Two SAS® Macros on Backward Elimination in Firth's penalized partial likelihood Approach: Firth's Logistic and Cox Models

Chao Zhang, Manali Rupji, Yuan Liu, and Jeanne Kowalski*, Emory University

ABSTRACT

The observation of monotone likelihood or separation may occur in the fitting process of a Cox or a logistic model and is indicated by at least one parameter estimate diverging to ±∞. This observation is typically seen in small sample sizes with rare events. Firth's correction is an approach to reduce the bias of maximum likelihood estimates in the setting of either logistic or cox regression models. While such a correction may be implemented using current SAS procedures (version 9.4), there is no option for performing variable selection within the context of a multivariable analysis. One type of variable selection in particular, backward elimination, has an advantage over other methods since it is possible for a set of variables to have considerable predictive capability even though a subset of them does not. To address this limitation, we developed two SAS macros in which to apply Firth's correction with backwards elimination in the setting of either logistic or Cox regressions. In addition to including the option for backward elimination, our macros generate summary reports in the form of tables displaying results that include, Hazard Ratio in the case of cox regression or Odds Ratio in the case of logistic regression, with its 95% confidence intervals, sample size, and p-values.

INTRODUCTION

The phenomenon of monotone likelihood occurs in the fitting process of a Cox model when at least one parameter estimation diverges to ± infinite. Monotone likelihood is mainly observed in small samples with rare events (Heinze, 2001). To address this monotone likelihood issue, the Firth's penalized partial likelihood Cox regression approach (Firth, 1993) was recommended, as it has been shown to decrease bias in parameter estimates on survival dataset with rare events (Lin et al., 2013). Firth's correction approach was also used to logistic regression models to solve the separation issue (Heinze, 2006), which is similar to the monotone likelihood problem in survival analysis.

As biostatistician our goal is to set out to produce high quality, professional looking analysis reports to enhance communication and collaboration with investigators (Nickleach D, 2013). Those SAS macros were created to produce tables and/or graphs in a rich text format (RTF) file with the goal of succinct information. The macros conduct backward elimination in an automatic fashion which has been seen to overcome drawbacks with other selection approaches such as forward selection which may render one or more of the already included variables non-significant. Backward elimination has advantage over other methods as it is can retain a set of variables that have considerable predictive capability over those which do not. These macros can be extremely helpful in saving time for statisticians to process many independent variables.

The paper is organized as bellows: First, we briefly introduce the each of the macro-parameter, and then in the following section we showcase the usability of the macro with an example dataset along with the output summary table.

DEFINING MACRO PARAMETERS AND EXAMPLES

1 THE MACRO %FirthPhreg_sel

The macro has 13 parameters in which 7 parameters must be provided by the users when the macro is called: The parameters are listed in Table 1.1

Macro variable	Description	required
DSN	The name of the data set to be analyzed.	Yes
EVENT	Variable name of time to event outcome.	Yes
CENSOR	Variable name of censoring indicator variable. Values of 0 indicate censored.	Yes
VAR	The list of variables on interest in the initial model that would be eliminated during the backward selection procedure separated by spaces. The order of variables in this list will be preserved in the final report.	Yes
CVAR	The list of categorical variables that are in VAR, separated by space. If need to change the reference group, you can follow each variable name by (DESC) or by (ref = "Ref level in formatted value") where needed and separate terms by *.	Yes
INC	Number of variables to be forced in the model. The first n variables in the VAR parameter will be included in every model. The default value is 0.	Optional
ALPHA	The significance level for removing variables from the model. The default value is .2.	Optional
REPORT	Set it to T if a results summary table is desired. Otherwise check Log for variable selected by the backward elimination.	Optional
TYPE3	Set to F to suppress type III p-values from being reported in the table (optional). The default value is T. This only has an effect if REPORT = T.	Optional
ORIENTATION	Orientation of the output Word table. Default is portrait, can be changed to landscape.	
FILENAME	File name for output table. This is necessary if report=T.	Yes
OUTPATH	File path for output table to be stored. This is necessary if report=T.	Yes
DEBUG	Set to T if running in debug mode (optional). Work datasets will not be deleted in debug mode. This is useful if you are editing the code or want to further manipulate the resulting data sets. The default value is F.	Optional

Table 1.1 Parameters and the description for % FirthPhreg_sel macro.

