

Customer Value Management Using the Cox Additive Regression Model

by

Jan van Wyk de Vries

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Department of Statistics

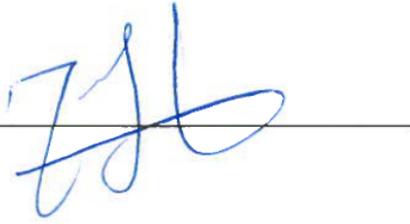
In the Faculty of Natural and Agricultural Sciences

University of Pretoria

December 2013

I, Jan van Wyk de Vries, declare that the thesis / dissertation, which I hereby submit for the degree Msc Mathematical Statistics at the University of Pretoria is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature

A handwritten signature in blue ink is written over a horizontal black line. The signature is stylized and appears to be 'JvdV'.

Date: 23 April 2014

ABSTRACT

The Cox-Proportional Hazards model is introduced as a potential tool to understand customer behavior relating to churn or disconnections in the telecommunications space. An overview of Survival Analysis is provided along with its associated quantities and metrics with examples to better illustrate concepts. The derivation of the classical Cox-Proportional Hazards model is discussed in detail and applied to network behavioural data. The development of additive models and generalised additive models are traced and described as a prelude to the additive Cox-Proportional Hazards Regression.

The cubic splines are used as a tool to automatically detect trends in the customer data and this is compared to the findings of the classic Cox-Proportional Hazard.

It is shown that using the cubic splines, trends are automatically detected in the data and the cubic spline functions themselves can be easily derived and implemented using the RCS macro. Insights and recommendations are reported on and made available to the Network for use in informing future retention strategies and in general, to better understand customer specific behaviour.

TABLE OF CONTENTS

ABSTRACT.....	4
TABLE OF CONTENTS.....	5
LIST OF FIGURES.....	8
LIST OF TABLES.....	9
LIST OF ABBREVIATIONS	10
1. Chapter 1 Introduction and Background	11
1.1. Motivation.....	11
1.2. Questions and expected findings.....	12
1.3. Approach.....	14
1.4. Notation	14
2. CHAPTER 2 – Survival Models and Hazard Functions	15
2.1.1. Introduction	15
2.1.2. The Survival Function.....	16
2.1.3. The Lifetime Distribution Function and Event Density	17
2.1.4. Estimating the Survival Function.....	18
2.1.4.1. Empirical Survival Function	18
2.1.4.2. Life Table Estimate of the Survival Function.....	18
2.1.4.3. Example using Life Table Estimate of the Survival Function.....	19
2.1.5. Hazards.....	21
2.1.5.1. Background to Hazards	21
2.1.5.2. Example of Hazards using Life Tables	22
2.1.5.3. Examples and Interpretation of Hazards	23
2.1.5.4. Examples of Hazard Calculation on the Myeloma Data	25
2.1.5.5. Properties of Hazard Functions.....	26
2.1.6. Censoring	28
2.1.6.1. Background on Censoring	28
2.1.6.2. Types of Censoring.....	28
2.1.6.2.1. Examples of Survival and Hazard Calculation with Censoring	29
2.1.6.3. Other types of Censoring	30

3.	CHAPTER 3 – The Cox Proportional Hazards Model for Survival Data.....	32
3.1.	Introduction	32
3.2.	Proportional Hazard Models.....	33
3.2.1.	Background	33
3.2.2.	The General Proportional Hazards Model	34
3.2.3.	Fitting the Cox Proportional Hazards Model.....	36
3.2.4.	Dealing with Ties in the Data	39
3.2.5.	Confidence Intervals and Hypothesis testing for β	40
3.2.6.	Example of Fitting the Proportional Hazards Model	42
3.2.7.	Model Selection	44
3.2.8.	Example using Selection Procedures	46
3.2.9.	Interpretation of the Parameter Estimates	47
3.2.10.	Estimating the Hazard and Survival Functions.....	47
3.2.10.1.	Estimating the Baseline Hazard	48
3.2.10.2.	Model Validation.....	49
3.2.10.2.1.	Introduction	49
3.2.10.2.2.	Cox-Snell Residuals.....	50
4.	CHAPTER 4 – A Background on Smoothers and Additive Models	53
4.1.	Smoothers and Smoothing	53
4.1.1.	Introduction	53
4.1.2.	Neighbourhoods	54
4.1.3.	Moving Average / Running-mean and Running Line Smoothers.....	55
4.1.4.	Kernel Smoothers.....	57
4.1.5.	Regression Splines.....	59
4.1.6.	Cubic Smoothing Splines.....	61
4.2.	Additive Models.....	63
4.2.1.	Introduction	63
4.2.2.	Review of Multiple Linear Regression.....	63
4.2.3.	Additive Models	65
4.2.4.	Fitting Additive Models.....	66
4.3.	Generalised Additive Models.....	68
4.3.1.	Introduction	68
4.3.2.	Review of Generalised Linear Models.....	69
4.3.3.	Fisher Scoring of Generalised Linear Models.....	69

