Prediction of Recovery Time from Liver Cirrhosis

Department of Statistics
The University of Burdwan

Arjun Samanta

Acknowledgment

I owe a great many thanks to great many people who helped me and support me during completion of this project.I express my gratitude to my department of Statistics for the extended support . My deepest thanks to the Faculty members' for guiding and correcting my work with attention and cane.My project guide Prof. Dr.Arindam Gupta for his innovative ideas and valuable guidance.I would like to thanks department staff for extending their support. I thank my fellow mates who took pain to go through the project and made necessary corrections wherever needed.

Contents

List of Figures

List of Tables

Abstract

Cirrhosis is a late stage of scarring(fibrosis) of the liver caused by many forms of liver diseases and conditions, such as hepatitis and chronic alcoholism.Now I am working for predicting the possible survival time to end the study with either incurred with censor or death for a randomized placebo control trial of the drug D-penicillamine The following data contains the information collected from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984

Chapter 1

Introduction

Survival Analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs.*

By **time**,we mean years, months,weeks,or days from the beginning of follow-up of an individual until an event occurs;alternatively,time can refer to the **age** of an individual when am event occurs.
By **event**,we mean death,disease incidence,relapse from remission, recovery (e.g.,return to work) or any designated experience of interest that may to an individual.
Although more than one event may be considered in the same analysis, we will assume that only one event is of designated interest. When more than one event is considered (e.g., death from any of several causes), the statistical problem can be characterized as either a recurrent event or a **competing risk** problem.
In a survival analysis, we usually refer to the time variable as survival time, because it gives the time that an individual has "survived" over some follow-up period. We also typically refer to the event as a failure, because the event of interest usually is death, disease incidence, or some other negative individual experience. However, survival time may be "time to return to work after an elective surgical procedure," in which case failure is a positive event.Due to presence of censoring,which is data whose event is not occurred yet,survival analysis model require special consideration.

Cox proportional hazard(Cox PH) and accelerated failure time model(AFT) are widely used to handle right censored data.Yet the

assumptions made by these model are violated in the real worlds .Recent studies showed that the Ordinary Differential Equation(ODE) modeling framework unifies many existing Survival analysis models including Cox PH and AFT. They also showed that the ODE modeling framework is flexible and widely applicable.
However, naively applying the ODE framework to survival analysis problems may result in wildly oscillating density function that may worsen the model's performance. Regularization techniques that can regularize this undesirable behavior are understudied.The cluster assumption from semi-supervised learning states that the decision boundaries should not cross high-density regions. Likewise, survival analysis models need hazard functions that slowly change in high-density regions.
In this paper ,we propose Cox Proportional Hazard Model to predict exact Survival Time where the individuals are either incurred with death or censored.Our method has several advantages 1)The model is computationally efficient.2)The model is theoretically sound.3)It is easy to implement.4)The model is applicable to any Survival analysis problem containing censored data.

Chapter 2

Methods and Materials

2.1 Study Design

The Survival analysis was performed using Kaplan-Meier and Cox Proportional Hazard methods. Kaplan-Meier model was used to determine the survival probability of Liver Cirrhosis patients.Then,Log-Rank test was used to determine the significance difference between survival expression of patient.Cox proportional hazard was used to determine the difference ratio of prognostic factors which were including age,stages,Drug,Sex,Ascites and other factors like Bilirubin,Edima,Albumin,Copper,SGOT etc.

2.2 The Survival Function

Individual opportunities to survive for time x are expressed by $S(x) = p(X > x)$. Let X be the continuous random variables, then the survival function is the complement of the cumulative distribution function $S(x) = 1 - F(x)$ where $F(X) = P(X \leq x)$. The survival function is the integral of the probability density function $f(x)$:

$$S(x) = P(X > x) = \int_x^{\infty} f(t)dt$$

$$f(x) = -\frac{dS(x)}{d(x)}$$

Then if X is the discrete random variables, and can be obtained x_j, $j = 1, 2, 3, \ldots$ with the probability mass function (p.m.f) $p(x_j) =$

