Paper 4907-2020

Don't Just Survive, Thrive with This Multi-Stratified Cox Proportional Hazards Model Macro

Katelyn J. Ware, Spectrum Health Office of Research and Education & Grand Valley State University; Paul Egeler, Spectrum Health Office of Research and Education

ABSTRACT

Survival analysis is a commonly used set of techniques for applied data analysis where the outcome variable is the time until an event. One of the most frequently used techniques for modeling this type of data is the Cox proportional hazards model, which can be implemented in SAS® with the PHREG procedure. This model assumes that the ratio of hazards for any two individuals is constant over time, that is, they are proportional. If this assumption is not met for any particular covariate, stratification is one method to still include and control for the effects in the model. It then must be decided if there is an interaction involved between the stratum levels and predictors in the model. However, it can be cumbersome to manually code the interactions with all levels of stratum in SAS when there are multiple stratum variables involved. The new and improved lrt_strat_cox_ph macro simplifies this procedure by allowing for multiple stratum variables in its strata_vars parameter. The new feature makes the macro a powerful tool in applied survival analysis. This paper discusses the use of multiple stratifying variables, how they are implemented in SAS, and includes a practical example using the macro to tie it all together.

INTRODUCTION

During survival analysis modeling, multiple independent variables at times fail to satisfy the proportional hazards assumption, but it might still be appropriate to include them in the model. You can control for these variables in the Cox Proportional Hazards (PH) model with stratification, but not as independent covariates. As described in the Survival Analysis textbook by Kleinbaum and Klein (2012), a stratified Cox PH model identifies variables that increase the likelihood of the event of interest occurring while still controlling for the effect of variables that fail to pass the PH assumption.

When using this stratified version of the model, you need to determine if the stratum variables have an interaction with the covariates in the model. If they do, the model must account for this by including stratum-covariate interaction terms. As usual in statistics, the simpler model (without interaction terms) is desired as it is easier to interpret, but the interaction terms must be included if there truly is an interaction. Normally, you create the model with interaction terms and the model without interaction terms and compare log-likelihood statistics to decide which model to use. This comparison method is called the likelihood ratio test. Originally this was quite cumbersome to complete in SAS; therefore, the lrt_strat_cox_ph macro was created to automate the test for a quicker decision process in deciding between the interaction and no-interaction stratified Cox PH models (Ware & Baxter, 2019). However, this original version only allows for a single stratum variable.

A recently improved version of the macro simplifies checking the interaction assumption by allowing for multiple stratum variables in its strata_vars parameter. It is maintained in a GitHub repo at https://github.com/SpectrumHealthResearch/lrt-strat-cox-ph. This latest version allows for two or more stratum variables by automating the creation of a new stratum variable that is made up of every level of strata from each input stratum variable. It also creates the multiple interaction terms to be used in the interaction model. This paper will explain how multiple stratum variables are incorporated in the SAS macro code.

SURVIVAL ANALYSIS BACKGROUND

There are several fields where the outcome of interest for a statistical analysis is the time until an event occurs. Ecologists might be interested in time until a seed germinates; manufacturers could be interested in the time from production of a component to failure; physicians are often interested in time after a certain diagnosis until death. Survival analysis is the set of techniques which allow for the analysis and modeling of these situations. Variables to collect include whether the event of interest occurs (censoring); the time until the event, withdrawal, or study completion (whichever comes first); and any other covariates of interest.

Kleinbaum and Klein (2012) explain that censoring is a unique characteristic that survival analysis captures. Right censored observations are those that never experience the event of interest due to a wide variety of reasons (loss to follow-up, death, *etc*.). These observations are still included in the analysis rather than treated as missing because they still contribute useful information. Throughout this paper, typical conditions (*i.e.* non-informative censoring, sufficiently large sample sizes, and right censoring) are assumed to be met for likelihood based statistical inference for survival analysis.

