Bayesian Network for Censored Data

Priyanka Arumugam (20309046), Sneha Konoth (20301494), Vivek Kumar (20312171)

ABSTRACT:

The project 'Bayesian Network for Censored Data' aims to build a Bayesian Network on a censored dataset and perform the survival analysis based on the Bayesian Network. The Bayesian network will define the relationships between random variables and infer knowledge from the distributions. For our dataset, we have decided to perform survival analysis using Kaplan Meier model and Cox Proportional Model using Frequentist and Bayesian approach.

1. INTRODUCTION

1.1. Bayesian Network:

The Bayesian network defines the relationship between random variables.
Suppose $x_1, x_2, \ldots, x_n$ be the random variables.

Fundamental probability function is given by,

$$P(x_1, x_2, \ldots, x_n) = \Pi P(x_i \mid \text{parent of } x_i)$$

This is the probabilistic approach to Bayesian Network, that is, the conditional probability of each node based on their parent.

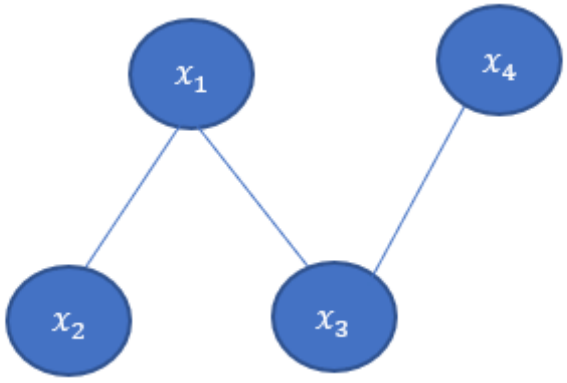


Fig. 1

In Fig.1, the state of node x3 depends on the state of its parent nodes x1 and x2.

The Bayesian network helps in identifying how we can simplify the probabilistic model that explains the state of each node in a complex system, that is, in how to solve the problem.

1.2. Survival Analysis:

Survival analysis lets you analyse the rates of occurrence of events over time, without assuming the rates are constant. Generally, survival analysis lets you model the time until an event occurs,1 or compare the time-to-event between different groups, or how time-to-event correlates with quantitative variables.

There are three important metrics in a survival analysis.

The survival probability

The survival function, is the probability an individual survives (or, the probability that the event of interest does not occur) up to and including time t. It's the probability that the event (e.g., death) hasn't occurred till a particular time. It looks like this, where T is the time of death, and Pr(T>t) is the probability that the time of death is greater than some time t. S is a probability, so $0 \leq S(t) \leq 1$, since survival times are always positive ($T \geq 0$).

$$S(t)=Pr(T>t)$$

The hazard rate

The hazard rate is the rate of immediate occurence of the event given that the patient has survived till time t.It gives the time period at which the event occurrence is highest or lowest.

$$h(t)=\lim \Delta t \rightarrow 0 P(T<t+\Delta t \mid T>t)/\Delta t$$

Cumulative hazard function

The cumulative hazard function is the integral of the hazard rate until a time t.

$$H(t)= \int t\ h(u)du$$

The cumulative hazard function can be expressed in the terms of cumulative hazard function as

$$S(t)= \exp(-H(t))$$

Censoring in Survival Analysis

Censoring is the key phenomenon of the Survival Analysis as it represents that a few subjects have not experienced the event till the study ends. This type of data is called Censored data. Censoring of data occurs due to following reasons - if the subject does not experience the event before the study ends, the subject is lost to follow-up during the study period, a subject withdraws from the study.There are three types of Censoring- Right Censoring, Left Censoring and interval censoring.When using censored data, the challenge lies in building a probabilistic

model that gives the state of the node based on the state of the parent nodes, of which some are missing or censored.

In this project, we will be performing survival analysis based on two methods.

Non- parametric (or) Kapler -Meier method

The Kapler-Meier method is a non-parametric statistical technique that estimates and approximates the true survival function of the censored data.The Kapler Meier estimates totally depend on the number of patients survived by the total number of patients who survive for a certain time after treatment.

Using a Kapler-Meier estimate for survival analysis , we need to make three assumptions
i)Survival probabilities are the same for the people who came early and for those who came late.
ii)Event occurrence is done at the specified time.
iii)Censoring data does not depend on the outcome.

Advantages

- The time to event is enough to calculate the survival function
- Gives an average brief about the event.

Disadvantages

- Lots of variables cannot be correlated.
- The model will be biased if the censored data is removed.

Semi-parametric (or) Cox Proportional Hazard model.

Cox Hazard Proportional Regression model is introduced by Cox. This technique is a very popular model used for survival analysis. It works on the hazard function from which the survival probability is calculated.

$$h(t|x) = b_0(t)\exp\{ß_i x_{ij}$$

The hazard function consists of two elements – one being the baseline hazard model i.e the hazard function when all the covariates are zero. and other is the exponential of the summation of the coefficients multiplied by the covariate parameters. These covariates are used to compare the survival of patient groups. This regression modelling technique is quite difficult than the Kapler-Meier model. However, the Cox Proportional Hazard model is far better than the Kapler-Meier model since it is more volatile with the data and features. Cox models are higher or low values and eventually decrease as the time t increases.

The table below summarises the difference between the frequentist approach and Bayesian Approach to the survival analysis.

Frequentist Approach	Bayesian Approach
The probability of the event is calculated based on the frequency of event under same repeatable conditions	Probability of the event is measured as a degree of belief.
Treats parameters as fixed	Treats parameters as random.
Calculates likelihood of the data given the parameter.	Calculates the likelihood of the parameter given the data.
Has properties of unbiasedness, minimum variance, efficient and sufficient and weakly robust.	Robust statistical models maintain stability when new samples are generated

2. DATASET

The dataset used in this project is the survival analysis of the kidney transplant patients of The Ohio State University Transplant Center during the period 1982 -1992. The follow-up time for the survival study is 9.47 years. The data used had been censored either due to the patients being moved from Columbus or if the patient was still alive on June 30, 1992.The dataset has 863 observations out of which there are 432 white males, 92 black males, 280 white females and 59 black females.Patient's age during the transplant ranges form 9.5 months to 74.5 years with a mean age of 42.8 years.Seventy-three of the white males, fourteen of the black males, thirty nine of the white females and fourteen of the black females died prior to the end of the study.

