Predicting Mortality due to CVD using Logistic Regression

University roll:193012-21-0400

Registration number:012-1111-1111-19

Sem:6

College roll:1002

Paper code : DSE-B2

Asutosh college

Contents

1 Introduction:

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

1.1 Objective:

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

2 About the data:

Our data is collected from www.kaggale.com

anaemia - anaemia is a condition in which you lack enough healthy red blood cells to carry adequate oxygen to your body's tissues. Having anaemia, also referred to as low haemoglobin, can make you feel tired and weak. (there is not anaemia - 0, there is anaemia - 1)

▲ ***creatine_phosphokinase: (CPK) - CPK is an enzyme in the body. It is found mainly in the heart, brain, and skeletal muscle. Total CPK normal values: 10 to 120 micrograms per litter (mcg/L)***

▲ ***ejection fraction: (EF) - EF is a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction. An ejection fraction of 60 percent means that 60 percent of the total amount of blood in the left ventricle is pushed out with each heartbeat. This indication of how well your heart is pumping out blood can help to diagnose and track heart failure. A normal heart's ejection fraction may be between 50 and 70 percent.***

▲ ***Platelets - platelets are colourless blood cells that help blood clot. Platelets stop bleeding by clumping and forming plugs in blood vessel injuries. Thrombocytopenia might occur as a result of a bone marrow disorder such as leukaemia or an immune system problem. The normal number of platelets in the blood is 150,000 to 400,000 platelets per microliter (mcL) or 150 to 400 × 109/L.***

▲ serum creatinine - The amount of creatinine in your blood should be relatively stable. An increased level of creatinine may be a sign of poor kidney function. Serum creatinine is reported as milligrams of creatinine to a decilitre of blood (mg/dl) or micromoles of creatinine to a litter of blood (micromoles/L). Here are the normal values by age: 0.9 to 1.3 mg/dl for adult males. 0.6 to 1.1 mg/dl for adult females. 0.5 to 1.0 mg/dl for children ages 3 to 18 years.

▲ serum_sodium - Measurement of serum sodium is routine in assessing electrolyte, acid-base, and water balance, as well as renal function. Sodium accounts for approximately 95% of the osmotic ally active substances in the extracellular compartment, provided that the patient is not in renal failure or does not have severe hyperglycaemia. The normal range for blood sodium levels is 135 to 145 mill equivalents per litter (meg/L).

high_blood_pressure - (True - 1, False - 0)

▲ age - between 40 – 95

▲ diabetes - (True - 1, False - 0)

▲ sex - (male - 1, female - 0)

▲ smoking - (True - 1, False – 0)

▲ DEATH_EVENT - (True - 1, False - 0)

3 Research Methodology:

Our research methodology primarily consists of the following steps:

1. Computing summary statistics, and performing exploratory analysis of the data
2. Discovering relationships between data in terms of graphical visualisations and correlations between variables.
3. Performing logistic regression and computing the accuracy of the model

3 Importing Data:

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	ser
63	1	61	1	40	0	221000	
65	1	305	0	25	0	298000	
75	0	582	0	45	1	263358	
80	0	898	0	25	0	149000	

42	0	5209	0	30	0	226000
60	0	53	0	50	1	286000
72	1	328	0	30	1	621000
55	0	748	0	45	0	263000
45	1	1876	1	35	0	226000
63	0	936	0	38	0	304000
45	0	292	1	35	0	850000
85	0	129	0	60	0	306000
55	0	60	0	35	0	228000
50	0	369	1	25	0	252000
70	1	143	0	60	0	351000
60	1	754	1	40	1	328000
58	1	400	0	40	0	164000
60	1	96	1	60	1	271000
85	1	102	0	60	0	507000
65	1	113	1	60	1	203000
86	0	582	0	38	0	263358
60	1	737	0	60	1	210000
66	1	68	1	38	1	162000
60	0	96	1	38	0	228000
60	1	582	0	30	1	127000
60	0	582	0	40	0	217000
43	1	358	0	50	0	237000
46	0	168	1	17	1	271000
58	1	200	1	60	0	300000
61	0	248	0	30	1	267000
53	1	270	1	35	0	227000
53	1	1808	0	60	1	249000
60	1	1082	1	45	0	250000
46	0	719	0	40	1	263358
63	0	193	0	60	1	295000
81	0	4540	0	35	0	231000
75	0	582	0	40	0	263358
65	1	59	1	60	0	172000
68	1	646	0	25	0	305000
62	0	281	1	35	0	221000
50	0	1548	0	30	1	211000
80	0	805	0	38	0	263358
46	1	291	0	35	0	348000
50	0	482	1	30	0	329000
61	1	84	0	40	1	229000
72	1	943	0	25	1	338000
50	0	185	0	30	0	266000
52	0	132	0	30	0	218000
64	0	1610	0	60	0	242000
75	1	582	0	30	0	225000

60	0	2261	0	35	1	228000
72	0	233	0	45	1	235000
62	0	30	1	60	1	244000
50	0	115	0	45	1	184000
50	0	1846	1	35	0	263358
65	1	335	0	35	1	235000
60	1	231	1	25	0	194000
52	1	58	0	35	0	277000
50	0	250	0	25	0	262000
85	1	910	0	50	0	235000
59	1	129	0	45	1	362000
66	1	72	0	40	1	242000
45	1	130	0	35	0	174000
63	1	582	0	40	0	448000
50	1	2334	1	35	0	75000
45	0	2442	1	30	0	334000
80	0	776	1	38	1	192000
53	0	196	0	60	0	220000
59	0	66	1	20	0	70000
65	0	582	1	40	0	270000
70	0	835	0	35	1	305000
51	1	582	1	35	0	263358
52	0	3966	0	40	0	325000
70	1	171	0	60	1	176000
50	1	115	0	20	0	189000
65	0	198	1	35	1	281000
60	1	95	0	60	0	337000
69	0	1419	0	40	0	105000
49	1	69	0	50	0	132000
63	1	122	1	60	0	267000
55	0	835	0	40	0	279000
40	0	478	1	30	0	303000
59	1	176	1	25	0	221000
65	0	395	1	25	0	265000
75	0	99	0	38	1	224000
58	1	145	0	25	0	219000
61	1	104	1	30	0	389000
50	0	582	0	50	0	153000
60	0	1896	1	25	0	365000
61	1	151	1	40	1	201000
40	0	244	0	45	1	275000
80	0	582	1	35	0	350000
64	1	62	0	60	0	309000
50	1	121	1	40	0	260000
73	1	231	1	30	0	160000
45	0	582	0	20	1	126000

77	1	418	0	45	0	223000
45	0	582	1	38	1	263358
65	0	167	0	30	0	259000
50	1	582	1	20	1	279000

```
'data.frame':   100 obs.  of  12 variables:
 $ age                : int  63 65 75 80 42 60 72 55 45 63 ...
 $ anaemia              : int   1 1 0 0 0 0 1 0 1 0 ..
 $ creatinine_phosphokinase:  int  61 305 582 898 5209 53 328 748 1876 936 ...
 $ diabetes             : int  1 0 0 0 0 0 0 0 1 0 ...
 $ ejection_fraction     : int  40 25 45 25 30 50 30 45 35 38 ...
 $ high_blood_pressure   : int  0 0 1 0 0 1 1 0 0 0 ...
 $ platelets            : num  221000 298000 263358 149000 226000 ...
 $ serum_creatinine      : num   1.1 1.1 1.18 1.1 1 2.3 1.7 1.3 0.9    1.1 ...
 $ serum_sodium        : int    140 141 137 144 140 143 138 137 138  133 ...
 $ sex                : int  0 1 1 1 1 1 0 0 1 1 1 ...
 $ smoking             : int   0 0 0 1 1 0 1 0 0 1 ...
 $ DEATH_EVENT         : int   0 0 0 0 0 0 1 0 0  0 ...
```

5 Summary of the data:

```
age               anaemia          creatinine_phosphokinase    diabetes

Min.   : 40.00     Min.   : 0.00          Min :  30.0              Min: 0.00

1st Qu.:50.00      1st Qu.:0.00           1st Qu.: 129.8           1st Qu.:0.00

Median: 60.00       Median: 0.00           Median: 316.5           Median: 0.00

Mean   : 60.42      Mean   : 0.48        Mean   : 635.0             Mean   :0.38

3rd Qu.:65.25      3rd Qu.:1.00           3rd Qu.: 723.5             3rd Qu.:1.00

Max. :86.00       Max.   :1.00           Max.   :5209.0            Max.   :1.00

ejection_fraction     high_blood_pressure          platelets          serum_creatinine

Min.   :17.00      Min.   :0.00                     Min.   : 70000        Min.   :0.600

1st Qu.:30.00      1st Qu.:0.00                    1st Qu.:219750       1st Qu.:0.900

Median :38.00       Median :0.00                  Median :255500       Median :1.100

Mean   :39.21       Mean   :0.34                  Mean   :259939       Mean   :1.302

3rd Qu.:45.00       3rd Qu.:1.00                   3rd Qu.:288250     3rd Qu.:1.325

Max.   :60.00       Max.   :1.00                    Max.   :850000     Max.   :6.100

serum_sodium       sex              smoking             DEATH_EVENT
```

```
Min:124.0          Min.:0.00          Min :0.00           Min :0.00
1st Qu:135.0       1st Qu.:0.00       1st Qu.:0.00        1st Qu.:0.00
Median : 137.0     Median :1.00     Median :0.00       Median :0.00
Mean   :137.1      Mean   :0.63      Mean   :0.33       Mean   :0.23
3rd Qu.:139.2      3rd Qu.:1.00      3rd Qu.:1.00       3rd Qu.:0.00
Max   :145.0       Max:1.00        Max :1.00          Max :1.00
```

5 Summary of the data:

Age:	anaemia:	creatinine_phosphokinase:	diabetes:
Sd=11.114	sd:0.500	sd: 882.265	sd:0.485
Pearsons' SK1:0.113	Pearsons' SK1: 2.880	Pearsons' SK1:1.083	Pearsons' SK1 : 2.350

ejection_fraction :	high_blood_pressure :	platelets:	serum_creatinine:
sd:11.789	sd:0.474	sd:98359.280	sd:0.747
Pearsons' SK1:0.254	Pearsons' SK1:2.152	Pearsons' SK1: 0.135	Pearsons' SK1:0.811

serum_sodium :	sex:	smoking:	DEATH_EVENT:
sd:3.799	sd:0.483	sd: 0.470	sd:0.420
Pearsons' SK1:0.790	Pearsons' SK1:-6.211	Pearsons' SK1:2.106	Pearsons' SK1:1.642

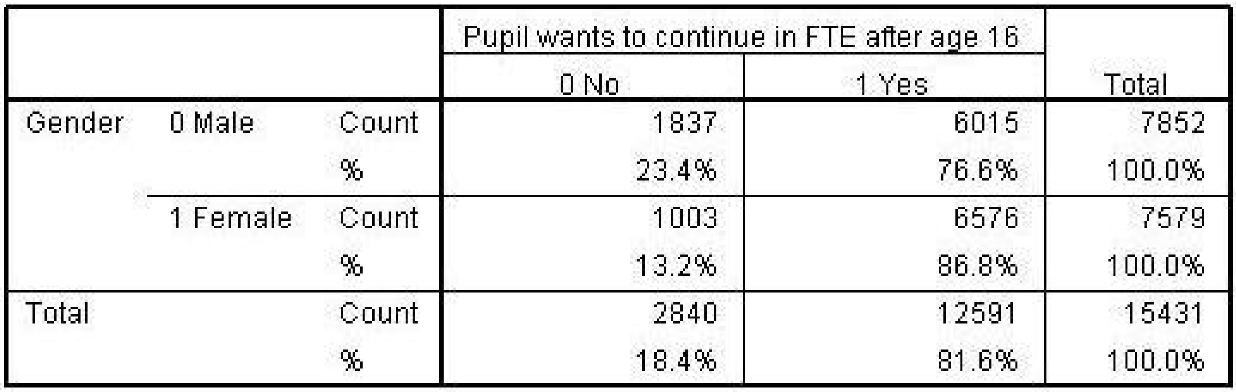
- comment: Frequency distribution of age,ejection farction,platelates are almost symmetric as they have a very low pearsons' skewness coefficient(>0);
  creatinine_phosphokinase,serum_creatinine,serum_sodium,DEATH_EVENT are moderately positively skewed, diabetes,high blood pressure,smoking are highly positively skewed and sex is negatively skewed

