Survival Analysis

B Srujana, MA17BTECH11001
Aparna Ambarapu, ME17BTECH11004

Introduction:

Survival Analysis, commonly called Time to Event analysis, is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems.

Objective:

→ Analyse the performance of various models in modelling survival times of PBC dataset
→ Introducing a new model - mixture model with a custom likelihood function.

Methodology :

1. Understanding Data and Pre Processing
2. Inferences using Kaplan Meier Plot
3. Cox Proportional Hazard Model
4. Random Forests [Appendix]
5. Parametric Regression Models [Weibull, Log Logistic, Exponential, Log Normal]
6. Mixture Models

Mixture Models:

A mixture model is a collection of probability distributions or densities $D_1, ..., D_k$ and mixing weights or proportions $w_1, w_2, ..., w_k$ where k is the number of component distributions

The mixture model, $f(x) = P(x|D_1, ..., D_k) = \sum_{j-1}^{k} w_j * P(x|D_j)$

Contribution towards likelihood of different observations -

C_r → Right censored Time
C_l → Left censored Time
L → Left interval incase of interval censored observations
R → Right interval incase of interval censored observations
Y_L → Left truncated value
Y_R → Right Truncated value
X → observed time

exact lifetimes : f (x)
right-censored observations : S (C_r)
left-censored observations : 1 - S (C_l)
interval-censored observations : [S (L) - S (R)]
left-truncated observations : f (x)/ S (Y_L)
right-truncated observations : f (x)/ [1 - S (Y_R)]
interval-truncated observations : f (x)/ [S (Y_L) - S (Y_R)

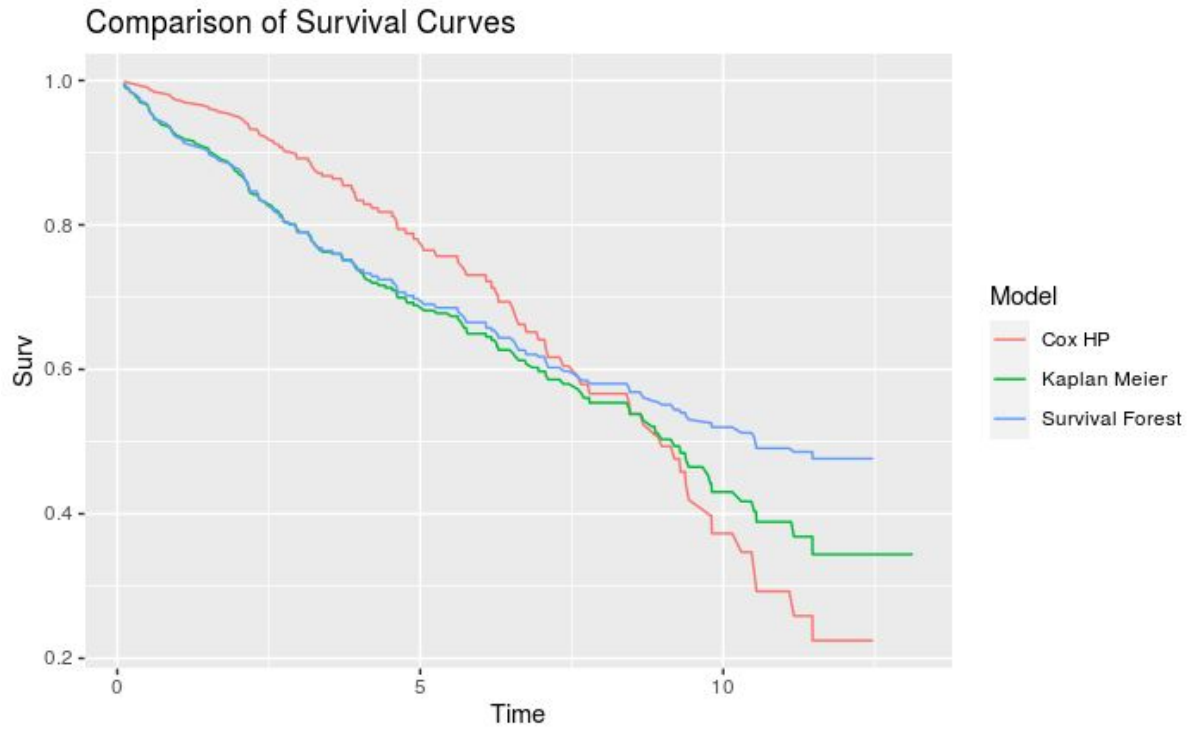
However, PBC Dataset consists of only Right Censored observations. Covariates are used to model the mean/shape parameter in the component distributions.