MYELOMA STUDY EXAMPLE

The example dataset uses multiple myeloma study data from Krall, Uthoff, and Harley (Krall et al., 1975) (SAS Institute Inc, 2017) in which 65 patients were treated with alkylating agents. Of those patients, 48 of 65 patients died during the study and 17 survived after the study. The key variables are shown as below.

Time: the survival time in months from diagnosis to death or last follow up.

VStatus: indicates survival status, i.e., 0 = Dead, 1 = Death.

And other covariates such as LogBUN, HGB, Platelet, Age, LogWBC, Frac, LogPBM, Protein, and SCalc at diagnosis. The detailed information can be found in SAS User's Guide . To show Firth's correction method by %FIRTHPHREG_SEL, a new independent variable, Contrived, was artificially created, which has the value 1 if the observed time is less than or equal to 65, otherwise has the value 0.

The following SAS program is used to perform the Firth correction survival analysis. And the summary report is shown as Table 1.2.

```
data Myeloma2;
     set Myeloma;
     Contrived= (Time <= 65);
run;

proc format;
  value Contrived  0='>65'  1='<=65';
run;

data Myeloma2;
  set Myeloma2;
  format Contrived Contrived. ;
run;


Title 'Table 1 Multivariate Cox Regression with Firth Correction for Survival ';

%FirthPhreg_sel (DSN =  Myeloma2,
                 EVENT= Time,
                 CENSOR=Vstatus,
                 VAR=LogBUN HGB Contrived,
                 CVAR=Contrived(desc),
                 INC=0,
                 ALPHA= 0.2,
                 TYPE3= t,
                 DEBUG=t,
                 OUTPATH=&dir.\,
                 FILENAME=Multivariable Firth);
Title;
```

		Survival Time		
Covariate	**Level**	**Hazard Ratio (95% CI)**	**HR P-value**	**Type3 P-value**
LogBUN		5.53 (1.79-17.57)	**0.003**	**0.003**
hgb		0.90 (0.79-1.01)	0.066	0.066
Contrived	<=65	15.60 (.-12193.44)	**0.006**	**0.006**
	>65	-	-	

* Number of observations in the original data set = 65. Number of observations used = 65.
** Firth correction Backward selection with an alpha level of removal of 0.2 was used. No variables were removed from the model.

Table 1.2 Multivariate Cox Regression with Firth Correction for Survival

2 THE MACRO %FirthLogistic_sel

The macro contains 14 parameters in which 7 parameters must be provided by the users when the macro is called: The parameters are listed in Table 2.1

Macro variable	**Description**	**required**
DSN	The name of the dataset to be analyzed.	Yes
OUTCOME	The variable name of the outcome. It must be binary or ordinal.	Yes
EVENT	The event category for the binary response model. Specify the value in quotes. This is the argument that will be passed to the event= option in the model statement. Leave this blank if you have an ordinal outcome with more than 2 levels.	Yes
DESC	Set to T to reverse the order of an ordinal outcome. The order will be based on the internal order. Only specify this if the EVENT is blank. The default value is F.	Optional
VAR	A list of variables of interest in the initial model, separated by spaces, which will be initially considered in the following backward selection procedure. List of variables to include in the model separated by spaces.	Yes
CVAR	List of categorical variables to include in the model separated by spaces. These should also appear in the VAR parameter. If you want to change the reference group you can follow each variable name by (DESC) where needed.	Yes

	However, you will need to separate terms with an asterisk instead of a space.	
INC	Number of variables to include in the model (optional). The first *n* variables in the var parameter will be included in every model. The default value is 0.	Optional
ALPHA	The significance level for removing variables from the model. The default value is .2.	Optional
REPORT	Set this to T if you want a table of the resulting model generated (optional). The default value is F.	Optional
TYPE3	Set to F to suppress type III p-values from being reported in the table (optional). The default value is T.	Optional
ORIENTATION	Orientation of the output Word table. It can be set as LANDSCAPE or PORTRAIT (default).	Optional
FILENAME	File name for output table. This is necessary if report=T.	Yes
OUTPATH	File path for output table to be stored. This is necessary if report=T.	Yes
DEBUG	Set to T if running in debug mode (optional). Work datasets will not be deleted in debug mode. This is useful if you are editing the code or want to further manipulate the resulting data sets. The default value is F.	Optional

Table 2.1 Parameters and the description for % FirthLogistic_sel macro.