6. Odds and odds ratios:

Odds express the likelihood of an event occurring relative to the likelihood of an event not occurring. Let us make it easy through an example;

Let's start by considering a simple association between two dichotomous variables (a 2 x 2 cross tabulation) drawing on the LSYPE dataset. The outcome we are interested in is whether students aspire to continue in Full-time education (FTE) after the age of 16 (the current age at which students in England can choose to leave FTE). We are interested in whether this outcome varies between boys and girls. We can present this as a simple cross tabulation

Figure 6.1.1: Aspiration to continue in full time education (FTE) after the age of 16 by gender: Cell counts and percentages:



			Pupil wants to continue in FTE after age 16		
			0 No	1 Yes	Total
Gender	0 Male	Count	1837	6015	7852
		%	23.4%	76.6%	100.0%
	1 Female	Count	1003	6576	7579
		%	13.2%	86.8%	100.0%
Total		Count	2840	12591	15431
		%	18.4%	81.6%	100.0%

We have coded not aspiring to continue in FTE after age 16 as 0 and aspiring to do so as 1. Although it is possible to code the variable with any values, employing the values 0 and 1 has advantages. The mean of the variable will equal the proportion of cases with the value 1 and can therefore be interpreted as a probability. Thus we can see that the percentage of all students who aspire to continue in FTE after age 16 is 81.6%. This is equivalent to saying that the *probability* of aspiring to continue in FTE in our sample is 0.816.

Odds and odds ratios

However another way of thinking of this is in terms of the *odds*. Odds express the likelihood of an event occurring relative to the likelihood of an event not occurring. In our sample of 15,431 students, 12,591 aspire to continue in FTE while 2,840 do not aspire, so the odds of aspiring are 12591/2840 = 4.43:1 (this means the ratio is 4.43 to 1, but we conventionally do not explicitly include the: 1 as this is implied by the odds). The odds tell us that if we choose a student at random from the sample they are 4.43 times more likely to aspire to continue in FTE than not to aspire to continue in FTE.

We don’t actually have to calculate the odds directly from the numbers of students if we know the proportion for whom the event occurs, since the odds of the event occurring can be gained directly from this proportion by the formula (Where p is the probability of the event occurring.):

$$odds = \frac{P}{1-P}$$

Thus the odds in our example are:

Odds= [p/(1-p)] = .816 / (1-.816)= .816 /.184 = 4.43.

The above are the *unconditional odds,* i.e. the odds in the sample as a whole. However odds become really useful when we are interested in how some other variable might affect our outcome. We consider here what the odds of aspiring to remain in FTE are separately for boys and girls, i.e. *conditional* on gender. We have seen the odds of the event can be gained directly from the proportion by the formula odds=p/(1-p).

So for boys the odds of aspiring to continue in FTE = .766/(1-.766)= 3.27

While for girls the odds of aspiring to continue in FTE = .868/(1-.868)= 6.56.

These are the *conditional odds*, i.e. the odds depending on the condition of gender, either boy or girl.

We can see the odds of girls aspiring to continue in FTE are higher than for boys. We can in fact directly compare the odds for boys and the odds for girls by dividing one by the other to give the *Odds Ratio* (OR). If the odds were the same for boys and for girls then we would have an odds ratio of 1. If however the odds differ then the OR will depart from 1. In our example the odds for girls are 6.53 and the odds for boys are 3.27 so the OR= 6.56 / 3.27 = 2.002, or roughly 2:1. This says that girls are *twice as likely* as boys to aspire to continue in FTE. 7

Seeing the relationship as a model

An interesting fact can be observed if we look at the odds for boys and the odds for girls in relation to the odds ratio (OR).

*For boys (our base group) the odds= 3.27 * 1 = 3.27*

*For girls the odds = 3.27 * 2.002 = 6.56.*

So another way of looking at this is that the odds for each gender can be expressed as a constant multiplied by a gender specific multiplicative factor (namely the OR).

*p/(1-p) = constant * OR.*

However there are problems in using ORs directly in any modeling because they are asymmetric. As we saw in our example above, an OR of 2.0 indicates the same relative ratio as an OR of 0.50, an OR of 3.0 indicates the same relative ratio as an OR of 0.33, an OR of 4.0 indicates the same relative ratio as an OR of 0.25 and so on. This asymmetry is unappealing because ideally the odds for males would be the opposite of the odds for females.

Note that the way odd-ratios are expressed depends on the baseline or comparison category. For gender we have coded boys=0 and girls =1, so the boys are our natural base group. However if we had taken girls as the base category, then the odds ratio would be 3.27 / 6.56= 0.50:1. This implies that boys are *half as likely* to aspire to continue in FTE as girls. You will note that saying "Girls are twice as likely to aspire as boys" is actually identical to saying "boys are half as likely to aspire as girls". Both figures say the same thing but just differ in terms of the base.
Odds Ratios from 0 to just below 1 indicate the event is *less likely* to happen in the comparison than in the base group, odds ratios of 1 indicate the event is *exactly as likely* to occur in the two groups, while odds ratios from just above 1 to infinity indicate the event is *more likely* to happen in the comparator than in the base group.

Odds, Log odds and exponents

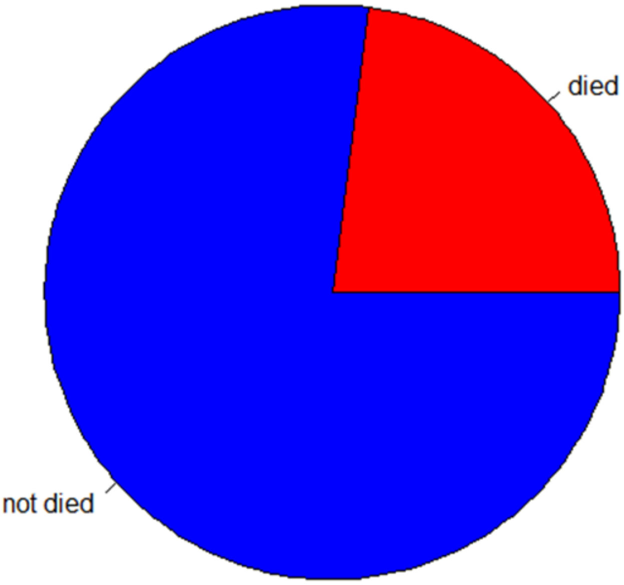
This asymmetry problem disappears if we take the „log" of the OR. „log doesn't refer to some sort of statistical deforestation… rather a mathematical transformation of the odds which will help in creating a regression model. Taking the log of an OR of 2 gives the value Log(2)= +0.302 and taking the log of an OR of 0.50 gives the value Log (0.5)= -0.302. See what's happened? The Log of the OR, sometimes called the logit , makes the relationships symmetric around zero (the OR"s become plus and minus .302). Logits and ORs contain the same information, but this difference in mathematical properties makes logits better building blocks for logistic regression.

7 Diagrammatic representation of the data :

The diagrammatic representation of the whole data is very useful not only for the analyst but also for layman to easily understand the facts. Here we will use pie charts, multiple bar diagrams to get picture how our response variable vary with the covariates.

7.1 Categorical analysis:

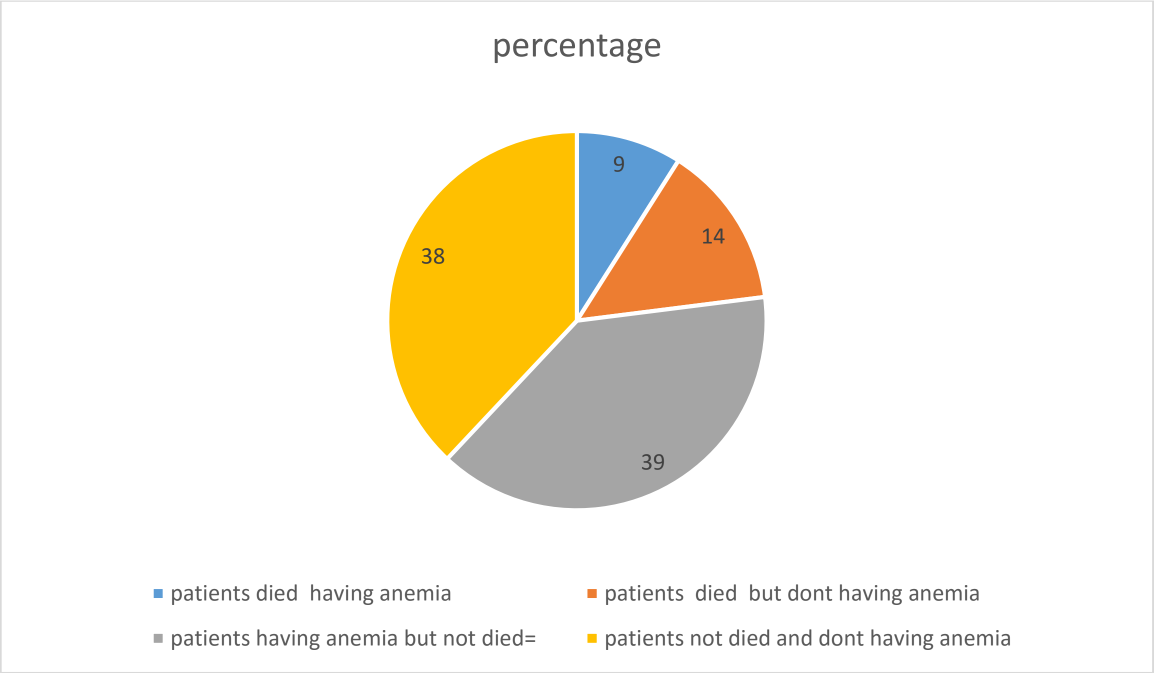
- We want to see diagrammatically the proportion of the patient died who have CVDs



Piechart shows that 23% of the patients died who have CVDS

- We want to see proportion of patients having anaemia died

#number of patients having anemia=48
#number of patients died =23
#number of patients died having anemia=9
#number of patients died but don’t having anemia=14
#number of patients having anemia but not died=48-9=39
#number of patients not died and don’t having anemia=100-(9+14+39)=



We will use yules method to find the association between anemia and death.

7.1.1 2x2 contingency table for death and anemia

	Deceased	Not deceased	total
Patients having anemia	fAB=9	faB=39	48
Patients not having anemia	fAb=14	fab=38	52
total	23	77	100

Yule’s Coefficient of association Q_{AB} =(fABfab −fAbfaB)/(fABfab+ fAbfaB)=-0.23

Death and anemia is occurring together less frequently than they would if they are two independent events.

