PharmaSUG China

A Macro to Automatically Select Covariates from Prognostic Factors and Exploratory Factors for Multivariate Cox PH Model

Yu Cheng, Eli Lilly and Company, Shanghai, China

ABSTRACT

Multivariate Cox PH model is a widely used analysis to estimate hazard ratio by constructing a model based on the selected variables among a large number of factors. The selection algorithm is normally simplified as: 1) use a full model to include all potential prognostic factors and exploratory variables, 2) select covariates which are significant at a pre-specified alpha level based on certain selection method, 3) fit a reduced model with only selected variables and variables forced into the final model, 4) repeat the above steps and construct other models on different combination of covariates. This paper presents a macro to automatically go through the process and generate final reports for clinical trial reporting use. From the examples described, the purpose is to provide a thought that other programmers can use to automatically generate a batch of analysis reports in the shortest possible time.

INTRODUCTION

Stepwise selection is an automatic procedure to choose prognostic or predictive variables. It can be used in Cox's proportional hazards model to analyze survival data. The selection algorithm is normally simplified as: 1) use a full model to include all potential prognostic factors and exploratory variables, 2) select covariates which are significant at a pre-specified alpha level based on certain selection method, 3) fit a reduced model with only selected variables and variables forced into the final model, 4) repeat the above steps and construct other models.

This theory can be used for any exploratory analysis. From a study level, there could be hundreds of such requests based on the different interests on data. A standardized output layout is in demand to provide all descriptive information about stepwise selection along with the final hazard ratios and p values together. A macro to automatically generate an informative output based on the results from fitting a reduced Cox model by using a selection criterion will help a programmer in simplifying SAS® code and improve work efficiency.

DESCRIPTION OF DATA USED IN THIS PAPER AND SAS® SOFTWARE

The data used in the following example are from Krall, Uthoff, and Harley (1975) who analyzed data from a study on multiple myeloma in which researchers treated 65 patients with alkylating agents (see SAS® 9.2 Help and Documentation – The PHREG Procedure Example 64.1 Stepwise Regression).

In order to simulate the analysis used in a clinical trial, a pseudo variable TRT (treatment) is manually added into the dataset with 0 = Placebo 1=Investigational Product (IP). This variable is added only for testing purpose and has nothing related to the original real data. A character variable PLATELET_ is also created containing the formatted character value of PLATELET.

As state by one paper (See SAS® Global Forum 2010 – Paul T. Savarese and Michael J. Patetta "An Overview of the CLASS, CONTRAST, and HAZARDRATIO Statements in the SAS® 9.2 PHREG Procedure), in SAS® 9.2, PROC PHREG has undergone significant additions, not the least of which is the new CLASS, CONTRAST, and HAZARDRATIO statements. The following SAS® codes are developed based on SAS® version 9.2 for windows.

MACRO PARAMETERS AND ASSUMPTIONS

The macro parameters listed as below.

Table 1. Macro Parameters

Name	Required?/ Default Value	Description
INDS	Y	Input dataset name with time to event information and covariates
SELECTION	Y	SELECTION = method in MODEL statement
SLENTRY	N	SLENTRY = value in MODEL statement

SLSTAY	N	SLSTAY = value in MODEL statement
PARAMDS	Y/PARAM	Parameter data set name with all covariates information
TIME	Y	Variable name of a response variable in INDS
CNSR	Y	Variable name of a censoring variable in INDS
CNSRY	Y/1	A list of censoring values
TRT	Y	Variable name of treatment in INDS
TRTN	Y	Variable name of treatment code in INDS
TRTREF	Y	Reference value of treatment variable TRT

The assumption for input dataset INDS for analysis is that it contains one record per subject and includes all covariates variables and treatment information. In a real clinical trial analysis, the data preprocessing before calling this macro may include 1) merge time to event dataset with a demographics dataset or a baseline characteristics dataset to retrieve covariates 2) subset the datasets to include only subjects in a predefined analysis population for a parameter e.g. overall survival or progressive free disease.

The assumption for PARAMDS parameter dataset is that it contains one record per covariate which will be used in the selection regression model and includes all covariates related information. It must include a variable NAME with a covariate variable name, a variable TYPE indicating the variable attributes C = character N = Numeric, a variable REF contains the reference value of a covariate variable (not necessary for continuous variables), a variable LABEL with the description information of a covariate variable.