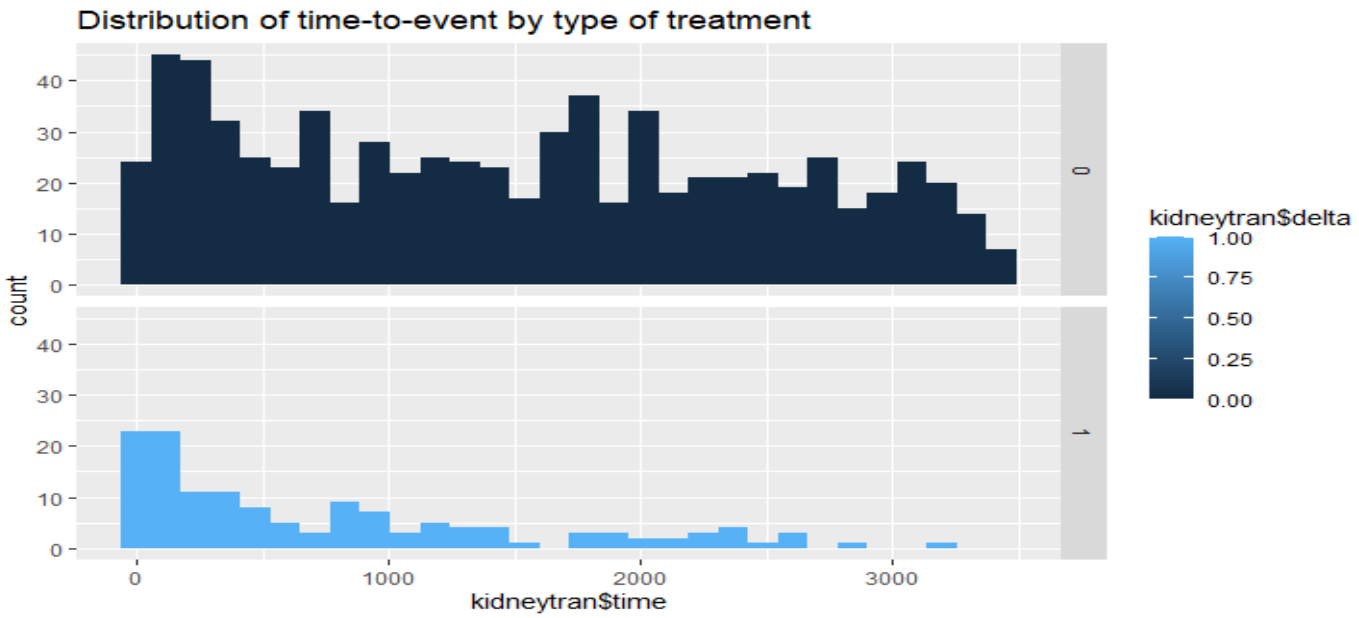
The overview of the dataset is listed below.

Column	Description	Value/ Value Type
Obs	Observation number	Number
Time	Time to death	Number

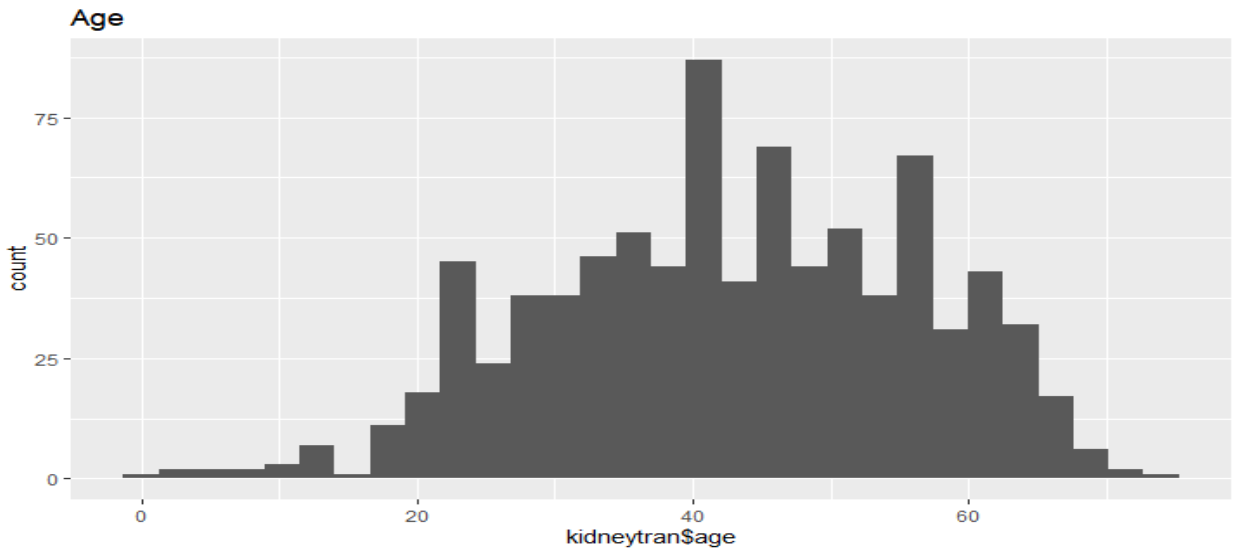
Delta	Event Indicator	Dead = 1 Alive = 0
Gender	Gender of the patient	Male = 1 Female = 2
Race	Race of the patient	White = 1 Black = 2
Age	Age in years	Number

Data Distribution of the Variables and Correlations

1. Time : As the time varies ,the number of events is decreasing slowly. However, for the “Alive” cases , rate is dropping faster if compared to the “Dead” cases.