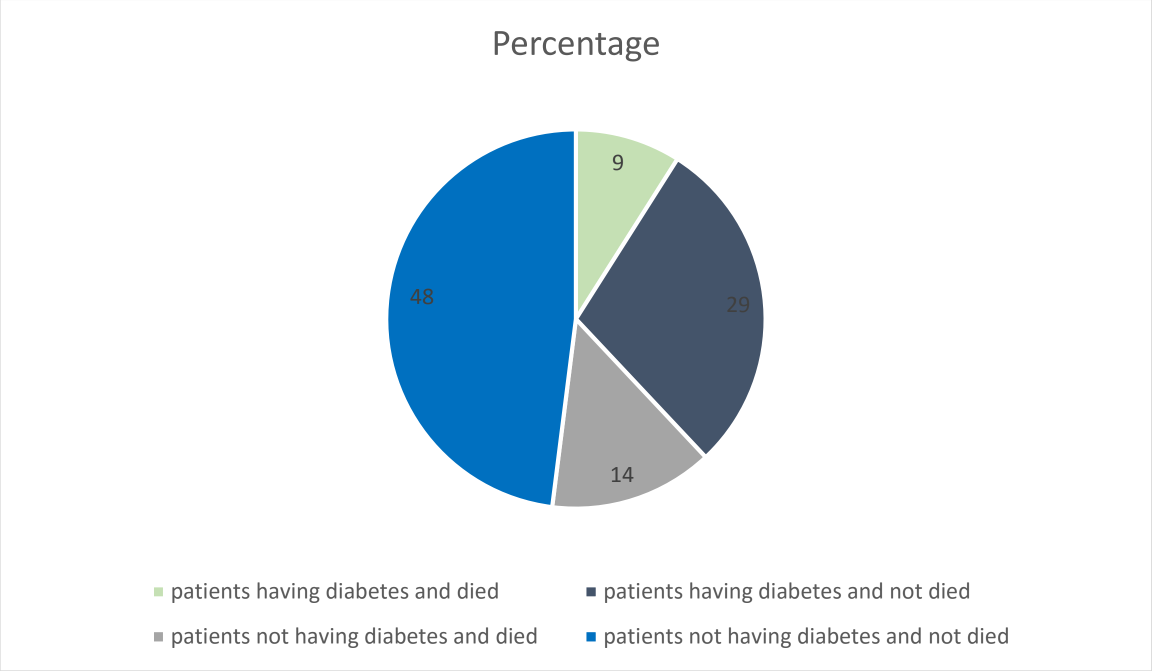
Odds of being deceased for a person having anemia is =fAB/faB=0.1875

Odds of being deceased for a person not having anemia is =fAb/fab=0.368

A person not having anemia is 1.9 times more likely to die than a person having anemia.

- We want to see proportion of patients having diabetes died

#number of patients having diabetes=38
#number of patients having diabetes and died=9
#number of patients having diabetes and not died=38-9=29
#number of patients not having diabetes and died=23-9=14
#number of patients not having diabetes and not died=100-(9+29+14)=48



We will use yules method to find the association between anemia and death.

7.1.1 2x2 contingency table for death and diabetes

	Deceased	Not deceased	total
Patients having diabetes	fAB=9	faB=29	38
Patients not having diabetes	fAb=14	fab=48	62
total	23	77	100

Yule's Coefficient of association Q_{AB} =(fABfab −fAbfaB)/(fABfab+ fAbfaB)=0.0310

Death and diabetes is occurring together more frequently than they would if they are two independent events.

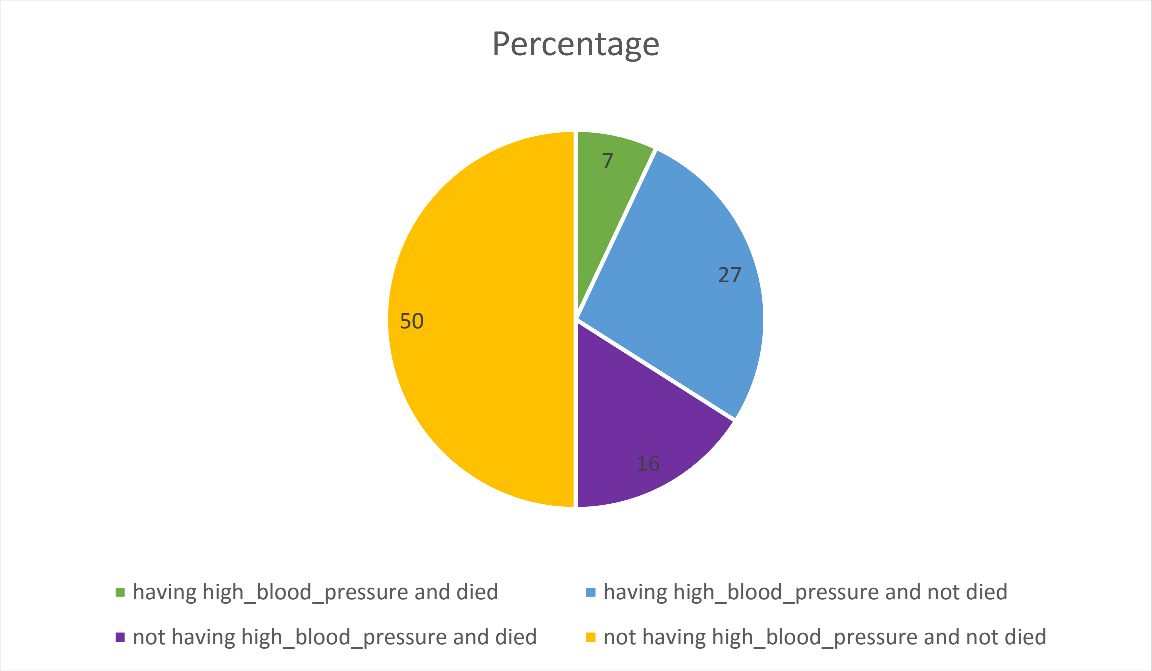
Odds of being deceased for a person having anemia is =fAB/faB=0.31

Odds of being deceased for a person not having anemia is =fAb/fab=0.29

A person not having diabetes is 0.935 times more likely to die than a person having diabetes

- We want to see proportion of patients having high blood pressure died

#of patients having high_blood_pressure=34
#of patients having high_blood_pressure and died=7
#of patients having high_blood_pressure and not died=34-7=27
#of patients not having high_blood_pressure and died=23-7=16
#of patients not having high_blood_pressure and not died=100-(7+27+16)=50



7.1.1 2x2 contingency table for death and highbloodpressure

	Deceased	Not deceased	total
Patients having highbloodpressure	fAB=7	faB=27	34
Patients not having highbloodpressure	fAb=16	fab=50	66
total	23	77	100

Yule's Coefficient of association Q_{AB} =(fABfab −fAbfaB)/(fABfab+ fAbfaB)=-0.105

Death and hbd is occurring together less frequently than they would if they are two independent events.

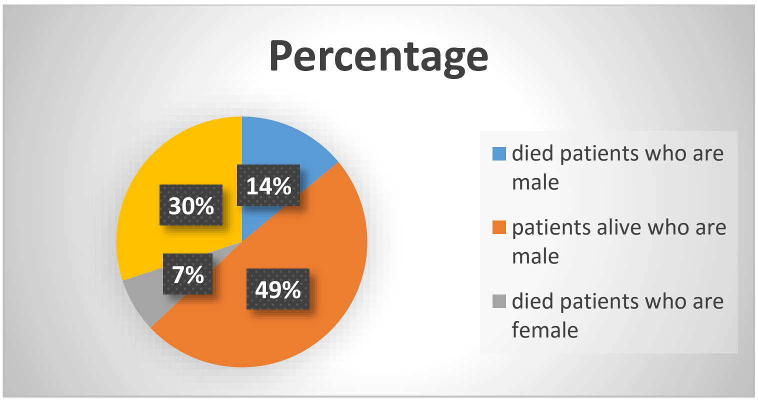
Odds of being deceased for a person having hbd is =fAB/faB=0.26

Odds of being deceased for a person not having hbd is =fAb/fab=0.32

A person not having highbloodpressure is 1.23 times more likely to die than a person having diabetes

- How death event varies with sex

#of male patients=63
#of died patients who are male=14
#patients alive who are male =63-14=49
# of died patients who are female =23-14=7
# patients alive who are male =100-(14+49+7)=32



7.1.3 2x2 contingency table for death and sex

	Deceased	Not deceased	total
male	fAB=14	faB=49	63
female	fAb=7	fab=30	37
total	23	77	100

Yule's Coefficient of association Q_{AB} =(fABfab −fAbfaB)/(fABfab+ fAbfaB)=0.1

Death and a patient being male is occurring together less frequently than they would if they are two independent events.

Odds of being deceased for a male is =fAB/faB=0.28

Odds of being deceased for a female is =fAb/fab=0.23

A female is 0.82 times more likely to die than a mae

- How death event varies with smoking

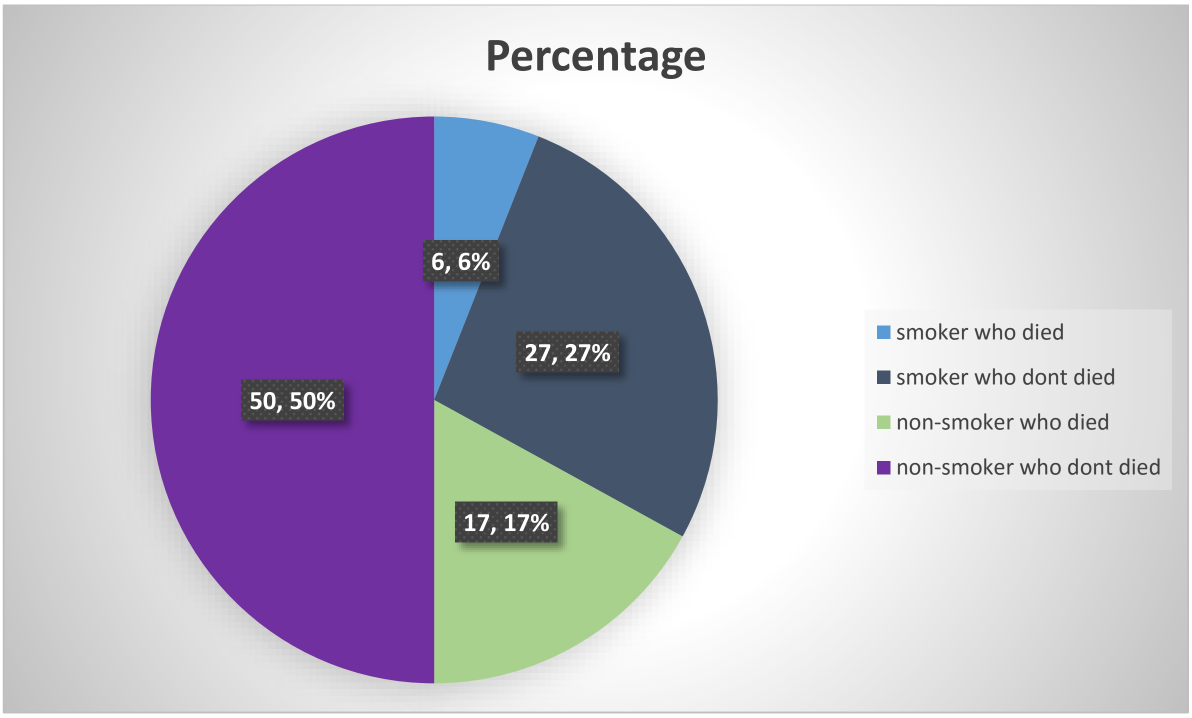
#of smokers=33

#of smoker who died=6

#of smoker who dont died=33-6=27

#of non-smoker who died=23-6=17

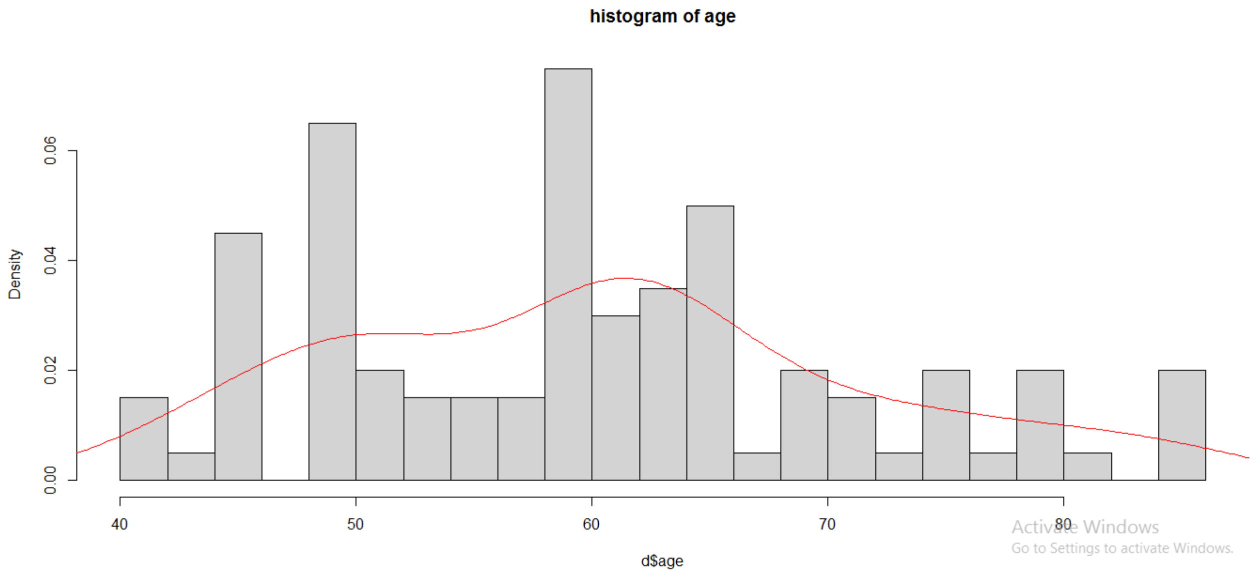
#of non-smoker who dont died=100-(6+27+17)=50