To reduce the complexity of the macro, the PHREG procedure and the REPORT procedure in the macro are constructed with some default settings. These settings are set as default because they do not need to be changed once after an initial modification done per study needs. Any changes to these settings would make it a different macro conceptually. In this example, treatment is not included in the original selection model but it is forced into the final model. Option TIES is set to EXACT in the MODEL statement of the PHREG procedure. A simple ODS RTF statement is used for testing purpose.

MODEL IMPLEMENTATION DETAILS

Before the macro run, it is required to build a parameter dataset with the information of all covariates used in the original selection model. This dataset is built by hands written codes and is the only required preprocessing before calling the macro %select. All information in this dataset will then be converted to several macro variables to enable automation.

NAME	TYPE	REF	LABEL
logBUN	N		Log(BUN) at diagnosis
HGB	N		Hemoglobin at diagnosis
Platelet_	C	Abnormal	Platelets at diagnosis
AGE	N		Age at diagnosis
LogWBC	N		Log(WBC) at diagnosis
Frac	N		Fractures at diagnosis: 0=none, 1=present
logPBM	N		Log percentage of plasma cells in bone marrow
Protein	N		Proteinuria at diagnosis
SCalc	N		Serum calcium at diagnosis

Figure 1. Parameter dataset of covariates

At the beginning of the macro, the values in the PARAMDS are converted to local macro variables. PN1-PNx is for variable names where x indicates the number of covariates. PL1-PLx is the description of covariates. PR1-PRx is the reference level for the model construction. PRR1-PRRx is the reference level of covariates for data selection. PM1-PMx is the missing value of covariates. PNUM is the number of covariates.

```
DATA _NULL_;
  SET &paramds END = eof;
  CALL SYMPUT("pn"||strip(put(_N_,best.)),STRIP(UPCASE(name)));
  CALL SYMPUT("pl"||strip(put(_N_,best.)),STRIP(label));
```

```
  IF ref^="" THEN DO;
    IF type='C' THEN
      CALL SYMPUT("prr"||STRIP(PUT(_N_,best.)),'"'||STRIP(ref)||'"');
    ELSE IF type='N' THEN
      CALL SYMPUT("prr"||strip(put(_N_,best.)),STRIP(ref));
  END;
  ELSE DO;
      CALL SYMPUT("prr"||STRIP(PUT(_N_,best.)),"");
  END;
  IF ref^="" THEN DO;
    IF type='C' THEN
      CALL SYMPUT("pr"||STRIP(PUT(_N_,BEST.)),'"'||STRIP(ref)||'"');
    ELSE IF type='N' THEN
      CALL SYMPUT("pr"||STRIP(PUT(_N_,BEST.)),'"'||STRIP(ref)||'"');
  END;
  ELSE DO;
      CALL SYMPUT("pr"||STRIP(PUT(_N_,BEST.)),"");
  END;
  IF type='C' THEN  CALL SYMPUT("pm"||STRIP(PUT(_N_,BEST.)),'""');
  ELSE IF type='N' THEN  CALL SYMPUT("pm"||STRIP(PUT(_N_,BEST.)),'.');
  IF eof THEN CALL SYMPUT("pnum",STRIP(PUT(_N_,BEST.)));
RUN;
```

The full model is fitted by using the defined local macro variables above. ParameterEstimates dataset is saved for further use in the reduced model. A macro variable INICOV is used to concatenate all covariates description together for further use in reporting.

```
ODS OUTPUT ParameterEstimates=pe;
PROC PHREG DATA=adtte ;
  WHERE
  %DO j=1 %TO &pnum;
    &&pn&j ^= &&pm&j and
  %END;
  1 ;
  class
    %DO j=1 %TO &pnum;
      %IF &&pr&j^= %THEN %DO;
        &&pn&j
     %END;
    %END;
   / ORDER=INTERNAL;
    MODEL &time*&cnsr(&cnsrv) =
   %LET inicov=;
   %DO j=1 %TO &pnum;
      &&pn&j
      %LET inicov=&inicov%STR(;)&&pl&j;
    %END;
   / TIES=EXACT SELECTION=&selection %IF &SLENTRY^= %THEN %DO; SLENTRY=&slentry
%END; %IF &SLSTAY^= %THEN %DO; SLSTAY=&slstay  %END;;
RUN;
```

