Survival Analysis and Cox Regression

Contents

1. Introduction to Survival Analysis
2. Time to Event and Censoring
3. Objective of Survival Analysis
4. Concept of Survival Function
5. Hazard Function
6. Cox Regression
7. Statistical Model
8. Case Study for Cox Regression

Introduction to Survival Analysis

- **Survival analysis is the study of time taken for an event to occur**.
- The **analysis variable** is the time between a **time origin and an end point**.
- The end point is either the occurrence of the event of interest, referred to as a death or failure, or the end of the subject's participation in the study.
- Two functions are of fundamental **interest—the survival function and the hazard function**
- One of the earliest applications of survival analysis was by Christiaan Huygens in 1669, showing how many out of 100 people survive until 86 years
- The name 'survival' analysis stems from the usage of this method for modeling 'time to death'; however the concept can be extended to several different areas and event can be defined as occurrence of a disease, lapse of a policy, etc.

Time to Event and Censoring

Consider that **time to occurrence of an event T is a random variable**. In order to define time-to-event following terms must be clearly defined:

Time Origin

The time origin must be specified such that individuals are as much as possible on an equal footing.
e.g. time point when treatment starts for a particular disease

Time Scale

Usually, observation time is used as the time scale for both clinical and observational studies
e.g. months, years, age

Definition of an Event

Based on the study objective, the event should be defined
e.g. death, disease occurrence etc.

> If rate of occurrence of an event is λ then the expected time - to – event is $1/\lambda$

Subjects are said to be **censored** in case of either of the following outcomes:

- If they are lost to follow up
- If they drop out of the study
- If the study ends before they have an outcome of interest

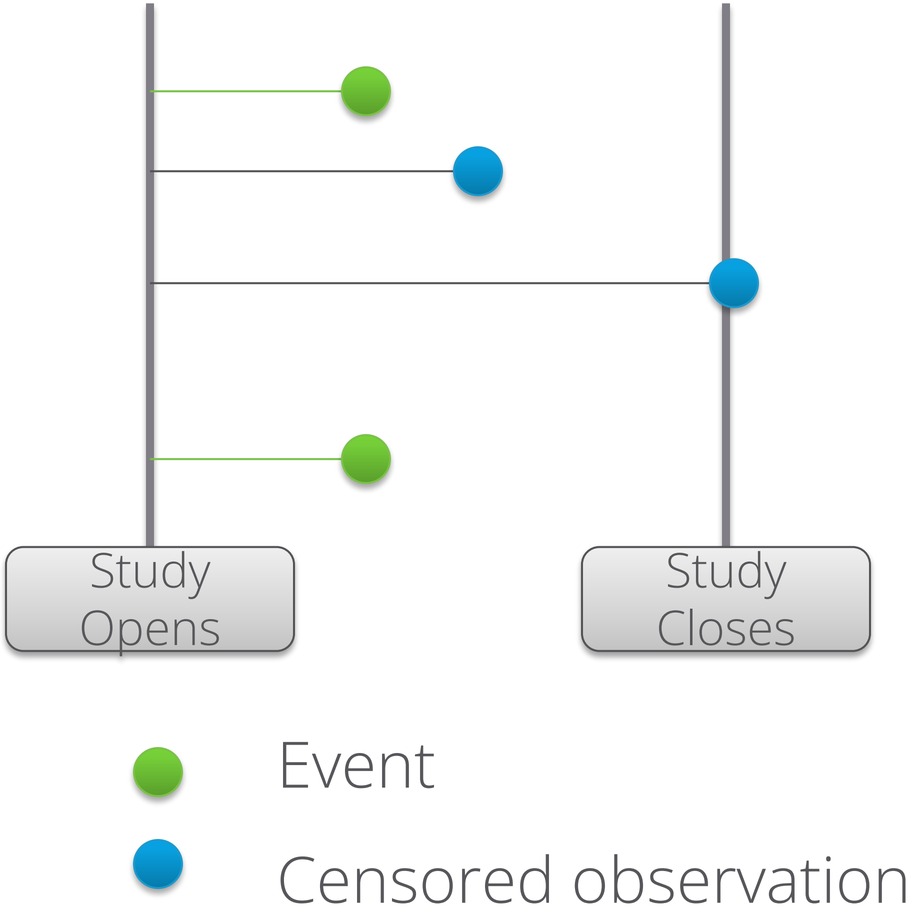
Concept of Censoring

Two-variable outcome :

Time variable: t_i = time at event, is a random variable having probability distribution

Status variable: c_i =1 if event occurred;

c_i =0 no event by time *t(Censored)*



Objectives of Survival Analysis

To estimate time-to-event for a group of individuals

- Eg. Time until recovery from low back pain

To assess the impact of factors/covariates on time-to-event

- Eg. Age, Gender, Occupation and treatment

Statistical Modeling

To compare time-to-event between two or more groups

- Eg. Time until recovery from low back pain in active vs. placebo groups

Statistical Inference

Concept of Survival Function

The goal of survival analysis is to estimate and compare **survival experiences of different groups.**

Survival experience is described by the cumulative survival function:

$$
\begin{aligned}
S(t) &= P(T > t) \\
&= 1 - P(T \leq t) \\
&= 1 - F(t)
\end{aligned}
$$

F(t) is the Cumulative Distribution Function (CDF)

Example: If t=40 years, S(t=40) = Probability of surviving beyond 40 years.