7.2 Diagrammatic representation of Continuous and discrete variables:

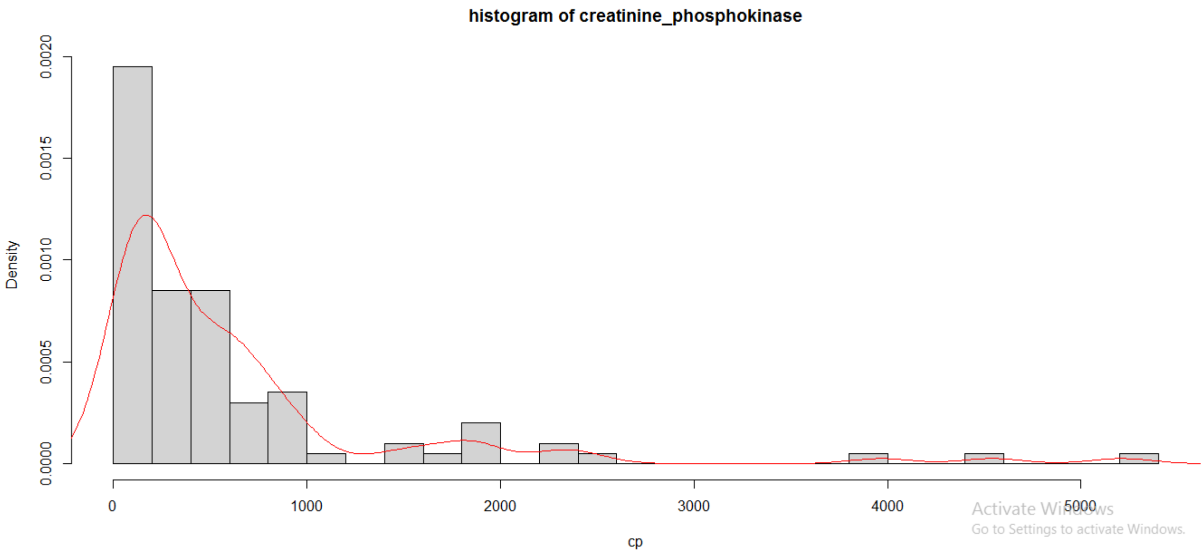
In this section we will see the frequency distributions of continuous and discrete variables through histograms and fit density curve to them. In histograms our x axis will be variables and y axis will be density=frequency/class width.

- Age:

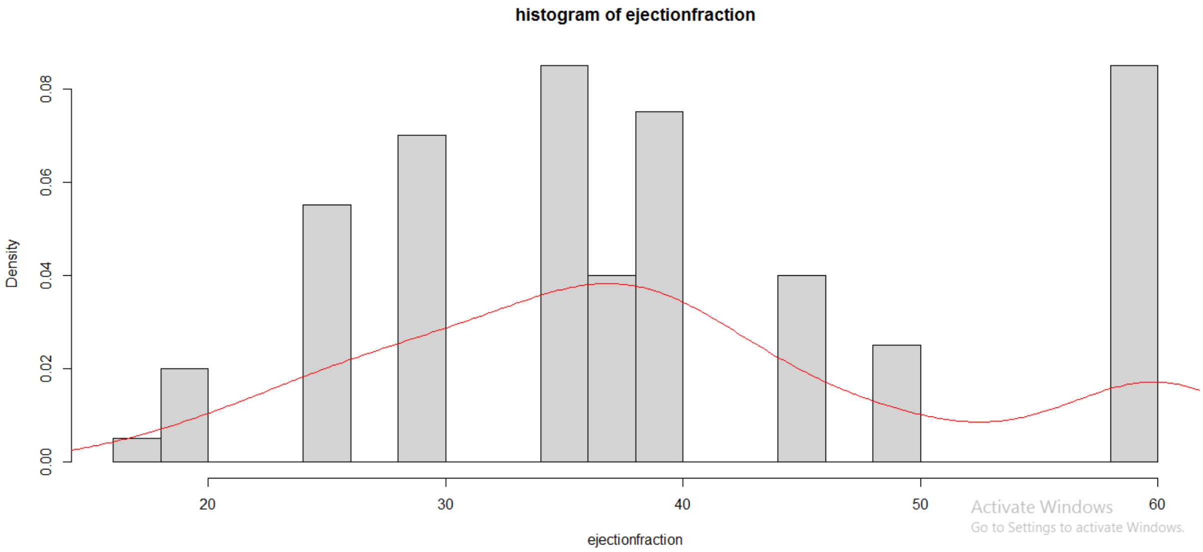


Age is in x axis and the density is in y axis.From the above diagram we can see that the modal of the distribution by histogram is 48-60 whereas the density curve shows us the mode is nearly 62. Age is almost symmetrically distributed, approximation of normalcy is highly appreciated. It's a mesokurtic distribution and the moderate sd of age supports the kurtosis comment.

creatinine_phosphokinase



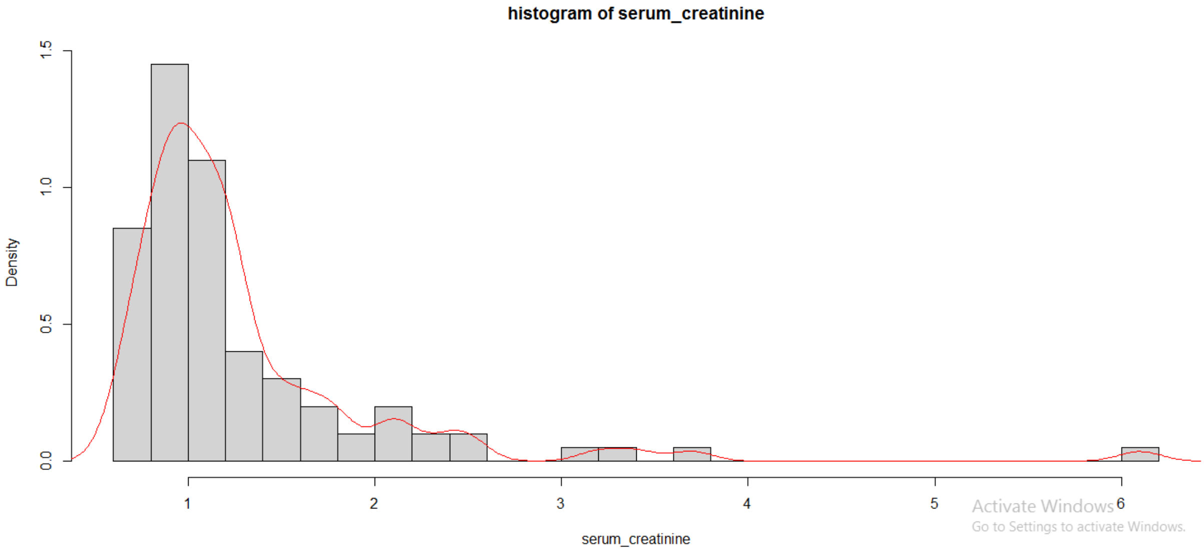
- creatinine_phosphokinase is in x axis and the density is in y axis.From the above diagram we can see that the modal of the distribution by histogram is 3-260 . creatinine_phosphokinase is highly positively skewed so chances of having higher creatinine_phosphokinase is very low
- ejection_fraction:



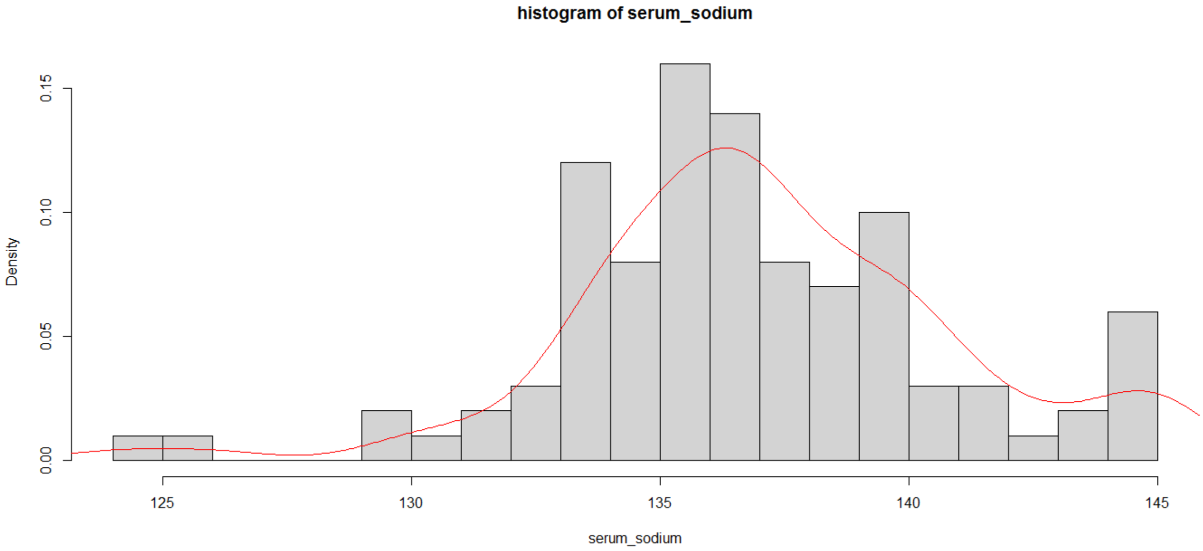
Ejection fraction is in x axis and the density is in y axis.From the above diagram we can see that the modal of the distribution by histogram is 32-34 whereas the density curve shows us the mode is nearly 35. Age is almost symmetrically distributed,. It’s almost mesokurtic distribution and the moderate sd of age supports the kurtosis comment.

