PharmaSUG China 2018 – Paper SP-75

Several Methods to assess proportional hazard assumption when applying COX regression model

Ying Yao, Boehringer Ingelheim Inc., Shanghai, China

ABSTRACT

Proportional Hazards Regression using a partial maximum likelihood function to estimate the covariate parameters in the presence of censored time to failure data (Cox, 1972) has become widely used for conducting survival analysis. When modeling a proportional hazard regression model with survival analysis data that right censored and time-to-event in a clinical trial, the researchers may have great interests to assess that if the data in different treatment groups or with certain covariates meet the criteria of proportional hazard assumption. In this paper, there would be some basic concepts of survival data analysis, cox model and proportional hazard ratio introduced. Then methods regarding how to implement the test of proportional hazard assumptions will be illustrated as a guideline when handling the survival data in continues or discrete. At the end of the paper there are several examples using these methods with outputs from SAS explanations.

INTRODUCTION

Basically the proportional hazards (PH) regression model has two assumptions, that when satisfied ordinarily allow one to rely on the statistical inferences and predictions the model yields. The first assumption is that the time independence of the covariates in the hazard function, that is, the ratio of the hazard function for two individuals with different regression covariates, does not vary with time, which is also known as the PH assumption. The second assumption is that the relationship between log cumulative hazard and a covariate is linear. In this paper we will focus on the methods assessing the first assumption that is PH assumption in continues or discrete survival analysis data.

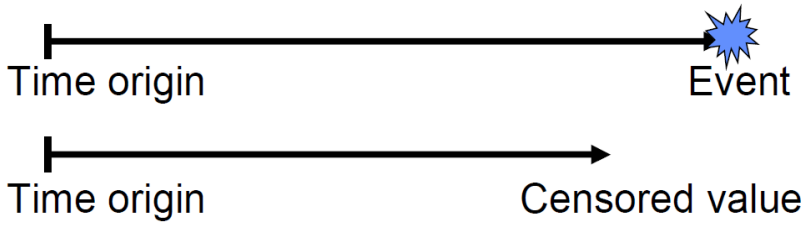
For the industry's approval, there are 3 different treating ways to evaluating the proportional hazards assumption. The first way is graphical analysis mainly includes Kaplan Meier curves and the estimated survival function with an ln-ln transform. The second way is the use of time-dependent covariates, the last way is the goodness of fit, a kind of statistical test. In the main part of this paper, all the 3 treating ways will be introduced with details and examples, and additionally the basic concepts and prerequisite knowledge of survival function, Cox hypothesis, and hazard function will also be presented as well.

BASIC CONCEPTS OF SURVIVAL ANALYSIS

In this paragraph, there will be some key components of survival analysis for introducing their definitions and internal relationships. First recall that time is continuous, and that the probability of an event at a single point of a continuous distribution is zero. To define the probability of these events over a distribution, the best described is by graphing the distribution of event times. To ensure the reader will start with the same fundamental tools of survival analysis, a brief descriptive section of these important concepts will follow. A more detailed description of the probability density function (pdf), the cumulative distribution function (cdf), the hazard function, and the survival function, can be found in any intermediate level statistical textbook.

THE CENSORING DATA

In statistics and medical research, censoring is a condition in which the value of a measurement or observation is only partially known. For example, suppose a study is conducted to measure the impact of a drug on mortality rate. In such a study, it may be known that an individual's age at death is at least 75 years (but may be more). Such a situation could occur if the individual withdrew from the study at age 75, or if the individual is currently alive at the age of 75.



s

Figure 1. Caption for censored value

Although in most cases when we talk about censoring data, they are all pointed to be right censored data. For displaying more information, most types of censoring data are listed below:

- *Left censoring* – a data point is below a certain value but it is unknown by how much.
- *Interval censoring* – a data point is somewhere on an interval between two values.
- *Right censoring* – a data point is above a certain value but it is unknown by how much.
- *Type I censoring* occurs if an experiment has a set number of subjects or items and stops the experiment at a predetermined time, at which point any subjects remaining are right-censored.
- *Type II censoring* occurs if an experiment has a set number of subjects or items and stops the experiment when a predetermined number are observed to have failed; the remaining subjects are then right-censored.

