2016

Extension of Cox PH Model When Hazards are Non-Proportional Applied to Residential Treatment for Drug Abuse

Prashant Narayan KC
Minnesota State University, Mankato

Minnesota State University, Mankato

Extension of Cox PH Model when Hazards are Non-Proportional Applied to Residential Treatment for Drug Abuse

A Master Thesis written and submitted to Minnesota State University Mankato in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics as STAT 699 - Thesis

Author
Prashant Narayan KC

Supervisor
Prof. Dr. Deepak Sanjel,
Minnesota State University, Mankato

November 2016

Declaration

We declare that this thesis paper has been examined and approved.

Mankato, MN, November , 2016

Approved by

Dr. Deepak Sanjel, Chairperson

Dr. Mezbahur Rahman, member

Dr. Namyong Lee, member

Abstract

Cox proportional hazard (PH) model (1972) is one of the most common methods used in survival analysis. CoxPH model has a strong assumption that the covariates have proportional effect on the hazard function of the lifetime distribution of an individual which need to be carefully verified before interpretation of parameters estimates. What to do if the proportional assumption is violated

This thesis discusses and presents ways of testing if the assumption of proportional hazard is satisfied and modification of Cox PH regression model in case when hazards are not proportional. The results are illustrated by an analysis of Drug remission data from UMASS Aids Research Unit IMPACT Study (UISSURV). Comparisons of cox regression model and proposed methods in case when hazards are not proportional is discussed. In particular the methods of incorporating time dependent variable and stratification method.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my thesis supervisors Dr. Deepak Sanjel for his continuous support for my project. His guidance and support provided during the whole period has helped me in all time of research. His knowledge in this field and infinite patience to deal with trivial issues is one of the factors for the completion of the work.

I would like to thank Prof. Dr. Mezbahur Rahman and Prof. Dr. Namyong Lee, members of the examination committee. They have agreed to be in the committe and examine the thesis. Their constant guidance, insightful recommendations and encouraging supervision has became a valuable tool for going forward with the thesis.

I would also like to express my deepest gratitude to my family for their constant support.I would like to thank my mother Nirmala KC, sister Prakriti KC and brother-in-law Rupesh Gartaula. Without their help and constant encouragement, I would not have reached to this stage. I am dedicating this thesis to my father Late Bharat Bahadur KC. I miss you a lot wish you were here.

In addition, I would like to thank the faculty and staff from Department of Mathematics and Statistics, Minnesota State University Mankato, USA for their help throughout my graduate studies.

Contents

List of Figures

List of Tables

1. Introduction

Cox proportional hazard model (PH) is one of the most common methods used in modeling censored survival data. Cox proportional hazard model has a strong assumption that the covariates have proportional effect on the hazard function of the lifetime distribution of an individual which need to be carefully verified before interpretation of parameters estimates.

This thesis discusses and presents ways of testing if the assumption of proportional hazard is satisfied and modification of Cox PH regression model in case when hazards are not proportional. The results are illustrated by an analysis of Drug remission data from UMASS Aids Research Unit IMPACT Study (UISSURV). Comparisons of cox regression model and proposed methods in case when hazards are not proportional is discussed. In particular the methods of incorporating time dependent variable and stratification method.

In addition, the modeling of the data using survival anlaysis models is presented and discussed. Analysis has been done with the paper review and discussions. The tool used for the simulation is SAS® Version 9.4 . LaTeX was used for the writing the report

1.1. Objectives

The Main Goals in this Project are as following :

- Application of Survival models on Drug Data.
- Extension of Cox Proportional Hazard Model (Cox DR, 1972) [2].
- Investigate the need and method for the Extension of Cox Proportional Hazard Model (borucka 2013) [3].

1.2. Outline

The concepts of Survival analysis follow a sequenced progressive approach. With the objective of providing an introduction to these concepts and provide generated results to validate a proposed model, this thesis is arranged in the following manner.

- Chapter 1 provides an overview of the previous research that has been conducted in the field of extending Cox proportional hazard model. The literature review provides an overview into the research that has been conducted over the years that are important in understanding the Extension of Cox proportional Hazard model and why they are important.It also provides the design methods that have been used, common statistical measurement and factors that are influential to calculate and analyze the Extension of Cox proportional hazard method. The literature provides an explanation as to what has been previously done to lead up to why the study on extension of Cox Proportional

Hazard using survival analysis technique should be conducted and why it is relevant to the statistical society at present.

- Chapter 2 discussed the Data set used in the explanation of these research.The data set and the variables used In analysis process are explained in detail.
- Chapter 3 introduces the study and theory used in Survival analysis technique. Overviews of common survival methods and distributions are described and the importance of the distributions to survival analysis technique is discussed. Information and descriptions of different types of methods used in analysis process and why it is important measures are explained.
- Chapter 4 presents the theory and development of Cox method. analysis and performance evaluation of the modeled system based on Cox Proportional Hazard method.
- Chapter 5 results evaluating results from extension of the Cox Proportional Hazard method. This chapter applies the methods that were discussed in Chapters 3-6 using Statistical Analysis Software. The interpretation is also given in this chapter.
- Chapter 6 concludes the project with the summary of the work done. In addition, the possibility for extending the project domain as a part of future work has been presented.

2. Data Description

2.1. Introduction

For the purpose of studying the research question of the thesis, we conducted several analysis. For the analysis, We conducted the survival analysis of the data from UMASS Aids Research Unit IMPACT Study [survival] (UISSURV.DAT). The data was provided by University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study (UIS) by Drs. Jane McCusker, Carol Bigelow and Anne Stoddard. The data on "Drug remission" is described in Table 2.1 for this thesis. We used this dataset to study the survival analysis model. We studied the parametric, non parametric and semi parametric model.

2.2. Drug Data

Name of the Data : UISSURV.DAT
Name of study : UMASS Aids Research Unit IMPACT Study
Size of the observation: 628 observations
Number of the column: 12 variables
Source of data: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2012) Applied Survival Analysis: Second Edition. These data are copyrighted by John Wiley & Sons Inc.

UISSURV was a 5-year (1989-1994) collaborative research project (Benjamin F. Lewis, P.I., National Institute on Drug Abuse Grant R18-DA06151) [4]. It comprised of two concurrent randomized trials of residential treatment for drug abuse. The purpose of the study was to compare treatment programs of different planned durations designed to reduce drug abuse and to prevent high-risk HIV behavior. The UIS sought to determine whether alternative residential treatment approaches are variable in effectiveness and whether efficacy depends on planned program duration. These data were used to illustrate model building in the first edition of this book and are being retained for use in the second edition primarily for end of chapter exercises. The small subset of variables from the main study we use in this text is described below.

Table showing UISSURV Data for the analysis

	Variables	Description	Type
1	id	Identification Code	1 - 628
2	age	Age at Enrollment	Years
3	beck	Beck Depression Score	0.000 - 54.001
4	hercoc	Heroin/Cocaine Use During 3 Months Prior to Admission	1 = Heroin & Cocaine 2 = Heroin Only 3 = Cocaine Only 4 = Neither
5	ivhx	IV Drug Use History at Admission	1 = Never 2 = Previous 3 = Recent
6	ndrugtx	Number of Prior Drug Treatments	0 - 40
7	race	Subject's Race	0 = White 1 = Non-White
8	treat	Treatment Randomization Assignment	0 = Short 1 = Long
9	site	Treatment Site	0 = A 1 = B
10	los	Length of Stay in Treatment (Admission Date to Exit Date)	Days
11	time	Time to Drug Relapse (Measured from Admission Date)	Days
12	censor	Event for Treating Lost to Follow-Up as Returned to Drugs	1 = Returned to Drugs or Lost to Follow-Up 0 = Otherwise

Table 2.1.: UISSURV Data for the analysis

Table showing first 10 observation of the data

Obs	id	age	beck	hercoc	ivhx	ndrugtx	race	treat	site	los	time	censor
1	1	39	9	4	3	1	0	1	0	123	188	1
2	2	33	34	4	2	8	0	1	0	25	26	1
3	3	33	10	2	3	3	0	1	0	7	207	1
4	4	32	20	4	3	1	0	0	0	66	144	1
5	5	24	5	2	1	5	1	1	0	173	551	0
6	6	30	32.55	3	3	1	0	1	0	16	32	1
7	7	39	19	4	3	34	0	1	0	179	459	1
8	8	27	10	4	3	2	0	1	0	21	22	1
9	9	40	29	2	3	3	0	1	0	176	210	1
10	10	36	25	2	3	7	0	1	0	124	184	1
11	11	35	.	.	.	12	1	1	0	2	5	1
12	12	38	18.9	2	3	8	0	1	0	176	212	1
13	13	29	16	3	1	1	0	1	0	79	87	1
14	14	32	36	3	3	2	1	1	0	182	598	0
15	15	41	19	1	3	8	0	1	0	174	260	1
16	16	31	18	1	3	1	0	1	0	181	210	1
17	17	27	12	2	3	3	0	1	0	61	84	1

Figure 2.1.: Study Data for 658 Subjects in the Drug Data

The SAS code to obtain the 10 observation of the data is given in appendix section A.