$P(X = x_j)$, $j = 1, 2, 3, \ldots$ where $x_1 < x_2 < x_3 \ldots$ then the survival function for the discrete variables X is given by:

$$S(x) = P(X > x) = \sum_{x_j > x} p(x_j)$$

2.3 Hazard Functions

The hazard function of the hold time X is denoted by h(x) and defined as individual probability fails in the time interval $(x, x + \Delta x)$ that the individual has lived for time x, the hazard function is expressed as:

$$h(x) = \mathbf{lim}_{x \to \Delta x} \left[\frac{P(x < X < x + \Delta x | X > x)}{\Delta x} \right]$$

The relationship between the hazard function and survival function is expressed by :

$$h(x) = \frac{f(x)}{S(x)}$$

2.4 Kaplan Meier's Survival curve and Log Rank Test:

Estimated Kaplan Meier survival function.expressed by:

$$\hat{S}(x_(j)) = \hat{S}(x_(j-1)\hat{P}[X > x_(j) | X \geq x_(j)]$$

A further log rank test is used to compare Kaplan Meier's survival curves formed by the following hypothesis: H_0: There is no difference between the survival curves. H_1: At least one difference between the survival curves:

$$LogRankStatistic = \frac{(O_i - E_i)^2}{Var(O_i - E_i}$$

m_ij denotes the number of individuals who experience the event at time x_j, and eij is the value of hope. The null hypothesis will be rejected if log rank statistics $\geq \chi_(a, df)$ α,df with degrees of freedom (df) = 1 or p value $< \alpha$

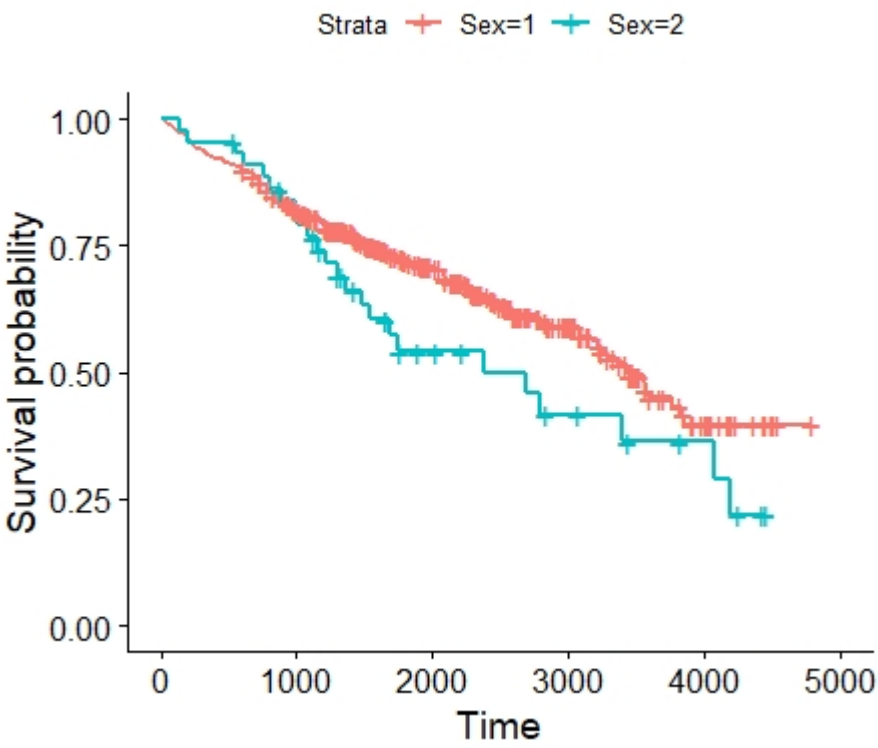


Figure 2.1: Kaplan Meier Curv

2.5 Cox Proportional Hazard Model:

The relationship between the hazard rate and the covariate set can be ecpressed using the model:

$$\ln[h(t) = \ln[h_0(t)] + \sum_{i=1}^{n} x_i \beta_i$$

or

$$h(t) = h_o(t) \exp^{\sum_{i=1}^{n} x_i \beta_i}$$

where $x_1, x_2, ..., x_n$ are covariates.$\beta_1, \beta_2, ..., \beta_n$ are the regression coefficient to be estimated .and$h_0(t)$ is baseline hazard rate when all the covariates are zero.