THE CUMULATIVE DISTRIBUTION FUNCTION

The cumulative distribution function is very useful in describing the continuous probability distribution of a random variable, such as time, in survival analysis. The cdf of a random variable T, denoted as $\mathbf{F_T(t)}$, is defined by $\mathbf{F_T(t)} = \mathbf{P_T}(\boldsymbol{T < t})$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties of a distribution function $\mathbf{F(t)}$ can be listed as a consequence of the knowledge of probabilities. Because $\mathbf{F(t)}$ has the probability $0 < \mathbf{F(t)} < 1$, then $\mathbf{F(t)}$ is a non-decreasing function of t, and as t approaches ∞, $\mathbf{F(t)}$ approaches 1.

THE PROBABILITY DENSITY FUNCTION

The probability density function is also very useful in describing the continuous probability distribution of a random variable. The pdf of a random variable T, denoted $\mathbf{f_T(t)}$, is defined by $\mathbf{f_T(t)} = \mathbf{dF_T(t)}/\boldsymbol{dt}$. That is, the pdf is the derivative or slope of the cdf. Every continuous random variable has its own density function, the probability $\mathbf{P}(\boldsymbol{a < T < b})$ is the area under the curve between times ***a*** and ***b***.

THE SURVIVAL FUNCTION

Let $\boldsymbol{T} > 0$ have a pdf $\mathbf{f(t)}$ and cdf $\mathbf{F(t)}$. Then the survival function takes on the following form:

$\mathbf{S(t)} = \mathbf{P_T}\ \{\mathbf{T > t}\} = 1 - \mathbf{F(t)}$

The survival function gives the probability of surviving or being event-free beyond time t. Because $\mathbf{S(t)}$ is a probability, it is positive and ranges from 0 to 1. It is defined as $\mathbf{S(0)} = 1$ and as t approaches ∞, $\mathbf{S(t)}$ approaches 0. The Kaplan-Meier estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step idea. The Kaplan-Meier estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. The survival curve describes the relationship between the probability of survival and time.

THE HAZARD FUNCTION

The hazard function h(t) is given by the following: $\mathbf{h(t)} = \mathbf{P_T}\ \{\mathbf{t < T < t + \Delta t | T > t}\} = \mathbf{f(t)}/(\mathbf{1} - \boldsymbol{F(t)}) = \mathbf{f(t)}/\boldsymbol{S(t)}$. The hazard function describes the concept of the risk of an outcome (e.g., death, failure, hospitalization) in an interval after time t, conditional on the subject having survived to time t. It is the probability that an individual dies somewhere between t and $(\mathbf{t + \Delta t})$, divided by the probability that the individual survived beyond time $\mathbf{t}$. The hazard function

seems to be more intuitive to use in survival analysis than the pdf because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time **t**.

COX PROPORTIONAL HAZARDS REGRESSION

Like many other models, the PH regression models the hazard function, as can be seen in equation 2.1.

David Cox's 1972 paper took a different approach to standard parametric survival analysis and extended the methods of the non-parametric Kaplan-Meier estimates to regression type arguments for life-table analyses. Cox advanced to prediction of survival time in individual subjects by only utilizing variables co-varying with survival and ignoring the baseline hazard of individuals. Cox did this by making no assumptions about the baseline hazard of individuals and only assumed that the hazard functions of different individuals remained proportional and constant over time.

When there are several explanatory variables, and in particular when some of these are continuous, it is much more useful to use a regression method such as Cox rather than a KM approach.

Here the hazard function for individual i is modeled as:

$$\mathbf{h}(\mathbf{t}) = \mathbf{h_0}(\mathbf{t}) \exp\left\{\sum_{i=1}^{p} \beta_i X_i\right\}$$

In the PH model, the hazard function is dependent on, or determined by, a set of *p* covariates [x_1, x_2, . . . , x_p], whose impact is measured by the size of the respective coefficients [β_1, β_2, . . . , β_p]. The 't' in $\mathbf{h}(\mathbf{t})$ reminds us that the hazard function varies over time. The term $\mathbf{h_0}(\mathbf{t})$ is called the baseline hazard, and is the value of the hazard if all the beta are equal to zero, since the quantity exp(0) equals 1. The proportional hazards model is considered semi-parametric because no assumption regarding the distribution of the baseline hazard is necessary.