Bargraph showing time (Data = UISSURV)

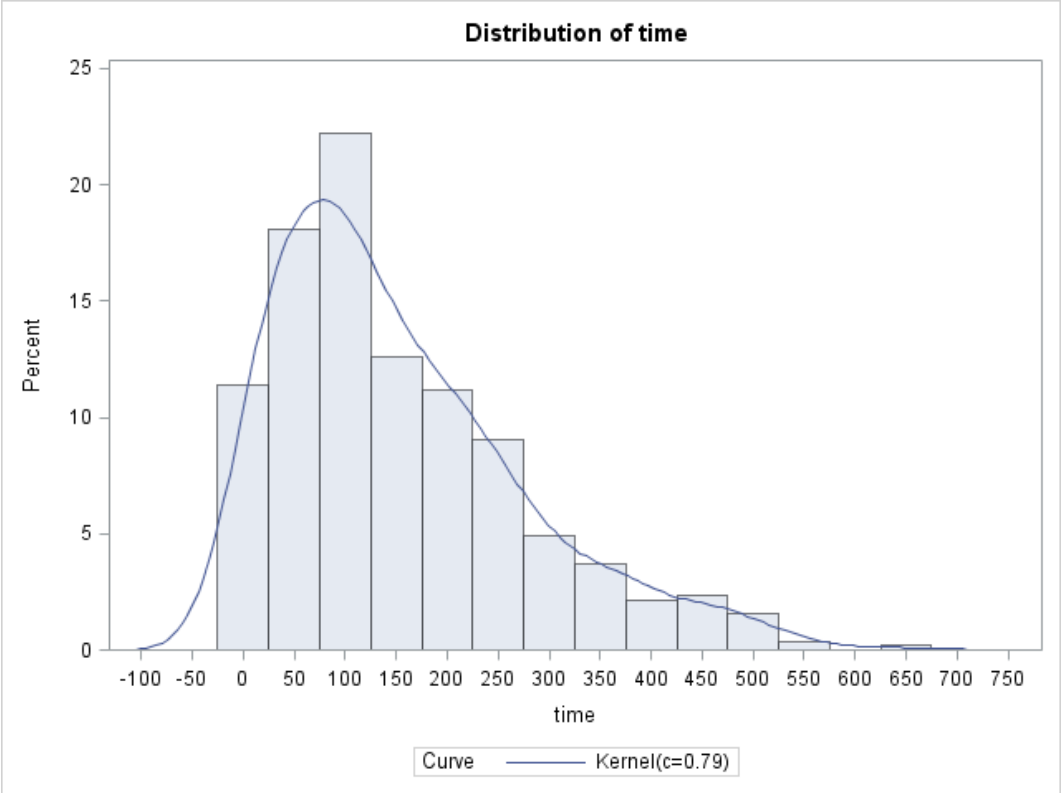

Figure 2.2.: Distribution of Time Variable

Figure 2.2 shows the bar graph time from UISSURV data.
The SAS code to obtain the figure 2.2 is given in appendix section A.

Analysis of scensor variable(Data = UISSURV)

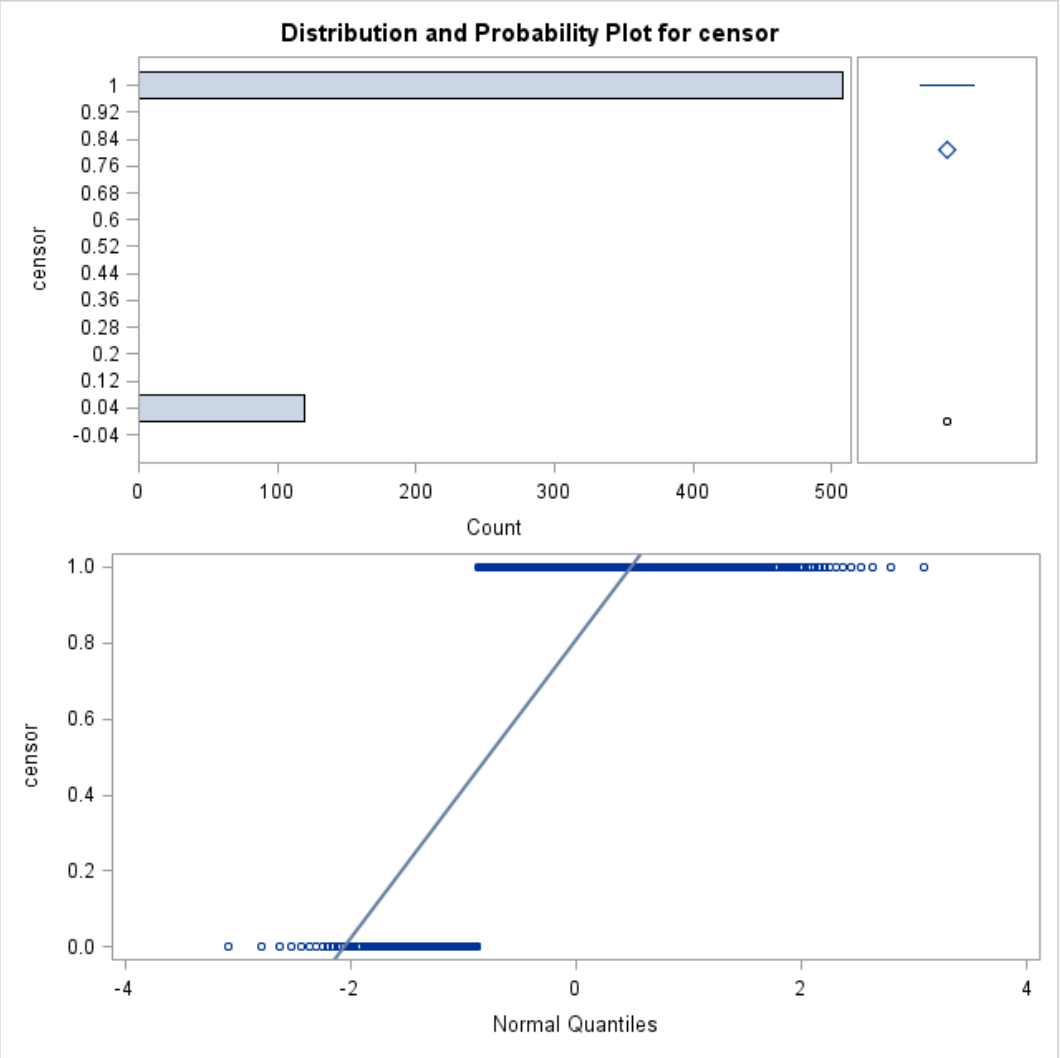

Figure 2.3.: Distribution of Censor Variable

Figure 2.3 shows the bar graph Length of follow up (days) time.

The SAS code to obtain the figure 2.3 is given in appendix section A.

Scatter plot of age covariate against time

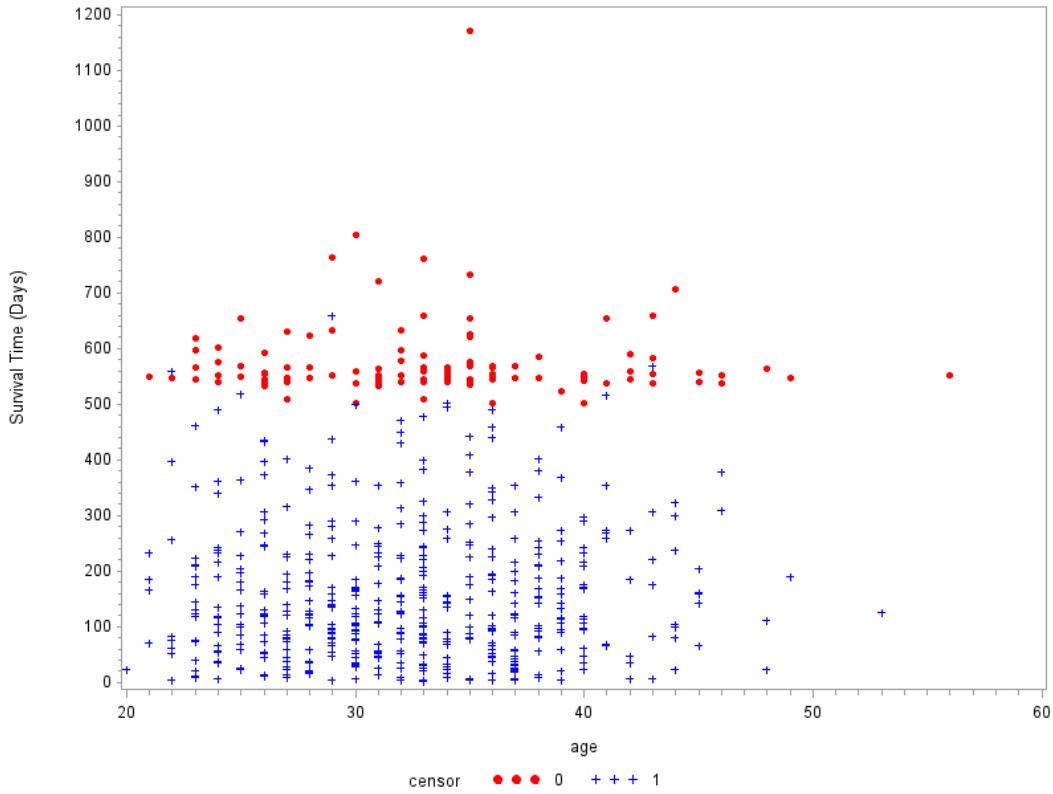

Figure 2.4.: Time vs censor for the different agegroup

Figure 2.4 shows the scatter plot of covariate "age" against the Length of follow up (days) time. With survival plot the shape of the plot is controlled by the nature of the distribution of the error and typically skewed to the right [4].

The cloud point in Figure 2.4 is densest for a short survival times and slowly trickles out to longer times with the plot truncated at the maximum length of follow up.
The SAS code to obtain the figure 2.4 is given in appendix section A.

3. Survival Analysis

3.1. Introduction

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc.

For example, if the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack [5].

In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Why not use linear regression to model the survival time as a function of a set of predictor variables? First, survival times are typically positive numbers; ordinary linear regression may not be the best choice unless these times are first transformed in a way that removes this restriction. Second, and more importantly, ordinary linear regression cannot effectively handle the censoring of observations [1] [5].

The method of survival analysis is used in other fields also.

- Deaths in Biological Science: (Survival Analysis)
- Mechanical Breakdown in Engineering: (Reliability Analysis)
- Insurance Claim in Actuarial Science (Time to Event Analysis)
- Events such as Divorce in Social Science: (Duration Analysis)

3.2. Significance of Survival Analysis studies

When we start any project, the significance of the project should be analyzed. For the purpose of conducting the Survival analysis, we have looked at some of its significance and the possible advantage due to many reasons [6].

Survival analysis attempts to answer questions such as:

1. What is the proportion of a population which will survive past a certain time?
2. Of those that survive, at what rate will they die or fail?
3. How do particular circumstances or characteristics increase or decrease the probability of survival? For example different types of treatments given to a group of patients.

