SURVIVAL ANALYSIS

EAS 509 - Final Project Report

Group #15 members:
Srikar Reddy | Abhinav Reddy Chintalapuri | Akhil Niranjan Devidi | Samhitha Gaddam

Abstract:

The aim of the project is to conduct survival analysis on Mayo Clinic's Primary Biliary Cirrhosis data to understand how various features in the dataset affect the patient's survival time. The patients in the data set are suffering from Cirrhosis and they are in various stages of the disease. By using Survival Analysis techniques such as Kaplan Meier Survival Probability method, Log-Rank test, and Cox-Proportional Hazard Model, we have identified the similarities/dissimilarities in survival times between various sub-populations in the study and also the most important features and their relative hazard rates. From our analysis, we find that patients who have been treated with placebo have slightly better survival rates (14 % better) compared to the patients who are treated with D-Penicillamine.

Data Description:

The Dataset is from the Mayo Clinic's Primary Biliary Cirrhosis data. This dataset is sourced from the Vanderbilt University Department of Biostatistics. It consists of 418 observations and 19 variables. The features of the dataset include Name, Bili, Albumin, Stage, Prothrombin time, Sex, Time of death, Age, status of edema, Cholestrol, Drug treatment, Edema, etc..

Data Preprocessing and Initial Analysis:

Understanding of the patient distribution via each stage revealed a skew towards severe cases. Also, majority of the observations are from Female sex. From a total of 258 observations, only 31 are male and 227 are female. Also the death rate in Males is 68% whereas in Females it is 40%. There are a lot of patients who were censored. These patients only consented to give their basic measurements but did not consent to follow up. Right handed censoring was observed where,
n = 258 and event(death) = 111 and censored cases = 147.

Survival Analysis:

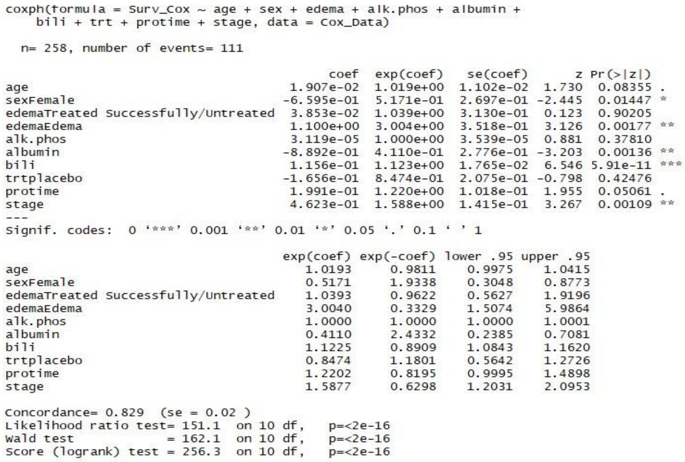
Kaplan Meier Probability Method:

From the Kaplan Meier Probability estimate process, we observed that the median survival time of the dataset is 9 years. From the plotted graphs for the survival time based on features, we found that there appears to be a significant difference between the survival time of males and females. However, there is no significant difference between the survival time of patients who have taken the treatment vs patients who were on placebo. In fact, the patients that were under the placebo have a median survival time of 9.5 years when compared to the median survival time of 8.5 years for the patients who have taken the treatment.

Log-rank Test:

We performed Log-rank Tests on various features to determine the significance of features in determining the survival rate of the patients. We found that the Edema feature (a condition where tiny blood vessels in the body (capillaries) leak fluid) significantly affects the survival time of the patients. The patients who have Edema have extremely low survival time compared to the patients who either don't have Edema or who have been successfully cured of Edema.

Cox-Proportional Hazard Test:



```
coxph(formula = Surv_Cox ~ age + sex + edema + alk.phos + albumin +
    bili + trt + protime + stage, data = Cox_Data)

  n= 258, number of events= 111 

                                     coef  exp(coef)   se(coef)      z Pr(>|z|)    
age                             1.907e-02  1.019e+00  1.102e-02  1.730  0.08355 .  
sexFemale                      -6.595e-01  5.171e-01  2.697e-01 -2.445  0.01447 *  
edemaTreated Successfully/Untreated  3.853e-02  1.039e+00  3.130e-01  0.123  0.90205    
edemaEdema                      1.100e+00  3.004e+00  3.518e-01  3.126  0.00177 ** 
alk.phos                        3.119e-05  1.000e+00  3.539e-05  0.881  0.37810    
albumin                        -8.892e-01  4.110e-01  2.776e-01 -3.203  0.00136 ** 
bili                            1.156e-01  1.123e+00  1.765e-02  6.546 5.91e-11 ***
trtplacebo                     -1.656e-01  8.474e-01  2.075e-01 -0.798  0.42476    
protime                         1.991e-01  1.220e+00  1.018e-01  1.955  0.05061 .  
stage                           4.623e-01  1.588e+00  1.415e-01  3.267  0.00109 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                    exp(coef) exp(-coef) lower .95 upper .95
age                                    1.0193     0.9811    0.9975    1.0415
sexFemale                              0.5171     1.9338    0.3048    0.8773
edemaTreated Successfully/Untreated    1.0393     0.9622    0.5627    1.9196
edemaEdema                             3.0040     0.3329    1.5074    5.9864
alk.phos                               1.0000     1.0000    1.0000    1.0001
albumin                                0.4110     2.4332    0.2385    0.7081
bili                                   1.1225     0.8909    1.0843    1.1620
trtplacebo                             0.8474     1.1801    0.5642    1.2726
protime                                1.2202     0.8195    0.9995    1.4898
stage                                  1.5877     0.6298    1.2031    2.0953

Concordance= 0.829  (se = 0.02 )
Likelihood ratio test= 151.1  on 10 df,   p=<2e-16
Wald test            = 162.1  on 10 df,   p=<2e-16
Score (logrank) test = 256.3  on 10 df,   p=<2e-16
```

Based on signs of regression coefficients in the Cox Model, we can say that a positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable.Here, **sexFemale, Albumin, trtplacebo** features have negative signs. Therefore, these features negatively affect the risk of death.

The exponentiated coefficients exp(coef), also known as *hazard ratios*, give the effect size of covariates. For example, **being female sex reduces the hazard by a factor of 0.51, or 49%.** Being female is associated with good prognostication. Also, the Cox model helped us identify which features are significant to the survival time. The star marks at the end of each row in the figure indicates the relative significance of the feature. Also, the Concordance (c-test) and other significance tests such as Likelihood Ratio Test, Wald Test, Log Rank test score tells us that the features fit the model well.

Conclusion:

After performing Kaplan-Meirer , Logrank test and cox proportional hazard , we can understand comparative feature importance relative to the event and realized that it is better off to not to take the treatment once diagnosed by the disease. Also, getting treatment for Edema would be extremely beneficial as Edema is the most detrimental feature that contributes to the death. From the Cox model, we have ranked the patients based on risk. From this ranking, the healthcare providers can focus their efforts on high-risk patients to increase their survival time.

References:

https://hbiostat.org/data/repo/Cpbc.html