- Serum_Creatinine:

  Serum_Creatinine is in x axis and the density is in y axis.From the above diagram we can see that the modal class of distribution is 0.87-1 . is Serum_Creatinine highly positively skewed so chances of having higher Serum_Creatinine is very low

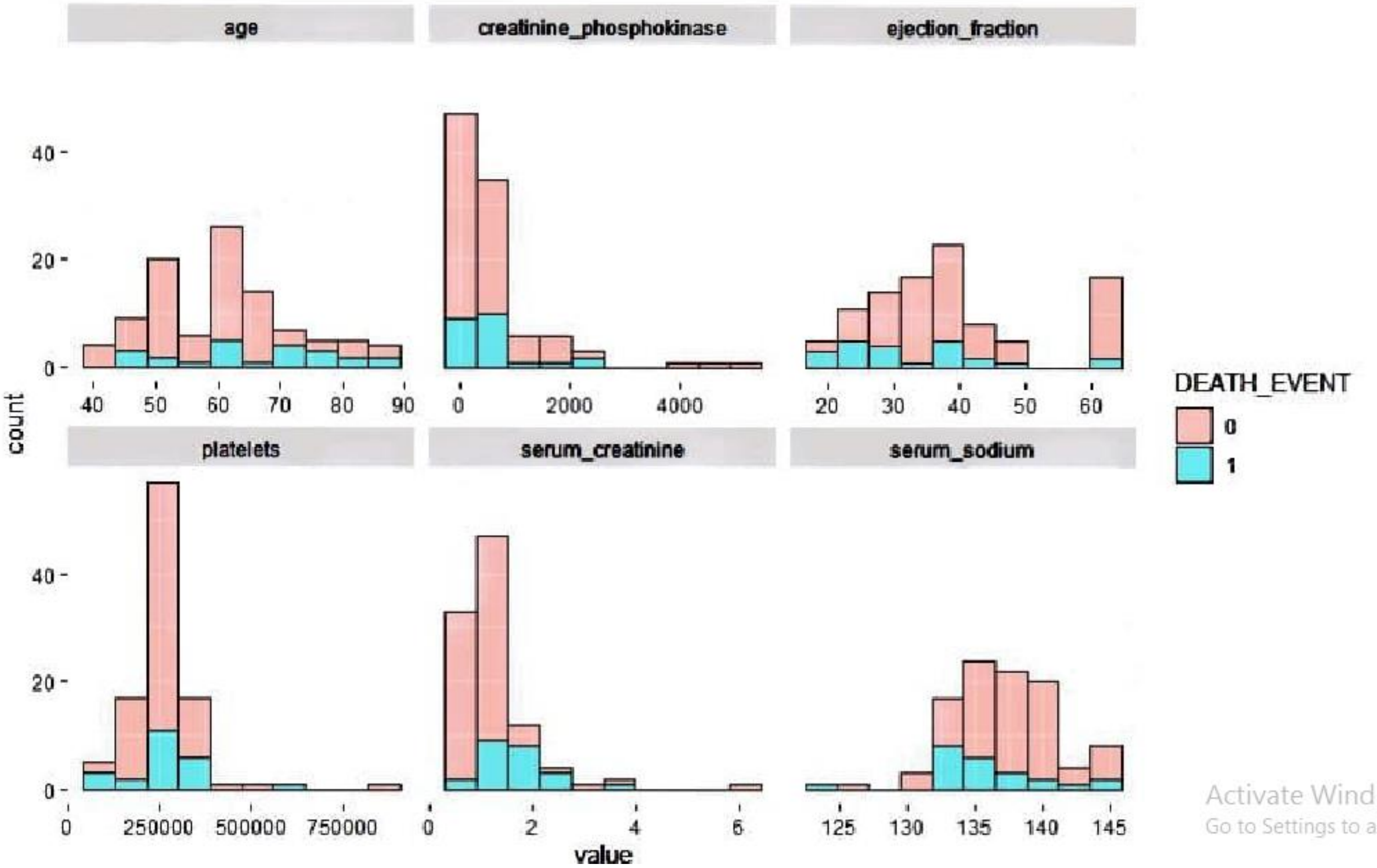


- Serum_sodium:



Serum_sodium is in x axis and the density is in y axis.From the above diagram we can see that the modal of the distribution by histogram is 135-137 whereas the density curve shows us the mode is nearly 137. Serum_sodium is almost symmetrically distributed,. It’s almost mesokurtic distribution and the moderate sd of age supports the kurtosis comment.

In the section 7.1 we have seen the proportion of certain categories are dying having cvds with pie diagramsd.In this section we will the proportion the discrete variables as well as continuous are dying for different values of them .



7.2.1 Multiple bar diagrams showing proportion of patients dying at different values for age, creatinine phosphokinase,ejection fraction,pleatelates,serum_creatinine,serum_sodium respectively

Analysis of the divided bar diagrams:

Age earlier than 60 has a low or moderate number of death events.The number of deaths get higher in a class as age increases though in 80-90 age gap shows a smaller number of deaths than 70-80.

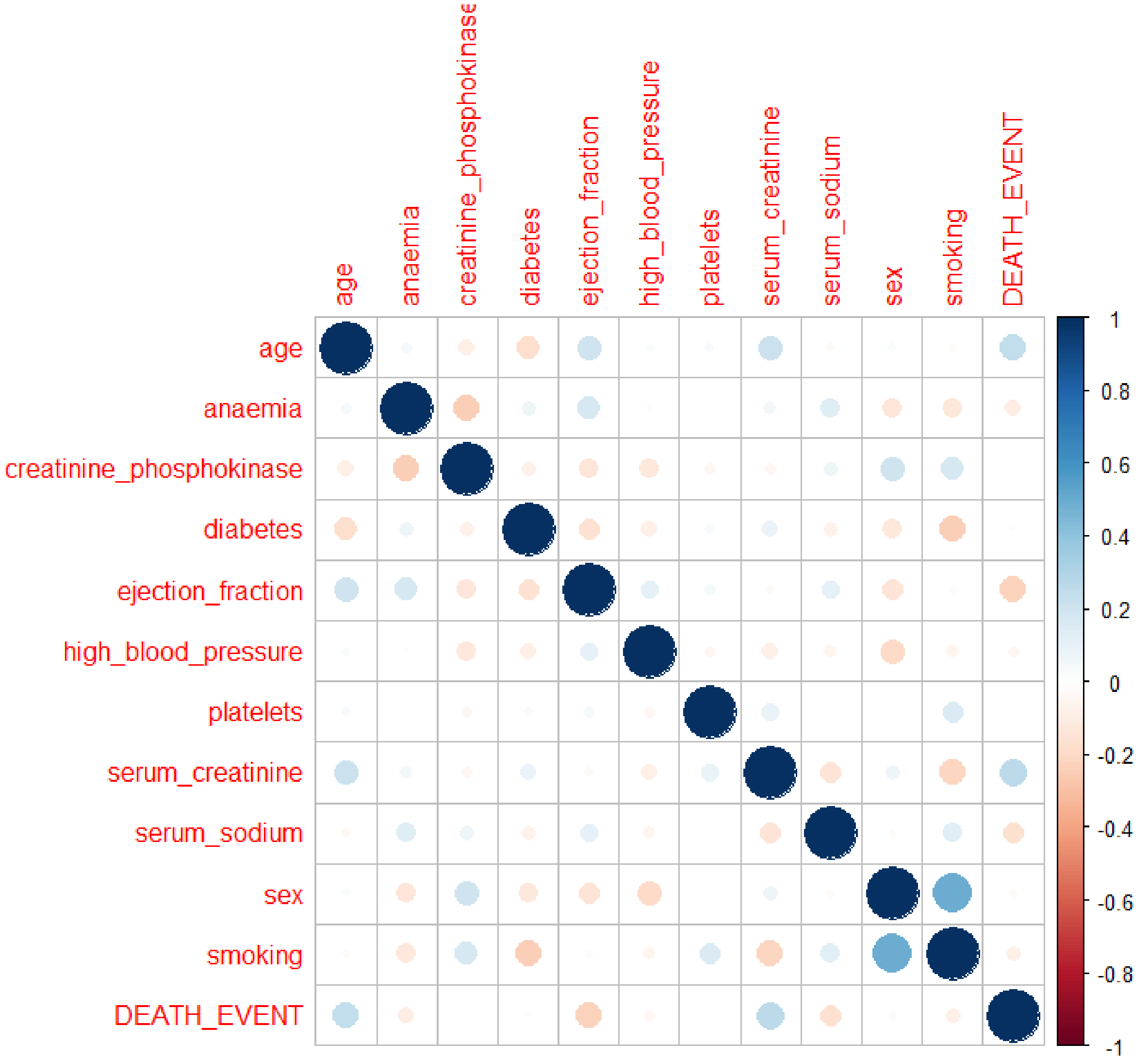
Patients having creatinine phosphokinase 0-500 mg in blood have higher number of death than other levels.

Proportion of death in patients having Ejection fraction 20-30 is quite larger than other levels.

Number of deaths are decreasing in higher classes of serum sodium.

8 Correlation between variables:

Correlation is measure of linear relationship between two variables. Here our main motto is to found the correlation matrix which describes linear association between each pair of variable. In logistic regression there's a provision that there should be **no multicollinearity**. Multicollinearity occurs when you have two or more independent variables that are highly correlated with each other. So correlation matrix plays a vital role here to check whether two explanatory variables are highly correlated or not.



8.1 Correlation matrix showing the correlation between variables

Anaemia,Ejection fraction,High blood pressure,serum sodium,smoiking all are negatively correlated with death event ;age and serum creatinine are positively corrrelated with death,creatinine phosphokinase ,diabetes,sex, and platelets are almost uncorrelated with death events.

9 Performing logistic regression:

9.1 What is Logistic regression?

Multivariable methods of statistical analysis commonly appear in general health science literature (Bagley, White, & Golomb, 2001). The terms "multivariate analysis" and "multivariable analysis" are often used interchangeably in the literature. In the strict sense, multivariate analysis refers to simultaneously predicting multiple outcomes and multivariable Analysis uses multiple variables to predict a single outcome (Katz, 1999). The multivariable methods explore a relation between two or more Predictor (independent) variables and one outcome (dependent) variable. The model describing the relationship expresses the predicted value of the outcome variable as a sum of products, each product formed by Multiplying the value and coefficient of the independent variable. The Coefficients are obtained as the best mathematical fit for the specified model. A coefficient indicates the impact of each independent variable on the outcome variable adjusting for all other independent variables.The model serves two purposes: (1) it can predict the value of the dependent variable for new values of the independent variables, and (2) it can help describe the relative contribution of each independent variable to the dependent variable, controlling for the influences of the other independent variables. The four main multivariable methods used in health science are linear regression, logistic regression, discriminant analysis, and proportional hazard regression.

The four multivariable methods have many mathematical similarities but differ in the expression and format of the outcome variable. In linear regression, the outcome variable is a continuous quantity, such as blood pressure. In logistic regression, the outcome variable is usually a binary event, such as alive versus dead, or case versus control. In discriminant analysis, the outcome variable is a category or group to which a subject belongs. For only two categories, discriminant analysis produces results similar to logistic regression. In proportional hazards regression, the outcome *variable is the duration of time to the occurrence of a binary “fail*ure” event (for example, death) during a follow-up period of observation.

The logistic regression is the most popular multivariable method used in health science (Tetrault, Sauler, Wells, & Concato, 2008). In this article logistic regression (LR) will be presented from basic concepts to interpretation. In addition, the use of LR in nursing literature will be examined by comparing the actual use of LR with published criteria for use and reporting.

- CONCEPTS RELATED TO LOGISTIC REGRESSION

Logistic regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed.

As an illustrative example, consider how coronary heart disease (CHD) can be predicted by the level of serum cholesterol. The probability of CHD increases with the serum cholesterol level. However, the relationship between CHD and serum cholesterol is nonlinear and the probability of CHD changes very little at the low or high extremes of serum cholesterol. This pattern is typical because probabilities cannot lie outside the range from 0 to 1. The relationship can be described as an ‘S’ -shaped curve. The logistic model is popular because the logistic function, on which the logistic regression model is based, provides estimates in the range 0 to 1 and an appealing S-shaped description of the combined effect of several risk factors on the risk of an event.