Answering these questions helps to plan and work out has a lot of applications. Some of the applications are listed below.

3.2.1. Application of Survival Analysis

Amongst the most promising application of Survival Analysis are .

1. Prospective cohort studies : Death Rate Among a Group of Patients (clinical trail) .
2. Business Planning : Profiling customers who has a higher survival rate and make strategy accordingly. Costumer Churn Rate or Attrition Rate in Business (subscriber -based business, cell phone company)
3. Lifetime Value Prediction : Engage with customers according to their lifetime value. for example :- Time to Life Insurance Claim, Mechanical Breakdown.
4. Active customers : Predict when the customer will be active for the next time and take interventions accordingly.
5. Campaign evaluation : Monitor effect of campaign on the survival rate of customers.

Following are some industrial specific applications of survival analysis [7] :

- Banking - customer lifetime and LTV
- Insurance - time to lapsing on policy
- Mortgages - time to mortgage redemption
- Mail Order Catalogue - time to next purchase
- Retail - time till food customer starts purchasing non-food
- Manufacturing - lifetime of a machine component
- Public Sector - time intervals to critical events

3.2.2. Three Main Goals in the Survival Analysis.

1. Estimating survival probability or hazard rate $\lambda(t)$ at time t.
2. Comparing survival rate for different groups (significance test). $H_0 : S_i(t) = S_j(t)$ Vs $H_1 : S_i(t) \neq S_j(t)$

   Insurance claim rate for different groups statistically significant? or Survival rate in groups statistically significant?
3. Fitting a survival models for given covariates:

   $T_i = exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i)$

   Or hazard models:

   $\lambda_i(t) = \lambda_0(t) exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$

3.3. Mathematics of Survival Analysis : Notations

Survival Time: Random outcome variable $T \geq 0$ and observe time t

Cumulative Distribution Function (CDF)

Cumulative distribution function (CDF) of a real-valued random variable T, or just distribution function of T, evaluated at t, is the probability that T will take a value less than

or equal to t.

For a continuous distribution, this can be expressed mathematically as :-
$F(t) = Pr(T \leq t)$

Survival Function :

The survival function, also known as a survivor function or reliability function, is a property of any random variable that maps a set of events, usually associated with mortality or failure of some system, onto time. It captures the probability that the system will survive beyond a specified time.
Survival Prob at time

Mathematically,
$t.\ S(t) = Pr(T > t) = 1 - F(t)$

Probability Density Function (P.D.F) : Rate of probability.

For a continuous function, the probability density function (pdf) is the probability that the variate has the value t. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points.

Mathematically,
$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$

Hazard Function: Failure Rate

The hazard function is not a density or a probability. However, we can think of it as the probability of failure in an infinitesimally small time period between t and t + Δt given that the subject has survived up till time T.

Mathematically,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Cumulative Hazard:

The cumulative hazard plot consists of a plot of the cumulative hazard H versus the time ti of the ith failure. As with probability plots, the plotting positions are calculated independently of the model and a reasonable straight-line fit to the points confirms that the chosen model

and the data are consistent.

Mathematically,

$$\Lambda(t) = \int_0^t \lambda(x)dx$$

Mean Life Time:

$E(T) = \int_0^\infty t\ f(t)dt$

Integrating by parts, and making use of the fact that $-f(t)$ is the derivative of $S(t)$, which has limits or boundary conditions $S(0) = 1$ and $S(\infty) = 0$,

gives
$\mu = \int_0^\infty S(t)dt$

The mean life time is simply the integral of the survival function.

3.4. Some Survival Models:

These are the some of the popular survival analysis model that we will consider today.

- Nonparametric Methods: Estimates and tests the difference between survival functions BUT does not allow covariate effects on the parameters.
- Parametric Models: Allows covariate effects on the model BUT does not allow time-dependent covariates.
- Semiparametric Model (Cox model): Most popular, no distributional assumptions needed and also allows time-dependent covariates. Suitable for both discrete (Poisson, logistic) and continuous time data.

3.5. Nonparametric estimate of $S(t)$

Kaplan-Meier (KM) method: method is used for estimating and comparing survival function S(t): The estimate of survival probabilities are computed using a product limit formula. It doesn't require distributional assumption. Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. The time starting from a defined point to the occurrence of a given event, for example death is called as survival time and the analysis of group data as survival analysis [8].

Product limit formula: Out of n_j individuals at risk of an event, let d_j individual died at time t_j

K-M estimator of S(t) is:

$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$ for $\quad t_1 \leq t \leq t_k$

KM analysis of Site Variable

Descriptive statistics with site as group

Table 3.1.: Estimated Survival Function Computed from the Survival Times for UISSURV Study

Product-Limit Survival Estimates					
time	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0	1	0	0	0	500
1	0.984	0.016	0.00561	8	492
2	0.968	0.032	0.00787	16	484
3	0.962	0.038	0.00855	19	481
4	0.958	0.042	0.00897	21	479
5	0.954	0.046	0.00937	23	477
6	0.944	0.056	0.0103	28	472
7	0.932	0.068	0.0113	34	466
10	0.926	0.074	0.0117	37	463
11	0.918	0.082	0.0123	41	459
< ... output omitted to save space ... >					
2160	0.4381	0.5619	0.0468	212	11
2350	0.2921	0.7079	0.1233	213	2
2353	0.146	0.854	0.1202	214	1
2358	0	1	.	215	0

Above we see the table of Kaplan-Meier estimates of the survival function produced by proc lifetest. Each row of the table corresponds to an interval of time, beginning at the time in the "time" column for that row, and ending just before the time in the "time" column in the first subsequent row that has a different "time" value.

Parameter Estimatation, Standard Error startified by Sites variable

The descriptive statistics by Site: B=1, A =0:

Mean	Standard Error
242.48	13.44

Summary of the Number of Censored and Uncensored Values					
Stratum	site	Total	Failed	Censored	Percent Censored
1	0	444	364	80	18.02
2	1	184	144	40	21.74
Total		628	508	120	19.11

Figure 3.1.: Descriptive statistics with site as group

Figure 3.1 Survival function representation of life-table estimate of the survival function for the uis data.

The SAS code to obtain the figure 3.1 is given in appendix section A.

Product Limit Survival Estimates stratified by Sites

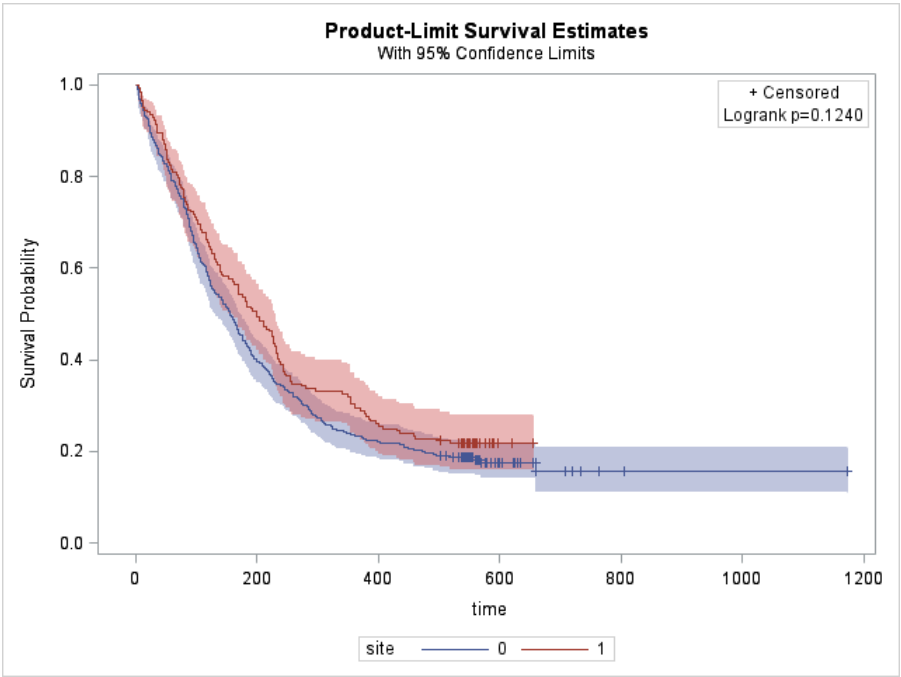

Figure 3.2.: Product Limit Survival Estimates

Figure 3.2 presents the Kaplan-Meier estimator of the survival function for the UIS study and the upper and lower pointwise 95 percent confidence bands computed. Stratified by Sites.

The step function form of the survival function is apparent in the graph of the Kaplan-Meier estimate. When a subject uses drug at a particular time point, the step function drops,

whereas in between failure times the graph remains flat. The survival function drops most steeply at the beginning of study, suggesting that the hazard rate is highest immediately after treatment ends during the first 200 days. Censored observations are represented by vertical ticks on the graph. Notice the survival probability does not change when we encounter a censored observation. Because the observation with the longest follow-up is censored, the survival function will not reach 0. Instead, the survival function will remain at the survival probability estimated at the previous interval. The survival function is undefined past this final interval at 1200 days. The blue-shaded and red shaded area around the survival curve represents the 95% confidence band, here Hall-Wellner confidence bands. This confidence band is calculated for the entire survival function, and at any given interval must be wider than the pointwise confidence interval (the confidence interval around a single interval) to ensure that 95% of all pointwise confidence intervals are contained within this band. Many transformations of the survivor function are available for alternate ways of calculating confidence intervals through the conftype option, though most transformations should yield very similar confidence intervals.

The SAS code to obtain the figure 3.2 is given in appendix section A.

Checking the survival goodness of fit

Testing H_0 : There is no significant difference between the survival curves.

All goodness of fit tests does not show significant difference in S(t) between two groups.

Log-Rank=$\frac{(O_i-E_i)^2}{Var(O_i-E_i)} \sim \chi^2_{(1)}$ where O_i is observed and E_i is expected values in group i

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	2.3658	1	0.1240
Wilcoxon	3.1073	1	0.0779
Tarone	2.8764	1	0.0899
Peto	3.1177	1	0.0774
Modified Peto	3.1181	1	0.0774
Fleming(1)	3.1293	1	0.0769

Figure 3.3.: Checking the survival goodness of fit for Sites

Looking at the figure 3.3, we can say that the difference in the site is not significant looking at the logrank test.

