Lecture

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: M.Sc. – SEM 2

Subject : Statistical and Risk Modelling 1

Chapter: Unit 2 Chapter 1

Chapter Name: Proportional Hazard Models

Today's Agenda

1. Introduction to Proportional Hazard Models
    1) What are Proportional Hazard (PH) Models?
    2) Applications of PH Models
    3) Common PH Models

2. The Cox PH Model

1. Covariates
    1) What are covariates?
    2) The coefficients of covariates

2. Ratio of Hazards

1. Partial Likelihood
    1) Maximizing Partial Likelihood
    2) Breslow's Approximation

1 Introduction to Proportional Hazard Models

1.1 What are Proportional Hazard Models?

Proportional hazards models are a class of survival models in statistics.

PH model consists of two parts: the underlying baseline hazard function, describing how the risk of event per time unit changes over time at *baseline* levels (hence the name proportional hazard); and the effect parameters, describing how the hazard varies in response to explanatory covariates.

A covariate is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, severity of symptoms and so on.

The most widely used regression model in recent years has been the *proportional hazards* model. Proportional hazards (PH) models can be constructed using both parametric and non-parametric approaches to estimating the effect of duration on the hazard function.

1.2 Applications of PH Models

- The proportional hazard models have good applications and inferences. The inferences about how each covariate affects the individual's mortality, gives information that can be used in different ways.

- Survival data is generally heterogeneous, and hence PH models prove effective in analyzing data for various groups. The model can be used to check the effects of gender, smoking habits, patients given treatment and patients who did not receive the treatment, geographical location etc. The effects of these covariates or factors can be assessed, in various combinations also to understand the various trends and effects on rate of mortality, morbidity and other sectors.

- Example: A life insurance company may wish to know how certain covariates affect mortality, so that it can charge premiums that accurately reflect the risk for an individual, *eg* higher premiums for smokers.

1 Introduction to PH Models

In PH models the hazard function for the i^{th} life, $\lambda_i(t, z_i)$, may be written:

$$\lambda_i(t, z_i) = \lambda_0(t) g(z_i)$$

where $\lambda_0(t)$ is a function only of duration t and $g(z_i)$ is a function only of the covariate vector.

1.3 Common PH Models

Fully Parametric models for the Hazard Function.
Here, strong assumption is made about the lifetime distribution and the hazard. We have already seen these models in the earlier chapter. They are:-

1. **Exponential model (constant hazard)**
2. **Weibull model (monotonically decreasing hazard)**
3. **Gompertz-Makeham model (exponentially increasing hazard)**
4. **Log – logistic model (humped hazard)**

2 The Cox PH Model

The **Cox PH model** (Cox, 1972) is essentially a regression model commonly used statistics in survival analysis for investigating the association between the survival time of patients and one or more predictor variables. It is a semi-parametric model. It allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time.

The Cox model is expressed by the *hazard function* denoted by h(t). It can be estimated as follow:

$$h(t) = h_0(t) \text{ x } e^{(\beta_1 z_1 + \beta_2 z_2 + \ \dots \ + \beta_p z_p)}$$

where:

- t represents time
- $h(t)$ is the hazard rate
- $h_0(t)$ is the baseline hazard. It corresponds to the value of the hazard if all the z_i are equal to zero. (the quantity $e^0 = 1$)
- the coefficients $(\beta_1, \beta_2 \ \dots \beta_p)$ measures the impact (i.e. the effect size) of the covariates
- There are a set of p covariates $(z_1, z_2 \ \dots z_p)$

3 Covariates

3.1 What are covariates?

A covariate is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, severity of symptoms and so on.

The covariates or the predictors can be of different types:-

- Direct Measurements [eg. Age]
- Indicator or Dummy Variables [eg. 0 for vaccine received and 1 for vaccine not received]
- Qualitative measurements that will be represented quantitatively. [eg. 1 to 10 ranking given to patients based on severity of symptoms]

Consider an example where three factors weight, place of residence – Mumbai or Delhi and the severity of symptoms from 1 to 5 are taken.
Then the covariates will have values like, Direct measurement for weight, Indicator for place of residence – 0 for Mumbai and 1 for Delhi and ranking based on the severity of symptoms being 1 for very mild increasing to 5 for extreme.