1. *Odds:*

Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is p, the probability of the event not occurring is (1-p). Then the corresponding odds is a value given by

odds of an event=p/1-p

Since logistic regression calculates the probability of an event occurring over the probability of an event not occurring, the impact of independent variables is usually explained in terms of odds. With logistic regression the mean of the response variable p in terms of an explanatory variable x is modeled relating p and x through the equation $p= \alpha+\beta x$. *Unfortunately, this is not a good model because extreme values of x will give values of* $\alpha+\beta x$ *that does not fall between 0 and 1. The logistic regression solution to this problem is to transform the odds using the natural logarithm (Peng, Lee & Ingersoll, 2002). With logistic regression we model the natural log odds as a linear function of the explanatory variable:*

$logit\ (y) = ln\ (odds)= ln(p/1-p) = a + \beta \chi$ *................(1)*

where p is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are α *and* β*. This is the simple logistic model.*
Taking the antilog of equation (1) on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome as
$p=P\ (Y=interested\ outcome/X= \chi,\ a\ specific\ value)= e^{a+bx}/1+e^{a+bx}$

Extending the logic of the simple logistic regression to multiple predictors, one may construct a complex logistic regression as

$logit\ (y)=ln(p/1-p) =a + \beta_1 \chi_1+ ... +\beta k\ \chi_k$
Therefore,

$p=P\ (Y=interested\ outcome/X1= \chi 1,\ ...\ Xk = \chi k)$
$=e^{a + \beta 1\ \chi 1+ ... +\beta k\ \chi k}/1+ e^{a + \beta 1\ \chi 1+ ... +\beta k\ \chi_k}$
Log odds can take values from $-- \infty$ *to* $+\infty$*. Thus we are not confining the range of the prediction.*

A better perseverance through our data :

let us regress our target variable DEATH_EVENT on age through linear regression,

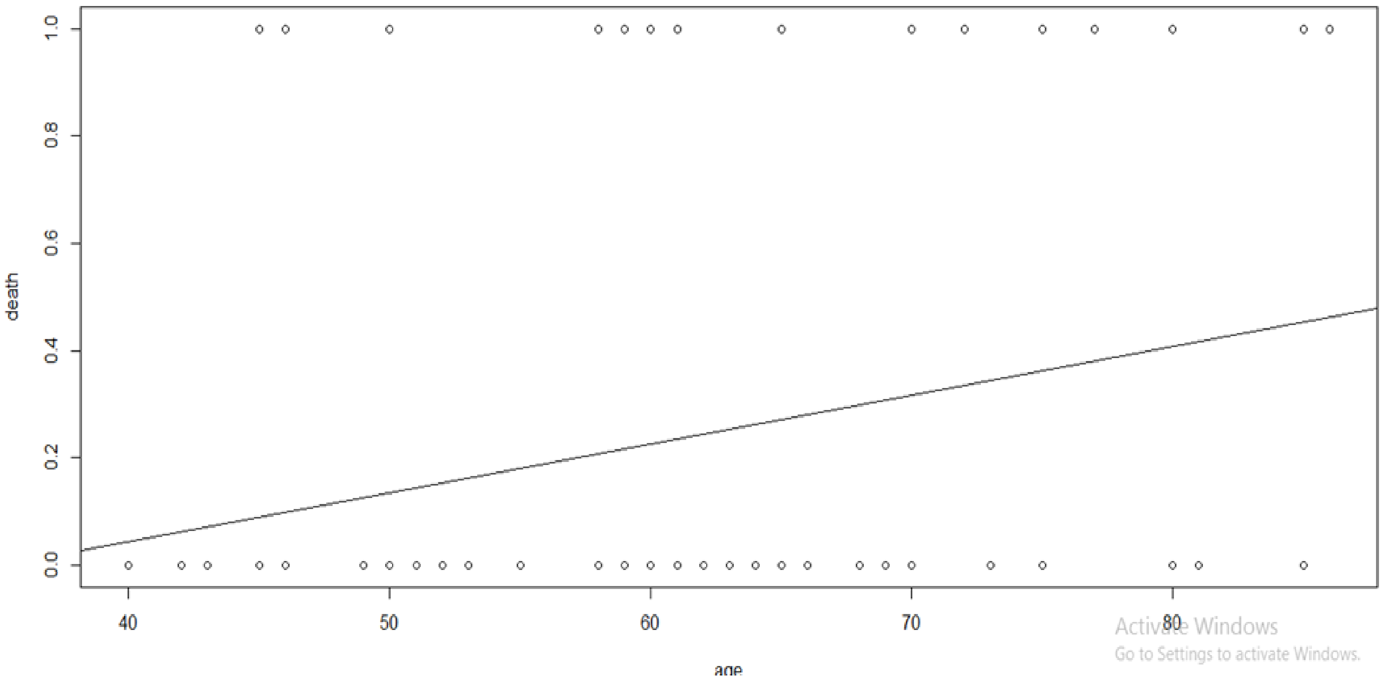
```
#code:
age=d$age
> plot(death,age)
> plot(age,death)
> lm(death~age)

Call:
lm(formula = death ~ age)

Coefficients:
(Intercept)        age
  -0.319497    0.009095
```

So, let y=DEATH_EVENT,x=age; then our linear regression equation of y on x is;

$$Y= -0.319497 + 0.009095*x$$

Let us plot it;

$$Y= -0.319497 + 0.009095*x$$



❖ *Regression equation of death on age*

So it is clearly seen that the fitted line can go further one and less than zero even can take any fractional value where the dependent variable is binary; so our fitted equation is of no use.

To get rid of this problem, at first we are finding the odds of a people will die,where the probability that a people will dies is p. After the logit transformation of the new target variable p, the range is not confined anymore ;let us regress our new variable log(p/1-p) on x;

```
#code:
>Glm(death~age,family=binomial)
Coefficients:
(Intercept)        age
   -4.40995      0.05159
```

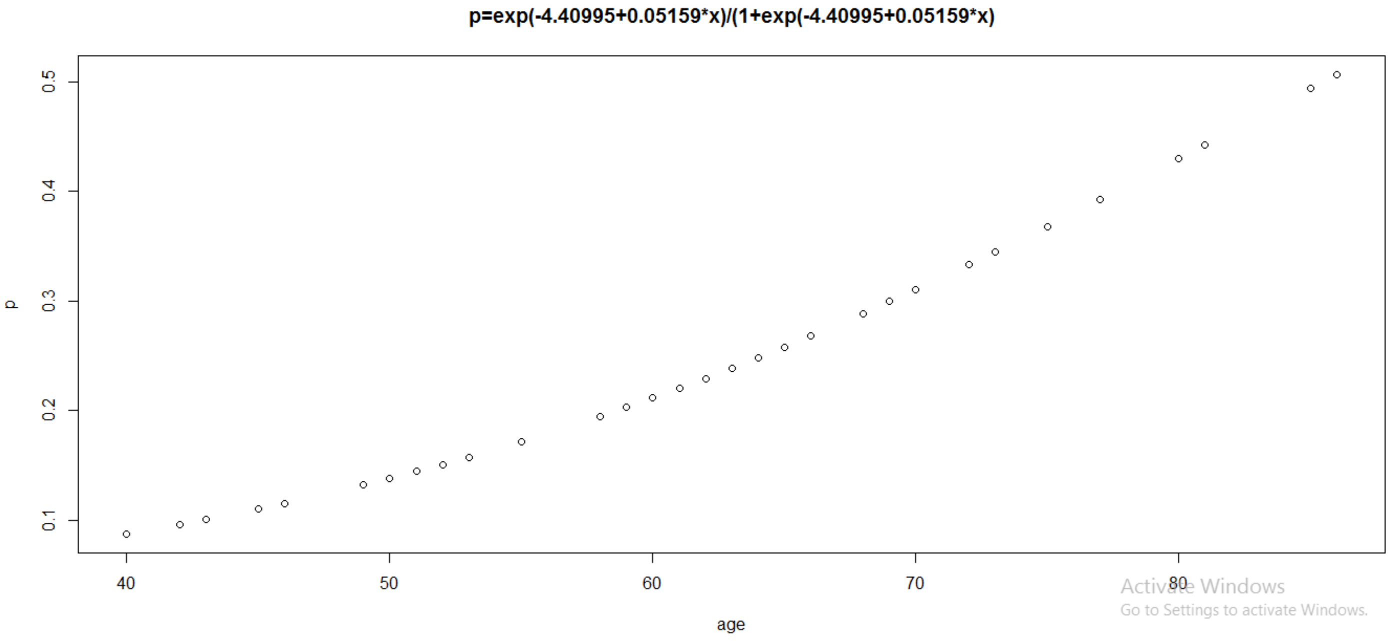
so our logistic equation looks like :

$$\log(p/1-p)= -4.40995+ 0.05159*X$$

by a suitable algebraic manipulation our equation turns into:

$$p=\exp(-4.40995+ 0.05159*X)/1+\exp(-4.40995+ 0.05159*X)$$

let us plot our data;



❖ Regression of odds on age

It is very evident from the above figure that our fitted equation is a increasing function of age which is quite consistent with the concept that probability of death gets increased with age.

9.2 Assumption in Multiple logistic regression:

One can see that the assumptions of linear regression and logistic regression are very familiar.

- Linearity: For linear regression the assumption is that the outcome variable has a linear relationship with the explanatory variables, but for logistic regression this is not possible because the outcome is binary. The assumption of linearity in logistic regression is that any explanatory variables have a linear relationship with the *logit* of the outcome variable. If the relationship between the log odds of the outcome occurring and each of the explanatory variables is not linear than our model will not be accurate.
- Independent error: Identical to linear regression, the assumption of *independent errors* states that errors should not be correlated for two observations. For example, if the whole sample of the patients having cvd can be divided into two groups, veg and non-veg where these binary variables are not included as the covariates,then characteristics of the vegetarian patients will be more likely than nonvegeterians, *though the homoscedasticity of the error is not necessary in logistic regression*
- Multicollinearity: In statistics, multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity may not reduce the predictive power or reliability of the model as a whole, it only affects calculations regarding individual predictors.

Consider a linear regression equation,

Physical strength=a+b*body weight+c*height+d*mental health ……..(2)

Now the regression coeff denotes the change in the Physical strength for unit change in the corresponding covariate keeping remaining covariates constant. But is very obvious height and weight will be highly correlated ,so while changing height we can't keep weight fixed ,so taking both variables as covariates often gives erroneous predictions. If two of the covariates are highly correlated whether positively or negatively, say 85% or above we will drop one of them. We checked that there are no such colinearity in between covariates .we are giving the values of cor between covariates

```
age       anaemia creatinine_phosphokinase     diabetes ejection_fraction high_blood_pressure
platelets serum_creatinine
age                  1.00000000  0.042934755            -0.08757407 -0.17417282        0.20569675
0.02036144  0.034803177      0.21178756
anaemia                0.04293476 1.000000000           -0.24199229  0.07257762        0.17134427
-0.01352123 -0.008281761      0.05973511
creatinine_phosphokinase -0.08757407 -0.241992289               1.00000000 -0.07835852        -
0.13319378       -0.12371878 -0.044788917      -0.04606500
diabetes              -0.17417282  0.072577622            -0.07835852  1.00000000       -0.15025344
-0.08350302  0.030037470      0.08623365
```

```
ejection_fraction      0.20569675  0.171344268          -0.13319378 -0.15025344       1.00000000
0.10002398  0.041611742     -0.03519936
high_blood_pressure          0.02036144  -0.013521232                   -0.12371878  -0.08350302
0.10002398        1.00000000 -0.040958314      -0.08249031
platelets              0.03480318 -0.008281761            -0.04478892  0.03003747       0.04161174
-0.04095831  1.000000000      0.09882428
serum_creatinine       0.21178756  0.059735112          -0.04606500  0.08623365      -0.03519936
-0.08249031  0.098824279      1.00000000
serum_sodium          -0.03373946  0.130451338           0.07787485 -0.06474957       0.10611462
-0.05234313  0.003245899     -0.14330522
sex                  0.02896042 -0.134323371           0.20519778 -0.12545517      -0.14271182       -
0.19325899  0.003717902      0.07552998
smoking               -0.02460773 -0.120893146            0.17221390 -0.24273240       -0.01971698
-0.05477142  0.156701454     -0.21691875
serum_sodium        sex    smoking
age                 -0.033739462  0.028960423 -0.02460773
anaemia               0.130451338 -0.134323371 -0.12089315
creatinine_phosphokinase  0.077874847  0.205197784  0.17221390
diabetes             -0.064749568 -0.125455166 -0.24273240
ejection_fraction      0.106114619 -0.142711820 -0.01971698
high_blood_pressure     -0.052343127 -0.193258990 -0.05477142
platelets             0.003245899  0.003717902  0.15670145
serum_creatinine       -0.143305217  0.075529982 -0.21691875
serum_sodium           1.000000000 -0.022843523  0.12712851
sex                 -0.022843523  1.000000000  0.49378751
smoking               0.127128509  0.493787515  1.00000000
```

9.3 Fitting the logistic model:

Although logistic regression model, logit (y)=$a+b_1 x_1 +b_2 x_2 +….b_n x_n$ looks similar to a multiple linear regression model, the underlying distribution is binomial and the parameters, a,b1,b2…bn cannot be estimated in the same way as for simple linear regression. Instead, the parameters are usually estimated using the method of maximum likelihood of observing the sample values.

