

## **Cox Proportional Hazards Model with Dichotomous Covariates**

Prepared by Michael Black  
Revised by Julia Soulakova

### **1. What research questions can be answered using the procedure?**

Cox Proportional Hazards Model can be used for testing if two or more groups have the same hazard rate. For example:

- Is there a significant difference in Survival rates for Men and Women who have cancer?
- Does the success rate to find professional employment after a junior year of college depends on one's race?
- Does your level of education effect your chances of bankruptcy in terms of the time to bankruptcy?

For this method we don't need to specify a distribution, because Cox model is a semi-parametric model.

### **2. Data considerations: restrictions on types of variables or variable coding or sample size**

All categorical variables of interest with more than 2 groups should be recoded as a series of dichotomous variables. For example, if original variable Education has 4 levels:

- High School
- College
- Grad School
- Drop out,

then instead of this single variable you have to create three dummy variables as follows:

$$X_1 = \begin{cases} 1, & \text{if High School,} \\ 0, & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1, & \text{if College,} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad X_3 = \begin{cases} 1, & \text{if Grad School,} \\ 0, & \text{otherwise} \end{cases}$$

Then “Drop out” is a reference group and it corresponds to  $X_1 = X_2 = X_3 = 0$ .

We assume parallel hazard rates for the groups, that is that the hazard rate curves don't cross at any specific time, but that if one is greater than another it is consistently greater (Allison 1995).

Ties in the data need to be addressed. That is, if 2 or more units experience the event of interest at the exact same time, some method has to be used to order the observations corresponding to such tied times. This is important in order to estimate the covariates' coefficients. There are several different methods used for dealing with ties in the data. SAS uses one of them (Breslow method) as a default one if none is specified in the codes.

### 3. Main ideas and statistics behind the procedure

The hazard rate is modeled as:

$$H(t) = H_0(t) \times \exp(b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k)$$

Where  $x_1 \dots x_k$  are a collection of predictor variables (covariates) and  $H_0(t)$  is the baseline hazard at time  $t$ , (usually the hazards are denoted by  $h$ , but we use  $H$  notation, although,  $H$  usually stands for the cumulative hazard) representing the hazard for a person with the value 0 for all the predictor variables.

By dividing both sides of the above equation by  $H_0(t)$  and taking logarithms, we obtain:

$$\ln\left(\frac{H(t)}{H_0(t)}\right) = b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k$$

The ratio  $H(t) / H_0(t)$  is called as the hazard ratio. The coefficients  $b_1 \dots b_k$  are estimated by Cox regression, and can be interpreted in a similar manner to that of multiple logistic regression.

Suppose the covariate (risk factor) is dichotomous and is coded 1 if present and 0 if absent. Then the quantity  $\exp(b_i)$  can be interpreted as the instantaneous relative risk of an event, at any time, for an individual with the risk factor present compared with an individual with the risk factor absent, given both individuals are the same on all other covariates.

Suppose the covariate is continuous, then the quantity  $\exp(b_i)$  is the instantaneous relative risk of an event, at any time, for an individual with a unit increase in the value of the covariate compared with another individual, given both individuals are the same on all other covariates. (Klein 1997)

## 4. Example

### 4.1. Data set

Consider the following data from ([Kalbfleisch and Prentice 1980](#)). Two groups of rats received different pretreatment regimes and then were exposed to a carcinogen. Investigators recorded the survival times of the rats from exposure to mortality from vaginal cancer. Four rats died of other causes, so their survival times are censored. Interest lies in whether the survival curves differ by pretreatment regimes.

The data set Rats contains the variable Days (the survival time in days), the variable Status (the censoring indicator variable: 0 if censored and 1 if not censored), and the variable Group (the pretreatment group indicator).