5.	CHAPTER 5 –Additive Modelling for Survival Times	71
5.1.1.	Introduction	71
5.1.2.	Estimation	72
5.1.2.1.	Further details on the computation.....	76
5.1.3.	Inference and Smoothing Parameter Selection.....	77
5.1.4.	Degrees of Freedom.....	78
5.1.5.	Selecting a Smoothing Parameter.....	79
5.1.6.	Tests of Hypothesis	79
5.1.7.	A Specific Application of a Mixed Model	85
6.	CHAPTER 6 –An Application of Additive Modelling for Survival Times.....	89
6.1.	Introduction	89
6.2.	Data preparation.....	90
6.2.1.	Data received	90
6.2.2.	Data Manipulation Steps.....	91
6.2.2.1.	Introduction	91
6.2.2.2.	Steps.....	92
6.2.3.	Exploratory Analysis.....	93
6.2.3.1.	Introduction	93
6.2.3.2.	Exploratory data analysis	94
6.2.4.	Classic Proportional Hazards Modelling.....	102
6.2.4.1.	Introduction	102
6.2.4.2.	Modelling Strategy and Results	102
6.2.5.	Proportional Hazards Modelling using Cubic Spline functions	106
6.2.5.1.	Introduction	106
6.2.5.2.	Background	106
6.2.5.3.	The RCS Macro	107
6.2.5.4.	Modelling Strategy and Results	108
6.2.5.5.	Final Model	114
6.2.5.6.	Discussion.....	116
7.	Chapter 7 – Conclusion	117
8.	REFERENCES.....	119
9.	APPENDIX – Multiple Myeloma Data Set.....	121
10.	APPENDIX - SAS CODE.....	123

LIST OF FIGURES

Figure 1 Life Table Estimate of Survival Time	20
Figure 2 Graphic display of the life table data showing the Bathtub-shaped hazard.....	23
Figure 3 Life Table Estimate of Survival Time	26
Figure 4 Survival Times of Five Individuals	39
Figure 5 Log Hazard Ratio for Age.....	61
Figure 6 Disconnection and Size Distribution for Average Value	97
Figure 7 Disconnection and Size Distribution for Average Voice.....	98
Figure 8 Disconnection and Size Distribution for Average Data	98
Figure 9 Disconnection and Size Distribution for Average Duration	99
Figure 10 Disconnection and Size Distribution for Average SMS Events.....	99
Figure 11 Disconnection and Size Distribution for Average Megabytes.....	100
Figure 12 Disconnection and Size Distribution for Age.....	100
Figure 13 Disconnection and Size Distribution for Time to Upgrade.....	101
Figure 14 Disconnection and Size Distribution for Tenure	101
Figure 15 Survival Function for Disconnections.....	105
Figure 16 Log Hazard Ratio for Age.....	111
Figure 17 Log Hazard Ratio for Average Value	111
Figure 18 Log Hazard Ratio for Average SMS.....	112
Figure 19 Log Hazard Ratio for Average Data	112
Figure 20 Log Hazard Ratio for Average SMS.....	113
Figure 21 Log Hazard Ratio for Average SMS Events.....	113

LIST OF TABLES

Table 1 Life Table Estimate of the Survival Function for the Multiple Myeloma Data.....	20
Table 2 Hazards for mortality in the United States in 2000, shown as a life table	22
Table 3 Life Table Estimate of the Hazard Function for the Multiple Myeloma Data	25
Table 4 Tenure Data for Several Customers	29
Table 5 Tracking Customers over Several Time Periods	29
Table 6 from Times to Hazards	30
Table 7 SAS Output on Fitting the Proportional Hazards Model to the Example Data	44
Table 8 SAS Output on Fitting the Proportional Hazards Model to the Example Data using Stepwise Selection.....	46
Table 9 Number of Disconnected Subscribers per Month.....	89
Table 10 Number of Disconnected Subscribers per Month.....	95
Table 11 Summary Statistics of Independent Variables	95
Table 12 Correlation of Independent Variables.....	96
Table 13 Stepwise Selection	103
Table 14 MLE for Stepwise Selection	104
Table 15 Values for Reference Setting.....	105
Table 16 Percentiles of Continuous Variables	109
Table 17 MLE for Cox Model with Cubic Splines.....	110
Table 18 MLE of Stepwise Regression.....	114
Table 19 Reference set of Covariates for Plotting	115

