Chapter -5

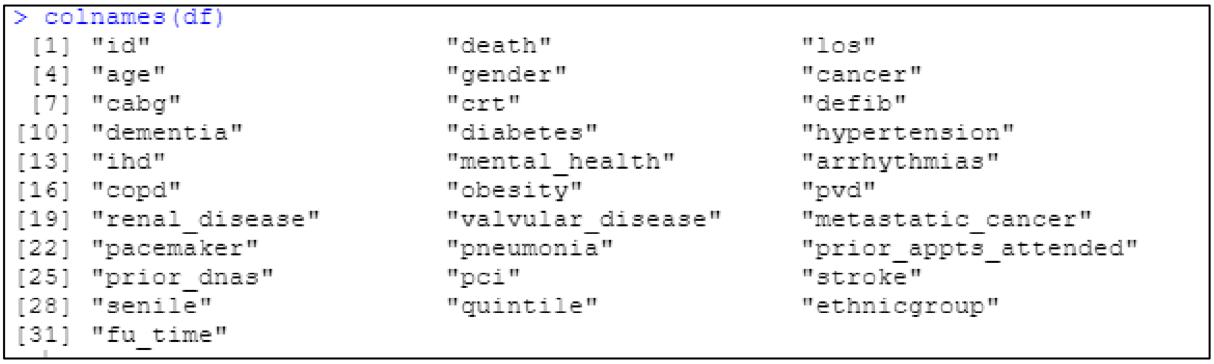
(Methodology/Procedure)

1) Heath Dataset

• The Kaplan Meier Plotting

One of the datasets where we can apply Survival analysis is from a hospital which recorded all the required events from a set of patients. We first performed EDA on it. Data preprocessing is not much required for this dataset as it deals with death or life, even a small wrong assumption will lead to wrong results.

Columns in the Health dataset:



```
> colnames(df)
 [1] "id"                   "death"                "los"
 [4] "age"                  "gender"               "cancer"
 [7] "cabg"                 "crt"                  "defib"
[10] "dementia"             "diabetes"             "hypertension"
[13] "ihd"                  "mental_health"        "arrhythmias"
[16] "copd"                 "obesity"              "pvd"
[19] "renal_disease"        "valvular_disease"     "metastatic_cancer"
[22] "pacemaker"            "pneumonia"            "prior_appts_attended"
[25] "prior_dnas"           "pci"                  "stroke"
[28] "senile"               "quintile"             "ethnicgroup"
[31] "fu_time"
```

- All the columns except the id, age, ethnicgroup,prior_appts_attended prior_dnas and fu_time(time in which the analysis is done) are binary.
- Age, prior_appts_attended prior_dnas and fu_time contain continuous values.
- Ethnicgroup is a categorical variable.

One of the applications where we use survival analysis is to evaluate the time of death. Here is the concept behind it.

The Life Tables:

Life tables are used to measure the probability of death at a given age and the life expectancy at varying ages. Actuarial science and of course life insurance companies need to know this in detail, but we in public health do too. There are two different kinds of life table:

- Cohort or generational life tables
- Current or period life tables

Cohort life tables take an actual set of people born at the same time, usually in the same year or even on the same day of the same year, and follow them up for their whole lives.

Period life tables take a hypothetical cohort of people born at the same time and use the assumption that they are subject to the age-specific mortality rates of a region or country.

How are life tables constructed?

In a common type of epidemiological study called a cohort study, a set or cohort of patients are enrolled at time zero and then followed up to see who gets the outcome of interest, such as death, and when they get it. The latter will often be measured in days since the study start, but not necessarily.

Let's suppose that we start off with 100 patients. Everybody makes it past time zero, so the probability of surviving at least to time t=0 is 1, or 100%. This probability is technically known as the survival function, one of two core concepts in survival analysis. Let's now say that two people die the day after they are enrolled. However, this assumes that everybody enters the study at the same time, t=0, and no one leaves it except by death. Let's now say that two people die the day after they are enrolled. The life table then looks like this:

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Probability of survival past time t
0 (study start)	100	0	1
1	100	2	0.98
2	98	??	??
3..	??	??	??

The calculations continue in that way. The technical term for dealing with the people that are lost to follow-up is that these people are censored. **Censoring** is a really important concept in survival analysis. The Kaplan-Meier table and associated plot is the simplest (but not the only) way of estimating the survival time when you have drop-outs.

How to calculate a Kaplan-Meier table and plot by hand?

The plot of the survival function versus time is called the survival curve. The Kaplan-Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution.Suppose we are monitoring patients after a particular treatment.