As mentioned in the introduction, a commonly used modeling technique for survival analysis is the Cox PH model. This models the instantaneous potential (risk) for the event to occur given that an individual has survived up to time *t* (Kleinbaum and Klein, 2012). This instantaneous potential is called the hazard rate. The model assumes that the hazard rates of each covariate are proportional across the duration of the study (the proportional hazards assumption). For example, females found to have twice the risk of a stroke than men at four months, will also have twice the risk at any other time. This assumption can be checked through various methods in SAS®, such as the observed versus expected plot, the log-log plot, and the goodness of fit test. Gillespie (2006) and Yao (2018) describe several ways to assess the PH assumption with corresponding SAS® code.

NEED FOR STRATA VARIABLES

One option when the hazard rates are not proportional across levels of a variable is to include the variable in the model but not test for its effect. This is called stratifying. The variable is defined as a stratum variable and the model is termed the stratified Cox PH model. If you know or discover that one or more variables do not satisfy the PH assumption, but you believe they affect the hazard rate (especially in relation to other covariates in the model), you should still include them through stratification. Stratification makes the most sense if the variable being stratified on is a nuisance variable.

As described by Liu (2012), stratification is essentially fitting a proportional hazards model for each stratified level. When subpopulations (*i.e.* strata levels) have significantly different distributions of the baseline hazard, the proportionality of their hazard rates becomes questionable. Thus, when you stratify, it allows the underlying hazard function to vary across each strata level and therefore produces more efficient regression coefficients for the other covariates. The effect of each covariate in the model will be estimated and adjusted for any other covariates in the model as well as the stratum variable(s). When there are multiple stratum variables, you essentially create a single new categorical variable whose categories are all combinations of the levels from the multiple stratum variables, with one of the levels being considered the reference category. If any of the stratum variables are quantitative, they must be categorized before utilizing them in the model.

There are instances where stratification is not the answer. If a stratum variable has a lot of levels, there probably will not be a sufficient number of events in each category to appropriately approximate the Cox PH model. This is especially true when you have more than one stratum variable, as the levels created from the single new stratum variable will be many. As described in the paragraph above, the number of levels from stratum variables

are multiplied. For example, if you have two stratum variables, one with 2 levels and another with 3, your single new stratum variable will have (2*3 = 6) six levels. What might have started out as a total of 120 events in a data set of 300 observations would be reduced to a maximum of 20 events for the stratum level with the least number of events. Of course, you can group levels together to create fewer categories altogether, and the more observations and events that you have in a dataset, the less of an issue this becomes.

It is important to note that there are other options besides stratification. In chapter 6 of their textbook, Kleinbaum and Klein (2012) note that you can create the interaction between the covariate and time or some function of time (such as the heavyside function). However, stratification offers the advantage of requiring less computational resources and not having to decide how the covariate varies with time if it is unclear.

INTERACTION

When using the stratified Cox PH model, it must be determined if the regression coefficients in the model vary over the strata levels. This applies if we suspect an interaction between covariates and the strata. If no interaction is assumed, there is only one set of coefficients no matter how many strata there are. On the other hand, if there is an interaction, we expect to obtain distinct regression coefficients for each of the strata. This interaction may be accounted for by including stratum-covariate interaction terms. Essentially, each level of the stratifying variable(s), except the reference level, must be multiplied with each independent covariate.

It must be decided which model (the interaction or no-interaction model) is more appropriate statistically. A commonly used method to test for this is the likelihood ratio test which uses a ratio comparison of maximum likelihood values to determine which model is the best (Kleinbaum et al., 2009). It tests all interaction terms as a whole set. The statistic has an approximate chi-square distribution with degrees of freedom equal to the difference in the number of coefficients between the two models being tested. The p-value associated with this test statistic gives statistical evidence for one of the two models. The test's null hypothesis states the extra (interaction) parameters are equal to 0. Thus, if the p-value is low, there is evidence to reject that the interaction parameters are 0 and it is concluded that there is an interaction effect. The updated macro handles this well for more than one stratum variable in the model.

EXPLANATION OF THE MACRO CODE

The macro models both the interaction and no-interaction model and computes the likelihood ratio test statistic and p-value. A more detailed explanation of this is included in the original paper on the macro. The following will discuss the new changes made.

THE MACRO PARAMETERS

The SAS macro has seven parameters. The changes from the original macro include the strata_vars parameter allowing for more than one covariate and the addition of the class_options parameter. Users can now choose which options to use in the class statement of PROC PHREG, such as the parameterization method or the reference level.