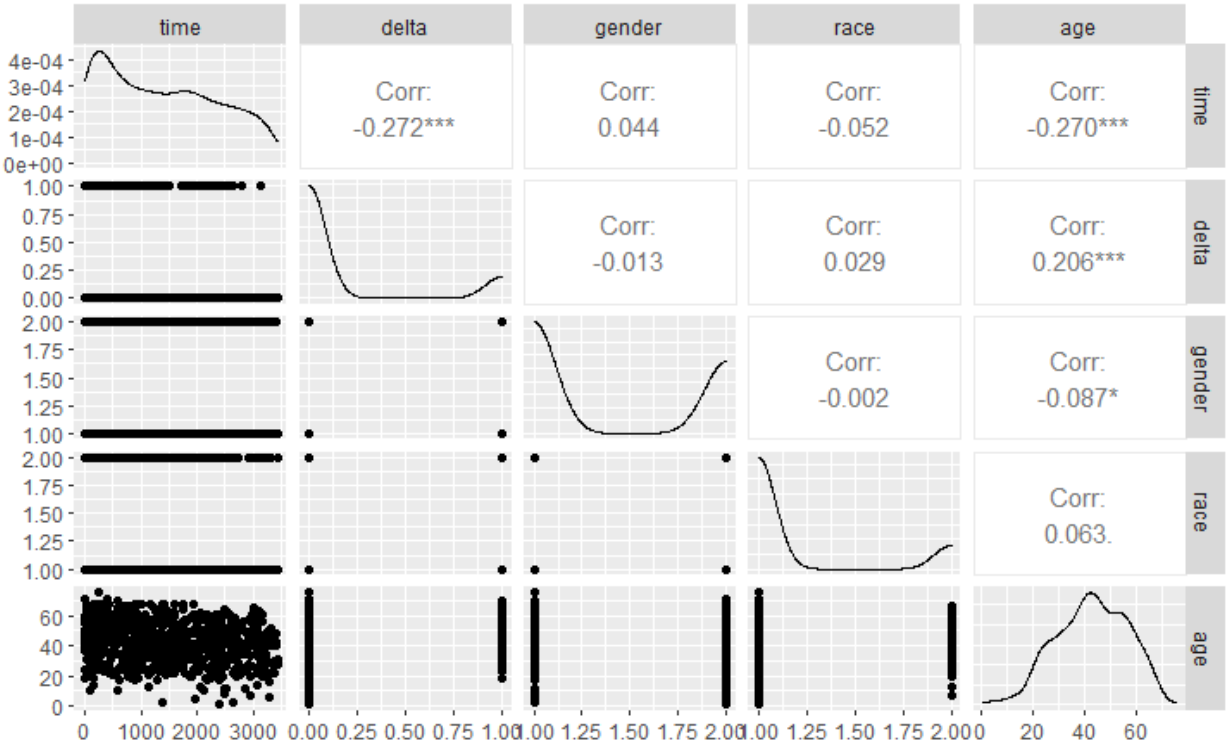


2. Age : As from the plot below , age seems following normal distribution and having max value at around 40 years and have high survivals between 40-60.



Correlation Matrix of all the variables

It is very evident from the below Correlation plot , time and age have maximum effect on Delta and are comparatively high correlation. Also , rest of the variables are quite independent.

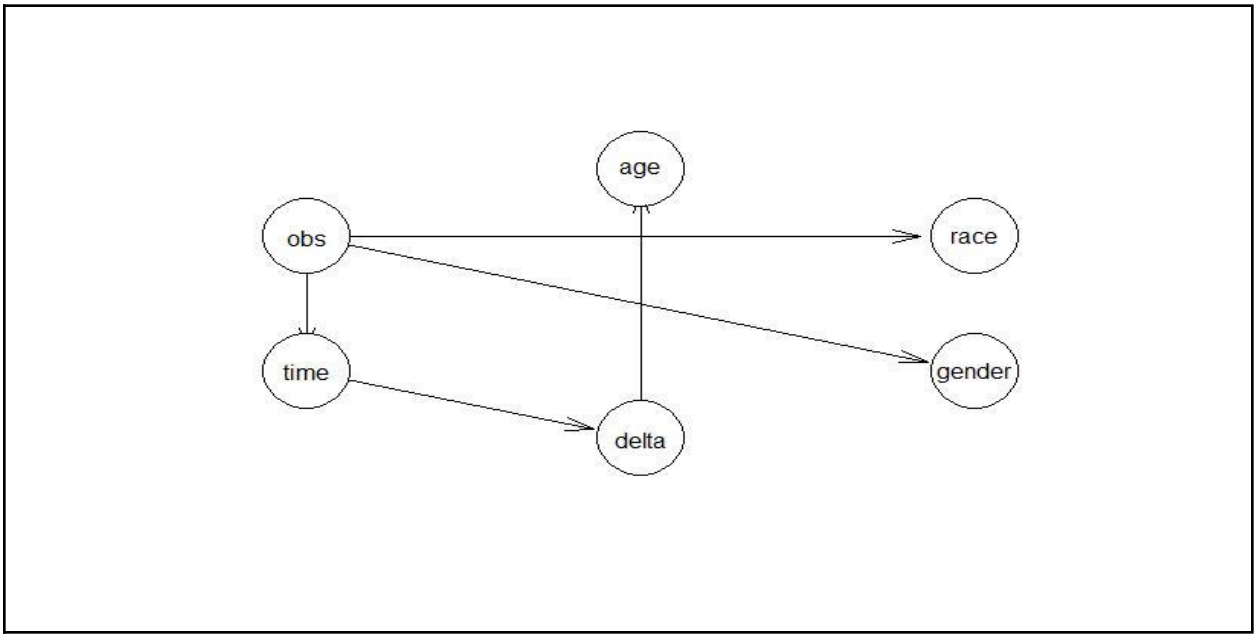


3. IMPLEMENTATION

3.1 Bayesian Network

Bayesian networks are the solid representation for knowledge and reasoning under uncertainty. They are probabilistic graphical models that encode the conditional dependence relationships between the variables via a directed acyclic graph. In the directed acyclic graph, the nodes are the variables , edges represent the conditional dependencies between the variables and if there is no edge between two variables, it represents the conditional independence between the variables.The network is used to update the probabilistic state of a child node when all the other nodes are observed.
As the dataset consists of mixed variables (categorical and continuous) , we have used different approaches to handle these to form a Bayesian Network.