<u>Estimating the parameters by log-likelihood</u>:

The likelihood function is used to estimate the probability of observing the data, given the unknown parameters (a,b1,b2,..bn). A "likelihood" is a probability that the observed values of the dependent variable may be predicted from the observed values of the independent variables. The likelihood varies from 0 to 1 like any other probabilities. Practically, it is easier to work with the logarithm of the likelihood function.

In logistic regression, we observe binary outcome and predictors, and *we wish to draw inferences about the probability of an event in the population. Suppose in a population from which we are sampling, each individual has the same probability p, that an event occurs. For each individual in our sample of size n, Yi =1 indicates that an event occurs for the ith subject, otherwise, Yi =0. The observed data are Y1, . . . , Yn and X1, . . . , Xn The joint probability of the data (the likelihood) is given by*

$L= \prod_{i=1}^{n} p^{yi}(1-p)^{1-yi}$

After taking log both side our equation turns into;
$LogL=\sum_{i=1}^{n} yi \ln p + \sum_{i=1}^{n}(1-yi)\ln(1-p)$

Where p= $e^{a+b1\,x1\,+b2\,x2\,+\ldots.bn\,xn}$ $/1+$ $e^{a+b1\,x1\,+b2\,x2\,+\ldots.bn\,xn}$.
We can't find the estimates explicitly. From here we will use iterative method to find our estimate starting with an arbitrary set of vector.

```
#codes:
> glm1=glm(DEATH_EVENT~.,data=d,family=binomial)
>summary(glm1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5531  -0.6769  -0.4251  -0.1909   2.2553

Coefficients:
                  Estimate         Std. Error                z value                Pr(>|z|)
(Intercept)      6.948e+00        1.064e+01                  0.653                 0.51381
age              6.475e-02        2.632e-02                  2.460                 0.01389 *
anaemia         -3.911e-01        5.778e-01                 -0.677                 0.49850
creatinine_p -5.037e-05            3.492e-04               - 0.144               0.88531
diabetes         -1.786e-01       6.072e-01                 -0.294                0.76870
ejection_fraction  -7.989e-02     3.059e-02                  -2.611               0.00902 **
high b d              -3.179e-01      6.250e-01             -0.509               0.61106
platelets        -9.244e-07       3.100e-06                 -0.298               0.76555
serum_creatinine    6.389e-01    3.476e-01                   1.838               0.06607 .
serum_sodium     -6.615e-02       7.565e-02                 -0.875                0.38185
sex              -7.636e-01       7.057e-01                  -1.082               0.27924
smoking        6.189e-02           7.287e-01                  0.085              0.93232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 107.855  on 99  degrees of freedom
Residual deviance:  86.142  on 88  degrees of freedom
```

9.4 Evaluation of a fitted logistic equation:

There are several parts involved in the evaluation of the logistic regression model. First, the overall model (relationship between all of the independent variables and dependent variable) needs to be assessed. Second, the importance of each of the independent variables needs to be assessed. Third, predictive accuracy or discriminating ability of the model needs to be evaluated.

9.4.1 ***Overall model evaluation:***

1)Likelihood ratio test: *Overall fit of a model shows how strong a relationship between all of the independent variables, taken together, and dependent variable is. One approach to testing for the significance of the coefficient of a variable in any model relates to the following question. Does the model that includes the variable in question tell us more about the outcome (or response) variable than a model that does not include that variable? This question can be answered by comparing the observed values of the response variable to those predicted by each of two models; the first with, and the second without, the variable in question. A logistic regression model with the n independent variables (the given model) is said to provide a better fit to the data if it demonstrates an improvement over the model with no independent variables (the null model). The overall fit of the model with n coefficients can be examined via a likelihood ratio test which tests the null hypothesis*

$H0 : \beta 1= \beta 2 =. . .= \beta n = 0.$

To do this, the deviance with just the intercept ***(-2 log likelihood of the null model)*** *is compared to the deviance when the n independent variables have been added* ***(-2 log likelihood of the given model).*** *Likelihood of the null model is the likelihood of obtaining the observation if the independent variables had no effect on the outcome. Likelihood of the given model is the likelihood of obtaining the observations with all independent variables incorporated in the model.*

The difference of these two yields a goodness of fit index ***G, $\chi 2$ statistic with n degrees of freedom (Bewick, Cheek, & Ball, 2005). This is a measure of how well all of the independent variables affect the outcome or dependent variable.***

G=$\chi 2$=(-2 log likelihood of given model)-(-2 log likelihood of null model)

The term ʻlikelihood ratio

test ʼ is used to describe this test. If the p-value for the overall model fit statistic is less than the conventional 0.05, then reject H0 with the conclusion that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

*In our data the null deviance=***(-2 log likelihood of given model) and the)** *and the residual deviance =***(-2 log likelihood of the given model)** *are given below;*

```
   Null deviance: 107.855  on 99  degrees of freedom
Residual deviance:  86.142  on 88  degrees of freedom
```

Now we will perform the chisquare(11) right tai test and if the p value is less than 0.05 we will reject H0 at 0.05 level of significance.

```
#codes:
>1-pchisq(107.855-86.142,11)
[1] 0.02669336
```

0.02<0.05, hence H0 is rejected at 0.05 significance level. It means at least some of the independent variables must have effect on response.

9.4.2 Statistical significance of individual regression coefficient :

If the overall model works well, the next question is how important each of the independent variables is. The logistic regression coefficient for the ith independent variable shows the change in the predicted log odds of having an outcome for one unit change in the ith independent variable, all other things being equal. That is, if the ith independent variable is changed 1 unit while all of the other predictors are held constant, log odds of outcome is expected to change bi units. There are a couple of different tests designed to assess the significance of an independent variable in logistic regression, the likelihood ratio test and the Wald statistic (Menard, 2001).

Wald statistic:

The Wald statistic can be used to assess the contribution of individual predictors or the significance of individual coefficients in a given model (Bewick et al., 2005). The Wald statistic is the ratio of the square of the estimated regression coefficient to the square of the standard error of the coefficient(squarer of the estimator of the variance of the estimated regression coefficient

The Wald statistic is asymptotically distributed as a Chi-square Distribution with df 1.

In our code z values are the square root of wald statistic and the Pr(>|z|) are the *corresponding p-values, if p value is less than 0.05 for any covariate then we say that effect of that covariate in prediction formula is insignificant.*

```
                  Estimate        Std. Error          z value          Pr(>|z|)
(Intercept)    6.948e+00       1.064e+01             0.653            0.51381
age             6.475e-02       2.632e-02             2.460            0.01389 *
anaemia        -3.911e-01       5.778e-01            -0.677            0.49850
creatinine_p -5.037e-05         3.492e-04          - 0.144            0.88531
diabetes       -1.786e-01      6.072e-01            -0.294            0.76870
ejection_fraction  -7.989e-02   3.059e-02            -2.611            0.00902 **
high b d              -3.179e-01    6.250e-01        -0.509            0.61106
platelets      -9.244e-07      3.100e-06            -0.298            0.76555
serum_creatinine   6.389e-01   3.476e-01             1.838           0.06607 .
serum_sodium    -6.615e-02     7.565e-02            -0.875            0.38185
sex            -7.636e-01      7.057e-01            -1.082            0.27924
 smoking       6.189e-02        7.287e-01             0.085          0.93232
```

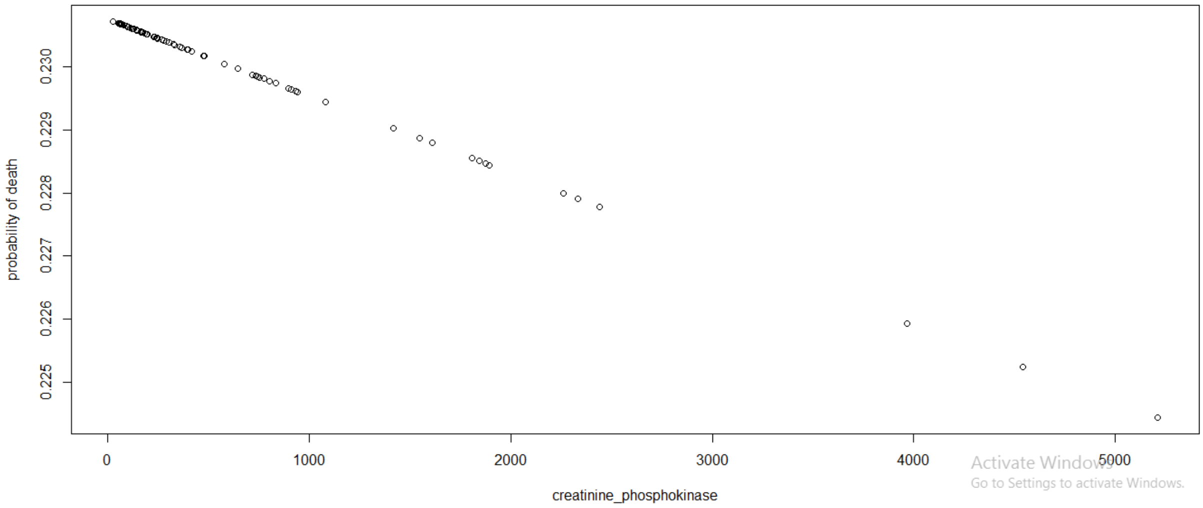
we can see that except anaemia and ejection fraction all other covariates are insignificant but it is very evident from the above result of z values that efficiency of wald statistic is highly questionable.

let us check how discrete and continuous covariates are effecting death :

- Effect of **Age** on death is previously checked and it is seen that probability of death increases with age.
- Effect of **creatinine Phosphokinase** is going to be checked through how logistic curve is fitted to it.

```
#codes:
>glm(DEATH_EVENT~creatinine_phosphokinase,data=d,family="binomial")
Call:  glm(formula = DEATH_EVENT ~ creatinine_phosphokinase, family = "binomial",
    data = d)

Coefficients:
         (Intercept)  creatinine_phosphokinase
```

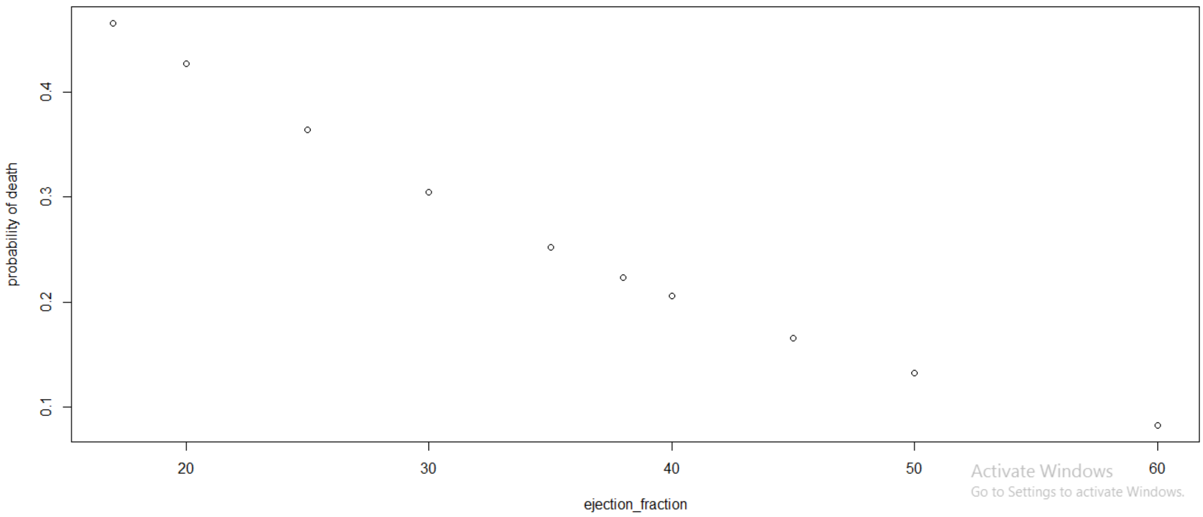
```
-1.204e+00           -6.910e-06
```



Probability of death is decreasing rapidly with increase in creatinine_Phosphokinase and in divided bar diagram we Have also seen that the number of deaths in high creatinine_Phosphokinase classes is very low so the fitted curve is consistent with diagrammatic representation but although the fitted curve is almost linear , creatinine_Phosphokinase is almost uncorrelated with death event. The main reason behind it is the binariness of DEATH_EVENT.