The seven parameters are outlined as follows:

- data = the data set name
- time_var = the event time variable
- censor_var = the censoring indicator variable

- censor_value = the censoring value (value that means an observation is censored)
    - This has a default value of 0
- strata_vars
    - Now allows for more than one stratum variable. Please be cautious to not include more than needed. Check the number of events in each level (included in macro output) to make sure a sufficiently large number is reached.
    - Note: it MUST be categorical, so if it is quantitative, you must categorize it first
- quant_covariates = the names of the numeric covariates in the model
- class_covariates = the names of the categorical covariates in the model
    - Note: the user can include the stratum variable in this parameter if it needs to be created into indicator variables
- class_options = options for the class statement in PROC PHREG

CHANGES TO THE MACRO CODE

Creating all combinations of the levels of each stratum variable

The new and improved macro includes the addition of iteratively creating the single multilevel stratum variable which includes all levels of every stratum variable in the model.

This new portion of the code is as follows:

```
%local strat_int
%let strat_int = %scan(&strata_vars,1);

%if %sysfunc(countw(&strata_vars, %str( )))>1 %then %do;
  %do r=2 %to %sysfunc(countw(&strata_vars));
    %if &r <= %sysfunc(countw(&strata_vars)) %then
    %let strat_int = %sysfunc(catx(|, &strat_int, %scan(&strata_vars,&r)));
  %end;
%end;
```

The first line initiates a macro variable called strat_int, which is set to the first stratum variable input by the user in the strata_vars parameter. The next part begins by filtering out instances where there is only one stratum variable in the model. If it finds that there are more than one, it will begin a loop starting at 2 and going though the number of stratum variables in the model. Each time through, it will be adding to the macro variable strat_int, utilizing the catx() function to tack on the next stratum variable with the '|' notation between each one. This creates a new single stratum variable that looks like <stratum_var_1>|<stratum_var_2>| … |<stratum_var_v>, where there are v stratum variables. Each of the stratum variables that compose this has their own number of levels with a minimum of 2 levels. Since the '|' symbol is essentially multiplication in this context, the number of levels in each strata will be multiplied together to be the total number of levels.

Creating the interaction terms

Each level of the above created stratum variable (minus the reference level) will be multiplied by each covariate in the model to create the interaction terms.

The following code is the updated version of creating these terms:

```
%local interaction_i interaction_vars all_covariates;
%let all_covariates = &quant_covariates &class_covariates;

%do interaction_i = 1 %to %sysfunc(countw(&all_covariates, %str( )));
```

```
    %let interaction_vars = &interaction_vars %sysfunc(catx(|, &strat_int,
          %scan(&all_covariates,&interaction_i)));
  %end;
```

The all_covariates macro variable is initiated as the list of all quantitative and categorical covariates specified by the user. A loop goes from 1 to the number of covariates in the model. Each iteration will create another interaction term comprised of the stratum macro variable strat_int and the i^{th} covariate.

EXAMPLE OF MACRO USE

The example data set, vets, has 137 patients from the Veteran's Administration Lung Cancer Trial and can be found at http://web1.sph.emory.edu/dkleinb/surv3.htm#data. The data set records the survival time in days, a censoring variable, treatment (standard and test), cell type (large, adeno, small, or squamous), prior therapy (none or some), disease duration in months, and age in years. The variable cell type is recorded as four separate binary variables in the original data set. For the purposes of this paper, the four binary variables for cell type were combined into one variable (denoted as ct) with four levels. However, it would also work if left in its original form.

Suppose we are interested in how the above variables affect the hazard rate of time until death and are particularly interested if the new treatment is any better in lengthening survival time. First, we must see if all the variables pass the PH assumption. After running the assumption checks, it was found that prior therapy (denoted as priortx) and ct do not pass. Thus, we will stratify by these two variables to still control for their effect. Note that we choose the class option PARAM=REF.