For Continuous variables:
Since all the variables initially have the same datatype of ‘int’, we have changed it to numeric . Then Discretize the variables , using the 'hartemink' method.
We have developed the network using the 'Hill-Climbing algorithm. It is a kind of greedy heuristic search in which it checks each attribute pair and tries to add , remove or reverse the direction of the arc .The network that minimises the score becomes the current candidate and then the process is repeated again. Process will stop when there is no change in score on changing the arc. The model has been fit using bn.fit ,
Since the observation is just a sequence number , it doesn't have any significance in the network.



Analysis of the Coefficients: From the below summary , It can be inferred that Age and time have a comparatively higher significance on Survival . Since ‘time’ has a negative coefficient , which means as time grows , no of events (Dead) , would decrease. On the other hand , Age

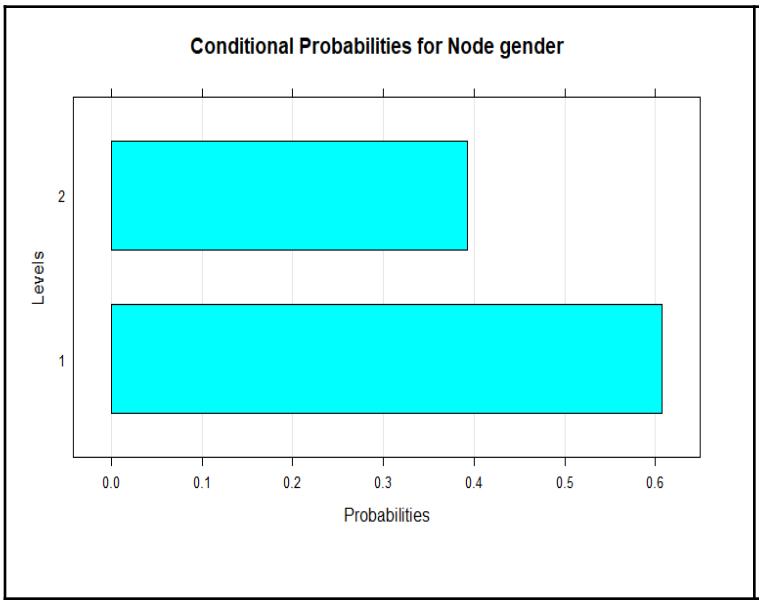
has a positive effect over no of events , and infers higher the age , higher would be the probability of getting more no of events (death).

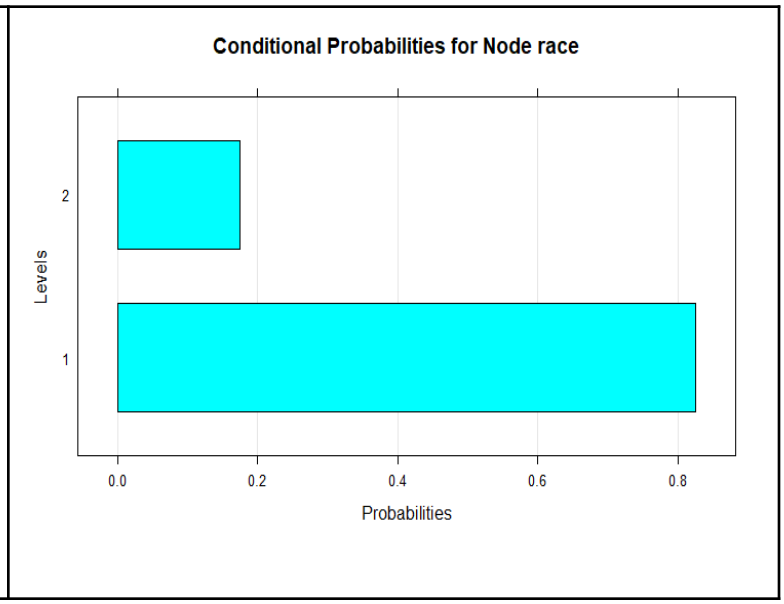
Coefficients of bayesian Network

$time
(Intercept)
1379.788
$delta
(Intercept) time
2.992284e-01 -9.929321e-05
$gender
(Intercept)
1.392816
$race
(Intercept)
1.174971
$age
(Intercept) time delta
46.249155500 -0.003091709 5.260377991

For Discrete Variables:

We have analysed the discrete variable on the basis of conditional probability. It can be suggested from below , Male has approx 60% of chances for occurrence of events over Female. Similarly , Black has below 20% probability.