The next step is to define a series macro variables used for the final model. The output datasets ParameterEstimates from the previous full model is used in this step. Similarly, VN1-VNy is for variable names where y indicates the number of selected covariates. VR1-VRy is point variable containing the sequence number of selected covariates in the original covariates. VNUM is the number of selected covariates.

```
DATA _NULL_;
  SET pe_ END=eof;
  CALL SYMPUT("vn"||STRIP(PUT(_N_,BEST.)),STRIP(UPCASE(parameter)));
  %DO j=1 %TO &pnum;
     IF STRIP(UPCASE(parameter)) = STRIP(UPCASE("&&pn&j")) THEN
        CALL SYMPUT("vr"||STRIP(PUT(_N_,BEST.)),"&j");
  %END;
```

```
  IF eof THEN CALL SYMPUT("vnum",STRIP(PUT(_N_,BEST.)));
RUN;
```

When it comes to the final reduced model, to get the final hazard ration and confidence interval, option RISKLIMITS is added to the MODEL statement and REF=reference level is added in the CLASS statement with the underlying default parameterization option PARAM=REFERENCE. Here all covariates information is from the previously defined original covariates macro variables pool (PN1-PNx, PR1-PRx, etc.) by using the point macro variable VR1-VRy. In this example, TRT is forced into the reduced model. Some codes to rearrange the data from ParameterEstimates added after this model for reporting purpose are required.

```
ODS OUTPUT ParameterEstimates=pe2;
PROC PHREG DATA=adtte;
  WHERE &trt ^=""
    %DO i=1 %TO &vnum;
      and  &&&&pn&&vr&i ^= &&&&pm&&vr&i
    %END;
  ;
  CLASS &trt(REF="&trtref")
    %DO i=1 %TO &vnum;
      %IF &&&&pr&&vr&i^= %THEN %DO;
      &&&&pn&&vr&i(ref=&&&&pr&&vr&i)
      %END;
    %END;
   / order=internal param=ref;
  MODEL  &time*&cnsr(&cnsrv) =&trt
    %DO i=1 %TO &vnum;
     &&&&pn&&vr&i
    %END;
  /TIES=EXACT RL;
RUN;
```

As discussed before at the beginning of this paper, the original selection regression information is expected to be showed in the final output. The following code is built to add title and footnotes for the final output. Here FOOTNOTE2 and FOOTNOTE3 are used to describe the selection method used in the selection model. FOOTNOTE4 is to list the original covariates in a readable format. For testing purpose, I use a simple ODS RTF statement here. FOOTNOTE4 then is going to be showed correctly although its length is longer than 262 by using the option NOQUOTELENMAX. Additional processing might be needed if the length of footnote is critical in your output reporting environment. FOOTNOTE5, FOOTNOTE6 and FOOTNOTE7 are for final reduced model description. Following the title and footnote setup, a PROC REPORT is used to generate the final output.

```
ODS LISTING CLOSE;
ODS RTF FILE="H:\pharmasug\test.rtf" STYLE=statistical BODYTITLE;
TITLE "Multivariate Cox Proportional Hazard model of Overall Survival";
FOOTNOTE1 "Abbreviations: N = total population size; CI = Confidence Interval.";
FOOTNOTE2 "Note: Hazard Ratio was estimated using a multivariate Cox Proportional
Hazard model by &selection selection method. The &selection";
%IF %lowcase(&selection)=stepwise %THEN %DO;
FOOTNOTE3 "selection used p-value <&slentry as the criterion for adding a variable
and p-value >= &slstay for dropping a variable.";
%END;
%ELSE %IF %lowcase(&selection)=backward %THEN %DO;
FOOTNOTE3 "selection used p-value >= &slstay for dropping a variable.";
%END;
%ELSE %IF %lowcase(&selection)=forward %THEN %DO;
FOOTNOTE3 "selection used p-value <&slentry as the criterion for adding a
variable.";
%END;
FOOTNOTE4 "Covariates in the inital model include &inicov";
FOOTNOTE5 "Treatment is not used for the &selection model, but is forced into the
final model. HR for treatment effect and corresponding";
FOOTNOTE6 "95% CI estimated from the final model.";
FOOTNOTE7 "a - Wald's p-value and exact method is used to handle ties.";
```

THE OUTPUT

The output from the macro is a RTF file containing the results from the final reduced model in the body section with the description of both full model and reduced mode in the footnote section. Factor is a column to describe the effects. For categorical variable or nominal variable or ordinal variable with a reference level specified in the parameter PARAMDS dataset REF column, this column shows a combination of covariates description (PL1-PLx), alternative level(ClassVal0 from ParameterEstimates) vs. reference level (PR1-PRx). For continuous covariates, this column shows covariates description (PL1-PLx) as continuous. N for Reference Level and N for Alternative level are two columns for counts included into the model which will be missing for continuous covariates. Hazard Ratio (95% CI) and p-value are two columns contain the final results we are going to look at. Because of page limits, codes to generate the first three columns and to handle the final datasets are not showed here.