Time (t) in days	Event
0 (study start)	8 patients recruited
1	2 patients die
4	1 patient dies

5.	1 patient dies
etc	etc

We’re interested in the event 'death'. We can easily determine how many patients were alive at any given day and how many died and when.

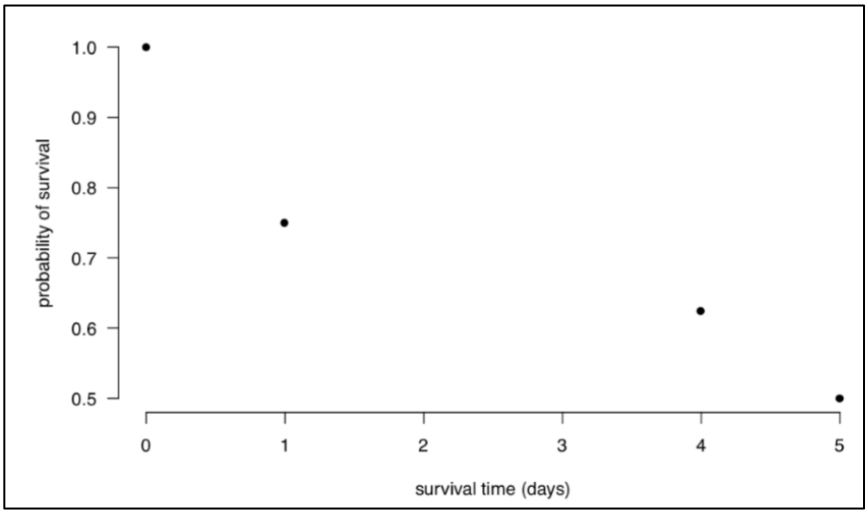
But how do we compute the probability of survival past time t? Start by computing the proportion of patients that survive day t, i.e. of those alive at the beginning of day t, what proportion make it to the next day alive? On day 0, the day the study begins, there are no deaths. Everybody survives. Hence the proportion surviving is 1. On the following day, 2 out of 8 patients don't make it; in other words, 75% survive the day.

With no deaths on day 0, the probability of surviving is 1. Computing the next probability is a bit trickier. The basic idea underlying Kaplan-Meier tables comes into play here: the probability of surviving past day t is simply the probability of surviving past day t-1 times the proportion of patients that survive on day t. Let's see it together:

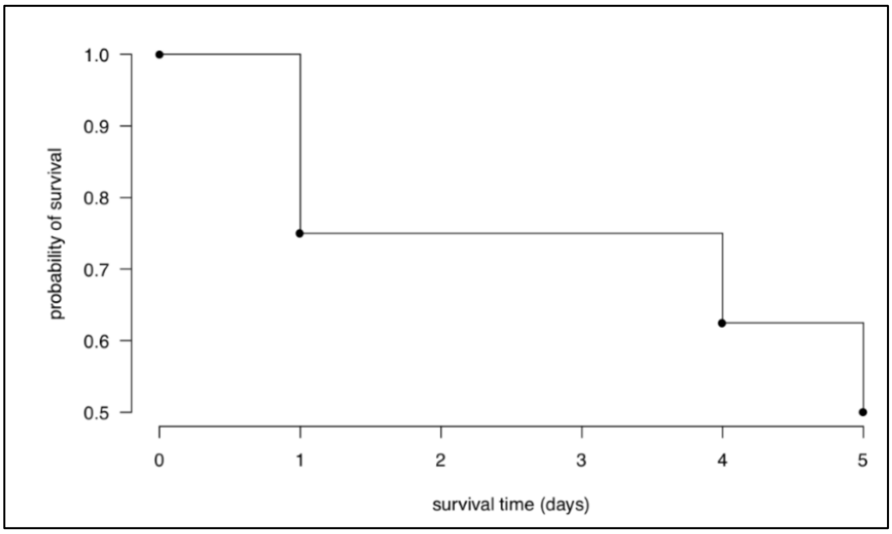
Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Proportion of patients surviving past time t	Probability of survival past time t
0	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	1 * 0.75 = 0.75
4	6	1	(6-1)/6=0.83	0.75*0.83 = 0.623
5	5	1	(5-1)/5=0.8	0.623*0.8 = 0.498

KM Plotting

If we now plot the time column against the probability column, we end up with a survival curve. We plot the time on the x-axis, running from 0 on the left to the highest day count, i.e. 5 in this example, on the right. The probability of survival goes on the y-axis, with 0 on the bottom and 1 as the maximum.



If we now connect the dots using steps, first horizontal then vertical, we have drawn our first survival curve.You might think why the “steps” involve a horizontal line followed by a vertical line and not the other way around. This is because the probability is assumed to be the same until the next death occurs. For example, there’s a death at day 1 but then no more deaths until day 4.