Class Practice

1. State whether the following co – variates can be Direct measurements, Indicator variables, or Qualitative aspects that can be represented quantitatively:

- Weight of a Person
- Profession of a Person
- Semester III Exam Grade scored by a Student
- Marks scored by a student in Semester III Exam
- Martial Status of a Person
- Numbers of Hours that a Person works on average in a day
- Severity of COVID Symptoms in a Patient

2. Should Gender as a co – variate be used in pricing life insurance or would that be gender discrimination?

3.2 The coefficients of covariates

- The coefficients of the covariates are the regression parameters in the model. Regression parameters measure the impact of covariates on the hazard rate.
- A value of β_i greater than zero, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases. There is **positive correlation** between hazard rate and covariate. Vice versa, if the value of β_i is less that zero, then as the value of i^{th} covariate increases, the event hazard decreases and length of survival increases. Thus a **negative correlation**.
- If the magnitude of the i^{th} regression parameter is large, the hazard rate is significantly affected by the i^{th} covariate, *ie* there is a **strong correlation** (positive or negative) between hazard rate and covariate. Where the magnitude is small, hazard rate is not significantly affected, there is **weak**

In summary
$\beta_i > 0$: Increase in Hazard
$\beta_i < 0$: Decrease in Hazard
$\beta_i = 0$: No effect

4 Ratio of Hazards

In general terms, let z_i is the vector of covariates for i^{th} life :$z_i = (Z_{i1}, Z_{i2} \ldots Z_{ip})$ and the vector of regression parameters is : $(\beta_1, \beta_2 \ldots, \beta_p)$.
Now consider the ratio of hazards for two lives z_1 and z_2, it will be given by:

$$\frac{\lambda(t, z_1)}{\lambda(t, z_2)} = \left(\frac{e^{\beta*(Z_1)^T}}{e^{\beta*(Z_2)^T}}\right) = \frac{e^{\sum_{j=1}^{p} \beta_j Z_{1j}}}{e^{\sum_{j=1}^{p} \beta_j Z_{2j}}} = e^{\sum_{j=1}^{p} \beta_j (Z_{1j} - Z_{2j})}$$

The ratio is finally independent of time t. Thus ratio is constant over time.
For example: If an individual has a risk of death at some initial time point that is twice as high as that of another individual, then at all later times the risk of death remains twice as high.

Question

An analysis into the survival lengths of lives who have received a vaccine with some treatment at 3 hospitals – K, E and M. The following data regarding covariates is recorded.

z_1 = 0 : for males.
= 1 : for females
z_2 = 1 : if vaccine and treatment given in hospital E
= 0 : otherwise
z_3 = 1 : if vaccine and treatment given in hospital M
= 0 : otherwise

The estimates of coefficients are β_1 = 0.041, β_2 = -0.022 and β_3 = 0.0123. To estimate the hazard rate at time t, the COX PH equation is used.

Making use of the information given, compare the hazard rate for a female patient who attended Hospital K with that of:

(i) a female patient who attended Hospital E
(ii) a male patient who attended Hospital M

Solution

The Cox PH formula is : $h(t) = h_0(t) \times e^{(\beta_1 z_1 + \beta_2 z_2 + \ \dots \ + \beta_p z_p)}$

i) Hazard for female patient who attended hospital K is = $\lambda_{female,K}$ = $\lambda_0(t) \cdot e^{(z_1\beta_1 + z_2\beta_2 + z_3\beta_3)}$

= $\lambda_0(t)$ .e^(1 x 0.041 +0 +0) = 1.041852 $\lambda_0(t)$

Hazard for female patient who attended hospital E is = $\lambda_{female,E}$ = $\lambda_0(t)$ . . $e^{(z_1\beta_1 + z_2\beta_2 + z_3\beta_3)}$

= $\lambda_0(t)$. e^(1x0.041 + 1 x (-0.022) + 0) = 1.01918165 $\lambda_0(t)$

The ratio of the hazard is then, $\frac{\lambda_{female,K}}{\lambda_{female,E}} = \frac{1.041852\ \lambda_0(t)}{1.01918165\ \lambda_0(t)} = 1.02224368$

Thus mortality of female who attended hospital K is more than the mortality of female who attended hospital E.