The SAS code to call the macro is as follows:

```
%lrt_strat_cox_ph( data= vets,
                   time_var = survt,
                   censor_var= status,
                   censor_vals = 0,
                   strata_vars = ct priortx,
                   quant_covariates = age dd,
                   class_covariates = tx,
                   class_opts = param=ref );
```

Output 1 outlines the composition of our new single stratum variable. There are 8 stratum levels made from the 2*4 combination of levels from ct and priortx. The number of events in each appears sufficiently large except for stratum 4 (ct=2 and priortx=10), which only has 5 events. For the purposes of explaining this macro, we will continue, but the user would need to decide if this was appropriate in their own analysis.

Summary of the Number of Event and Censored Values						
Stratum	CT	PRIORTX	Total	Event	Censored	Percent Censored
1	1	0	17	16	1	5.88
2	1	10	10	10	0	0.00
3	2	0	22	21	1	4.55
4	2	10	5	5	0	0.00
5	3	0	37	35	2	5.41
6	3	10	11	10	1	9.09
7	4	0	21	19	2	9.52
8	4	10	14	12	2	14.29
Total			137	128	9	6.57

Output 1. Summary of the Stratum Levels

The first portion of the results includes output from the PHREG procedure fitted for the interaction model. The Type 1 Test table is saved within the macro so that the -2Log*L* can be saved to be used in the computation of the test statistic. This is seen in Output 2.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	543.496	512.127
AIC	543.496	560.127
SBC	543.496	628.576

Output 2. Model Fit Statistics from the Interaction Model

The table of maximum likelihood estimates for the interaction model can be overwhelming the more stratum levels and variables. As seen in Output 3, our model has 24 parameters: 3 covariates and 21 interaction terms. The interaction terms are comprised of the 7 levels of the stratum (the 8th level is the reference category) times the 3 covariates. The parameter estimates of 0 are for the 8 stratum levels, as their effect is not measured in the model.

Analysis of Maximum Likelihood Estimates								
Parameter				DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
AGE				1	0.03270	0.06764	0.2337	0.6288
DD				1	0.05435	0.02652	4.1991	0.0404
TX	1			1	0.63266	0.95314	0.4406	0.5068
CT	1			0	0	.	.	.
CT	2			0	0	.	.	.
CT	3			0	0	.	.	.
PRIORTX	0			0	0	.	.	.
CT*PRIORTX	1	0		0	0	.	.	.
CT*PRIORTX	2	0		0	0	.	.	.
CT*PRIORTX	3	0		0	0	.	.	.
AGE*CT	1			1	-0.07436	0.08083	0.8464	0.3576
AGE*CT	2			1	3.22799	201.52999	0.0003	0.9872
AGE*CT	3			1	-0.01274	0.08193	0.0242	0.8764
AGE*PRIORTX	0			1	-0.01772	0.07285	0.0591	0.8079
AGE*CT*PRIORTX	1	0		1	0.01091	0.09262	0.0139	0.9063
AGE*CT*PRIORTX	2	0		1	-3.26575	201.52999	0.0003	0.9871
AGE*CT*PRIORTX	3	0		1	-0.00936	0.08873	0.0111	0.9160
DD*CT	1			1	0.00447	0.11694	0.0015	0.9695
DD*CT	2			1	1.33544	108.73768	0.0002	0.9902
DD*CT	3			1	-0.06274	0.03361	3.4852	0.0619
DD*PRIORTX	0			1	-0.01520	0.04023	0.1426	0.7057
DD*CT*PRIORTX	1	0		1	0.04791	0.13817	0.1202	0.7288
DD*CT*PRIORTX	2	0		1	-1.19495	108.73774	0.0001	0.9912
DD*CT*PRIORTX	3	0		1	0.01034	0.05474	0.0357	0.8501
TX*CT	1	1		1	-0.81885	1.31232	0.3893	0.5326
TX*CT	1	2		1	35.46872	2344	0.0002	0.9879
TX*CT	1	3		1	-1.41278	1.52874	0.8541	0.3554
TX*PRIORTX	1	0		1	-0.17064	1.12457	0.0230	0.8794
TX*CT*PRIORTX	1	1	0	1	-1.29280	1.61854	0.6380	0.4244
TX*CT*PRIORTX	1	2	0	1	-36.79967	2344	0.0002	0.9875
TX*CT*PRIORTX	1	3	0	1	0.34940	1.68964	0.0428	0.8362