A custom likelihood function has been used to compute the likelihood of the model. Negative of this value is subject to minimisation using techniques from convex optimisation. (Thus, a local minima is observed). This method is used with varying numbers of mixtures(i.e from 2 to 8) and optimal no of mixtures is determined by the mixture model with least log likelihood value.

Results and Observations :

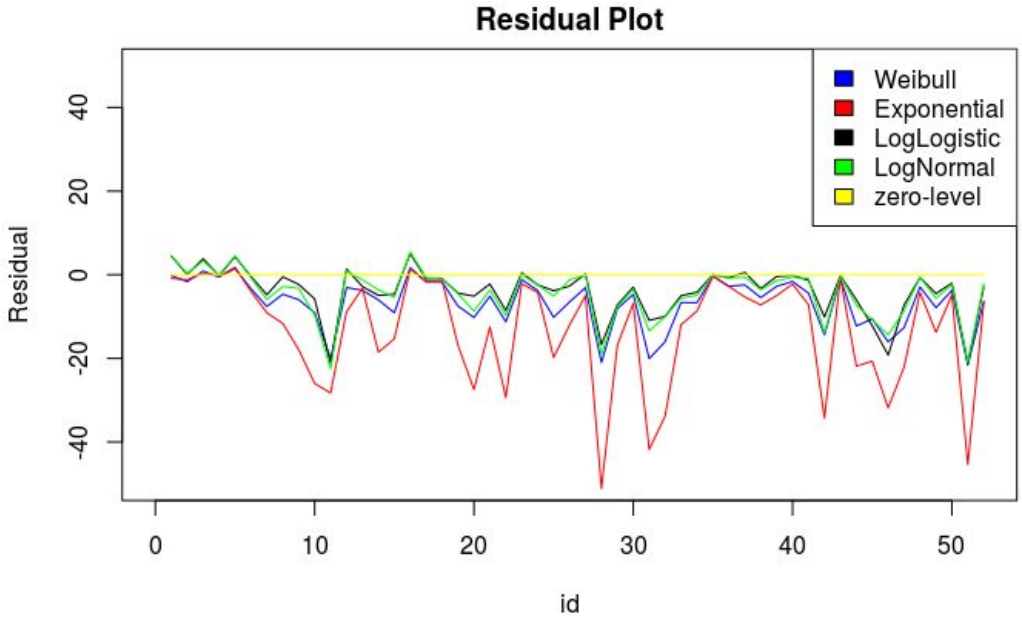
→ There is no significant evidence that the treatment succeeds in increasing the survival time.
→ Sex doesnt have an effect on survival times, however when coupled with age variable, interesting inferences are derived.

a) Survival plots:



→Random forest model estimates similar to the KM model in the beginning when compared to Cox PH model. Also as the KM plot becomes less accurate as the number of censored data increases ,we cannot conclude which model predicts the best as we move right in the graph.

b) Residual plot of the Parametric regression models:



→The parametric regression models are predicting a conservative event time as most of the residual plot is observed to lie below the zero-level.

→ The residual plot reflects the log likelihood values of the respective models. Loglogistic, Lognormal and Weibull performing similarly and the Exponential regression model being the worst of all.