3.2 *Survival Analysis*

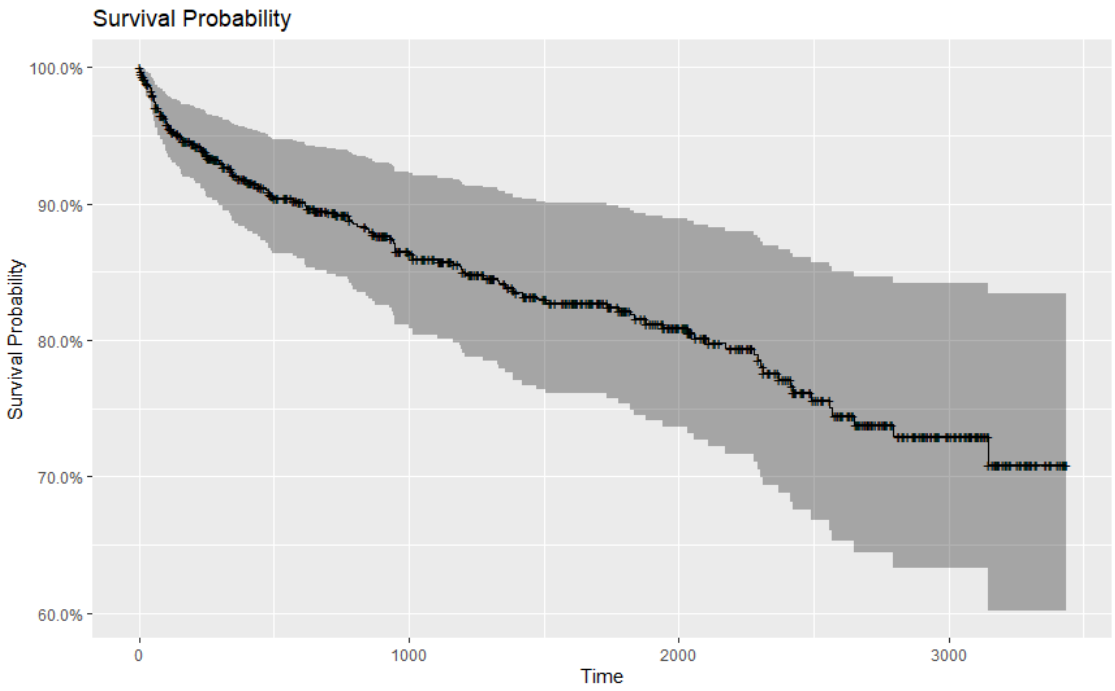
Frequentist Approach

Semiparametric Modelling or Cox Proportional Hazard Model

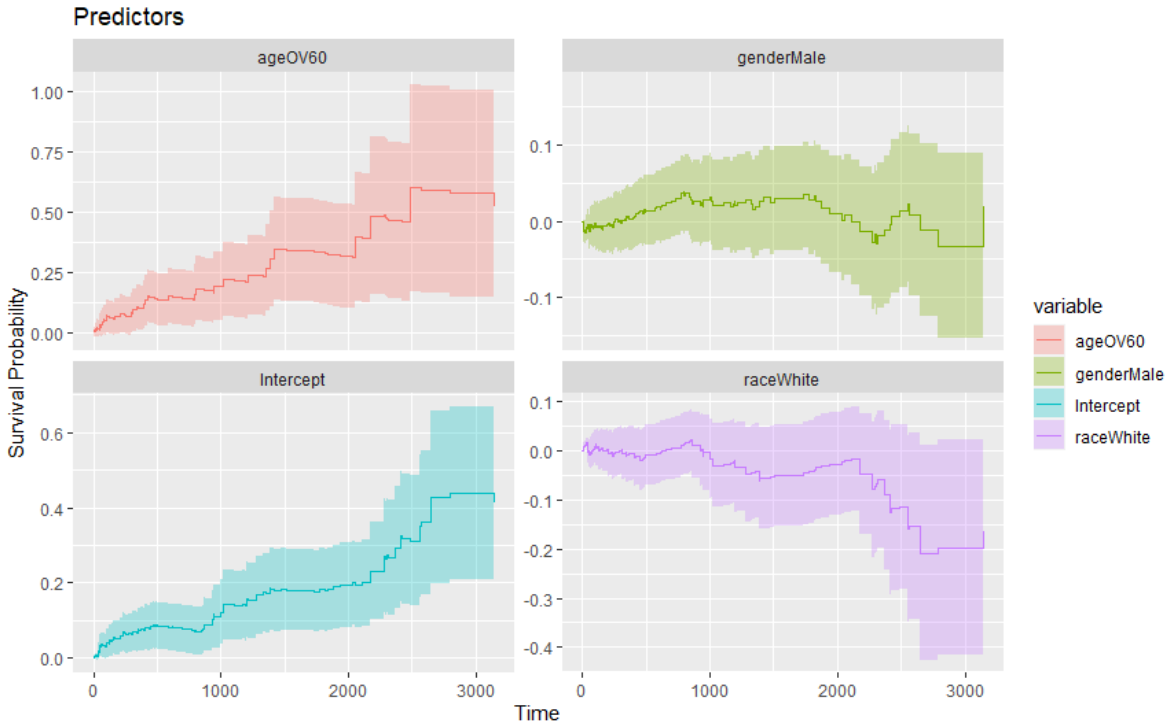
In the kidney transplant dataset, the significant covariates are age, gender and race that measures the event of death of the patient.We are generating a cox proportional hazard model using age, gender and race as the covariates.

```
Call:
coxph(formula = Surv(time, delta) ~ age + gender + race, data = kidneytran_mutated)

  n= 863, number of events= 140

                coef      exp(coef)   se(coef)      z      Pr(>|z|)
ageOV60      1.08219     2.95113     0.20869   5.186   2.15e-07 ***
genderMale   0.06214     1.06411     0.17465   0.356   0.722
raceWhite   -0.27583     0.75894     0.21208  -1.301   0.193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



              exp(coef)  exp(-coef)  lower .95 upper .95
ageOV60       2.9511       0.3389    1.9604    4.442
genderMale    1.0641       0.9398    0.7557    1.498
raceWhite     0.7589       1.3176    0.5008    1.150

Concordance= 0.589  (se = 0.024 )
Likelihood ratio test= 23.15  on 3 df,   p=4e-05
Wald test            = 28.14  on 3 df,   p=3e-06
Score (logrank) test = 30.74  on 3 df,   p=1e-06
```

The summary of the model indicates that the time-to-event depends majorly on the age. The age plays a significant role in the model than the gender and race.We can infer that the model performs better when the age of the patients is above 60.



The plot gives the survival probability across the time. As can be inferred from the plot, the survival probability decreases with time. It decreases from 100% to nearly 71%.