The quantities exp(β_i) are called hazard ratios. A value of β_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases. In other words, a hazard ratio above one indicates the covariate is positively associated with the event probability, and thus negatively associated with the length of survival.

The PH model is essentially a multiple linear regression of logarithm of the hazard on the variables, x_i, with the baseline hazard being an 'intercept' term that varies with time. The covariates then act multiplicatively on the hazard at any point in time, and this provides us with the key assumption of the PH model: the hazard of the event in any group is a constant multiple of the hazard in any other. This assumption implies that the hazards for groups should be proportional and cannot cross or diverge. This proportionality assumption is often appropriate for survival time data, but in some cases where it is inappropriate can lead to false conclusions.

PROPORTIONAL HAZARDS ASSUMPTION

The assumption is that the time independence of the covariates in the hazard function, that is, the ratio of the hazard function for two individuals with different regression covariates, does not vary with time, which is also known as the PH assumption. In a regression type setting this means that the survival curves for two strata (determined by the particular choices of values for the x-variables) must have hazard functions that are proportional over time (i.e. constant relative hazard). We have seen how this can be evaluated graphically using "log-log" plots in the two-sample comparison case. In that situation, and also for the Cox model, there are tests that can be applied to test proportionality.

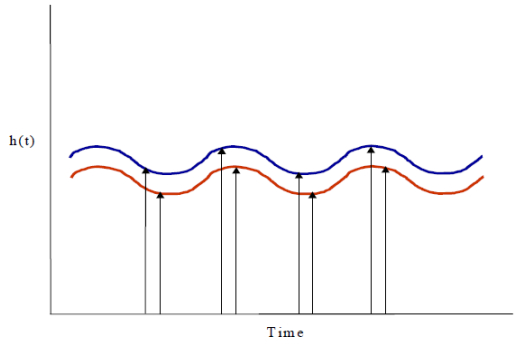


Figure 2. Graphical representation of proportional hazards over the follow-up period

METHODS TO ASSESS PROPORTIONAL HAZARDS ASSUMPTION

The methods to assess proportional hazards assumptions varies from different aspects. For the industry's approval, there are 3 different treating way to evaluating the proportional hazards assumption. The first way is graphical analysis mainly includes Kaplan Meier curves with and without an –ln-ln transform. The second way is the use of time-dependent covariates, the last way is the goodness of fit, a kind of statistical test.

The input datasets that used for programming are all from SAS learning help.

GRAPHICAL ANALYSIS

In this section we will use Kaplan-Meier survival curves and the curves of ln(-ln(S(t))) vs t curves to assess the proportional hazards assumption with an example dataset.

Kaplan Meier Curves

This method was initially raised by Cox in his scholarly article “Regression models and life table” ’in 1972, indicating the satisfaction of PH assumption, If the two curves show the same trend without crossover. Additionally it can be applied in continues, binary, and categorical variables. For binary and categorical covariates, plotting the Kaplan Meier curves within each subgroup and checking their trends will give you a rough picture of PH ratio between each group. For continuous covariates, the similar practices are useful as well after discretization of continues covariates, just be careful for the thresholds.

```
proc lifetest data=rats plots=survival(atrisk) ;
  time days*status(0);
  strata group;
run;
```