LIST OF ABBREVIATIONS

CBI – Critical Business Issue

CDF – Cumulative density function

CVM – Customer Value Management

HF – Hazard Function

PDF – Probability density function

RCS – Regression Cubic Splines

SIM – Subscriber Identifier Module

SF – Survival / Survivor Function

1. Chapter 1 Introduction and Background

1.1. Motivation

The question of “how to best look after one’s own customers” is a challenge that is becoming more and more common for businesses in the economic climate of South-Africa and even the rest of the world. This is especially true with regards to customers who subscribe to some service of a bank, a retailer or a service provider in general.

Before the advent of the National Credit Act in 2007 and the global financial crisis of 2009, service providers were more focused on acquiring new business. After the global financial crisis, “Customer Value Management” became more and more important and specific skill sets were recruited and cultivated within big organisations with the sole purpose of understanding a customer’s behavior and to apply analytical insights to client internal data to ensure that the customer remains loyal to the brand. These analytical teams mined customer behavioral data to look at what are drivers for; amongst others – new sales, closing an account and falling behind on payments or reduction in spend.

Where in general, the organisational focus has been on acquiring and expanding the business, it is now more on maintaining market share and ensuring that customers do not move to competitors. Understanding customer spend and transactional patterns now are more important than before as these data elements can be used to predict future behavior.

Specific emphasis now is also on the CVM department to set specific targets relating to customer attrition or churn as well as to cross and up sell to existing customers. In the credit space, this would translate as managing customers better to ensure they are well educated on the credit facility they have and that they do not fall behind on payments. The key performance indicator of the CVM team would be managing the so-called bad rate and provisioning.

A unique situation exists in the telecommunications landscape in South Africa. Only as recently as fifteen years ago, the mobile phone was a piece of technology only seen in movies. Then suddenly it arrived and South-Africans started to take to this new medium of communication. When the networks launched pre-paid tariff plans, the uptake was enormous and over the following years, the networks saw a cell phone placed in the hand of every resident. (Hutton 2011). Now we have

a mobile penetration rate in excess of 100% - there are more active SIM-cards in circulation than persons in the country (Smith 2013).

The shift in focus from acquiring new business to managing existing business also applies to the networks. In South-Africa, there are now four main networks. With the larger networks, acquisition and expansion of the business happened automatically as the population adopted mobile phone technology. South-Africa now has a saturated market and a user population that has become very informed and cell-phone savvy to the extent that they make use of services such as mobile banking and even applying for credit via the mobile device.

These conditions are forcing the networks to look into their own customers' behavioural data to better built a relationship with that customer and to understand what factors or indicators could possibly influence that customer to disconnect from the network.

In this essay, the Cox-Proportional Hazards model will be introduced as a potential tool to understand customer behavior relating to churn or disconnections in the telecommunications space. Chapter two provides an overview of Survival Analysis, its associated quantities and metrics as well as some examples to better illustrate key concepts. Chapter three introduces and unpacks the classical Cox-Proportional Hazards model. In Chapter Four, the development of additive models is traced and described as a prelude to Chapter Five, where the additive Cox-Proportional Hazards model is discussed. An application on client data testing the questions discussed below is reported in Chapter Six, followed by a conclusion that contains key findings, next steps as well as insights and recommendations.

Varying coefficient modelling and analysis, accelerated time to failure models, and coding of routines in IML[®] are out of scope for this essay.

1.2. Questions and expected findings

The Cox-Proportional Hazards Model has long been a powerful tool in the analysis of survival or "time to event" data. This model describes the relationship of several covariates on the hazard rate under study. These covariates or independent variables are often modelled in a "linear fashion" i.e. the untransformed variable is included in the model.

Recent developments have pushed the area of Proportional Hazards modelling to include non-parametric or smooth functions as covariates to allegedly better capture underlying trends in the data.

In this essay, the following questions will be posed and described:

- By using the traditional Cox-Proportional Hazards model is it possible to develop a tool or gain insight into customer behavior with regards to disconnection?
- By using an additive Cox-Proportional Hazards model with cubic splines, is it possible to develop a tool or gain insight into customer behavior with regards to disconnection?
- Can these tools be used for future prediction?
- What are the advantages and disadvantages with each?
- Does the additive model really do a better job of uncovering trends in the data than the original Cox –Proportional hazards model?

