown below to create all needed terms in the analysis data set, which the macro then uses to run PROC LOGISTIC. The program consists of a data step that sets the original analysis SAS data set “HLA3”, creates all new needed variables, and outputs the result to the “Temp” SAS data set.

```
options nodate ls=132  ;
  libname x '[abdellat.backward]';
  data x.temp;
  set  x.hla3;
  format _NUMERIC_ _CHARACTER_;

  one001=a1;
  ...
  one008=v23152;
  LABEL
  one001='a1'
  ...
  one008='v23152';

  two001=a1*b27;
  ...
  two028=dq7*v23152;
  LABEL
  two001='a1*b27'
  ..
  two028='dq7*v23152’;

  three001=a1*b27*v23152;
  ...
  three021=dq2*dq7*v23152;
  LABEL
  three001='a1*b27*v23152'
  ...
  three021='dq2*dq7*v23152';
```

Then the macro constructs and runs the TEMP2.SAS program shown below.

```
  options nodate ls=132;
  libname x
  '[abdellat.backward]' ;
  proc logistic
  data=x.temp;
  where axper=1;
  model respon=
  one001
   ...
  one008
  two001
   ...
  two028
  three001
   ...
  three021
  ;
  title1 ’...';
  title2 ’...’;
```

This program uses a PROC LOGISTIC step to get the initial logistic regression run using all terms. Its output is stored in LIST1.OUT [not shown here]. Next, accessing that List1.OUT file, the macro constructs a new list of main effects and interactions based on the hierarchical backward elimination approach. Then it constructs and runs TEMP3.SAS (example shown below) with the new list of effects.

```
options nodate ls=132;
  libname x
  '[abdellat.backward]';
  proc logistic
  data=x.temp;
  where axper=1;
  model respon=
  one001
  one003
  one004
  one005
  one007
  one008
```

```
two004
two007
two015
two025
three004
;
title1 ’...'  ;
title2 ’...’  ;
```

Finally, the macro enters a loop in which it accesses the output of TEMP3.SAS, reconstructs and re-runs TEMP3.SAS program to get subsequent logistic regression models after excluding non-significant terms until there are no more non-significant terms remaining in the model. At this point, the final iterative execution of the TEMP3.SAS program contains only the final logistic model. The output is sent to LIST2.OUT. In the iterative model-fitting phase, intermediate PROC LOGISTIC steps and their outputs are overwritten.

The output obtained from fitting the final model is stored in LIST2.OUT and is shown in Figure 3. Notice that some interaction terms with non-significant p values are kept in the model because they contain terms that belong to significant higher order interaction terms. The macro output BACKWORD.LIS contains the removed terms and their p values. The first two terms removed are shown in figure 4.

```
                              Analysis of Maximum Likelihood Estimates

               Parameter  Standard     Wald        Pr >      Standardized   Odds   Variable
Variable  DF   Estimate    Error    Chi-Square  Chi-Square    Estimate     Ratio   Label
INTERCPT  1    -0.0380    0.2059     0.0341     0.8534          .            .     Intercept
ONE001    1    -0.1820    0.3396     0.2872     0.5920      -0.044520    0.834     a1
ONE003    1     0.1480    0.2408     0.3778     0.5388       0.035698    1.160     dr1
ONE004    1    -1.1317    0.5011     5.1007     0.0239      -0.137328    0.322     dr8
ONE005    1    -2.1868    0.8182     7.1431     0.0075      -0.402332    0.112     dr14
ONE007    1    -0.4203    0.2315     3.2961     0.0694      -0.104946    0.657     dq7
ONE008    1     0.8103    0.2499    10.5122     0.0012       0.223590    2.249     v23152
TWO004    1     2.9106    1.0491     7.6977     0.0055       0.312369   18.369     a1*dr14
TWO007    1     -0.6597   0.4807     1.8829     0.1700      -0.120419    0.517     a1*v23152
TWO015    1     2.9343    1.3872     4.4744     0.0344       0.189543   18.808     dr1*dr14
TWO025    1     1.7648    0.9694     3.3145     0.0687       0.223134    5.841     dr14*v23152
THREE004  1     -4.7032   2.0304     5.3656     0.0205      -0.277666    0.009     a1*dr14*v23152
```

Figure 3 the final logistic regression model

```
                    Removed Term
OBS          L              MAXP       V         ID
1      dq2*dq7*v23152     0.9986    THREE003    3003
OBS          L              MAXP       V         ID
1      dq2*dq7*v23152     0.999     THREE017    3017
```