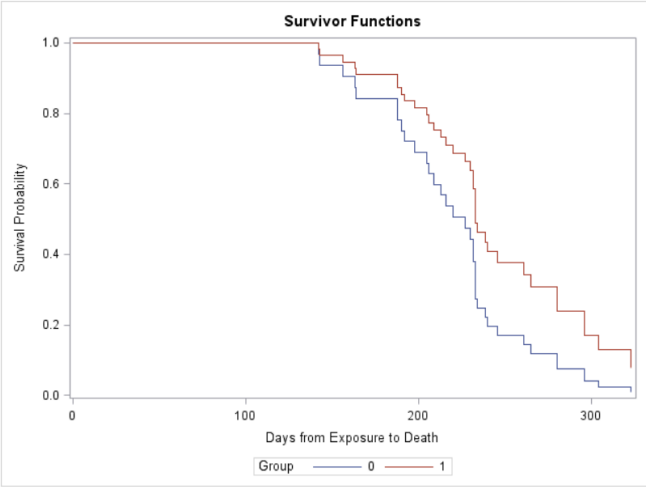


Figure 2. Kaplan Meier Curves with PH assumption not violated

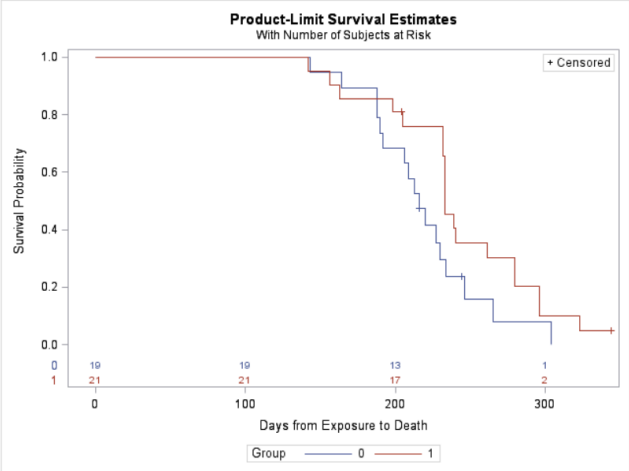


Figure 2. Kaplan Meier Curves with PH assumption violated

ln (-ln(S(t))) vs t Curves

For the binary covariates 0-1(two-sample survival data), when the data satisfied PH assumption, the following expressions are true:

$$\boldsymbol{h(t)} = \boldsymbol{h_0(t)}\, exp(\beta x) \qquad (1\text{-}1)$$

$$\boldsymbol{H(t)} = \boldsymbol{H_0(t)}\, exp(\beta x) \qquad (1\text{-}2)$$

$$\boldsymbol{log(H(t))} = \boldsymbol{log(H_0(t))} + \beta x \qquad (1\text{-}3)$$

$$\boldsymbol{log\big(-logS(t)\big)} = \boldsymbol{log\big(-log\big(S_0(t)\big)\big)} + \beta x \qquad (1\text{-}4)$$

The plot of ln (-ln(S(t))) curves of 2 groups plot has reasonably parallel lines there is not a significant problem with the assumption for the model.

```
Proc lifetest data=Myeloma plots=(lls);
 time Time*VStatus(0);
 strata platelet;
run;
```