1.3. Approach

Using behavioural data from a major South-Africa telecommunications network, the classic Cox-Proportional hazards model is applied as well as the Cox-Proportional additive model using SAS®.

Both models are fitted and evaluated using SAS and the results compared and discussed in order to answer the questions stated in the previous section.

1.4. Notation

The purpose of this section is to provide an overview of some of the mathematical notation used in this essay.

Random variables are represented in the upper case, for example: Y, X_1 , and the observed values in lower case e.g. y_{i1}, x_{i1} .

The $n \times 1$ column vector is denoted in bold as, $\mathbf{X}_1 = (X_{11}, X_{21}, \dots, X_{n1})^T$

The $n \times p$ matrix, denoting p random variables, is written as:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$$

2. CHAPTER 2 – Survival Models and Hazard Functions

2.1.1. Introduction

Survival analysis is a branch of statistics that deals with death in biological organisms or failure in mechanical systems. In engineering, survival analysis is known as reliability theory and in biology simply as survival analysis.

Survival analysis deals with the analysis and modelling of “time to event” data – i.e. it looks to answer the question: how long does it take for something to happen? Survival analysis can also answer questions such as which factors or covariates affect survival time and how, making it an incredibly powerful technique in applied statistics.

Some examples of questions that survival analysis seeks to answer are:

- What fraction of a population will survive past a certain time?
- Of those that survive, at what rate will they die / fail?
- Can multiple causes of death or failure be taken into account?
- How do particular circumstances or characteristics increase or decrease the odds of survival?

Survival analysis is by no means limited to only biological or mechanical applications. It can also be used to answer questions in the business setting, specifically in marketing or CVM that are faced with problems such as:

- When does the business need to start worrying about their customers leaving / closing their account?
- When is the next time that a customer is likely to migrate to a new customer segment?
- When is the next time that a customer is likely to broaden or narrow a customer relationship?
- Which are the factors in the customer relationship that increase or decrease the likely tenure of the customer?
- What is the quantitative effect of various factors on customer tenure? E.g. what could possibly cause customers to leave?

The answers to some of the above critical business issues (CBI) may be used to inform a customer retention strategy e.g. if one knows when a customer is likely to leave and how many are likely to go, the budget for a specific retention strategy can be set.

Also, studying customer behavior and understanding which factors impact customer tenure and affect it negatively can be used to define a retention strategy and offer different types of incentives to increase tenure to customers who will respond favourably.

These insights into customers feed directly into the marketing process e.g. how long will particular segments of customers stay and how profitable will they be?

In this section, the concept of survival analysis is explored as well as related concepts such as hazard probabilities and survival curves. Censoring, which is a problem often encountered in survival analysis, is also discussed in detail.

The next chapter will build on the concepts developed here and provide an overview of parametric survival analysis and modelling, e.g. the proportional hazards model by Cox and concludes by summarising and illustrating the value of extending the above concepts to the non-parametric scene.

For a more detailed discussion on survival analysis, refer to Collett D (2003).

2.1.2. The Survival Function

Key to survival is the *survival function*, also known as the *survivor function* or *reliability function* or *complementary cumulative distribution function*.

If a random variable X represents the lifetime of a unit, then the survival function of the unit at time t is defined to be:

$$S(t) = P[X > t] = 1 - F_X(t).$$

This function describes the probability that the system will survive beyond a specified time t .

Typically, the time $t = 0$ represents some origin such as the beginning of the study or the start of the operation of some system.

$S(0) = 1$ commonly but can be less to represent the probability that the system fails immediately upon operation.

The survival function is non-increasing i.e. $S(u) \leq S(t)$ for all $u > t$, since

$$S(t) = 1 - F_X(t).$$

The above reflects the concept that survival to a later age is only possible if all the younger ages have been attained. From this property, it follows that the lifetime and event density distributions are well defined. (Refer to the next section.)

Because X is nonnegative, distributions such as the Weibull, Gamma, Exponential and Lognormal are of interest.

The survival function usually approaches zero as time increases to infinity in the limit, although the limit could be greater if eternal life is possible. As an example, when studying a mixture of stable and unstable carbon isotopes, unstable isotopes would decay sooner, but stable isotopes would last indefinitely.