Model	Residual	Log Likelihood
Log Logistic	248.68818620837	-252.5
Log Normal	270.55415342	-255.3
Weibull	353.745565367937	-255.4
Exponential	688.976054806401	-268.5

c) Log Likelihood values of the Mixture models:

Model	LogLikelihood (BestCox Covariates)	No.of mixtures	LogLikelihood (Random Forests Covariates)	No.of Mixtures
Mixture of Normals	-349.5825	5	-343.0989	7
Mixture of LogNormals	-363.2742	8	-362.9548	8
Mixture of Weibulls	-651.8244	2	-725.052	2
Mixture of LogLogistics	-683.1873	2	-765.6595	2

→Theoretically Gaussians are flexible to the input data and hence widely used as the component distribution of the Mixture models. Though the Loglogistic regression model was the best among the parametric regression models, when mixture models are concerned, mixture of gaussians perform better than any other component distributions and the same is observed here.

Conclusion:

→ Optim function used in mixture models converge to a local minimum rather than the global minimum. Thus the result obtained is subject to change depending on the initialisation of the parameters.There is also a constraint on no of mixtures considered as stated above.

However, it is noteworthy to consider that the model(Mixture of Gaussians) is performing not worse when compared with the other models.

→ As a summary, in this study, we have identified some significant covariates on survival probability of patients. We have tried understanding the survival dependencies of the covariates by fitting various models such as Kaplan Meier, Cox PH, Random forests,Parametric Regression Models. We also introduced a new model - Mixture Models having a custom likelihood function using Mixtures of various probability Distributions

References:

1. (Statistics for Biology and Health) David G. Kleinbaum, Mitchel Klein - Survival Analysis A Self Learning Text-Springer (2005)
2. John P. Klein, Melvin L. Moeschberger - Survival analysis_ Techniques for censored and truncated data-Springer (2003)
3. Kevin J. Carroll MSc - On the use and utility of the Weibull model in the analysis of survival data (2003)
4. Q. Zhang, H. Dai, B. Fu, A proportional hazards model for time-to-event data with epidemiological bias, Journal of Multivariate Analysis (2016),
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194196/#:~:text=A%20random%20forest%20is%20a,results%20of%20many%20survival%20trees.

Appendix

Mayo Clinic Primary Biliary Cirrhosis Data

Description

This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants.

A nearly identical data set found in appendix D of Fleming and Harrington; this version has fewer missing values.

Covariates

Age	:	in years
Albumin	:	serum albumin (g/dl)
Alk.phos	:	alkaline phosphatase (U/liter)
Ascites	:	presence of ascites

Ast : aspartate aminotransferase, once called SGOT (U/ml)
Bili : serum bilirubin (mg/dl)
Chol : serum cholesterol (mg/dl)
Copper : urine copper (ug/day)
Edema : 0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy
Hepato : presence of hepatomegaly or enlarged liver
Id : case number
Platelet : platelet count
Protime : standardised blood clotting time
Sex : m/f
Spiders : blood vessel malformations in the skin
Stage : histologic stage of disease (needs biopsy)
Status : status at endpoint, 0/1/2 for censored, transplant, dead
Time : number of years between registration and the earlier of death, transplantation, or study analysis in July, 1986
Trt : 1/2/NA for D-penicillamine, placebo, not randomised
Trig : triglycerides (mg/dl)

Source - T Therneau and P Grambsch (2000), *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York. ISBN: 0-387-98784-3.

Random forests

A random forest is a nonparametric machine learning strategy that can be used for building a risk prediction model in survival analysis. In survival settings, the predictor is an ensemble formed by combining the results of many survival trees. The general strategy is as follows:

Step 1. Draw B bootstrap samples.

Step 2. Grow a survival tree based on the data of each of the bootstrap samples $b = 1, \ldots, B$:

(a) At each tree node select a subset of the predictor variables.

(b) Among all binary splits defined by the predictor variables selected in **(a)**, find the best split into two subsets (the daughter nodes) according to a suitable criterion for right censored data, like the log-rank test.

(c) Repeat **(a)-(b)** recursively on each daughter node until a stopping criterion is met.

Step 3. Aggregate information from the terminal nodes (nodes with no further split) from the B survival trees to obtain a risk prediction ensemble.