Output 3. Maximum Likelihood Estimates from the Interaction Model

Lastly, the output from the no-interaction model is included in Output 4 and 5 below.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	543.496	542.051
AIC	543.496	548.051
SBC	543.496	556.607

Output 4. Model Fit Statistics from the No-Interaction Model

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
AGE		1	0.00106	0.01027	0.0107	0.9175
DD		1	0.01030	0.00988	1.0855	0.2975
TX	1	1	-0.12256	0.21092	0.3377	0.5612

Output 5. Maximum Likelihood Estimates from the No-Interaction Model

The results in Output 7 below summarize the likelihood ratio test. The p-value = 0.0935 indicates there is not significant evidence to reject the null hypothesis that the interaction parameters are not needed. It can be concluded that the interaction model is not significantly better than the no-interaction model, so the interaction terms are not needed.

```
================================================================================
    Stratified Cox Proportional Hazards Model Likelihood Ratio Test
                        Summary of results
--------------------------------------------------------------------------------
      -2LogLikelihood of the Reduced Model    =                        542.05
      -2LogLikelihood of the Full Model       =                        512.13
      Degrees of Freedom of the Reduced Model =                             3
      Degrees of Freedom of the Full Model    =                            24
      Model Degrees of Freedom                =                            21
      Difference                              =                         29.92
      Chi-Square p-value                      =                        0.0935
================================================================================
```

Output 7. Likelihood Ratio Test Output

CONCLUSION

Depending on the survival analysis data being analyzed, it may be necessary to include multiple stratum variables in a stratified Cox PH model. As was the intent of the original version of the lrt_strat_cox_ph macro, the decision of whether to include interaction terms in a stratified Cox PH model should not have to involve hard coding or hand calculating the likelihood ratio test. The goal of the improved macro is to enable users to seamlessly incorporate two or more stratum variables in their model. Utilizing this macro allows the user to see the interaction versus no interaction model decision making information in one easy to read table, regardless of the number of strata.

REFERENCES

Kleinbaum, David G et al. 2009. Applied Regression Analysis and Other Multivariate Methods. 4th ed., Thompson Higher Education, pp.595-598

Kleinbaum, David G et al. 2012. *Survival Analysis*. 3rd ed. Springer, NY.

Liu, X. 2012. *Survival analysis: Models and applications*. Chichester, West Sussex, United Kingdom: Wiley.

Ware, K. J. and R. Baxter. 2019. "Surviving Survival Analysis 101: Making the Likelihood Ratio Test Easier Using a SAS® MACRO". *Proceedings of the Midwest SAS Users Group Conference*. Available at https://www.mwsug.org/proceedings/2019/RF/MWSUG-2019-RF-100.pdf

ACKNOWLEDGMENTS

The entire Scientific Analysis and Scholarly Support team at Spectrum Health for their feedback on this paper. A special thanks to Jessi Parker for her mentorship and Rachel Baxter for her contributions to the original macro.

Dr. Daniel Frobish at Grand Valley State University for his expertise on the topic.

RECOMMENDED READING

https://github.com/SpectrumHealthResearch/lrt-strat-cox-ph

CONTACT INFORMATION

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

Katelyn J. Ware
Spectrum Health Office of Research and Education
Grand Valley State University
warek@mail.gvsu.edu