2.1.3. The Lifetime Distribution Function and Event Density

The *lifetime distribution* and *event density* functions are quantities closely related to the survival function.

The lifetime distribution function is the distribution function used to derive the survival function and is related to the survival function as follows i.e.

$$F_X(t) = P[X \leq t] = 1 - S(t).$$

The Lifetime distribution function, being the complement of the survival function, gives the probability of dying / failing before a certain time t .

The derivative of the Lifetime distribution function f , is known as the event density function

$$f(t) = \frac{d}{dt} F_X(t).$$

This is the rate of death or failure per unit time.

Similarly, a survival event density function can be defined as

$$s(t) = \frac{d}{dt} S(t) = \frac{d}{dt} [1 - F_X(t)] = -f(t).$$

2.1.4. Estimating the Survival Function

Several methods exist for estimating the survival function and two of these will be discussed in this section.

2.1.4.1. Empirical Survival Function

Given the definition of the survival function (the probability that a unit will survive beyond a specific point in time) and assuming that there are no censored observations, the survival function can be estimated by the following definition:

$$\hat{S}(t) = \frac{\sum_{j=1}^n I_j(t)}{n}.$$

where n is the number of individuals in the dataset and $I_j(t)$ is an indicator variable for individual i with values one if that individual has a survival time $\geq t$ and zero otherwise.

The above definition is a generalisation assuming that there are no censored observations in the dataset. The topic of censoring is addressed in section 2.1.6.

2.1.4.2. Life Table Estimate of the Survival Function

The life table estimate, which is also known as the actuarial estimate of the survival function, is calculated by first dividing the data into several intervals. The intervals need not be of equal length but often are in practice. The number of intervals will depend on the size of the sample and will usually range between 5 and 15 intervals.

Let the survival times be divided into m intervals such that the j^{th} interval is the time window t'_j to t'_{j+1} . Further assume, that in this interval the following quantities can be observed:

- n_j the number of individuals who are alive and therefore at risk of death at the start of the j^{th} interval.
- d_j the number of deaths in the j^{th} interval

- c_j the number of censored observations in the j^{th} interval

The life table approach assumes that the censored survival times occur uniformly during the j^{th} interval so that the average number of individuals at risk during this interval is:

$$n'_j = n_j - c_j/2.$$

This assumption is also known as the *actuarial assumption*. In the j^{th} interval, the probability of death can be estimated by d_j/n'_j and therefore the corresponding probability of surviving is

$$\frac{n'_j - c_j}{n'_j}.$$

The probability of an individual surviving beyond time t_k , $k = 1, 2, \dots, m$ that is until sometime after the start of the k^{th} interval, will be the product that the individual survives beyond the start of the k^{th} interval and through each of the $k - 1$ preceding intervals. Therefore the life table estimate of the survival function is given by:

$$\hat{S}(t) = \prod_{j=1}^k \left(\frac{n'_j - c_j}{n'_j} \right),$$

for $t_k \leq t < t_{k+1}$, $k = 1, 2, \dots, m$.

A graphical representation of the life table survival function will be a step function with constant values of the function at each interval.

The estimation of surviving to the start of the first interval will be one, as intuitively expected and the estimation of surviving beyond the last interval, e.g. t_{m+1} will be zero.

As expected, the estimates obtained are very sensitive to the selection of the intervals.

2.1.4.3. Example using Life Table Estimate of the Survival Function

In his book, *Modelling survival data in medical research (2003)*, David Collett makes use of data from a study carried out at the Medical Center of the University of West Virginia, USA. The data is on survival times of patients with multiple Myeloma, which is a malignant disease characterised by the accumulation of abnormal plasma cells in the bone marrow. The aim of the above study

was to determine the association between certain explanatory variables and survival times. For the purpose of illustration, this data set will be used in examples throughout this essay.

In the construction of the life table to estimate the survival function, the survival times were divided into intervals of 12 and the below results obtained:

Table 1 Life Table Estimate of the Survival Function for the Multiple Myeloma Data

Time Period	d_j	c_j	n_j	n'_j	$(n'_j - d_j)/n'_j$	$S^*(t)$
0-	16	4	48	46.0	0.6522	0.6522
12-	10	4	28	26.0	0.6154	0.4013
24-	1	0	14	14.0	0.9286	0.3727
36-	3	1	13	12.5	0.7600	0.2832
48-	2	2	9	8.0	0.7500	0.2124
60-	4	1	5	4.5	0.1111	0.0236

A graphical representation of the above estimate of the survival function yields the following:

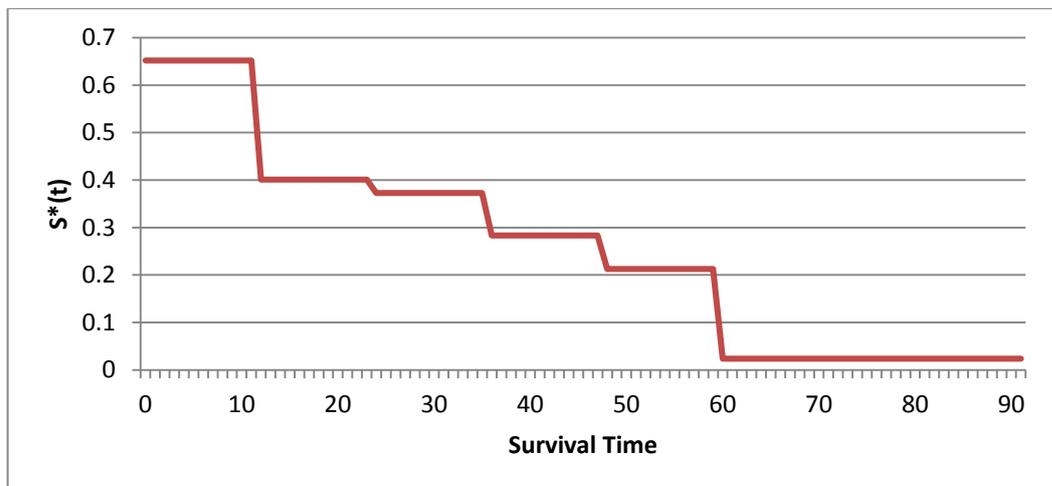


Figure 1 Life Table Estimate of Survival Time

2.1.5. Hazards

2.1.5.1. Background to Hazards

Vital to the analysis of survival data is the concept of hazards or hazard probabilities and hazard functions. These quantities can be directly linked back to the survival curve and will be explored in this section.

Theoretically, the hazard function or failure rate function for a PDF is defined as follows:

$$h(x) = \frac{f(x)}{1-F(x)} = \frac{\frac{dS(x)}{dx}}{S(x)} = \frac{d}{dx} [\ln(S(x))]. \quad \text{from (Bain and Engelhardt, 1992)}$$

From the above definition, it can be seen that the hazard function is a property of a distribution, much like the mean and the variance.

Force of mortality in Actuarial Science is a synonym for the hazard function.

In simpler terms, the hazard function can be interpreted as the probability of having survived up to time x and then failing in time $x + \Delta$, where Δ is a small positive quantity. In the context of marketing, this would equate to a customer having survived (has tenure) until a time x and the probability of that customer leaving before time $x + \Delta x$.

Another way of stating this is to say; the hazard at time t is the risk of losing customers between time t and time $t + \Delta t$.

Mathematically, the hazard function can be expressed as a conditional probability and interpreted as the instantaneous failure rate, or conditional density of failure at time x , given that the unit has survived until time x .

The conditional probability of an event A given the event B is given by:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}.$$

From Bain and Engelhardt (1992), the following derivation for the hazard function is given:

$$h(x) = f(x|X \geq x)$$

$$\begin{aligned}
&= \frac{dF(x|X \geq x)}{dx} \\
&= \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x|X \geq x) - F(x|X \geq x)}{\Delta x} \\
&= \lim_{\Delta x \rightarrow 0} \frac{P[x \leq X \leq x + \Delta x|X \geq x]}{\Delta x} \\
&= \lim_{\Delta x \rightarrow 0} \frac{P[x \leq X \leq x + \Delta x, X \geq x]}{\Delta x P[X \geq x]} \\
&= \lim_{\Delta x \rightarrow 0} \frac{P[x \leq X \leq x + \Delta x]}{\Delta x [1 - F(x)]} \\
&= \frac{f(x)}{1 - F(x)}.
\end{aligned}$$

The hazard probability summarises the survival data and reflects the information already present in the data. No special function is fitted to the data to aid interpretation.

2.1.5.2. Example of Hazards using Life Tables

To better understand the idea of hazard probabilities and functions and illustrative example using US life tables is given below.