Hazard Plot for the Site variable

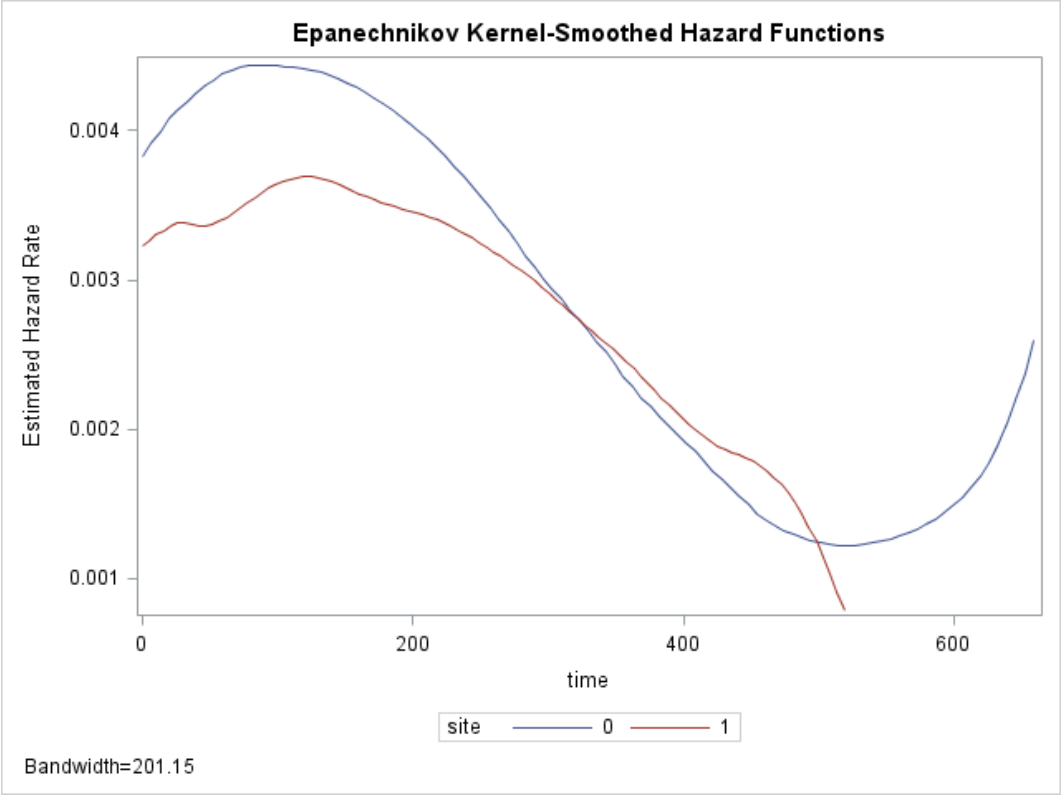

Figure 3.4.: Hazard Plot for Site variable

Figure 3.4 shows the hazard function. The lines in the graph are labeled by the midpoint site in each group. From the plot we can see that the hazard function indeed appears higher at the beginning of follow-up time and then decreases until it levels off at around 200 days and stays low and increases for Site 1. The SAS code to obtain the figure 3.4 is given in appendix section A.

KM analysis of Race Variable

Descriptive statistics with Race as group

The descriptive statistics by Site: Non- White =1, White =0:

Mean	Standard Error
293.06	19.04

Summary of the Number of Censored and Uncensored Values					
Stratum	race	Total	Failed	Censored	Percent Censored
1	0	467	388	79	16.92
2	1	155	116	39	25.16
Total		622	504	118	18.97

Figure 3.5.: Descriptive statistics with race as group

Figure 3.5 Survival function shows the parameter estimation of the KM estimator for the data stratified by race.

The SAS code to obtain the figure 3.5 is given in appendix section A.

Product Limit Survival Estimates stratified by Race

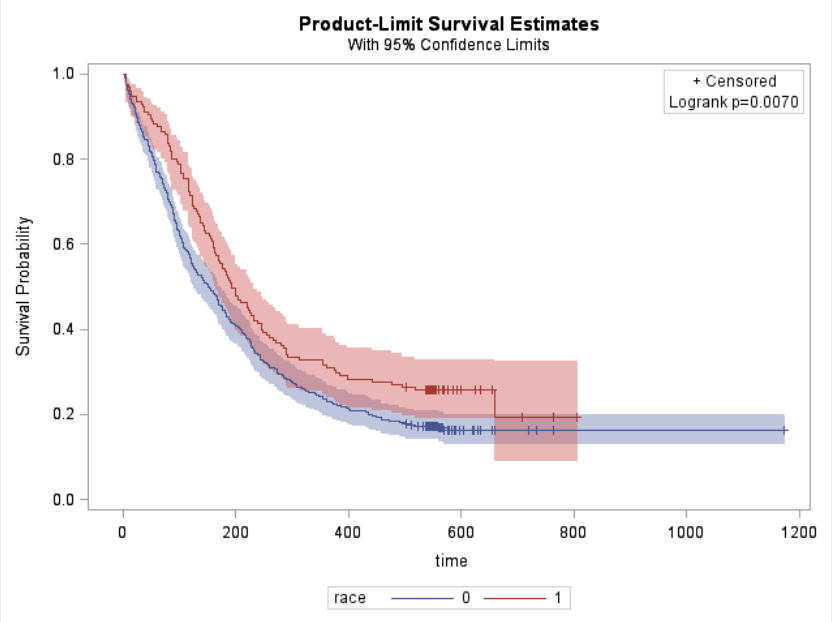

Figure 3.6.: Product Limit Survival Estimates for Race

Figure 3.6 In the graph of the Kaplan-Meier estimator stratified by race below, it appears that White people generally have a worse survival experience. The difference between this plot and the previous plot is that it shows the 95% CI. The SAS code to obtain the figure 3.6 is given in appendix section A.

Checking the survival goodness of fit

Testing H_0 : There is no significant difference between the survival curves.

All goodness of fit tests show significant difference in S(t) between two groups.

Log-Rank=$\frac{(O_i-E_i)^2}{Var(O_i-E_i)} \sim \chi^2_{(1)}$ where O_i is observed and E_i is expected values in group i

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	7.2836	1	0.0070
Wilcoxon	9.1336	1	0.0025
Tarone	8.4098	1	0.0037
Peto	9.1410	1	0.0025
Modified Peto	9.1495	1	0.0025
Fleming(1)	9.1121	1	0.0025

Figure 3.7.: Checking the survival goodness of fit for Race

Looking at the figure 3.7, we can say that the difference in the race is significant looking at the logrank test

Hazard Plot for the Race variable

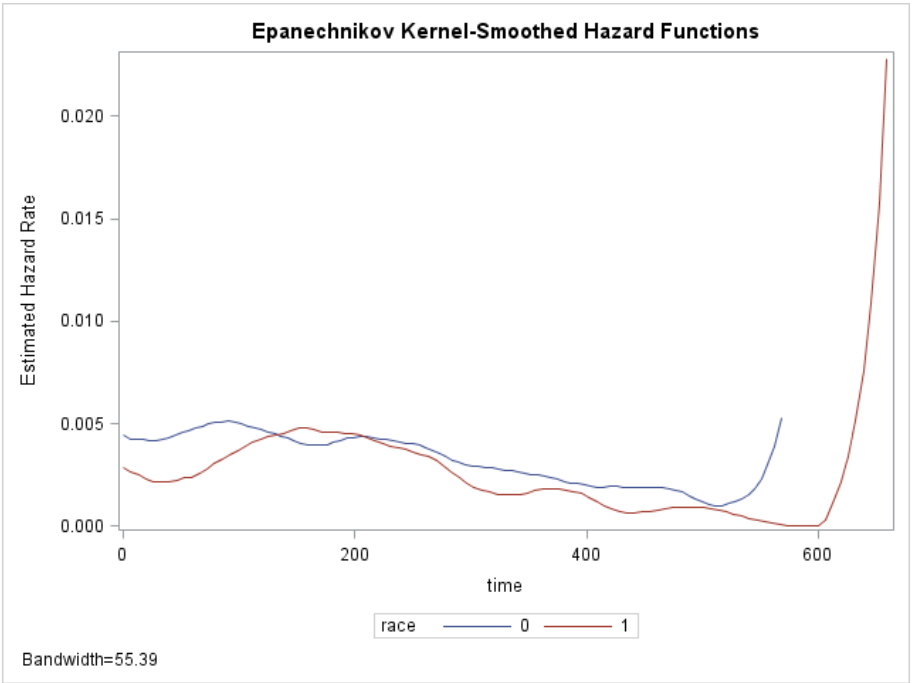

Figure 3.8.: Hazard Plot for Race variable

Figure 3.8 shows the hazard function. The lines in the graph are labeled by the midpoint site in each group. From the plot we can see that the hazard function indeed appears lower at the beginning of follow-up time and then decreases until it levels off at around 400 days and stays low and increases for both race. However increases dramatically for the White race. The SAS code to obtain the figure 3.8 is given in appendix section A.

3.6. Parametric estimate of $S(t)$

We will learn the parameteric using the Accelerated failure time (AFT) model. AFT model: Effect of a covariate is to accelerate or decelerate the life time by some constant. In the statistical area of survival analysis, an accelerated failure time model (AFT model) is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. This is especially appealing in a technical context where the 'disease' is a result of some mechanical process with a known sequence of intermediary stages [8].