Figure 4 BACKWARD.LIS

The following summarizes the major steps in the macro:

1. Define parameters
2. Construct terms dataset
3. Build, include, and run TEMP1.SAS and TEMP2.SAS
4. Get terms with 0 D.F. from the output of the initial run
5. Construct interactions data set from TERMS data set removing terms with 0 D.F.
6. Build TEMP3.SAS.
7. Include and execute TEMP3.SAS
8. Read terms and their p-values from LIST2.out
9. Decide which term to remove and print the removed term
10. Terms to remove? If yes continue, otherwise stop
11. Reconstruct terms list
12. Rebuild TEMP3.SAS
13. Include and run TEMP3.SAS. Loop to step 8

DISCUSSION

This macro handles two functions. The first is to generate the needed interactions in the analysis data set. The second is to automate the process of the hierarchical backward elimination method in PROC LOGISTIC to build several logistic models. We developed the macro using SAS Release 6.12, which did not support any of these functions. However, these functions are supported by SAS Release 8 and above [7]. Interactions can be specified in the MODEL statement by joining main effects with asterisks. For example, the three main effects A, B, and C give three second order interaction and one third order interaction, which can be specified in the MODEL statement as A*B B*C A*C and A*B*C. Interactions can be specified using the @ notation followed by the order number. For example, A B C and all their second order interactions can be obtained by specifying A | B | C@2. Nevertheless, there is no automated method of defining interactions that involve selected main effect terms. This option is implemented in the macro for the second and third order interactions. The hierarchical backward elimination method can be specified using the SELECTION= BACKWARD option and the HIERARCHY= SINGLE option.

In PROC REGRESSION, interaction terms must be variables in the analysis data set. For example, to use the second and third order interactions of VAR1, VAR2, and VAR3, you need to create new three variables in the data set; one for each interaction (INTER1A=VAR1*VAR2; INTER2B=VAR1*VAR3; INTER3A=VAR1*VAR2*VAR3). Those who have not upgraded yet, or do not intend to upgrade to version 8 in the near future might find this macro appealing. An electronic version of the macro can be obtained upon request to abdellat@research.hines.med.va.gov.

CONCLUSION

It was reasonable to use the macro to model treatment response on the HLA antigen variables and their interactions based on their significance level because they are similar in terms of their importance, complexity, and collection cost. We ran selected sets manually to compare their final models to those obtained from the macro. In all cases, the final models were identical. While it took an average of three hours to obtain a final model manually, it took an average of ten minutes using this macro.

While the automation process of defining interaction terms and achieving a final hierarchical logistic regression model is now supported in PROC LOGISTIC of SAS version 8 and higher, the macro part that builds interaction terms can still be used to define terms in the analysis data set for PROC REGRESSION, which does not support defining interactions in its MODEL statement. It could also be handy for PROC LOGISTIC, when only interactions that involve specific main effects are desired.

REFERENCES

1. SAS/MACRO is a registered trademark or a trademark of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration.

2. SAS/STAT® is a registered trademark or a trademark of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration.

3. Clegg, DO, Reda, DJ, Weisman MH, et al. “Comparison of sulfasalazine and placebo in the treatment of ankylosing spondylitis: A Department of Veterans Affairs Cooperative Study.” Arthritis and Rheumatism, 39:2004-2012, 1996.

4. Clegg, DO, Reda, DJ, Mejias E, et al. “Comparison of sulfasalazine and placebo in the treatment of psoriatic arthritis: A Department of Veterans Affairs Cooperative Study.” Arthritis and Rheumatism, 39:213-2020, 1996.

5. Clegg, DO, Reda, DJ, Weisman MH, et al. “Comparison of sulfasalazine and placebo in the treatment of reactive arthritis (Reiter’s Syndrome): A Department of Veterans Affairs Cooperative Study.” Arthritis and Rheumatism, 39:2021-2027, 1996.

6. Clegg, DO, Reda, DJ, Abdellatif, M “Comparison of sulfasalazine and placebo for the treatment of axial and peripheral articular manifestations of the seronegative spondyloartropathies. A Department of Veterans Affairs Cooperative Study.” Arthritis and Rheumatism, 42:11:2325-2329, November 1999.

7. SAS OnlineDoc®, Version 8 is a registered trademark or a trademark of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:
Abdellat@research.hines.med.va.gov