Neuralgia Study Example

Firth correction method can achieve the almost identical results when the number of events are large enough. To show how to implement %FIRTHLOGISTIC_SEL, we use a well-known data example, Neuralgia Study Example, as SAS document (SAS Institute Inc, 2017). The data set Neuralgia have five variables, i.e., Treatment, Sex, Age, Duration, and Pain. Pain is the outcome, and other 4 variables are covariates. The below SAS program is used to show how to conduct the Firth correction logistic regression. And the summary report is shown as Table 2.2.

```
Title 'Table 3 Multivariate Logistic Regression with Firth Correction Model for Pain';
%FirthLogistic_sel(DSN=Neuralgia,
                   OUTCOME=pain,
                   EVENT=,
                   DESC= F,
                   VAR=Treatment Sex Age Duration,
                   CVAR=Treatment Sex,
                   INC=0,
                   ALPHA=.2,
                   TYPE3=T,
                   REPORT=T,
                   ORIENTATION = portrait,
                   OUTPATH=&dir.\,
                   FILENAME=Firth logistic regression, debug=t);
```

Covariate	Level	Pain Odds Ratio (95% CI)	OR P-value	Type3 P-value
Treatment	A	15.79 (3.08-121.11)	**0.003**	**0.003**
	B	24.60 (4.18-246.95)	**0.002**	
	P	-	-	
Sex	F	4.83 (1.28-23.02)	**0.031**	**0.031**
	M	-	-	
Age		0.80 (0.66-0.93)	**0.009**	**0.009**

* Number of observations in the original data set = 60. Number of observations used = 60.
**Firth correction backward selection with an alpha level of removal of .2 was used. The following variables were removed from the model: Duration.

Table 2.2 Multivariate Logistic Regression with Firth Correction for Pain

CONCLUSION

Firth's correction is an approach to reduce the bias of maximum likelihood estimates in the setting of either logistic or cox regression models when a dataset has too small events. Backward elimination, has an advantage over other methods since it is possible for a set of variables to have considerable predictive capability even though a subset of them does not. The two macros on backward elimination fill in the gap of current SAS procedures. And they enable users to generate tables and plots that may be directly used for publications, presentations. They provide results in a format easy for researchers to understand thus enabling efficient communication and collaboration. The syntax of the macros is too detailed to be provided in appendix of this paper. Please free contact the author directly to get the two macros.

REFERENCES

SAS Institute Inc, 2017 SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc. User's Guide the LOGISTIC Procedure.

SAS Institute Inc, 2017 SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc. User's Guide the PHREG Procedure.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometricka*. 80:27-38.

Heinze, G. 2001. The application of Firth's procedure to Cox and logistic regression, Technical Report 10/1999, update in January 2001, Section of Clinical Biometrics, Department of Medical Computer Sciences University of Vienna, 2001.

Heinze, G. 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine*. 25:4216-4226.

Krall, J.M., V.A. Uthoff, and J.B. Harley. 1975. A step-up procedure for selecting variables associated with survival. *Biometrics*. 31:49-57.
Lin, I.F., W.P. Chang, and Y.N. Liao. 2013. Shrinkage methods enhanced the accuracy of parameter estimation using Cox models with small number of events. *Journal of clinical epidemiology*. 66:743-751.
Nickleach D, L.Y., Shrewsberry A, Ogan K, Kim S, Wang Z. 2013. Macros to Conduct Common Biostatistical Analyses and Generate Reports. *SESUG*.

CONTACT INFORMATION

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

Dr. Jeanne Kowalski
718 Gatewood Road NE, Atlanta, GA 30322
Biostatistics and Bioinformatics Shared Resource
Winship Cancer Institute, Emory University
Work-Phone: 404-778-5305
Email: jeanne.kowalski@emory.edu