These tables describe the probability of someone dying at a particular age for the U.S population in 2000 from (Berry and Linoff 2004)

Table 2 Hazards for mortality in the United States in 2000, shown as a life table

Age	Percent Of Population That Dies In Each Age Range
0 – 1 yrs	0.73%
1 – 4 yrs	0.03%
5 – 9 yrs	0.02%
10 – 14 yrs	0.02%
15 – 19 yrs	0.07%
20 – 24 yrs	0.10%
25 – 29 yrs	0.10%
30 – 34 yrs	0.12%
35 – 39 yrs	0.16%

40 – 44 yrs	0.24%
45 – 49 yrs	0.36%
50 – 54 yrs	0.52%
55 – 59 yrs	0.80%
60 – 64 yrs	1.26%
65 – 69 yrs	1.93%
70 – 74 yrs	2.97%
75 – 79 yrs	4.56%
80 – 84 yrs	7.40%
85 yrs or more	15.32%

Life tables are good examples of hazards. From the above table it can be seen that infants have a 0.73% chance of not surviving past one year of age. Thereafter, the mortality rate decreases and then starts to steadily increase after age 45. This is an example of a “bathtub” shape for hazards where hazards start high, drop quickly and then gradually increase.

The below figure illustrates the characteristic ‘bathtub’ shape of this data:

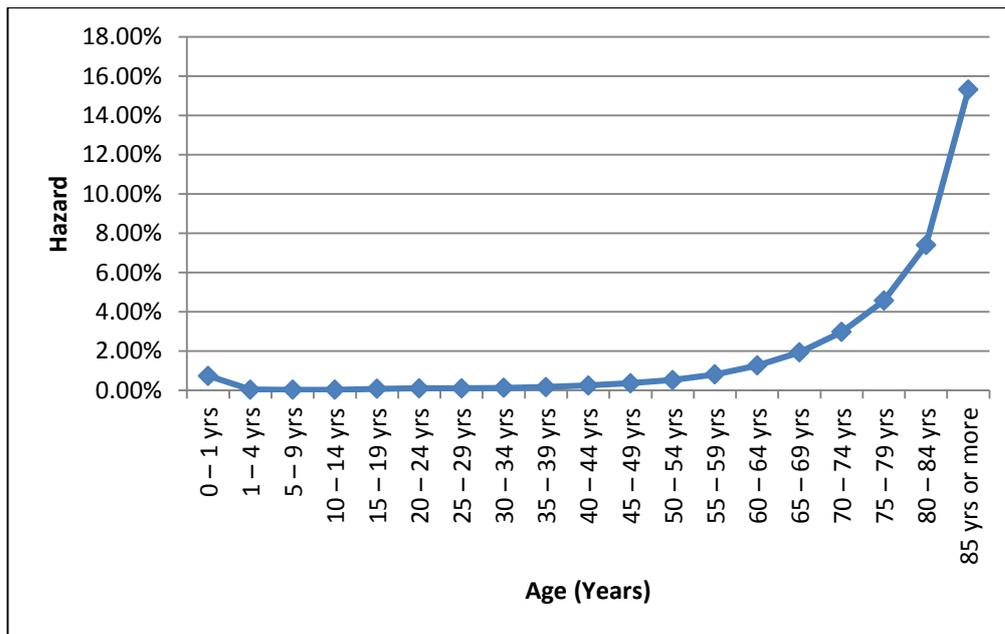


Figure 2 Graphic display of the life table data showing the Bathtub-shaped hazard

2.1.5.3. Examples and Interpretation of Hazards

In practice there are some typical examples of hazard functions. Each of these is discussed briefly below.

- The Bathtub hazard

The bathtub hazard, described above, is an example of a hazard function often encountered in practice. As described earlier, a bathtub hazard starts with high hazards which then drop quickly and eventually then gradually increase.

Customers who are on cell-phone contracts typically cause a bathtub like hazard – early in the contract, the customer cancels the contract because the service is bad or they do not know that they have actually taken a contract or they do not pay their account. Thereafter, they tend to stay with their service provider. Towards the end of the contract lifetime, customers then churn again, having fulfilled the obligation of the contract, and move to another network.

- A Constant Hazard

A constant hazard function is typically more appropriate in physics rather than business sciences e.g. the half-life property of decaying uranium. What the constant hazard describes is a hazard that is always the same – no matter what happened in the past, e.g. in customer lifetime analysis, the chance of that customer leaving is always the same, no matter how long that customer has been around. This is in effect the “no memory” property of the hazard function derived from the Exponential distribution.

For example, if $X \sim EXP(\theta)$ then the hazard function of X is

$$\begin{aligned} h(x) &= \frac{f(x)}{1 - F(x)} \\ &= \frac{\frac{1}{\theta} e^{-\frac{x}{\theta}}}{e^{-\frac{x}{\theta}}} \\ &= \frac{1}{\theta}. \end{aligned}$$

The above quantity does not depend on time or the age of the unit.