APPENDIX: MACRO CODE

```
/*************************************************************************/
/** Stratified Cox Proportional Hazards Model Likelihood Ratio Test Macro **/
/*************************************************************************/
/*                                                                       */
/* Authors: Rachel R. Baxter, B.S.; Katelyn J. Ware, B.A.;               */
/*          Paul W. Egeler, M.S., GStat                                  */
/*                                                                       */
/* Required parameters:                                                  */
/*                                                                       */
/*  data             = The data set name                                 */
/*  time_var         = The event time variable                           */
/*  censor_var       = The censoring indicator variable                  */
/*  censor_vals      = The value(s) for censored individuals             */
/*  strata_vars      = The stratifying variable(s) (MUST BE CATEGORICAL) */
/*  quant_covariates = The names of numeric covariates in the model      */
/*  class_covariates = The names of categorical covariates in the model  */
/*  class_opts       = Options for the class statement in PROC PHREG     */
/*                                                                       */
/*************************************************************************/

%macro lrt_strat_cox_ph(
  data             = /* The data set name                              */,
  time_var         = /* The event time variable                        */,
  censor_var       = /* The censoring indicator variable               */,
  censor_vals      = /* The value(s) for censored individuals          */,
  strata_vars      = /* The stratifying variable(s) (MUST BE CATEGORICAL) */,
  quant_covariates = /* The names of numeric covariates in the model   */,
  class_covariates = /* The names of categorical covariates in the model */,
  class_opts       = /* Options for the class statement in PROC PHREG  */
  );

/* Local variables */
%local error strat_int interaction_i interaction_vars all_covariates;
%let error = 0;
%let all_covariates = &quant_covariates &class_covariates;
%let strat_int= %scan(&strata_vars,1);

/* User Input Processing */
%if ~%sysfunc(countw(&strata_vars, %str( ))) %then %do;
  %put ERROR: strata_vars requires at least one variable;
  %let error = 1;
%end;

%if &error = 1 %then %goto finish;

/* If more than one strata variable, create new term of their combinations
with the '|' */
%if %sysfunc(countw(&strata_vars, %str( )))>1 %then %do;
  %do r=2 %to %sysfunc(countw(&strata_vars));
    %if &r <= %sysfunc(countw(&strata_vars)) %then
    %let strat_int = %sysfunc(catx(|, &strat_int, %scan(&strata_vars,&r)));
  %end;
%end;

/* Create all interactions between strata_vars and covariates */
%do interaction_i = 1 %to %sysfunc(countw(&all_covariates, %str( )));
```

```
  %let interaction_vars = &interaction_vars %sysfunc(catx(|, &strat_int,
%scan(&all_covariates,&interaction_i)));
%end;

/* FULL MODEL */
proc phreg data=&data;
  class &class_covariates &strata_vars / &class_opts;
  model &time_var*&censor_var(&censor_vals) = &all_covariates
      &interaction_vars / type1;
  strata &strata_vars;
  ods output Type1 = lrt_strat_cox_ph_type1_full;
run;

data lrt_strat_cox_ph_type1_full (keep=neg2ll_full df_full);
  set lrt_strat_cox_ph_type1_full (rename=(Neg2LogLike=neg2ll_full))
      end=last;
  retain df_full 0;
  df_full = sum(df_full, DF);
  if last;
run;

/* REDUCDED MODEL */
proc phreg data=&data;
  class &class_covariates &strata_vars / &class_opts;
  model &time_var*&censor_var(&censor_vals) = &all_covariates / type1;
  strata &strata_vars;
  ods output Type1 = lrt_strat_cox_ph_type1_red;
run;

/* Final processing and output results */
data _null_;
  set lrt_strat_cox_ph_type1_red (rename=(Neg2LogLike=neg2ll_red)) end=last;
  retain df_red 0;
  df_red = sum(df_red, DF);
  if last then do;
    set lrt_strat_cox_ph_type1_full;
    diff   = neg2ll_red - neg2ll_full;
    df     = df_full - df_red;
    pvalue = 1-probchi(diff, df);
    file print;
    HBAR1 = REPEAT("=",80);
    HBAR2 = REPEAT("-",80);
    put
      HBAR1
    / @5 "Stratified Cox Proportional Hazards Model Likelihood Ratio Test"
    / @25 "Summary of results"
    / HBAR2
    / @7 "-2LogLikelihood of the Reduced Model    = " neg2ll_red  31.2
    / @7 "-2LogLikelihood of the Full Model       = " neg2ll_full 31.2
    / @7 "Degrees of Freedom of the Reduced Model = " df_red      31.
    / @7 "Degrees of Freedom of the Full Model    = " df_full     31.
    / @7 "Model Degrees of Freedom                = " DF          31.
    / @7 "Difference                              = " diff        31.2
    / @7 "Chi-Square p-value                      = " pvalue      31.4
    / HBAR1;
  end;
run;
```

```
/* Clean up datasets used in MACRO */
proc datasets lib=work memtype=data noprint;
  delete lrt_strat_cox_ph_type1_red lrt_strat_cox_ph_type1_full;
  quit;
run;

%finish:

%mend lrt_strat_cox_ph;
```