Hazard Function

Hazard Function (Instantaneous Failure Rate)

It is the ratio of conditional probability that the failure/death will occur in the interval t+Δt given that it has not occurred before time t and width of the interval (Δt).

Hazard function, $h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T > t)}{\Delta t}$, T is a random variable denoting 'Survival Time'

$$= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

$$= \frac{f(t)}{S(t)}$$

Hazard vs Density Function

- When one is born, there exists a certain probability of dying at any age; that is the probability density (Marginal probability)

  Example: A baby girl born today has, say, a 1% chance of dying at 80 years.

- However, as one survives for a while, the probabilities keep changing (Conditional probability)

  Example: A woman who is 79 today has, say, a 5% chance of dying at 80 years.

Cox Regression

DEPENDENT VARIABLE

Time to Event

Time variable must be quantitative

INDEPENDENT VARIABLES

Categorical or Continuous

Cox regression produces a survival function that predicts the probability of survival till time t for given values of the predictor variables

Statistical Model

$$h(t|x) = h_0(t)\exp(b_1x_1 + b_2x_2 + \cdots + b_kx_k)$$

$$\ln\left(\frac{h(t|x)}{h_{0(t)}}\right) = b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where

$h_{0(t)}$: Baseline hazard function (All x variables = 0)

$x_1, x_2, \ldots, x_k$: Independent variables

$b_1, b_2, \ldots, b_k$: Unknown parameters of the model

Cox Regression model is semi-parametric method

(No assumption about specific distribution but parametric form of the model)

Case Study – Predicting Time Taken to Default

Background

- The bank possesses demographic and transactional data of its loan customers. If the bank has a robust model to predict defaulters it can undertake better resource allocation.

Objective

- To predict whether the customer applying for the loan will be a defaulter and to identify early defaulters.

Available Information

- Sample size is 700
- Age group, Years at current address, Years at current employer, Debt to Income Ratio, Credit Card Debts, Other Debts are the independent variables
- Status and Time are used to create survival objects. Status =1 if customer defaulted before 36 months, and 0 if no default was observed in 36 months

Data Snapshot

BANK LOAN (COX)

Independent variables

Survival objects

Columns	Description	Type	Measurement	Possible values
AGE	Age Groups 1 (<28 years), 2(28-40 years), 3 (>40 years)	Factor	1,2,3	3
EMPLOY	No. of Years the Customer is Employed	Numerical	Years	positive value
ADDRESS	No. of Years the Customer is Staying at their Current Address	Numerical	Years	positive value
DEBTINC	Debt to Income Ratio	Numerical	-	positive value
CREDDEBT	Credit to Debt Ratio	Numerical	-	positive values
OTHERDEBT	Other Debt	Numerical	-	Positive value
STATUS	Whether the Customer Defaulted on the Loan (1) or 0 (Censored at 36 Months)	Binary	0,1	2
TIME	Indicates Time of 'Default'	Numerical	In months	positive value

Model Fitting in R

```
#Importing the Data
bankloan<-read.csv("BANK LOAN (COX).csv",header=TRUE)
bankloan$AGE<-as.factor(bankloan$AGE)
```

- **AGE** is converted to factor.

```
#Creating a Survival Object
library(survival)

surv.object<-Surv(bankloan$TIME,bankloan$STATUS)
```

Surv() creates a survival object which will be used as the response variable in Cox regression. It requires time to event and event variable

```
#Model Fitting
timemodel<-coxph(surv.object~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT
                             +OTHDEBT, data=bankloan, x=TRUE)

summary(timemodel)
```

- **coxph()** from package **survival** fits a Cox Regression.
- Dependent variable (Survival object) is followed by a tilde and independent variables are separated by plus signs.
- x logical value: if TRUE, the x matrix is returned in component x