Dealing with censored data:

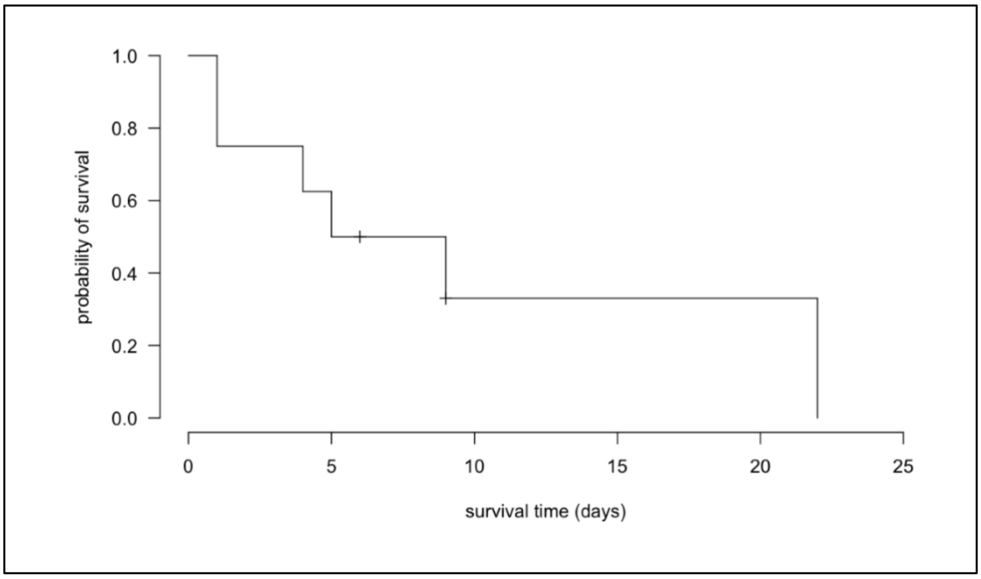
The patients that are left out from the survey i.e. censored should be treated differently from patients that die. When a patient is censored at time t, we know the patient was alive at time t, but we don't know whether the patient has died or survived. For this reason, censored patients are classified neither as 'survived' nor as 'died' on any given day. We simply deduct them from the number of patients alive. When there are censored patients at the same time as patients that die, we deal first with patients that die. Then we add a new line, mark it with a little '+' right after the time count and denote the censored patient(s) by taking them off the count of patients alive at time t.

Let's follow up our patients for another two weeks. This is what we have:

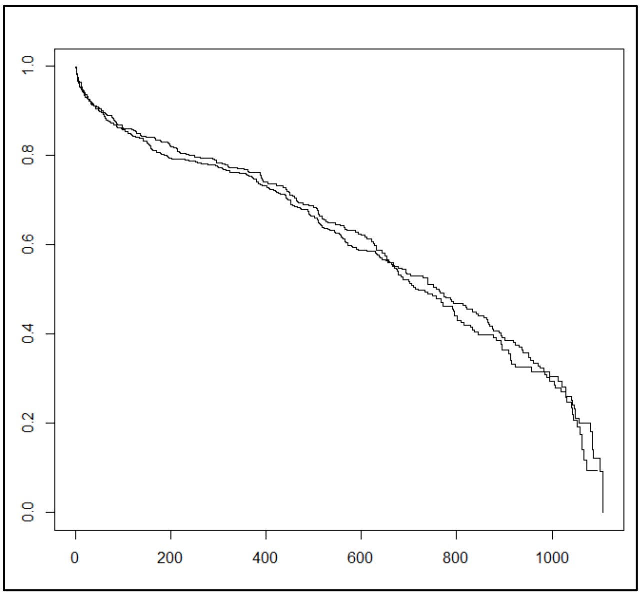
Time (t) in days	Event
0 (study start)	8 patients recruited
1	2 patients die
4	1 patient dies
5.	1 patient dies
6	1 patient drop out
9	1 patient dies and 1 drops out
22	1 patient dies

Time (t) in days	Number of patients alive at time t	Number of patients who died at time t	Proportion of patients surviving past time t	Probability of survival past time t
0	8	0	1	1
1	8	2	0.75	0.75
4	6	1	0.83	0.75*0.83 = 0.623
5	5	1	0.8	0.623*0.8 = 0.498
6+	4	0	4/4=1	0.498*1 = 0.498
9	3	1	(3-1)/3=0.667	0.498*0.667 = 0.332
9+	2	0	2/2=1	0.332*1 = 0.332
22	1	1	0/1=0	0

Here's the complete survival curve relating to these data. You indicate that a patient was censored from the study with a little '+' on the curve on the day they were lost to follow-up:



Now lets deal with the actual dataset where we use the KM plot.