Try the second one yourself!

5 Partial Likelihoods

- To estimate suitable values of the parameters βj, for use in cox model, we gather data on lives with different characteristics, and use this information to write down the partial likelihood function.
- The partial likelihood estimates the regression coefficients but avoids the need to estimate the baseline hazard.
- For each life we need to record the covariate values, the time spent in the investigation and the reason for leaving the investigation (i.e. whether due to hazard being considered or otherwise).
- Each time a life is observed to leave the investigation due to the hazard being considered , we get the contribution to the partial likelihood of the form:

$$\frac{\text{hazard for the life leaving the investigation at that time}}{\text{total hazard for all lives still under observation just before that time (at risk group)}}$$

5.1 Maximizing Partial Likelihood

We will now consider how to estimate the regression parameters $\beta = (\beta_1, \beta_2, \ldots\ldots, \beta_p)$. To estimate b in the Cox model it is usual to maximise the *partial likelihood*. Let $R(t_j)$ denote the set of lives which are at risk just before the j^{th} observed lifetime, and for the moment assume that there is only one death at each observed lifetime, that is $d_j = 1$ $(1 \leq j \leq k)$.

The Partial Likelihood is: $L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta Z_j^T}}{\sum_{i \in R(t_j)} e^{\beta Z_i^T}}$

Intuitively, each observed lifetime contributes the probability that the life observed to die should have been the one out of the $R(t_j)$ lives at risk to die, conditional on one death being observed at time t_j .The maximisation procedure is usually carried out using a computer.

5.2 Breslow's approximation to the partial likelihood

If there are ties in the data, *ie* the death times are not distinct, then Breslow's approximation to the partial likelihood can be used:

$$L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta s_j^T}}{\left(\sum_{i \in R(t_j)} e^{\beta Z_i^T}\right)^{d_j}}$$

where:
s_j is the sum of the covariate vectors z of the d_j lives observed to die at time t_j .

- The maximum partial likelihood estimator of the vector of parameters b , which we denote by $\tilde{b}$, will have asymptotic multivariate normal distribution.
- This approximation tells us that all lives leaving the investigation due to the hazard at given time should be included in the 'at risk' group used for denominator of each life's contribution.

Question

An investigation was carried out into the survival times (measured in months) of patients in hospital following liver transplants. The covariates are z_{1i} = 0 for placebo, 1 for treatment X, and z_{2i} = weight of patient (measured in kg).
The observed lifetimes (with weights in brackets) were as follows:

Placebo	Treatment X
3 (83)	6*(58)
9 (68)	11 (73)
14 (75)	14 (68)
16 (86)	14* (49)

Observations with an asterisk represent censored observations.
Using Breslow's assumption, determine the contribution to the partial likelihood that is made by the deaths at time 14.

Solution

Just before time 14, there were four lives at risk. The total force of mortality for these four lives at time 14 is:

$$\mu_0(14)e^{75\beta_2} + \mu_0(14)e^{\beta_1+68\beta_2} + \mu_0(14)e^{\beta_1+49\beta_2} + \mu_0(14)e^{86\beta_2}$$

where $\mu_0(t)$ denotes the baseline hazard at time t , measured in months since the transplant operation.