Model Fitting in R

Output

```
> summary(timemodel)
Call:
coxph(formula = surv.object ~ AGE + EMPLOY + ADDRESS + DEBTINC +
    CREDDEBT + OTHDEBT, data = bankloan)

  n= 700, number of events= 183

             coef exp(coef) se(coef)       z Pr(>|z|)
AGE2      0.30668   1.35891  0.18701   1.640   0.1010
AGE3      0.54006   1.71611  0.25293   2.135   0.0327 *
EMPLOY   -0.24177   0.78524  0.02238 -10.803  < 2e-16 ***
ADDRESS  -0.09825   0.90643  0.01634  -6.011 1.84e-09 ***
DEBTINC   0.05859   1.06034  0.01308   4.478 7.53e-06 ***
CREDDEBT  0.58482   1.79468  0.05020  11.649  < 2e-16 ***
OTHDEBT   0.06465   1.06679  0.03166   2.042   0.0411 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
AGE2        1.3589     0.7359    0.9419    1.9605
AGE3        1.7161     0.5827    1.0453    2.8173
EMPLOY      0.7852     1.2735    0.7515    0.8204
ADDRESS     0.9064     1.1032    0.8779    0.9359
DEBTINC     1.0603     0.9431    1.0335    1.0879
CREDDEBT    1.7947     0.5572    1.6265    1.9802
OTHDEBT     1.0668     0.9374    1.0026    1.1351

Concordance= 0.833  (se = 0.014 )
Likelihood ratio test= 336.9  on 7 df,   p=<2e-16
Wald test            = 282.4  on 7 df,   p=<2e-16
Score (logrank) test = 322  on 7 df,   p=<2e-16
```

Interpretation :
The **coef** column gives the estimates of the parameters in the model.

Individual Testing in R

```
# Output
```

```
              coef exp(coef) se(coef)       z Pr(>|z|)
AGE2       0.30668   1.35891  0.18701   1.640   0.1010
AGE3       0.54006   1.71611  0.25293   2.135   0.0327 *
EMPLOY    -0.24177   0.78524  0.02238 -10.803  < 2e-16 ***
ADDRESS   -0.09825   0.90643  0.01634  -6.011 1.84e-09 ***
DEBTINC    0.05859   1.06034  0.01308   4.478 7.53e-06 ***
CREDDEBT   0.58482   1.79468  0.05020  11.649  < 2e-16 ***
OTHDEBT    0.06465   1.06679  0.03166   2.042   0.0411 *
```

Interpretation:

- Except AGE2, all variables are significant and have an impact on time taken by a customer to default (p-values <0.05)
- Higher the number of years spent at one address or employee, lesser is the probability to default (as the coefficients are negative)
- Higher the amount of liabilities, higher is the probability to default (as the coefficients are positive)

Predicted Probabilities in R

```
#Importing New Data for Predictions & check if the structure is same as
#the train data
```

```
bankloantest<-read.csv("BANK LOAN (COX) TEST.csv",header=TRUE)
bankloantest$AGE <- as.factor(bankloantest$AGE)
```

```
#Predicted Probabilities
```

```
install.packages("pec")
library(pec)
bankloantest$prob24<-predictSurvProb(timemodel,bankloantest,times=24)

head(bankloantest)
```

- **predictSurvProb()** extracts probability predictions from different modeling approaches, most commonly used for Cox regression.
- **times=** is a vector of times in the range of the response variable, e.g. times when the response is a survival object, at which to return the survival probabilities
- Here, it is used to give probability that the customer will survive (Remain non-defaulter) for at least 24 months.

Predicted Probabilities in R

#Output

	SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	prob24
1	701	3	17	12	9.4	11.38	5.01	0.1262493
2	702	2	10	6	17.3	1.36	4.00	0.9341567
3	703	3	15	13	5.5	0.86	2.17	0.9957209
4	704	2	15	14	2.9	2.66	0.82	0.9930838
5	705	1	2	0	17.6	1.79	3.06	0.4630305
6	706	1	5	5	10.2	0.35	2.16	0.9416758

Interpretation :
Predicted probabilities that the customer will default before 24 months

Proportional Hazards Model

For any two cases, the ratio of hazard function at any time point is constant

Consider simple example in which only independent variable is X=1 for group 1 and X=0 for group 2

For X=1,hazrad function is

$$h(t|x)=h_0(t)\ \exp^{(b_1)}$$

For X=0,hazrad function is

$$h(t|x)=h_0(t)$$

Therefore, hazard ratio is

$$\exp^{(b_1)} = \text{Constant}$$

Quick Recap

Survival Analysis,	• Survival analysis is the study of time taken for an event to occur
Cox regression	• Cox regression is used to model "Time to Event" variable
Cox Regression Model Fitting,	• `coxph()` from package **survival** fits a Cox regression • Survival Object is the response variable in `coxph()` • `summary()` of `coxph()` object returns Likelihood Ratio test results and p-values for checking variable significance
Predictions on New Data	• `predictSurvProb()` from package **pec** generates predicted probabilities
Proportional Hazards Model	• For any two cases, the ratio of hazard function at any time point is constant