Chapter 3

Results and Discussion

3.1 Kaplan Meier Analysis

time	n.risk	n.event	survival	std. error	lower 95 CI	Upper 95 CI
533	382	1	0.99738	0.00261	0.992271	1.00
691	370	1	0.99203	0.00458	0.983096	1.0000
837	350	1	0.98101	0.00711	0.967167	0.9950
839	349	1	0.97820	0.00763	0.963362	0.9933
994	330	1	0.95807	0.01061	0.937507	0.9791
1022	326	1	0.95513	0.01097	0.933865	0.9769
1030	325	1	0.95219	.01133	0.930251	0.9747
1055	323	1	0.94925	0.01167	0.926649	0.9724
1234	290	1	0.88920	0.01698	0.856529	0.9231
1236	288	1	0.88611	0.01720	0.853030	0.9205
1250	287	1	0.88303	0.01742	0.849539	0.9178
1260	286	1	0.87994	0.01763	0.846057	0.9152
2301	143	1	0.52903	0.02882	0.475452	0.5886
2318	142	1	0.52531	0.02886	0.471681	0.5850
2330	141	1	0.52158	0.02889	0.467915	0.5814

Table 3.1: Summary table of Kaplan Meier fit

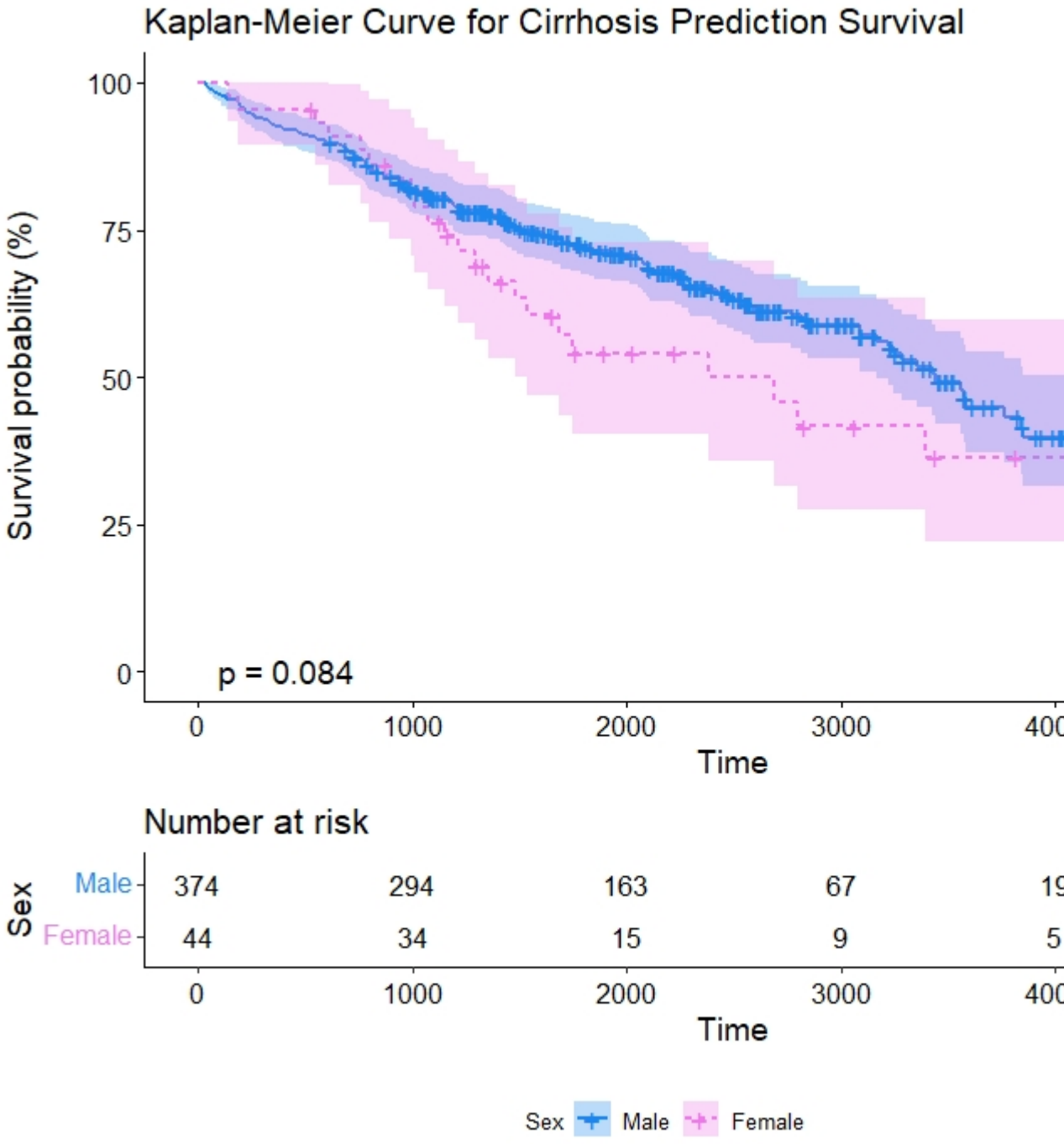


Figure 3.1: Decreasing Number at Risk with Time

3.2 Rank Log Test

To determine whether there are difference between the survival curvethen used the log rank test.

Variable	Log rank$_{Chi-Squre}$	df	P-Value
Bilirubin	0.51384	1	0.47348
Albumin	11.065669	1	0.00088
Copper	0.36833	1	0.54391
Alk$_P hos$	1.12439	1	0.28898
Prothrombin	0.07198	1	0.78848

Table 3.2: Rank Log Test Factors affecting Survival of Liver Cirrhosis

Based on table 3.2, it can be seen that the survival pf Liver cirrhosis patients based on variables Bilirubin,Albumin,Alk-Phos,Copper were statistically significantly different with p-value¡0.05.

3.3 Cox Proportional Hazard Model

Let,Bilirubil(x_1) ,Albumin(x_2), Copper(x_3) ,Alp-Phos(x_4), Prothrombin(x_5) allegedly affecting survival of Liver cirrhosis patients are generally modeled as:

$$h(t) = h_0(t) \exp \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{3.1}$$

Based on table 3.3 the obtained model:

$$h(t) = h_0(t) \exp 1.87 x_1 - 17.31 x_2 + 4.03 x_3 - 7.19 x_4 - 15.52 x_5 \tag{3.2}$$