Life time is linear effect of covariates and a random error (log linear). Random error has certain life time distributions

Failure time of ith individual (T_i):

$log(T_i) = \beta_0 + \beta_1 X_{i1} + \cdots \beta_k X_{ik} + \sigma\varepsilon_i$

or

$T_i = exp[\beta_0 + \beta_1 X_{i1} + \cdots \beta_k X_{ik} + \sigma\varepsilon_i]$

3.6.1. Common distributions in ACT survival analysis

These distributions are called life time distributions. Exponential dist. $f(t) = \lambda e^{-\lambda t}$ and $S(t) = e^{-\lambda t}$

Distribution of ε	Distribution of T
extreme value (2 par.)	Weibull
extreme value (1 par.)	exponential
log-gamma	gamma
logistic	log-logistic
normal	log-normal

Figure 3.9.: Common Distributions for AFT [1]

Figure 3.9 shows some common distribution for the non-parametric model of the acceleration failure time model.

3.6.2. Graphical Method for Model Check

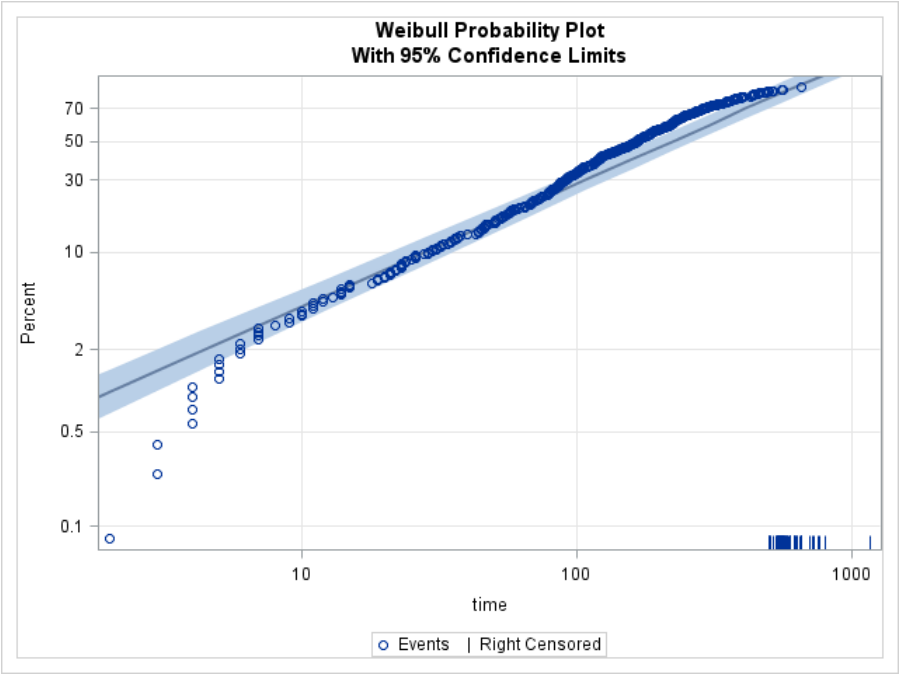

Figure 3.10.: Graphical Model Fitting check using Weibull

Figure 3.10 displays the probability plot for the Weibull model. Here not all the non-parametric estimates fall within the 95% confidence bands.

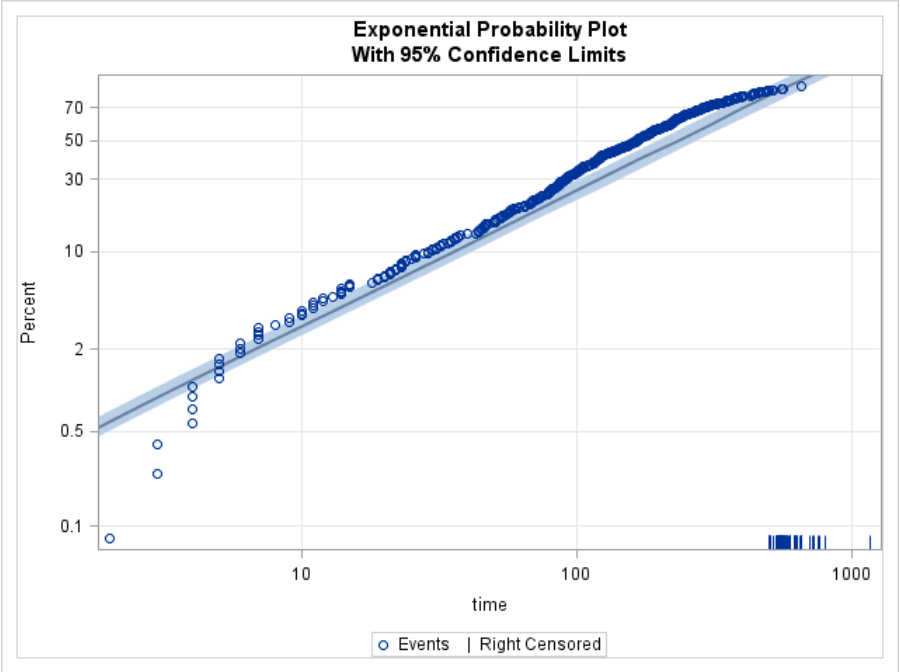

Figure 3.11.: Graphical Model Fitting check using exponential

Figure 3.11 displays the probability plot for the exponential model. Again, we see evidence that the model doesn't fit the data well, although it's not nearly as bad as the Weibulll model. Here not all the non-parametric estimates fall within the 95% confidence bands.

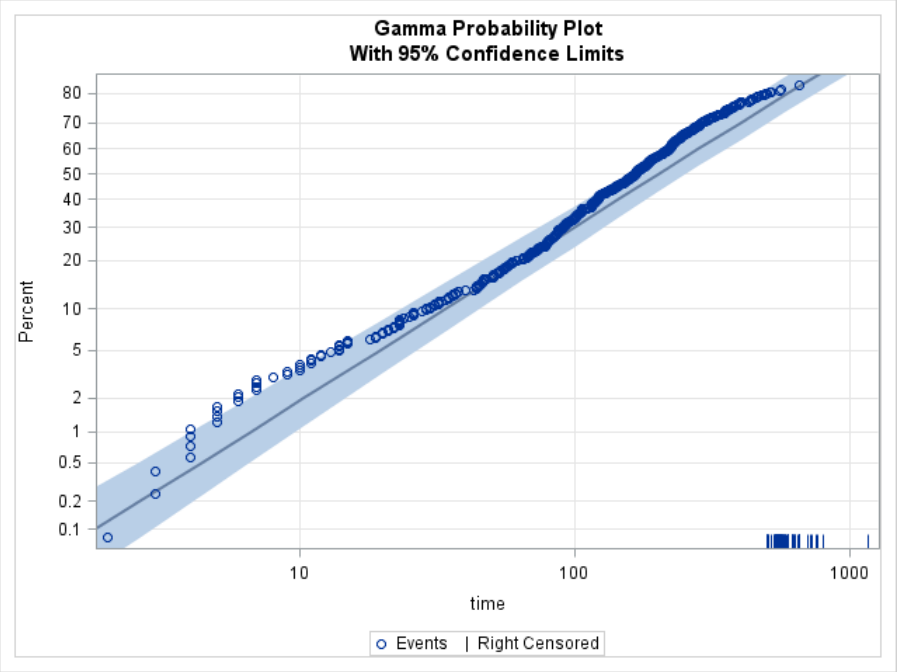

Figure 3.12.: Graphical Model Fitting check using gamma

Figure 3.12 displays the probability plot for the gamma model. We see evidence that the model fit the data well visually. Here all the non-parametric estimates fall within the 95% confidence bands.

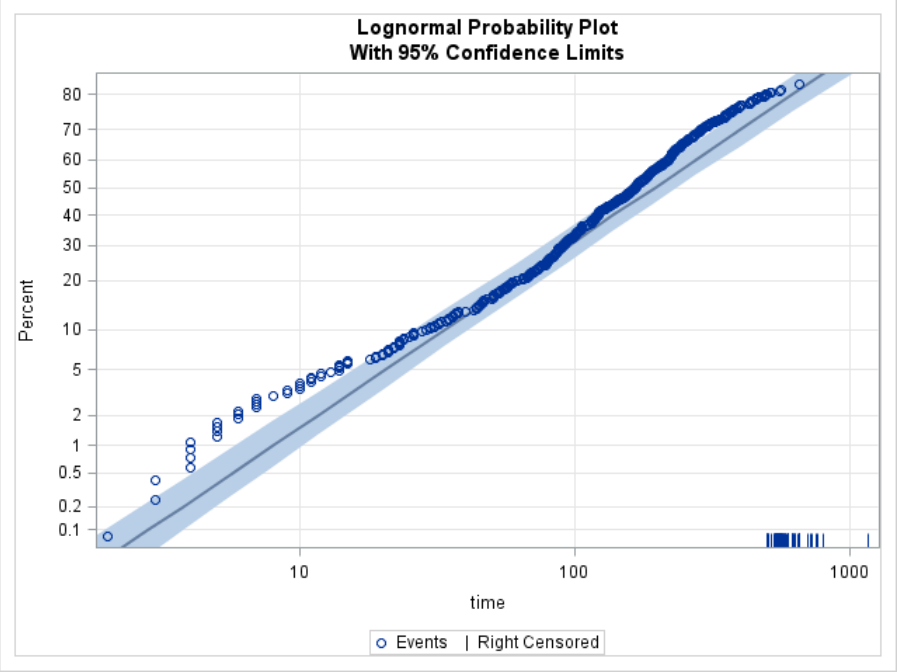

Figure 3.13.: Graphical Model Fitting check using lognormal

Figure 3.13 displays the probability plot for the logNormal model. We see evidence that the model fit the data well visually. Here almost all the non-parametric estimates fall within the 95% confidence bands.