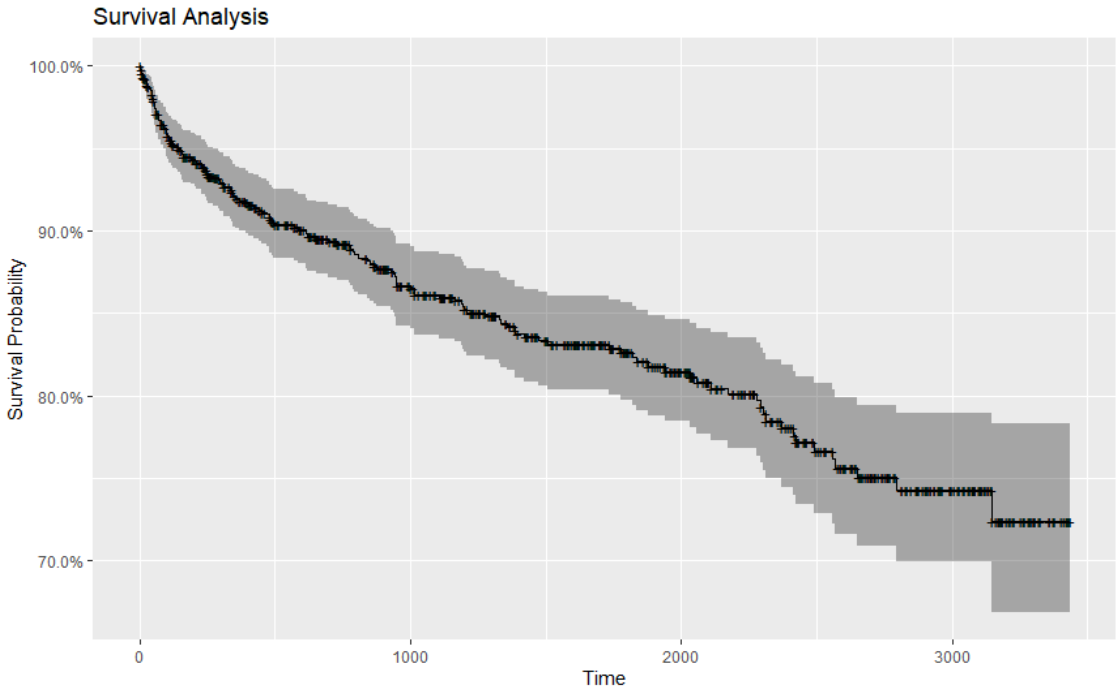
The above plot gives the survival probability of predictors. The significant parameters are plotted in this graph. The age above 60 is considered significant since the survival probability tends to zero.

Non parametric Modelling (or) Kaplan-Meier Model

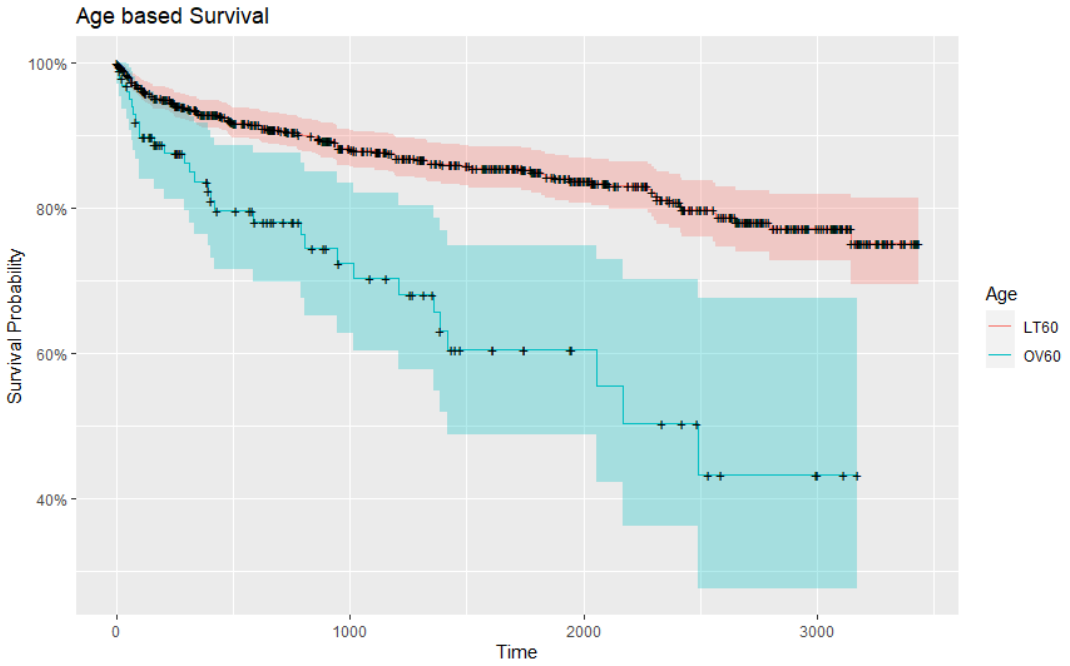
Kaplan-Meier Coefficients

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	863	0	1.000	0.00000	1.000	1.000
30	839	11	0.987	0.00385	0.980	0.995
60	816	12	0.973	0.00557	0.962	0.984
90	797	8	0.963	0.00647	0.951	0.976
180	748	15	0.945	0.00791	0.929	0.960
270	704	9	0.933	0.00870	0.916	0.950
360	668	10	0.920	0.00958	0.901	0.939
450	637	6	0.911	0.01009	0.892	0.931
540	615	5	0.904	0.01051	0.884	0.925
630	595	5	0.897	0.01095	0.875	0.918
720	566	2	0.893	0.01112	0.872	0.916
810	541	6	0.884	0.01169	0.861	0.907
900	523	4	0.877	0.01206	0.854	0.901