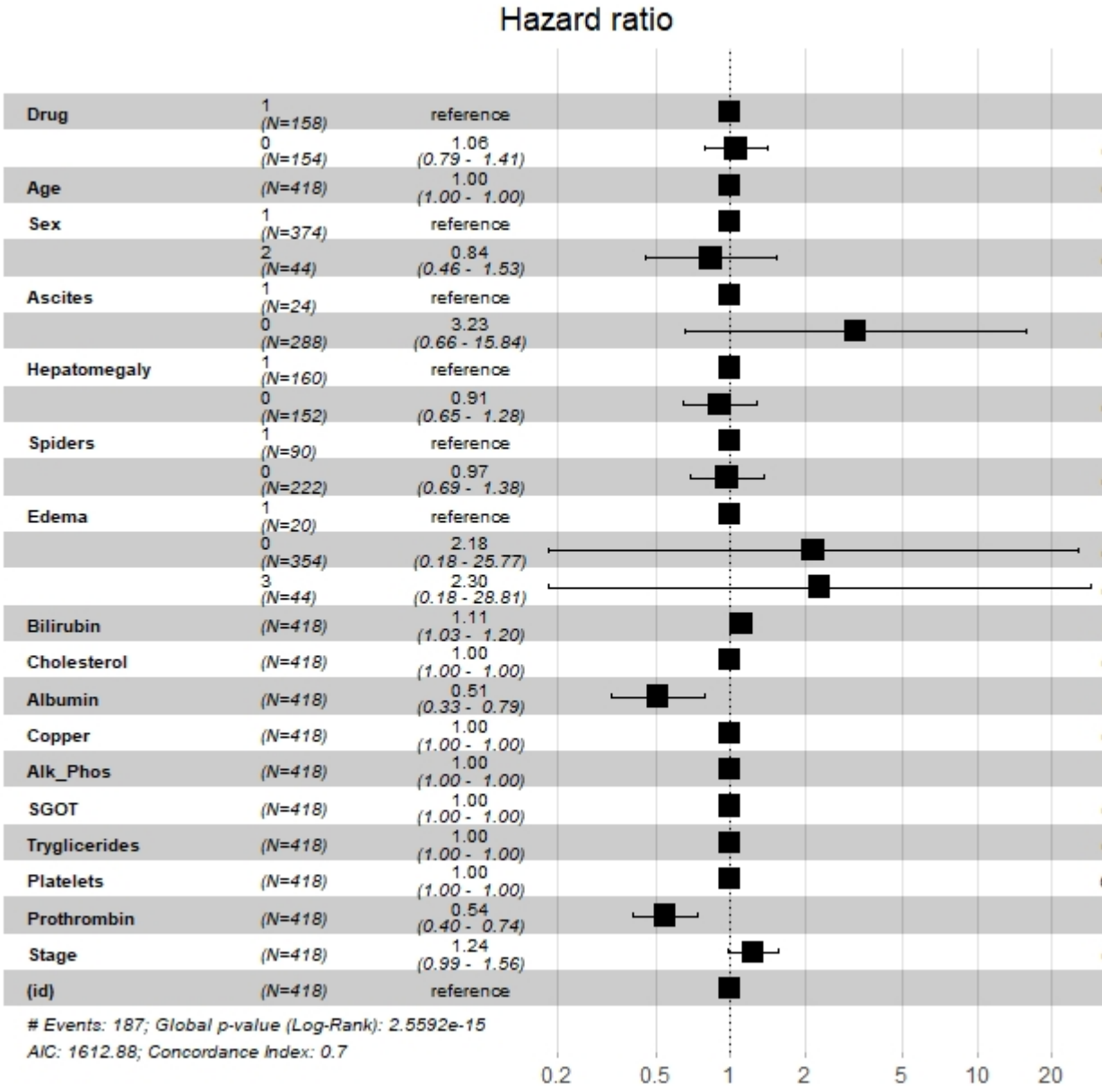