Parameter estimates check for different models

In the previous section, we could only see the fitting on the visual basis. Now we will look at the fitting of the model and specifically Parameter estimates and standard error (for significance test): Dependent variable is $Log(T_i)$

Par/Dist	Weibull	Exponential	Gamma	LogNormal
Intercept	5.1153 (0.2726)	5.1583 (0.2452)	4.6796 (0.3157)	4.5511 (0.3235)
Age	0.0152 (0.0081)	0.0142 (0.0073)	0.0164 (0.0089)	0.0322 (0.0114)
site	-0.2421 (0.1117)	-0.2296 (0.1010)	-0.2448 (0.1203)	-0.2435 (0.1239)
race	0.3444 (0.1196)	0.3262 (0.1079)	0.3743 (0.1286)	0.3848 (0.1324)
treat	0.2607 (0.0995)	0.2502 (0.0898)	0.2657 (0.1097)	0.2487 (0.1137)

On the basis of the parameter estimate from the above table, we have calculated the average survival time of the individual with same covariates value:

- Weibull Model:
  $\hat{T}_i = exp(5.1153 + 0.0152 - 0.2421 + 0.3444 + 0.2607) = 243.13$

- Exponential model:
  $\hat{T}_i = exp(5.1583 + 0.0142 - 0.22960.3262 + 0.2502) = 249.46$

- Gamma Model
  $\hat{T}_i = exp(4.6796 + 0.0164 - 0.2448 + 0.3743 + 0.2657) = 162.58$

- LogNormal Model
  $\hat{T}_i = exp(4.5511 + 0.0322 - 0.2435 + 0.3848 + 0.2487) = 144.50$

Goodness of Fit check

Goodness of fit tests with the likelihood ratio Statistic (Smaller is better)

Dist/Test	-2 LogL	AIC	BIC
Weibull	2003.069	2015.069	2041.618
Exponential	2011.059	2021.059	2043.183
Gamma	1971.103	1985.103	2016.077
LogNormal	1975.542	1987.542	2014.091

On the basis of the table above we can choose the suitable models. The value of the AIC , BIC and -2LogL should be the lowest. In this case, the lognormal and Gamma are better than exponential and Weibull. We can also verify this with the help of graph. If we refer to the Figure 3.10, 3.11, 3.12 and 3.13

4. Cox Proportional

4.1. Introduction

Cox (1972) proposed a model which does not require assumption that survival times follow certain probability distribution. As a consequence, Cox model is considerably more robust. Cox regression (or proportional hazards regression) is method for investigating the effect of several variables upon the time a specified event takes to happen [2]. In the context of an outcome such as death this is known as Cox regression for survival analysis. The method does not assume any particular "survival model" but it is not truly nonparametric because it does assume that the effects of the predictor variables upon survival are constant over time and are additive in one scale [6].

Cox PH model is called semi-parametric model because hazard distribution is not specified but proportional assumption is there. Cox PH model can be expressed as

$h_i(t) = h_0(t)exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik})$

where $h_i(t)$ is the hazard function for individual i at time t.

$h_0(t)$ is the blue baseline hazard function = hazard function for an individual whose covariates $x_1, \cdots x_k$ all have values of 0.
Cox model is called the proportional hazards model because the hazard for any individual is a fixed proportion of the hazard for any other individual.

$HR = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t)exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{h_0(t)exp(\beta_1 x_{j1} + \cdots + \beta_k x_{jk})}$

$= exp(\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_k(x_{ik} - x_{jk}))$

Baseline hazard $h_0(t)$ cancel out and HR is constant with respect to time.
Semi-Parametric Model: No distributional assumptions but Some assumptions about the hazard function

Pros: Covariates easily incorporated. Most popular survival model. No distributional assumptions needed and also allows time-dependent covariates. Suitable for both discrete and continuous time data.

Cons: Does not provide the baseline hazard and can only interpret in terms of relative differentials.

4.2. Cox Model Using purposeful selection of co-variates

The results of the univariable analysis of each covariate in relation to survival time (in days) following admission to a rehabilitation centre after an drug abuse was calculated for the discrete covariates. All the covariates over the 20-25 % were discarded. The categorical value is differentiated into different indicator levels.

4.2.1. Full model in purposeful selection

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	-0.02887	0.00817	12.4833	0.0004	0.972	age
beck	1	0.00834	0.00498	2.8109	0.0936	1.008	beck
ndrugtx	1	0.02837	0.00831	11.6692	0.0006	1.029	ndrugtx
hercoc2	1	0.06532	0.15001	0.1896	0.6633	1.067	
hercoc3	1	-0.09362	0.16547	0.3201	0.5715	0.911	
hercoc4	1	0.02798	0.16028	0.0305	0.8614	1.028	
ivhx2	1	0.17439	0.13864	1.5822	0.2084	1.191	
ivhx3	1	0.28071	0.14693	3.6501	0.0561	1.324	
race	1	-0.20289	0.11669	3.0232	0.0821	0.816	race
treat	1	-0.23995	0.09437	6.4648	0.0110	0.787	treat
site	1	-0.10249	0.10927	0.8798	0.3483	0.903	site

Figure 4.1.: purposeful selection of co-variates Full Model

In Figure 4.1 the covariates in the full model was presented. All the variable except age categorized in four groups were sigificant at 20% level. Therefore they were included in the model.

The table present the result of fitting the multivariable PH model containing all the covariates. This analysis only inculded 575 subjects for which all the information were available for all the covariates.

Looking at the p-values of the covariates we can see the heroine level are insignificant , the intervention is also insignificant. Next we remove the insignificant and make the reduced model as shown in Figure 4.2

4.2.2. Reduced models and selection processes

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.0669	6	<.0001
Score	34.0129	6	<.0001
Wald	33.8680	6	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	-0.02055	0.00784	6.8717	0.0088	0.980	age
beck	1	0.00849	0.00489	3.0196	0.0823	1.009	beck
ivhx3	1	0.32422	0.10293	9.9213	0.0016	1.383	
race	1	-0.23562	0.11365	4.2979	0.0382	0.790	race
treat	1	-0.21908	0.09292	5.5595	0.0184	0.803	treat
site	1	-0.07631	0.10694	0.5092	0.4755	0.927	site

Figure 4.2.: purposeful selection of co-variates Reduced Model

Figure 4.2 shows the reduced modal after removing the insignificant co-variates from the previous model. The next thing we need to do is to check the proportionality assumption of each of this co-variates. The results of checking is shown in Figure 4.3 and Figure 4.4

Testing Proportional Hazard Assumptions

Testing Proportional Hazard Assumptions for Site Variable

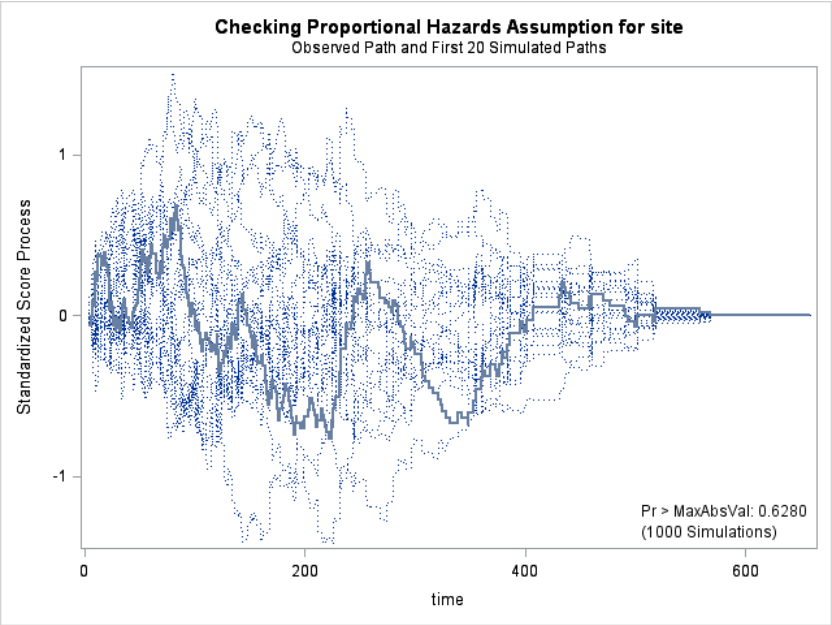

Figure 4.3.: Checking Proportional Hazard assumption for Site Variable

Figure 4.3 shows the site co-variate accepts the PH assumption. The p-value is 0.62.

Testing Proportional Hazard Assumptions for Treat Variable

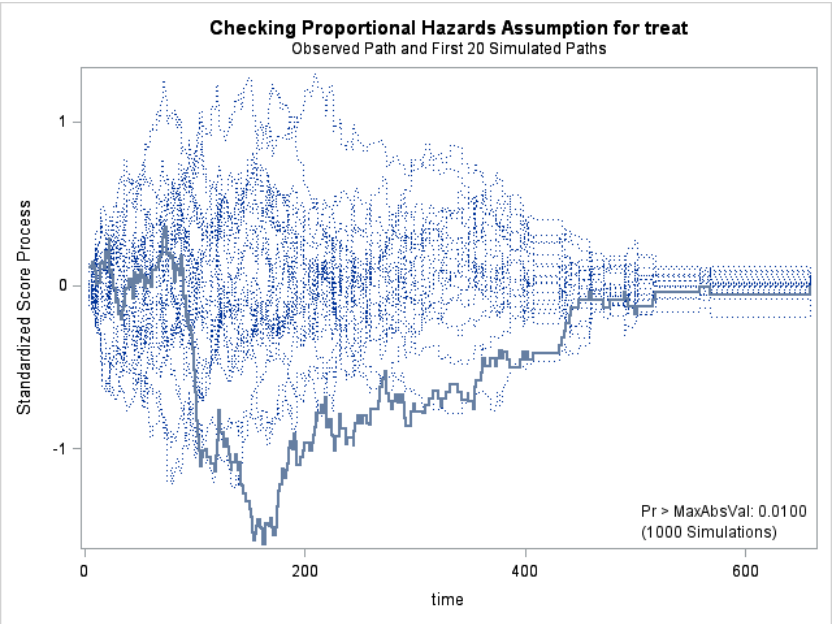

Figure 4.4.: Checking Proportional Hazard assumption for treat variable

Figure 4.4 shows the treat co-variate which does not accepts the PH assumption. The p-value is less than 0.05. So we must remove it from the model.