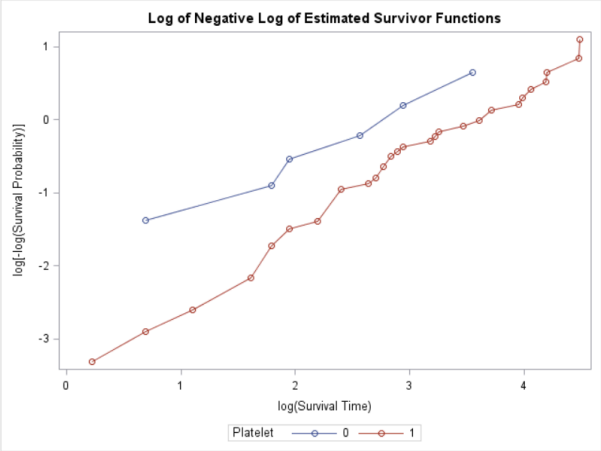


Figure 3. Log of Negative Log plot with PH assumption not violated

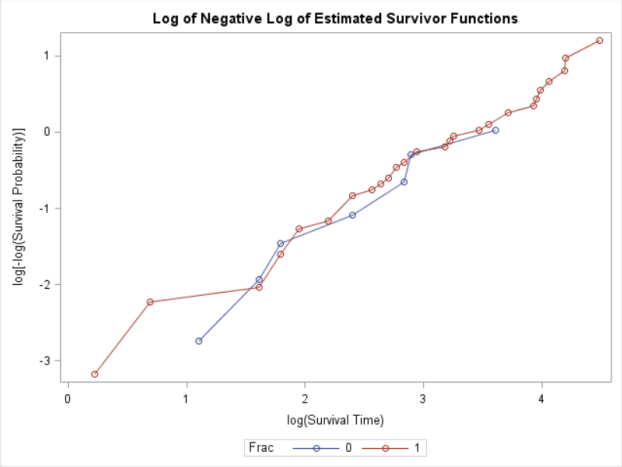


Figure 4. Log of Negative Log plot with PH assumption violated

MODEL USING TIME-DEPENDENT EXPLANATORY VARIABLES

The second method is establish the COX regression model using time-dependent explanatory variables to assess the PH assumption. When Cox(1972) raise the COX regression model, at the same time it was introduced that adding a time interaction e.g. x*log(t) into the model then test its significance. The null hypothesis is that the survival data satisfy PH assumption. If the p value is less than the significant level(usually alpha=0.05) than PH assumption fails.

For 1 predictor X as example, the hazard function including time-dependent exponential term as:

$\boldsymbol{h(t)}= \boldsymbol{h_0(t)}\, exp(\beta x + \gamma X * g(t))$

where $g(t) = t$

or $g(t) = \log(t)$

or $g(t) = 1(\mathrm{t} \geq \mathrm{t}_0)$

When modeling COX regression, just add a time-dependent variable in the sas statement as below:

```
proc phreg data=valung;
  model time*status(0)=age aget;
  aget=age*time;
run;

or

proc phreg data=valung;
 model time*status(0)=age aget;
 aget=age*log(time);
run;
```

The null hypothesis are:

"$\gamma = 0$" for 1 predictor being assessed.

"$\gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$" for several predictors being assessed.

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
logbun_T	1	-1.55775	0.58834	7.0103	0.0081	0.211
LogBUN	1	4.72274	1.29836	13.2312	0.0003	112.476
HGB	1	-0.13143	0.05928	4.9153	0.0266	0.877
Frac	1	0.47033	0.39546	1.4145	0.2343	1.601

Display 1. Maximum likelihood estimates for logbun_T with PH assumption violated

The constructed statistics of the hypothetical test method are Wald or likelihood ratio statistics from chi-square distribution with 1degree of freedom for 1 predictor under null hypothesis and log-likelihood ratio statistic from chi-square distribution with p degree of freedom under null hypothesis.

There's another kind of time-dependent covariates that have only binary value 0 or 1, or the time-dependent covariates have no interactive effect with other static effects in the cox model. They can be created by both counting process and programming process and will result to the same output. For more details please explore in the example from SAS Support "Model Using Time-Dependent Explanatory Variables" and some papers also give more information about the methods, e.g. "Time-Dependent Covariates "Survival" More in PROC PHREG" and "Your 'Survival'Guide to Using Time - Dependent Covariates".

GOOD OF FITNESS

Schoenfeld Residuals

Another method to assess the ph assumption is to examine the Schoenfeld residuals. The Schoenfeld residual computed with one per observation per covariate is defined at the observed event times as the difference between covariate for observation and the weighted average of the covariate values for all subjects still at risk when observation experiences the event.

For the i^{th} subject and k^{th} covariate, the estimated Schoenfeld residual r_{ik}, is given by (notation from Hosmerand Lemeshow)

$$\hat{r}_{ik} = x_{ik} - \hat{\bar{x}}_{w_i k}$$

Where x_{ik} is the value of the k^{th} covariate for individual i, and $\hat{\bar{x}}_{w_i k}$ is a weighted mean of covariate values for those in the risk set at the given event time. If the data meet the ph assumption then the schoenfeld residual do not depend on

the survival time, i.e the schoendeld residual has no correlation with the rank of survival time.