Above, is the KM plot to show the distribution of probability over two different genders which are indicated by the two curves. For more understanding we should refer the code.

2) Cox Model

Cox model is the most commonly used survival analysis method for incorporating not just one but multiple predictors of survival: Cox proportional hazards regression modelling. Above we saw the KM plot for the probability distribution over days from two different genders. So, to compare them numerically we use Cox model.With this method you will be able to compare the survival of multiple groups of patients at the same time.

A Cox model is a well-recognized statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death, or other event of interest, for individuals, given their prognostic variables.Interpreting a Cox model involves examining the coefficients for each explanatory variable. A positive regression coefficient for an explanatory variable means that the hazard for a patient having a high positive value on that particular variable is high.

Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable.Cox’s method does not assume any particular distribution for the survival times, but it rather assumes that the effects of the different variables on survival are constant over time and are additive in a particular scale.

Hazard Function:

Like the name suggests, the model is formulated around the concept of hazards. The hazard function h(t) is the probability of the event happening at time t, given that it has not yet happened. In other words, h(t) is the probability of dying at time t having survived up to time t.

Risk set:

An important concept involved in the calculation of the hazard is the risk set. Just like the risk of dying (or experiencing some specific event) changes over time, so the number of patients that are subjected to that risk change over time as people die or drop out. The risk set at time t is defined as the set of patients at time t that are at risk of experiencing the event.

Hazard Ratio:

Usually in survival analysis, we are interested in the difference between survival curves of different groups of patients. Earlier you saw the log-rank test, which gives a p value for comparing the survival curves between different groups of patients with a Kaplan-Meier plot. The p value tells you nothing about the size of the difference between the survival curves, however.

This is done by dividing one hazard by another to give a hazard ratio. For example, dividing the hazard for females by the hazard for males gives you a hazard ratio for females compared with males. It tells you how much more likely female patients will die than male patients.

- In R we need to fit our constraints in coxph function which evaluates the above hazard function and risk set ade finally the hazard ratio.
- Here we will evaluate the Cox model using the predictors age and ethnicgroup individually.

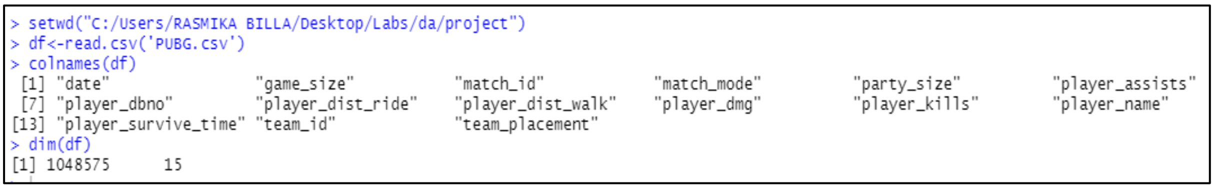
3) Multiple Cox Model

Earlier you loaded the data set for the course and ran Kaplan-Meier analysis and then simple Cox regression, but you only looked at the outcome variable – death – and a single predictor – age and then ethnic group. The latter had some missing values. Now we want to incorporate more variables into the Cox model, so we need to summarise each of them first, to see if they too have any hidden traps. Having run the Cox model with just one predictor and become acquainted with its code and output, you'll want to move on to include multiple predictors in a multiple Cox regression model.

2) PUBG Dataset

The theory behind solving this dataset and predicting the survival rate is same as the health dataset. We have included further more concepts such as making a web application using shiny package here.

The Columns in PUBG dataset are:



```
> setwd("C:/Users/RASMIKA BILLA/Desktop/Labs/da/project")
> df<-read.csv('PUBG.csv')
> colnames(df)
 [1] "date"               "game_size"          "match_id"           "match_mode"         "party_size"         "player_assists"
 [7] "player_dbno"        "player_dist_ride"   "player_dist_walk"   "player_dmg"         "player_kills"       "player_name"
[13] "player_survive_time" "team_id"           "team_placement"
> dim(df)
[1] 1048575      15
```

The same model: KM plot is applied here to observe the player survival time with respect to the three main columns i.e., game size, Party size and Player distance of riding. The game size column tells about the size of the play which is been divided into three major sections to get results and at the same time the party size means player size and in PUBG the players are single, duo and squad and lastly whether a player rides and walks in the game is determined by the player distance ride column.

So, with these we do various plotting and observe the values to draw certain conclusions. Also, we use the shiny package in order to make a user understand his/her decisions on how to win the PUBG game.