4.2.3. Model Using purposeful selection of co-variates after removing non proportional

Final Model selected after removing the non PH co-variates from the model. We removed the "Treat" covariate and fitted the model which is shown in Figure 4.5. The interpretation of the some of the covariate is shown in the following section.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.2004	4	<.0001
Score	28.0624	4	<.0001
Wald	27.9249	4	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	-0.02041	0.00780	6.8527	0.0089	0.980	age
beck	1	0.00875	0.00487	3.2318	0.0722	1.009	beck
ivhx3	1	0.35595	0.09901	12.9250	0.0003	1.428	
race	1	-0.23123	0.11126	4.3192	0.0377	0.794	race

Figure 4.5.: Final Model chosen under Cox PH

Estimating Hazard Ratio ($\widehat{HR}$)

From Cox PH Model:

$\widehat{HR} = \frac{h_i(t)}{h_j(t)} = exp(\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_k(x_{ik} - x_{jk}))$

For indicator variables (Dummy) with values 0 and 1. HR is the estimated hazard for those with values of 1 to those with values 0 controlling for other covariates.

For example: Hazard of remission of drug person with race White is 20% less than the Non White.

For quantitative covariates, subtract 1 from HR and multiply by 100.

For example: 1 year increase in age hazard of the death will go down by about 2%

4.3. Comparing Different Survival Models

The advantages of different model are given in the following table.

(Nonparametric Survival Methods) Kaplan-Meier (KM) method	(parametric models) Accelerated Failure Time	(Semi-parametric) Cox-Proportional Hazard
No distribution assumptions is required	Covariate effects easily incorporated on the estimates	Can works with very little distribution assumptions is required
Easy to estimate and interpret	easily modelled for the standard distributions.	Covariate effects easily incorporated on the estimates

Disadvantages of different model

(Nonparametric Survival Methods) Kaplan-Meier (KM) method	(parametric models) Accelerated Failure Time	(Semi-parametric) Cox-Proportional Hazard
It is mostly descriptive	Need distribution assumptions for survival time may not be accurate	assumes the hazard to be constant over the time
cannot handle covariates	cannot handle time dependent covariates	cannot handle time dependent covariates

Looking at both the advantage and the disadvantage of the different models, we can say that there is no model which can address the time dependent variable. Therefore, it is the motivation for us to continue with the extension of the Cox model in the following section.

5. Extension of Cox Proportional Hazard Discussion

Cox regression model is applicable only to time- invariant predictors with time-constant effects only. we can extend a linear regression model in a variety of ways, so, too, can we extend the Cox regression model. Some naive extensions technique: the inclusion of statistical interactions terms are identical to extensions we can make to any statistical model [6]. (interaction term extension is skipped in this thesis). Some a set of novel extensions techniques are :-

- including time-varying predictors in the Cox regression model
- stratified Cox regression model.

5.1. Extending Cox PH Model using time varying co-variates

Until now we have assumed that the values of all covariates were determined at the point when follow-up began (time zero) and that these values did not change over the period of observation. one or more of the covariates may be measured during the period of follow up and their values change. We may be the value of the hazard for the event depends on the current values of rather than on their values at time zero [4].

We must pay close attention to the definition of when the analysis time is zero (i.e when the clock starts). The earliest applications of time-varying covariates was in analysis of the Stanford heart transplant data.

5.1.1. The method to include time varying co-variates

We need to generalize the notation to include time-varying covariates in the model. For example, suppose we define a time-varying covariate as the length of stay in a hospital. Subject 1 may be admitted to a hospital on May 1 and stay in the hospital for 60 days. Subject 2 may be admitted to a hospital on July 1 and also stay in the hospital for 60 days. It is important that both subjects were in the hospital for 60 days, not that they were admitted at different times.

$$h_i(t, x_{in}(t)) = h_0(t)exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \beta_j x_{in}(t))$$

where $h_i(t)$ is the hazard function for individual i at time t.

$\beta_1 x_{i1} + \cdots + \beta_k x_{ik}$: are still time invariant co-variates

$\beta_j x_{in}(t)$:- Time dependent co-variates.

$h_0(t)$ is the baseline hazard function = hazard function for an individual whose covariates $x_1, \cdots x_k$ all have values of 0.

5.1.2. The method's application to Data

o examine the "under treatment" hypothesis, we create a time-varying dichotomous subject specific time varying covariate $OFF\_TRT()$, where

$$OFF\_TRT = \begin{cases} 0 & if \qquad t \leq los \\ 1 & if \qquad t > los \end{cases}$$

We discovered that treatment appeared to have a time dependent effect. We hypothesized that the treatment effect may simply be housing a subject where he/she has no access to drugs

Checking if Treatment is significant

Analysis of Maximum Likelihood Estimates									
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treat	1	-0.23109	0.08899	6.7430	0.0094	0.794	0.667	0.945	treat

Figure 5.1.: Testing the significance of Treatment

Based on this model, we would conclude that longer duration of treatment significantly reduces the rate of returning to drug use

Fitting the model with Time variant variable

tot = interaction between $OFF\_TRT()$ and Treatment.

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
treat	1	-0.52388	0.22583	5.3814	0.0204	0.592	treat
off_trt	1	2.27033	0.18650	148.1847	<.0001	9.683	
tot	1	0.62097	0.24630	6.3563	0.0117	1.861	

Figure 5.2.: Fitting the model with Time variant variable

The results in Table demonstrate a dramatic effect due to the time-varying covariate, $OFF\_TRT(t)$. We note that all three coefficients are significant.

Estimating Extra sets of HR

Using the model in previous section, we can estimate four different hazards ratios:

Hazard Ratio for	Within Those	HR
Long vs. Short	On Treatment	0.59
Treatment Assignment	Off Treatment	1.1
Off vs. On	Shorter Tx Duration	9.68
Treatment	Longer Tx Duration	18

5.2. Extending Cox PH Model with stratified model

A Cox regression model invokes a proportionality assumption, that the hazard function for each individual in the population is a constant multiple of a common baseline function [4]. Although this assumption often holds, you may encounter data sets in which it does not. If exploratory analyses, theory, or regression suggest that subgroups of individuals have different baseline hazard functions, you have to fit a stratified model, which posits explicitly the existence of the multiple baseline hazard functions.

5.2.1. The method to include statification in Cox Model

We should split the whole sample into subgroups on the basis of categorical variable (stratification variable) and re-estimate the model. Then we let the baseline hazard function differ between these subgroups. It makes sense to choose covariate if it interacts with time (i.e. proportional hazard assumption is not satisfied for this covariate) [3]. We now describe the stratified proportional hazards model using constant slopes.

The proportional hazard function for stratum s is

$$h_s(t, x) = h_{s0}(t)exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik})$$

where s = 1, 2, ..., S and S is the total number of subgroups created on the basis of stratification variable.

5.2.2. The stratified method's application to Data

Let us consider TREAT for which it is known that proportional hazard assumption is violated as a stratification variable. It will be not possible to obtain parameter estimates for TREAT, however using BASELINE statement enables to estimate survival and cumulative hazard function estimates for each treatment separately, adjusting for age .

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.0226	1	0.0449
Score	3.9742	1	0.0462
Wald	3.9720	1	0.0463

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	-0.01438	0.00722	3.9720	0.0463	0.986	age

Figure 5.3.: Extending Cox PH Model with stratified model

Figure 5.3 show the fitted model with age covariate for the startified proportional Hazard model with Treat. In previous chapter, we learnt that treat was not following PH assumption. But with the stratification model we can see Treat can be useful covariate and may help in important decision and prediction for the future.
The use of the extension of the stratified model in Cox helped us to include the Treat which otherwise would have been discarded from our model.

Cumulivite Hazard for the stratified Cox PH

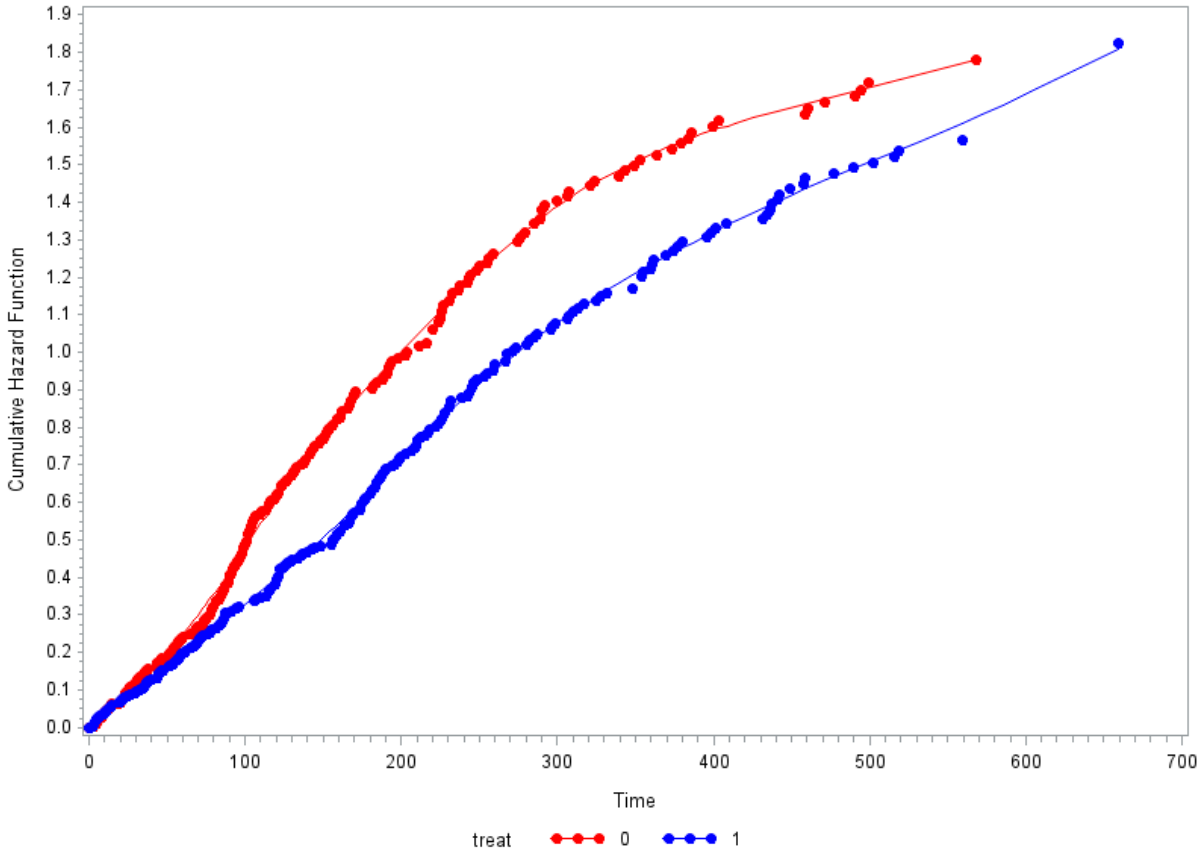

Figure 5.4.: Survival Plot of Extending Cox PH Model

Figure 5.4 shows the plots of cumulative hazard. As it is expected on the basis of previous results, subjects from Treat 1 tend to have relatively longer survival times as compared with subjects from Treat 0. Cumulative hazard function for almost all time points is higher for treat 0 which means that expected number of events till the given time point is usually higher for subjects treated with this treatment method.