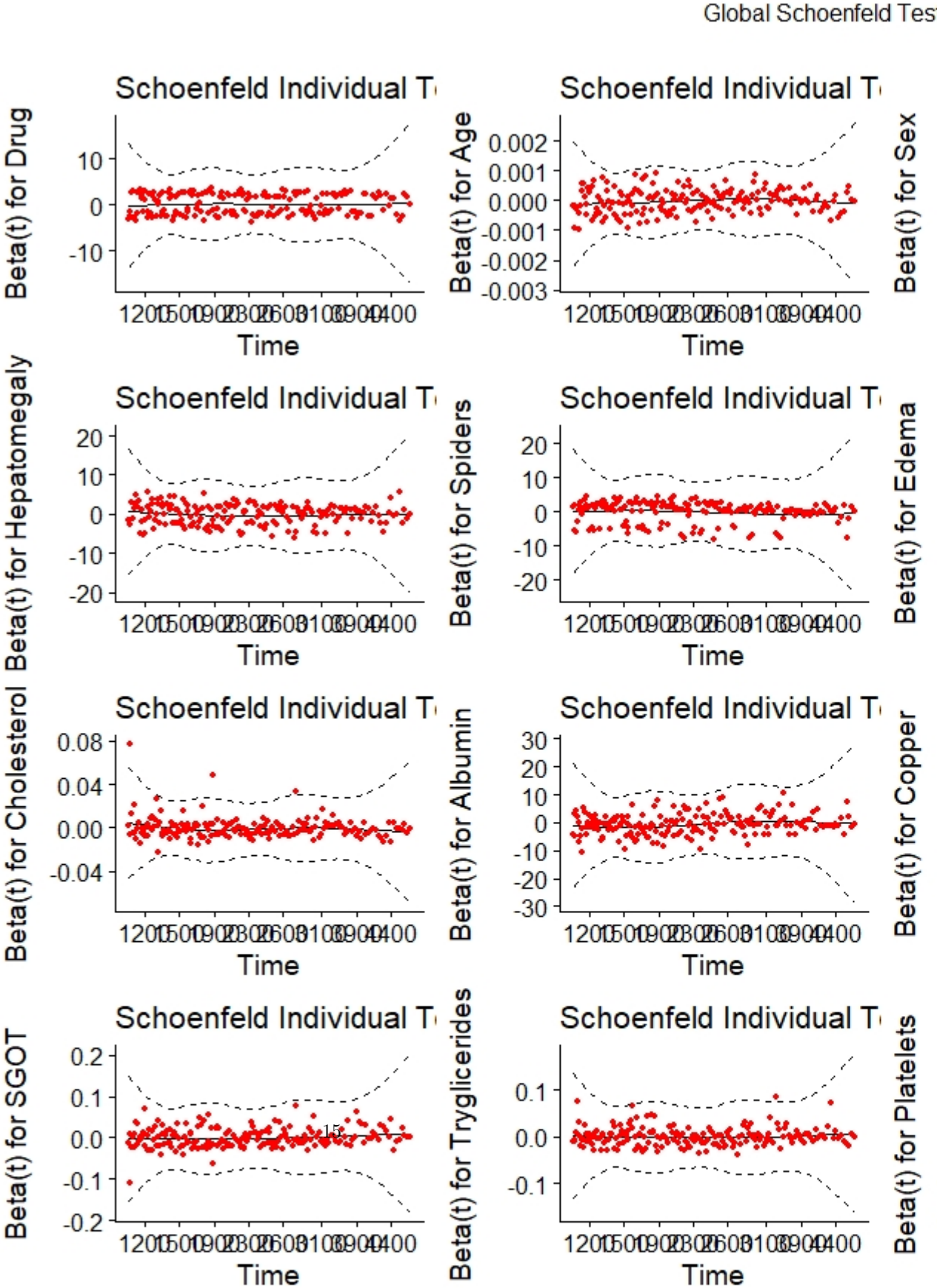