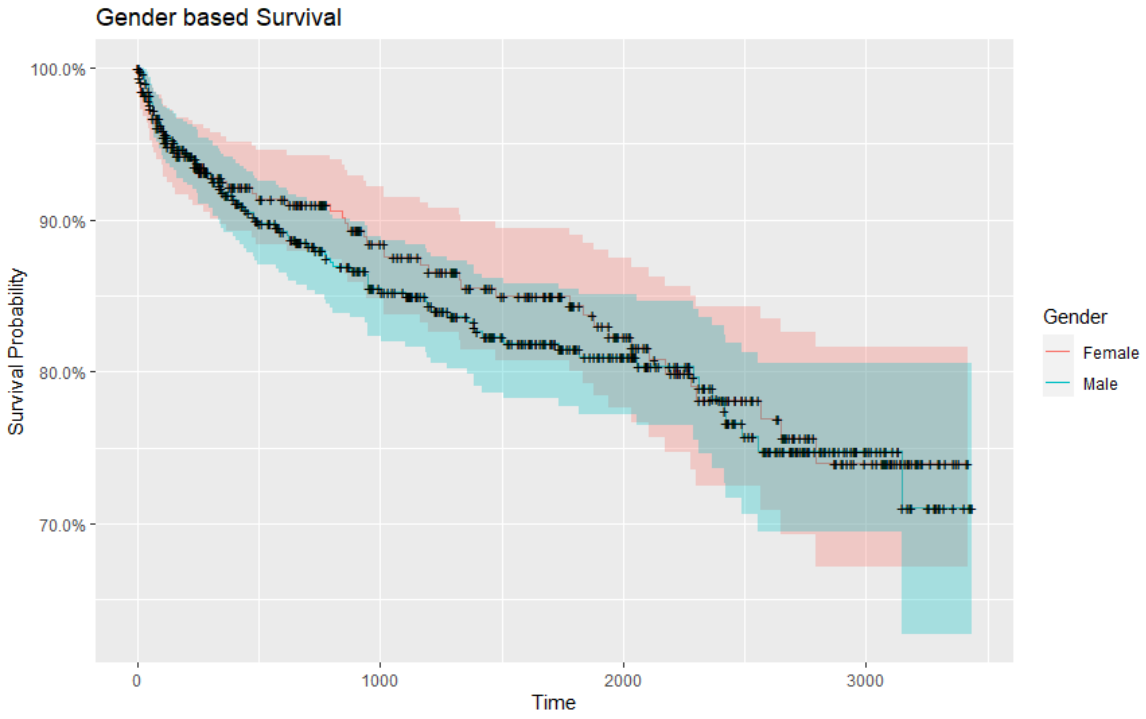
The table above gives the output of the Kaplan-Meier model. It shows the time at which the event has occurred, number of subjects after each event, cumulative survival rate or probability, standard error associated with corresponding probability and lower 95% Confidence Interval and upper 95% Confidence Interval at regular intervals of time. Initially the survival rate is at its maximum. However, as the time passes, the survival rate reduces indicating the decline in survival probability with the time.



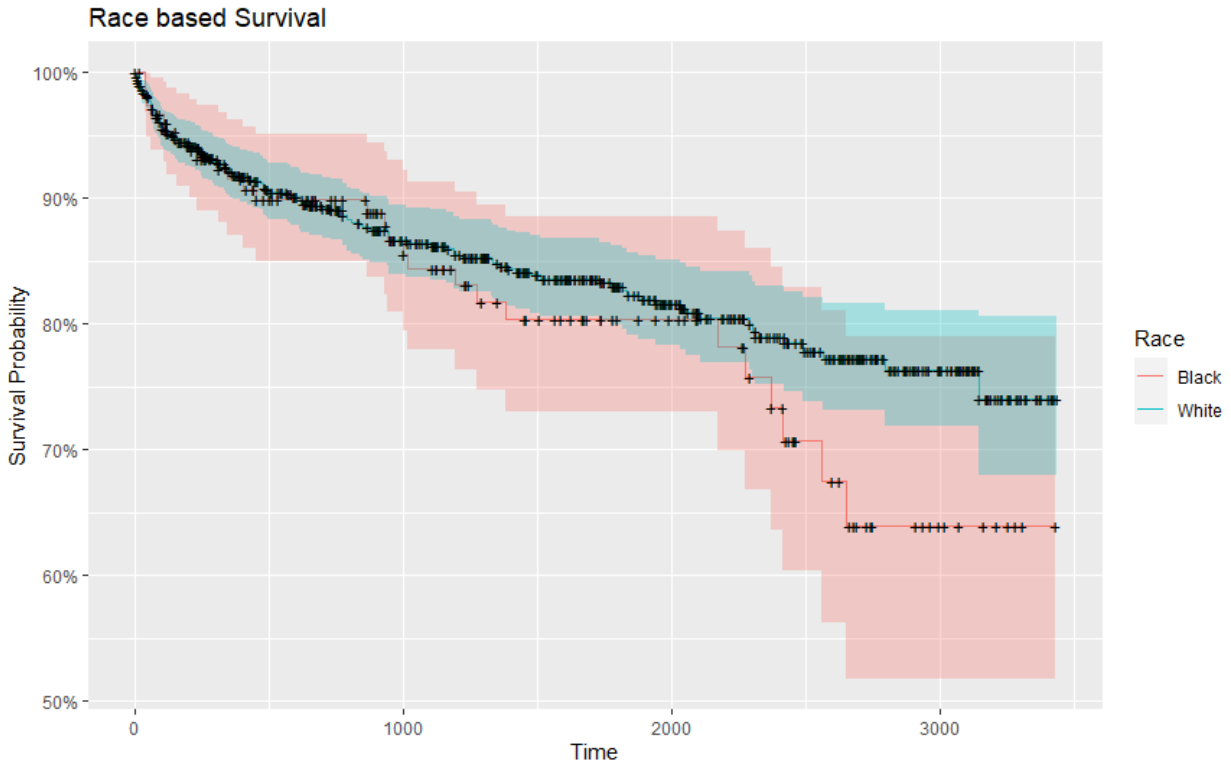
The above plot gives the survival analysis for the dataset. The plot gives the Kaplan Meier Survival curve in which the survival probability is plotted across the time. As can be seen, probability decreases with time from 100% to nearly 72.5%.



The above plot shows the survival probability based on age. It can be observed that the decrease in survival probability is more prominent in age over 60.