6. Conclusion and Future Steps

6.1. Conclusion

After the analysis and simulation of a UISURGH data , we understood several survival analysis model. Most importantly, we learnt how to apply survival analysis model on a set of censored data. We also learnt how to interpret the co-efficients and the hazard ratios. The conclusions are sorted below:

- We can apply the one of the four models of Surival analysis.
- We could perform the analysis of data using non parametric regression methods for survival analysis.
- We could perform the analysis of data using semi parametric regression methodsfor survival analysis.
- We could perform the analysis of data using parametric regression methods for survival analysis.
- We can use the purposeful selection of the covariates to come up with one or model suitable Models for our analysis.
- We learnt analysis of Extension of the Cox Proportional Hazard methods.

6.2. Future Steps

There possibility for extending the project domain as a part of future work . We could work on different sets of data and perform the analysis.
We could also investigates the other methods of survival analysis which are covered by our literature but not covered by this project.

A. Appendix

A.1. SAS Codes

Getting Data into the File

```
proc import
  datafile="C:\Users\PNKC\Downloads\uissurv"
  dbms=xls
  out=WORK.uissurv
  replace;
run;
```

Getting histogram of censor data

```
proc univariate censorata = uissurv(where=(censor=1));
var time;
histogram time / kernel;
run;
```

Getting univariate of time variable

```
proc lifetest data=uissurv(where=(censor=1)) plots=survival(atrisk);
time time*censor(0);
run;
```

Plotting Age vs censor data

```
goptions reset=all;
symbol1 v=dot c=red h=.8;
symbol2 v=plus c=blue h=.8;
axis1 label=(a=90 'Survival Time (Months)');

proc gplot data=uissurv;
  plot time*age=censor / vaxis=axis1;
run;
```

KM plots for Sites Variable

```
proc lifetest data=uissurv PLOTS= ( S(TEST CL), H(BW=89.64));
time time*censor(0);
strata site / TESTS=ALL;
run;
```

KM plots for race Variable

```
proc lifetest data=uissurv PLOTS= ( S(TEST CL), H(BW=55.39));
time time*censor(0);
strata race / TESTS=ALL;
run;
```

Parametric modelling for Weibull

```
PROC LIFEREG DATA=uissurv;
        CLASS site;
        MODEL time*censor(0)= age       site    race
        treat / D=WEIBULL;
    PROBPLOT;
RUN;
```

Parametric modelling for exponential

```
PROC LIFEREG DATA=uissurv;
        CLASS site;
        MODEL time*censor(0)= age       site    race
        treat / D=exponential;
    PROBPLOT;
RUN;
```

Parametric modelling for gamma

```
PROC LIFEREG DATA=uissurv;
        CLASS site;
        MODEL time*censor(0)= age       site    race    treat / D=Gamma;
    PROBPLOT;
RUN;
```

Parametric modelling for Log Normal

```
PROC LIFEREG DATA=uissurv;
        CLASS site;
        MODEL time*censor(0)= age        site     race     treat / D=LNORMAL;
    PROBPLOT;
RUN;
```

Preparing variable for the purposeful

```
data uis;
  set uissurv;
  hercoc1 = (hercoc=1);
  hercoc2 = (hercoc=2);
  hercoc3 = (hercoc=3);
  hercoc4 = (hercoc=4);
  ivhx1 = (ivhx = 1);
  ivhx2 = (ivhx = 2);
  ivhx3 = (ivhx = 3);
  agecat = 0;
  if age < 28 then agecat = 1;
  else if age < 33 then agecat = 2;
  else if age < 38 then agecat = 3;
  else agecat = 4;
  agecat1 = (age < 28);
  agecat2 = (age >= 28 & age < 33);
  agecat3 = (age >= 33 & age < 38);
  agecat4  = (age >= 38);
  beckcat = 0;
  if beck < 10 then beckcat = 1;
  else if beck < 15 then beckcat = 2;
  else if beck < 25 then beckcat = 3;
  else beckcat = 4;
  beckcat1 = (beck < 10);
  beckcat2 = (beck >= 10 & beck < 15);
  beckcat3 = (beck >= 15 & beck < 25);
  beckcat4 = (beck >= 25);
  drugcat = 0;
  if ndrugtx < 2 then drugcat  = 1;
  else if ndrugtx < 4 then drugcat  = 2;
  else if ndrugtx < 7 then drugcat  = 3;
  else drugcat = 4;
  drugcat1 = (ndrugtx < 2);
  drugcat2 = (ndrugtx >= 2 & ndrugtx < 4);
  drugcat3 = (ndrugtx >= 4 & ndrugtx < 7);
  drugcat4 = (ndrugtx >= 7);
run;
```

```
proc print data = uis;
run;
```

Purposeful selcetion of data -1

```
proc phreg data=uis;
  model time*censor(0) = age beck  ivhx3 race site;
run;
```

Purposeful selcetion of data -2

```
proc phreg data=uis;
  model time*censor(0) = age beck ndrugtx hercoc2
                         hercoc3 hercoc4 ivhx2 ivhx3
                          race treat site;
run;
```

Purposeful selcetion of data -3

```
proc phreg data = uis;
model time*censor(0) = age site treat / ties = exact;
output out = schoen ressch = age_s site_s ;
run;
```

Purposeful selcetion of data -4

```
proc phreg data=uis;
  model time*censor(0) = age beck ndrugtx ivhx2 ivhx3 race treat site;
run;
```

```
proc phreg data=uis;
  model time*censor(0) = age beckt ivhx3 race treat site;
run;
```

PH test

```
proc phreg data=uis;
  model time*censor(0) = age beck ndrugtx hercoc2 hercoc3
                                                hercoc4 ivhx2 ivhx3
                          race treat site;
```

```
                                                    Assess PH / RESAMPLE;
run;
```

Negative ln check

```
proc lifetest data = UIS plots = (s, lls);
strata site;
time time*censor(0);
run;
```

Shoenfel Residual

```
proc gplot data = schoen;
symbol1 v = dot c = red width = 1 i = sm80s;
plot site_s*time / haxis = axis1 vaxis = axis2;
axis1 label = ('Time');
axis2 label = (a = 90 'Schoenfeld Residual for Site');
run;
```

Shoenfel Residual for age and site

```
proc phreg data = uis;
model time*censor(0) = age beck ivhx3 race treat site    / ties = exact;
output out = schoen ressch = age_s site_s ;
run;
```

Extented PH model with interaction

```
proc phreg data = uis;
model time*censor(0) = age site site_t/ ties = exact;
site_t = site*time;
test: test site_t;
run;
```

Extented PH model with interaction for site a only

```
proc phreg data = a.site;
format site site.;
model time*censor(0) = age site site_t/ ties = exact;
site_t = site*time;
test: test site_t;
run;
```

Extented PH model with interaction for age

```
proc phreg data = uis;
model time*censor(0) = age site treat age_t/ ties = exact;
age_t = age*time;
test: test age_t;
run;
```

Extented PH model with stratified on sites

```
proc phreg data = uis;
baseline out = base survival = surv cumhaz = cumhaz;
strata site;
model time*censor(0) = age / ties = exact;
run;
```

Extented PH model with stratified on treatment

```
proc phreg data = uis;
baseline out = base survival = surv cumhaz = cumhaz;
strata treat;
model time*censor(0) = age / ties = exact;
run;
```

Extented PH model with stratified on treatment (Cumulitive Hazard Function)

```
proc gplot data = base;
symbol1 v = dot c = red width = 1 i = sm50s;
symbol2 v = dot c = blue width = 1 i = sm50s;
plot cumhaz*time = site / haxis = axis1 vaxis = axis2;
axis1 label = ('Time');
axis2 label = (a = 90 'Cumulative Hazard Function');
run;
```

Extented PH model including time dependent variable

```
ods output ParameterEstimates=out1;
proc phreg data = uis;
model time*censor(0) = treat off_trt tot;
if time <= los then off_trt = 0;
else off_trt = 1;
tot = treat*off_trt;
run;
```

Bibliography

[1] D. G. Kleinbaum and M. Klein, *Survival analysis: a self-learning text*. Springer Science & Business Media, 2006.

[2] D. R. Cox, "Regression models and life-tables," in *Breakthroughs in statistics*, pp. 527–541, Springer, 1992.

[3] J. Borucka, "Extensions of cox model for non-proportional hazards purpose," *PhUSE, Paper SP07. Warsaw: PAREXEL*, 2013.

[4] W. David, S. Hosmer, L. Stanley, and M. Susanne, "Applied survival analysis: regression modeling of time to event data," *Massachusetts: A Wiley interscience publication*, p. 167, 1999.

[5] R. G. Miller Jr, *Survival analysis*, vol. 66. John Wiley & Sons, 2011.

[6] J. D. Singer and J. B. Willett, *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.

[7] A. C. Leon, R. A. Friedman, J. A. Sweeney, R. P. Brown, and J. J. Mann, "Statistical issues in the identification of risk factors for suicidal behavior: the application of survival analysis," *Psychiatry research*, vol. 31, no. 1, pp. 99–108, 1990.

[8] P. D. Allison, *Survival analysis using SAS: a practical guide*. Sas Institute, 2010.