Figure 3.2: Hazard Ratio


Global Schoenfeld Tes
Schoenfeld Individual T
Beta(t) for Drug
Beta(t) for Age
Beta(t) for Sex
Beta(t) for Hepatomegaly
Beta(t) for Spiders
Beta(t) for Edema
Beta(t) for Cholesterol
Beta(t) for Albumin
Beta(t) for Copper
Beta(t) for SGOT
Beta(t) for Tryglicerides
Beta(t) for Platelets
Time

Values	coef	exp(coef)	se(coef)	robust se	z	$P_r(>\|z\|)$
Bilirubin	1.056e-01	1.111e+00	4.247e-02	3.902e-02	2.77	0.00678
Albumin	-6.738e-01	5.098e-01	2.568e-01	2.205e-01	-3.056	0.00224
Copper	2.587e-03	1.003e+00	1.365e-03	1.010e-03	2.561	0.01044
Alk-Phos	-4.129e-04	9.996e-01	7.444e-05	6.163e-05	-6.699	2.10e-11
Cholesterol	-5.673e-04	9.994e-01	5.900e-04	7.974e-04	-0.711	0.47684
Prothrombin	-6.083e-01	5.443e-01	1.301e-01	1.534e-01	-3.966	7.32e-05

Table 3.3: Value of the Coeff. with Z score

Tests	Value	df	P-Value
Lik.Ratio	110.5	18	3e-15
Wald	88.83	18	2e-11
Score(logrank)	82.78	18	3e-10
Robust	87.83	-	4e-11

Table 3.4: Tests for Cox PH

Chapter 4

Goodness of fit test

4.1 Test of Exponentiality

To test the hypothesis that the observed residuals $\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3, ..., \hat{\xi}_n$ are realizations from a standard exponential distribution, we use a number of goodness-of-fit tests for exponentiality. The approaches used to construct these tests can broadly be categorized into two groups: Classical tests based on the EDF and new tests based on empirical transforms. We also consider a recent moment-based test for the exponential distribution.