The individual forces of mortality for the two lives that die at time 14 are:

$$\mu_0(14)e^{75\beta_2} \text{ and } \quad \mu_0(14)e\text{^}(\beta_1 + 68\beta_2)$$

So the contribution to the partial likelihood from the deaths that occur at time 14 is:

$$\frac{\mu_0(14)e^{75\beta_2}}{\mu_0(14)e^{75\beta_2} + \mu_0(14)e^{\beta_1+68\beta_2} + \mu_0(14)e^{\beta_1+49\beta_2} + \mu_0(14)e^{86\beta_2}} \text{ x } \frac{\left(\mu_0(14)e^{\beta_1+68\beta_2}\right)}{\mu_0(14)e^{75\beta_2} + \mu_0(14)e^{\beta_1+68\beta_2} + \mu_0(14)e^{\beta_1+49\beta_2} + \mu_0(14)e^{86\beta_2}}$$

$$= \frac{\mu_0(14)e^{75\beta_2} \text{ x } \mu_0(14)e^{\beta_1+68\beta_2}}{[\mu_0(14)e^{75\beta_2} + \mu_0(14)e^{\beta_1+68\beta_2} + \mu_0(14)e^{\beta_1+49\beta_2} + \mu_0(14)e^{86\beta_2}]^2}$$

$$= \frac{\left(e^{\beta_1+143\beta_2}\right)}{\left[e^{75\beta_2} + e^{\beta_1+68\beta_2} + e^{\beta_1+49\beta_2} + e^{86\beta_2}\right]^2}$$

since all the baseline hazard terms cancel.

Recap

- Proportional Hazard(PH) Models consist of a underlying hazard function and effect parameters
- We've learnt about fully parametric models in earlier chapters
- Cox Model is a semi-parametric model which allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time
- **Cox Model:** $h(t) = h_0(t)\, x\, e^{(\beta_1 z_1 + \beta_2 z_2 + \ \ldots\ + \beta_p z_p)}$
- A covariate is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, etc… and it's coefficient is the regression parameter in the model
- Ratio of Hazards is independent of time.
- Partial likelihood estimates the regression coefficients but avoids the need to estimate the baseline hazard.
- **The Partial Likelihood is:** $L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta Z_j^T}}{\sum_{i \in R(t_j)} e^{\beta Z_i^T}}$
- If there are ties in the data, *ie* the death times are not distinct, then Breslow's approximation to the partial likelihood can be used:
- $L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta s_j^T}}{\left(\sum_{i \in R(t_j)} e^{\beta Z_i^T}\right)^{d_j}}$

Homework Question

Subject CT4 September 2009 Question 11

A study was undertaken into the length of spells of unemployment among young people in a certain city. A sample of young people was monitored from the time they started to claim unemployment benefit until either they resumed work, or they moved away from the city. None of the members of the sample died during the study.

The study investigated the impact of age, sex and educational qualifications on the hazard of returning to work using the following covariates:
A: a young person's age when he or she started claiming benefit (measured in exact years since his or her 16th birthday)
S: a dummy variable taking the value 1 if the person was male and 0 if the person was female
E : a dummy variable taking the value 1 if the person had passed a school leaving examination in mathematics, and 0 otherwise
with associated parameters β_A , β_S and β_E .

The investigators decided to use a Cox proportional hazards regression model for the study.

Continued

i. Explain what is meant by a proportional hazards model.
ii. Explain why the Cox model is a popular model for the analysis of survival data.
iii. Parts:
    a. Write down the equation of the model that was estimated, defining the terms you use (other than those defined above).
    b. List the characteristics of the young person to whom the baseline hazard applies.

The results showed:

- The hazard of resuming work for males who started claiming benefit aged 17 years exact and who had passed the mathematics examination was 1.5 times the hazard for males who started claiming benefit aged 16 years exact but who had not passed the mathematics examination.
- Females who had passed the mathematics examination were twice as likely to take up a new job as were males of the same age who had failed the mathematics examination.
- Females who started claiming benefit aged 20 years exact and who had passed the mathematics examination were twice as likely to resume work as were males who started claiming benefit aged 16 years exact and who had also passed the mathematics examination.

iv. Calculate the estimated values of the parameters β_A , β_S and β_E