There are 3 steps to implement the hypothesis test. The first step is to call proc phreg to calculate Schoenfeld residual. The second step is to call proc rank to re-sort the survival time variable and get its ranking by survival time. The third step is call proc corr to test the correlation between ranked survival time and Schoenfeld residual.

```
proc phreg data= Heart;
      model Time*Status(0)= trans prevsurg Acc_Age ;
      output out=out atrisk=atrisk resmart=mresides resdev=dresids ressch=rtrans rsurg raccage;
       assess var=(acc_age) ph NPATHS=20 CRPANEL;
       id id ;
 run;

data event;
  set out;
  if status=1;
run;

proc rank data=event out=ranked ties=mean;
 var time;
 ranks timerank;
run;
```

The CORR Procedure

1 With Variables:	timerank
1 Variables:	rtrans

Pearson Correlation Coefficients, N = 75
Prob > |r| under H0: Rho=0

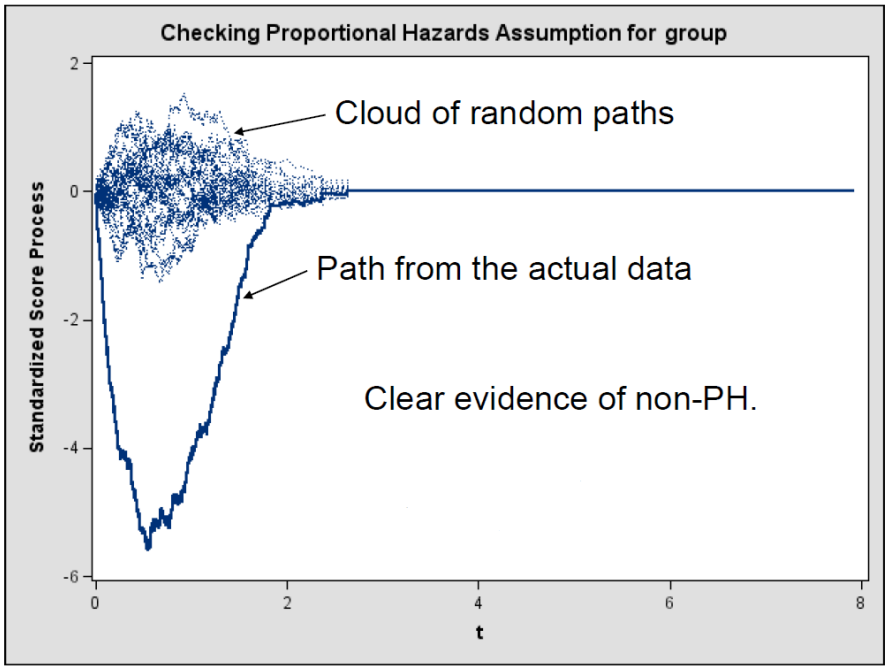
	rtrans
timerank	0.29932
Rank for Variable Time	0.0091

Display 2. P value of Correlation Coefficients= 0.0091 with PH assumption not violated

The ASSESS statement in PROC PHREG

The ASSESS statement performs the graphical and numerical methods of Lin, Wei, and Ying (1993) for checking the adequacy of the Cox regression model. The methods are derived from cumulative sums of martingale residuals over follow-up times or covariate values. You can assess the functional form of a covariate or you can check the proportional hazards assumption for each covariate in the Cox model. PROC PHREG uses ODS Graphics for the graphical displays.

```
assess var=(acc_age) ph
```



Display 3. Check the ph assumption by ACCESS statement

CONCLUSION

In this paper we introduce some of the classical methods to assess the ph assumption before apply COX model. They are graphical analysis, involving time-dependent covariates, and the goodness of fitness. For the readers they can choose one or two methods to test and see if their trial data satisfy the ph assumption and if not, they may need to adjust the model, it's another topic anyway. Checking assumptions takes time and can be never ending, so balance is need. Please bear in mind that without testing ph assumption the survival data could not be used in the COX model without any concern.

REFERENCES

Time-Dependent Covariates "Survival" More in PROC PHREG Fengying Xue Michael Lai, Sanofi .

Your "Survival" Guide to Using Time - Dependent Covariates Teresa M. Powell, Melissa E. Bagnell, SAS Global Forum 2012.

CONTACT INFORMATION

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

Name: Ying Yao
Enterprise: Boehringer Ingelheim Inc.
E-mail: ying_2.yao@Boehringer-Ingelheim.com