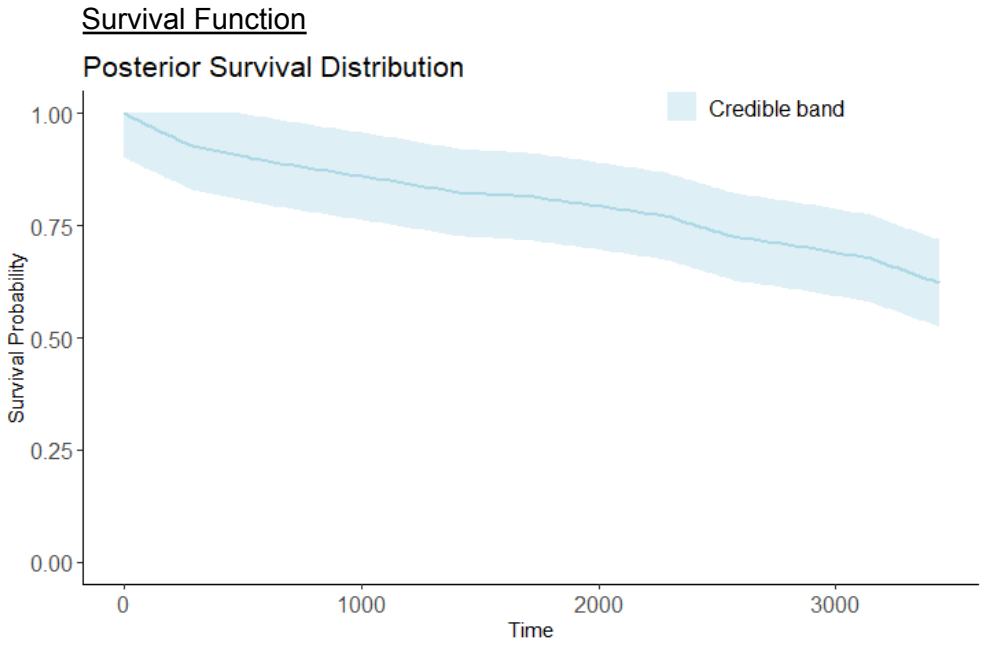
The above plot shows the survival analysis based on Gender. The decrease in survival probability is almost equal for both Male and Female. However, relatively less survival probability can be seen for male gender but the difference is not significant.

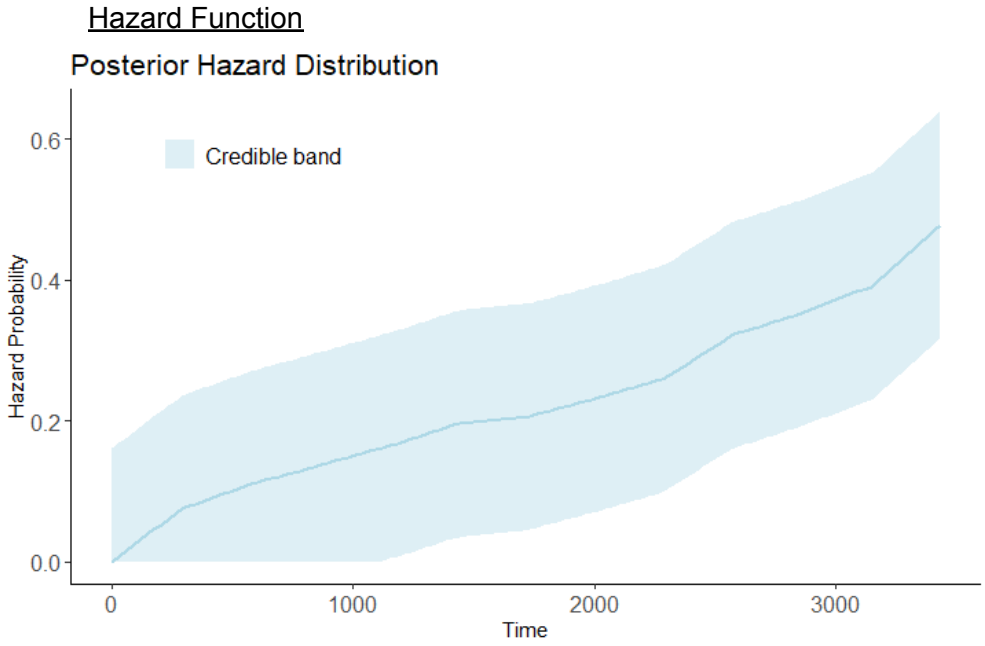


The above plot shows the survival based on the Race. The decrease in survival probability is almost equal initially. However, as time passes the black race tends to have a lesser chance of survival when compared to the white race. A difference of nearly 10% in survival probability can be observed towards the end of the plot.

Bayesian Approach

We have implemented the survival analysis with Bayesian Approach using BayesSurv package.BayesSurv function uses default priors as “Independent” , for N = 1000 with hyperparameter value alpha as 0.05. They are producing very similar results as the Kaplan Meier model and need to explore more by changing the hyperparameters to find out the best confidence interval , and check if it could produce better results. We have plotted two graphs - one graph depicts the posterior survival probability against the time and other illustrates the posterior hazard distribution





4. CONCLUSION

We are discussing about the frequentist approach and Bayesian approach to the survival analysis of kidney transplant dataset available in R.Of 863 patients listed for the first time to receive a single kidney transplant during 1982 to 1992 in Ohio State Transplant Center, the data was censored for 723 patients. We implemented the survival analysis using Cox Proportional Models and Kaplan-Meier models. Later , we created a Bayesian network for the censored data and produced posterior distributions from the network. The patients having age above 60, have a significant role in Bayesian model framework.In conclusion, we affirm that Bayesian modelling approach to predict future survival estimates will assist the healthcare researchers and providers and can be extended to the various applications such as early diagnosis, intervention planning, making statistical inference about the parameters and future disease patterns that affects the health.

5. REFERENCES

[1]I. Štajduhar and B. Dalbelo-Bašić, "Learning Bayesian networks from survival data using weighting censored instances", *Journal of Biomedical Informatics*, vol. 43, no. 4, pp. 613-622, 2010. Available: 10.1016/j.jbi.2010.03.005.

[2]J. Klein and M. Moeschberger, *Survival analysis*. New York: Springer, 2011.

[3]D. Helsel, *Statistics for censored environmental data using minitab and r*. Hoboken, N.J.: Wiley, 2013.

[4]J. Kraisangka and M. Druzdzel, "A Bayesian network interpretation of the Cox's proportional hazard model", *International Journal of Approximate Reasoning*, vol. 103, pp. 195-211, 2018. Available: 10.1016/j.ijar.2018.09.007.

[5]B. Honari, J. Donovan and E. Murphy, "Using Bayesian Networks in reliability evaluation for an -out-of-:F distributed communication system", *Journal of Statistical Planning and Inference*, vol. 139, no. 5, pp. 1756-1765, 2009. Available: 10.1016/j.jspi.2008.05.042.
[6]https://becarioprecario.bitbucket.io/inla-gitbook/ch-survival.html#non-parametric-estimation-of-the-survival-curve
[7]https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/