- Effect of **Ejection fraction** is going to be checked through how logistic curve is fitted to it:

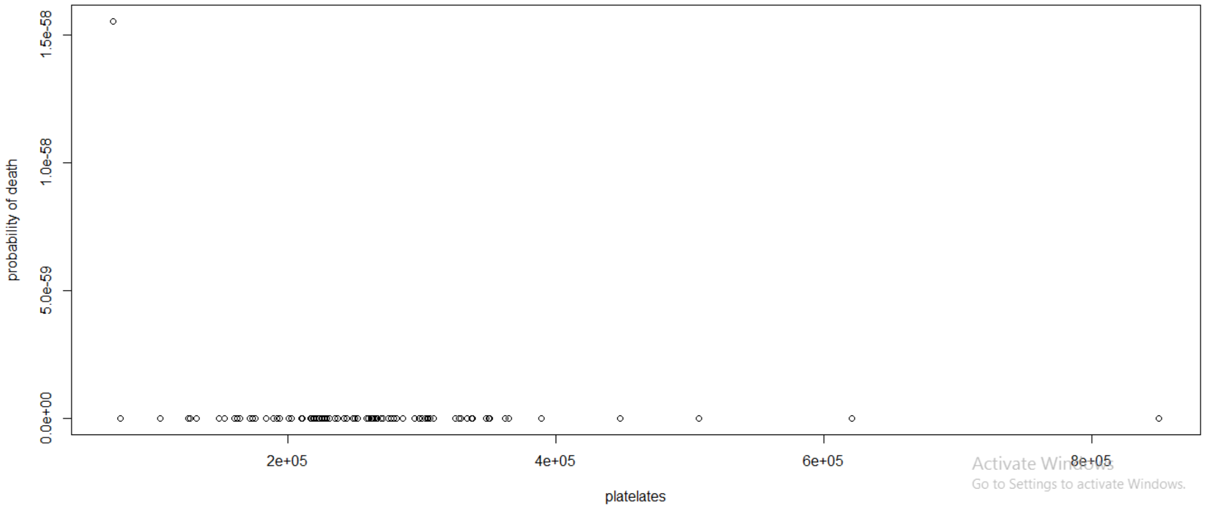
```
#codes:
ejection_fraction=d$ejection_fraction
> glm(DEATH_EVENT~ejection_fraction,data=d,family="binomial")

Call:  glm(formula = DEATH_EVENT ~ ejection_fraction, family = "binomial",
    data = d)

Coefficients:
      (Intercept)  ejection_fraction
          0.75591          -0.05266
```



Probability of death is decreasing rapidly with increase in ejection fraction but in divided bar diagram we Have also seen that the number of deaths in low ejection fraction classes and medium ejection fraction classes are almost same so the fitted curve is not fitted well with diagrammatic representation it is consistent with the perspective of negative correlation with death event.

- Effect of **platelets** is going to be checked through how logistic curve is fitted to it:

```
#codes:
> platelets=d$platelets
> glm(DEATH_EVENT~platelets,data=d,family="binomial")
Call:  glm(formula = DEATH_EVENT ~ platelets, family = "binomial", data = d)

Coefficients:
(Intercept)    platelets
 -1.159e+00   -1.885e-07
```

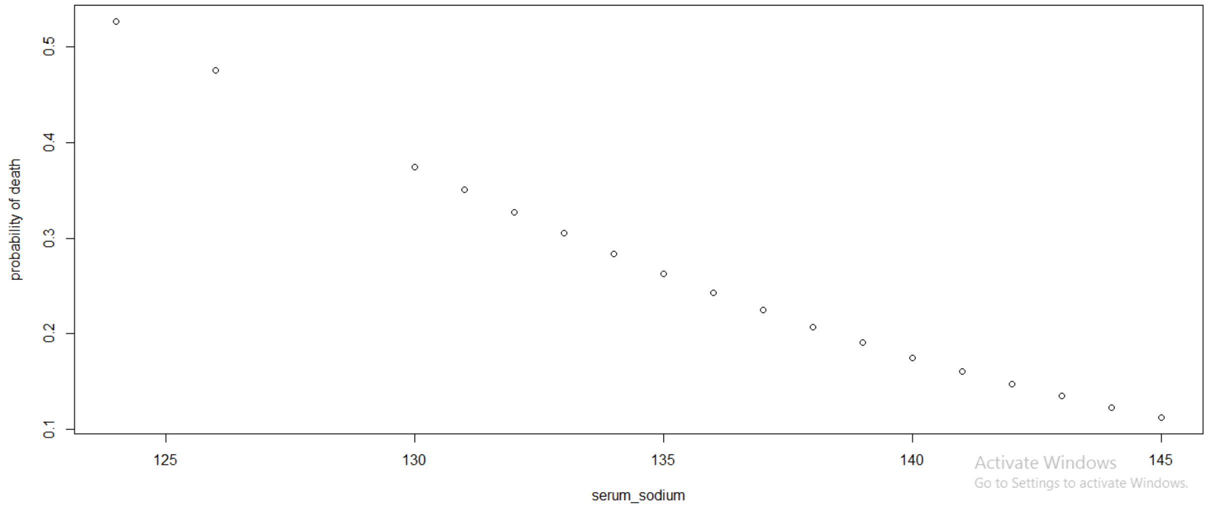


probability of death is remained almost constant with increase of platelets it is consistent with the perspective that it is actually almost uncorrelated with deaths.

- Effect of **serum sodium** is going to be checked through how logistic curve is fitted to it:

```
#codes:
> serum_sodium=d$serum_sodium
> glm(DEATH_EVENT~serum_sodium,data=d,family="binomial")
Call:  glm(formula = DEATH_EVENT ~ serum_sodium, family = "binomial",
    data = d)

Coefficients:
 (Intercept)  serum_sodium
     12.9276       -0.1034
```

Serum sodium os



Probability of death is a decreasing function of death .

9.4.3 Goodness of fit:

In linear regression, residuals can be defined as
$yi - \hat{y}i$, where yi is the observed dependent variable for the ith subject, and
$\hat{y}i$ the corresponding prediction from the model. The same concept applies
to logistic regression, where yi is equal to either 1 or 0, and the corresponding
prediction from the model is as

$$\hat{y}i = \exp(\alpha + \beta 1xi1 + \ldots + \beta kxik)/(1 + \exp(\alpha + \beta 1xi1 + \ldots + \beta kxik)$$

Chi-square test can be based on the residuals, $yi - \hat{y}i$
A standardized residual can be defined as

$$ri = yi - \hat{y}i/\sqrt{\hat{y}i}(1 - \hat{y}i)$$

where the standard deviation of the residuals is $\sqrt{\hat{y}i}(1 - \hat{y}i)$. Now our main motive is to test whether our fitted model is fitting with our expectation or not.

H0: Fitted model is correct vs H1: Fitted model is incorrect

One can then form a χ2 statistic as

$$\chi 2 = \sum_{i=1}^{n} ri^2$$

This statistic follows a χ2 distribution with n−(k+1) degrees of freedom .

```
#codes
>et=predict(glm1,d,type="response")
>diff=death-et
>diff2=(diff)^2
>diff2/(sqrt(et)*sqrt(1-et))
>um(diff2/(sqrt(et)*sqrt(1-et)))
>1-pchisq(35.05043,100-12)
```

```
[1] 0.9999999
```

It is clearly seen that the p value is greater than 0.05; so there is not enough evidence to reject H0 against H1. P.S:-but greater p value never justify to accept the H0 blindly. Basic difference between accuracy and goodness of fit is accuracy is itself a measure how good the data is fitted and on the other hand goodness of fit decides whether there is enough evidence to reject the fitted model or not.

9.4.3 Accuracy:

Here we are using the classification table to predict the accuracy of the model.The classification table is a method to evaluate the predictive accuracy of the logistic regression model . In this table the observed values for the dependent outcome and the predicted values (at a user defined cut-off value) are cross-classified. For example, if a cutoff value is 0.5, all predicted values above 0.5 can be classified as predicting an event, and all below 0.5 as not predicting the event. Then a two-bytwo table of data can be constructed with dichotomous observed outcomes, and dichotomous predicted outcomes.
The table has following form.

	predicted	
observed	*1*	*0*
1	*a*	*b*
0	*c*	*d*

If the logistic regression model has a good fit, we expect to see many counts in the a and d cells, and few in the b and c cells.
Accuracy=(a+d)/(a+b+c+d)

```
#codes:
> accuracy=mean(et==death)
[1] 0.82
```

So our model is fitted with 82% accuracy.

Codes;

For bringing data

```
>read.csv(file.choose())
>d= read.csv(file.choose())
```

For descriptive ;

```
>Summary(d)
```

For histogram we used;

```
hist(d$age,prob=T,breaks=20,main="histogram of age")
> lines(density(d$age),col="red")
hist(serum_creatinine,prob=T,breaks=20,main="histogram of serum_creatinine")
>   lines(density(serum_creatinine),col="red")
> hist(serum_creatinine,xlim=c(0.6,6.0),prob=T,breaks=20,main="histogram of serum_creatinine")
> serum_sodium=d$serum_sodium
>  hist(serum_creatinine,xlim=c(0.6,6.0),prob=T,breaks=20,main="histogram of serum_creatinine")
> hist(serum_sodium,prob=T,breaks=20,main="histogram of serum_sodium")
> lines(density(serum_sodium),col="red")
> hist(serum_creatinine,prob=T,breaks=20,main="histogram of serum_creatinine")
>   lines(density(serum_creatinine),col="red")
> hist(serum_creatinine,xlim=c(0.6,6.0),prob=T,breaks=20,main="histogram of serum_creatinine")
> serum_sodium=d$serum_sodium
>  hist(serum_creatinine,xlim=c(0.6,6.0),prob=T,breaks=20,main="histogram of serum_creatinine")
> hist(serum_sodium,prob=T,breaks=20,main="histogram of serum_sodium")
> lines(density(serum_sodium),
```

Acknowledgements:

I would sincerely like to thank *Dr.SankhaBhattachary*, Asutosh College, Kolkata for her valuable inputs during the completion of this paper. Without his valuable inputs and insights, the paper would not have made it to completion.

Krishanu Mukherjee,

Department of statistics

References

1. Logistic Regression for malignancy prediction in cancer - Luca Zammatoro, Towards Data Science, Dec 23, 2019
2. Hosmer,D.,Sturdivant,R.andLemeshow,S.,2013. Appliedlogistic regression. New York , Toronto: Wiley.
3. Wikipedia

Declaration: I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Krishanu Mukherjee