4.2 Tests Based on EDF

The EDF-based tests are expreterms of distances between the EDF

$$G_n(t) = \frac{1}{n}\sum_{j=1}^{n}(I(\hat{\xi}_j))$$

of the residuals $\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3, ..., \hat{\xi}_n$ defined in Eq. (4) and the corresponding population quantity for the standard exponential distribution $G(t) = 1 - \exp(t)$ The following EDF-based test statistics are considered: The Kolmogorov-Smirnov (KS) test statistic:

$$\hat{KS_n} = sup_{t\geq 0}[|G_n(t) - G(t)|]$$

the Cramer-von Mises (CM) test statistic:

$$CM_n = \int_0^\infty [G_n(t) - G(t)]^2 dG(t)$$

and the Anderson-Darling test (AD) statistic:

$$AD_n = \int_0^\infty \frac{[G_n(t) - G(t)]^2}{G(t)[1 - G(t)]} dG(t)$$

There are computationally efficient formulae for the EDF-based statistics that can be found in D'Agostino and Stephens (1986). Specifically, the KS, the CM and the AD test statistics simplify to

$$\hat{KS}_n = max KS_n^+, KS_n^+$$

$$KS_n^+ = max_{1 \le j \le n} [\frac{j}{n} - (1 - e^{\hat{\xi}_(j))}]$$

$$KS_n^- = max_{1 \le j \le n} [(1 - e^{\hat{-\xi}_(j))} - \frac{j-1}{n}]$$

$$CM_n = \frac{1}{12n} + \sum_{j=1}^{n} [(1 - e^{-\hat{\xi}_(j)})] - \frac{2j-1}{2n}]$$

and

$$AD_n = -n - \sum_{j=1}^{n} \frac{2j-1}{n} [\ln(1 - \exp^{-\hat{\xi}_(j)}) - \hat{\xi}_(n + 1 - j)]$$

respectively, and where $\hat{\xi}_(1) \le \hat{\xi}_(2) ... \le \hat{\xi}_(n)$ denotes the ordered residuals. For the KS test statistic we use the Bolshev correction that rejects the null hypothesis for large values of

$$KS_n = \frac{6n.\hat{KS}_n + 1}{6\sqrt{n}}$$

Chapter 5

Conclusion

The main factor causing low survival time was because the patient comes for treatment already in an advanced stage even accompanied by comorbidities (such as diabetes, anemia and hypertension). It is recommended that health workers conduct promotions to motivate women at risk for early selfexamination if they know of any signs of Liver Cirrhosis earlier.

5.1 Estimated Result

As per prediction and tests : After 3000 Days the number of patient at Risk is minimizes drastically.
[1]

[1] Cirrhosis Prediction :Arjun Samanta:github:https://github.com/Arjun392/Prediction-of-Recovery-Time-from-Liver-Cirrhosis

Chapter 6

References :

** Data Set Link :https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset

1. Cox Proportional Hazard Survival Analysis to Inpatient Breast Cancer Cases : M. Nadjib Bustan et al 2018 J. Phys.: Conf. Ser. 1028 012230

2. Statistics for Biology and Health:Survival Analysis,A Self-Learning Text:M. Gail, K. Krickeberg, J.M. Samet, A. Tsiatis, W. Wong http://www.springer.com/series/2848

3. Goodness-of-fit tests in the Cox proportional hazards model:Marike Cockeran, Simos George Meintanis James S. Allison https://www.tandfonline.com/loi/lssp20

4. Chi-Squared Goodness-of-Fit Tests for the Proportional Hazards Regression Model:David Schoenfeld :TrustChi-Squared Goodness-of-Fit Tests for the Proportional Hazards Regression ModelAuthor(s): David SchoenfeldSource: Biometrika, Vol. 67, No. 1 (Apr., 1980), pp. 145-153Published by: Biometrika TrustStable http://www.jstor.org/stable/2335327Accessed