*You can copy and paste the following data lines into SAS editor and then click "Run current"*

```
data Rats;
  label Days = 'Days from Exposure to Death';
  input Days Status Group @@;
  datalines;
143 1 0   164 1 0   188 1 0   188 1 0
190 1 0   192 1 0   206 1 0   209 1 0
213 1 0   216 1 0   220 1 0   227 1 0
230 1 0   234 1 0   246 1 0   265 1 0
304 1 0   216 0 0   244 0 0   142 1 1
156 1 1   163 1 1   198 1 1   205 1 1
232 1 1   232 1 1   233 1 1   233 1 1
233 1 1   233 1 1   239 1 1   240 1 1
261 1 1   280 1 1   280 1 1   296 1 1
296 1 1   323 1 1   204 0 1   344 0 1
;
run;
```

Since Group takes only two values, the null hypothesis for no difference between the two groups is identical to the null hypothesis that the regression coefficient for Group is 0. (SAS Help and Documentation, the PHREG Procedure)

#### 4.2. SAS codes to run the procedure and obtain the graphs

PHREG is used for the analysis and the Hazard Rate graphs are provided from LIFETEST.

```
proc phreg data=Rats;
  model Days*Status(0)=Group;
run;

symbol1 c=blue; symbol2 c=orange;
title 'Rats! Foiled again';
proc lifetest data=rats plots=(h) method=life outtest=Test;
  time Days*Status(0);
  strata Group;
run;
```

### 4.3. Explanation of SAS output

Preceding SAS codes result in the following output.

```

The PHREG Procedure

Model Information

Data Set          WORK.RATS
Dependent Variable Days          Days from Exposure to Death
Censoring Variable Status
Censoring Value(s) 0
Ties Handling     BRESLOW
    
```

```

Number of Observations Read    40
Number of Observations Used    40
    
```

#### Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
40	36	4	10.00

#### Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

#### Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	204.317	201.438
AIC	204.317	203.438
SBC	204.317	205.022

#### **A)** Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.8784	1	<b>0.0898</b>
Score	3.0001	1	<b>0.0833</b>
Wald	2.9254	1	<b>0.0872</b>

#### Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Group	1	<b>-0.59590</b>	0.34840	2.9254	0.0872	<b>0.551</b>

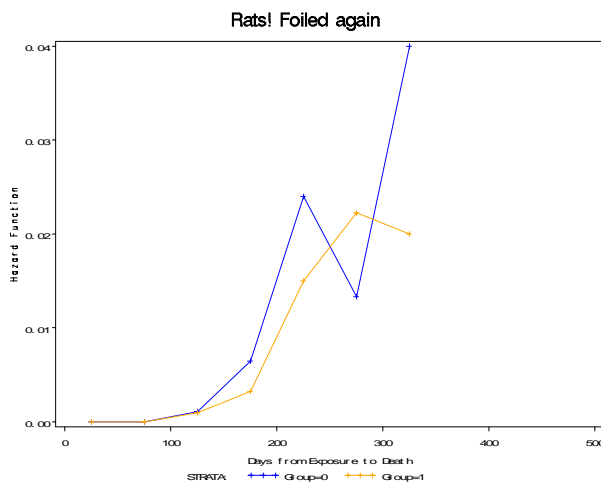
Output area **A)** presents the Chi-square scores for several tests with the null hypothesis stating that the regression coefficient for Group is 0. Since  $Pr > ChiSq$  values we Fail to reject the Null at 0.05 significance level, but we Reject it at 0.10 significance level. That is, at the  $\alpha=0.10$  level the regression coefficient for Group is significantly different from 0.

Output area **B)** contains the Cox Proportional Hazards Model estimates for the coefficient for group. The parameter estimate for the group coefficient is -0.59590 and hence, the hazard ratio is given by  $\exp(-0.59590)=0.551$ . In other words, pretreatment regime 1 results in 55% chance of dying (compared to pretreatment regime 0) from vaginal cancer at any given moment, i.e., rats receiving pretreatment regime 0 have almost twice the risk of dying from vaginal cancer as rates receiving regime 1.

If we had more than two groups we would need to use the differences in the Parameter estimates to evaluate the Hazard Ratio's between groups. Similarly if we looked at 2 or more covariates at a time we would need to use the Parameter estimates to compare the various combinations of responses.

#### 4.4. Illustrating conclusions using graphs

Graph below illustrates our conclusion that pretreatment regime 0 results in higher hazard rate of death than regime 1.



Allison, P. D. (1995), *Survival Analysis Using SAS: A Practical Guide*, SAS Institute Inc., Cary, NC, USA

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 20, 187 - 220.

Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269 - 276.

Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons, Inc.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis; Techniques for Censored and Truncated Data*. Springer-Verlag New York, Inc.