Multivariate Cox Proportional Hazard model of Overall Survival

Factor	N for Reference Level	N for Alternative Level	Hazards Ratio (95% CI)	p-value*a
Treatment IP vs. Placebo (reference)	32	33	1.253(0.671,2.342)	.4794
Hemoglobin at diagnosis as Continuous			0.900(0.800,1.013)	.0797
Log(BUN) at diagnosis as Continuous			5.342(1.572,18.16)	.0073

Abbreviations: N = total population size; CI = Confidence Interval.
Note: Hazard Ratio was estimated using a multivariate Cox Proportional Hazard model by stepwise selection method. The stepwise selection used p-value <0.05 as the criterion for adding a variable and p-value >= 0.10 for dropping a variable.
Covariates in the inital model include ;Log(BUN) at diagnosis;Hemoglobin at diagnosis;Platelets at diagnosis;Age at diagnosis;Log(WBC) at diagnosis;Fractures at diagnosis: 0=none, 1=present;Log percentage of plasma cells in bone marrow;Proteinuria at diagnosis;Serum calcium at diagnosis
Treatment is not used for the stepwise model, but is forced into the final model. HR for treatment effect and corresponding 95% CI estimated from the final model.
a - Wald's p-value and exact method is used to handle ties.

Figure 2. Sample output

EXAMPLE OF MACRO CALL

The following codes call the macro %select. The macro call the input data and fit the selection regression by using stepwise method with significant level for entering effects as 0.05 and removing effects as 0.10. Censor indicator in input dataset is 0. Another preprocessing required is a parameter dataset in PARAMDS. The sample codes to generate this dataset can be found in the previous section and is omitted here.

```
OPTIONS  NOQUOTELENMAX  NOMPRINT NOCENTER NOSYMBOLGEN ORIENTATION=LANDSCAPE
PS=50 LS=150 NODATE NONUMBER NOBYLINE MISSING = ' ' NOMLOGIC FORMCHAR='|----
|+|---+=|-/\<>*'  ;
%select(inds=Myeloma_, selection=stepwise, slentry=0.05,
slstay=0.10,time=time,cnsr=vstatus, cnsrv=0,trt=trtp, trtn=trt,
trtref=Placebo);
```

CONCLUSION

The macro has been developed to automatically generate an informative report for selection regression in COX PH model. Setting up the covariates dataset is not time consuming. The macro was developed to support a large number of exploratory analyses. It will accommodate the reviewer to look at the accurate result quickly and reduce the programming workload.

We note that this macro will not work as expected if interaction effects is asked to be in the final regression because option RISKLIMITS only produces confidence intervals for hazard ratios of main effects not involved in interactions or nesting. In this case, the macro will need an update on the final model to use HAZARDRATIO statement instead.

REFERENCES

Paul T. Savarese, Michael J. Patetta (2010) "An Overview of the CLASS, CONTRAST, and HAZARDRATIO Statements in the SAS® 9.2 PHREG Procedure
https://www.google.com/url?q=http://support.sas.com/resources/papers/proceedings10/253-2010.pdf&sa=U&ved=0CAQQFjAAahUKEwjKggn8gt_GAhWTGJIKHbXnBgo&client=internal-uds-cse&usg=AFQjCNGENhKIKA54rId0EJx5N71Ts4z_fA

Quan Jenny Zhou, Bala Dhungana (2012) " A SAS® Macro for Biomarker Analysis using Maximally Selected Chi-Square Statistic With Application in Oncology" http://www.lexjansen.com/pharmasug/2012/SP/PharmaSUG-2012-SP12.pdf

SAS Institute Inc. 2009 SAS/STAT® 9.2 User's Guide, Second Edition. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf

ACKNOWLEDGMENTS

The authors would like to thank all of my colleagues who reviewed this paper and provided insight comments

CONTACT INFORMATION

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

Name: Yu Ella Cheng
Enterprise: Eli Lilly and Company
Address: Lilly Suzhou Pharmaceutical Co., Ltd Shanghai Branch
City, State ZIP: Shanghai 200021 P.R. China
E-mail: cheng_yu_ella@lilly.com