Typically, an increasing hazard function at time x means that the unit is more likely to fail in the next increment of time i.e. the unit is wearing out with age.

A decreasing hazard function at time x can be interpreted as a unit that is getting better with time.

2.1.5.4. Examples of Hazard Calculation on the Myeloma Data

Suppose that the data again have been grouped into m intervals as in the life table example. An appropriate estimate of the average hazard of death per unit time over each interval is the observed number of deaths in that interval divided by the average time survived in that interval which is also the average number of persons at risk in that interval multiplied by the length of the interval.

Following on the quantities defined for the life table estimate in section 2.1.4.3, define

- τ_j the length of the j^{th} time interval.

Assuming that the death rate during the j^{th} interval is constant, the life table estimate of the hazard function is then given by:

$$h^*(t) = \frac{d_j}{(n'_j - d_j/2)\tau_j},$$

for $t_j \leq t < t_{j+1}$, $j = 1, 2, \dots, m$.

$h^*(t)$ will also be a step function.

Expanding on the table using the estimation of the survival function for the Myeloma data, the estimated hazard function is calculated as:

Table 3 Life Table Estimate of the Hazard Function for the Multiple Myeloma Data

Time Period	d_j	c_j	n_j	n'_j	$(n'_j - d_j)/n'_j$	$S^*(t)$	τ_j	$h^*(t)$
0-	16	4	48	46.0	0.6522	0.6522	12	0.0351
12-	10	4	28	26.0	0.6154	0.4013	12	0.0397
24-	1	0	14	14.0	0.9286	0.3727	12	0.0062
36-	3	1	13	12.5	0.7600	0.2832	12	0.0227
48-	2	2	9	8.0	0.7500	0.2124	12	0.0238
60-	4	1	5	4.5	0.1111	0.0236	36	0.0444

The estimated hazard function is plotted below. From the graph, it can be seen that the hazards remain roughly constant over the first two years of analysis, after which time it declines and then gradually increases again.

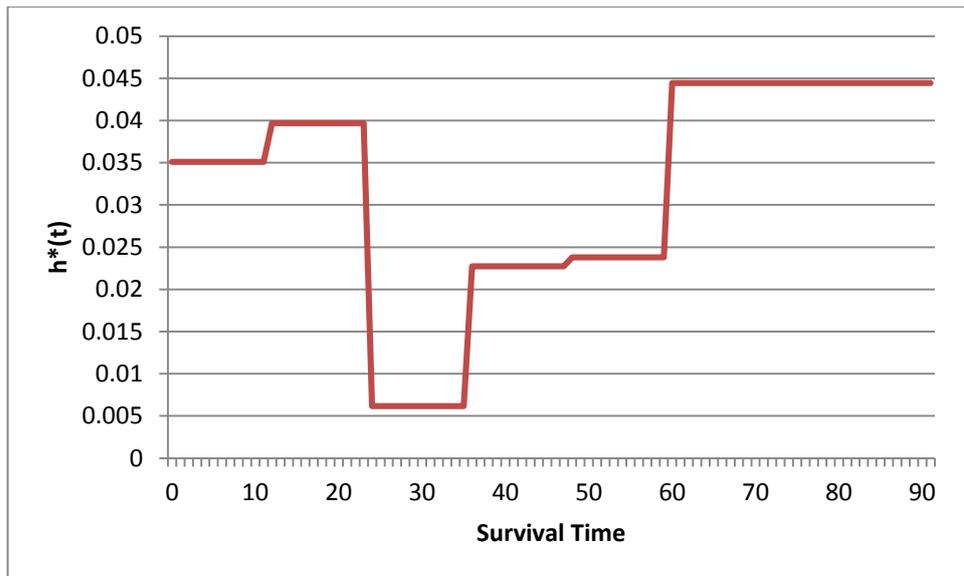


Figure 3 Life Table Estimate of Survival Time

2.1.5.5. Properties of Hazard Functions

The hazard function must be non-negative (because it is a rate) and its integral over $[0, \infty]$ must be infinite, but not otherwise constrained. The proof is given below in property 2.

It can be increasing or decreasing, non-monotonic or discontinuous. A hazard function completely defines a CDF and vice versa. The below properties are taken from Bain and Engelhardt (1992).

Property 1 - For any hazard function, $h(x)$, the associated CDF is determined by the relationship

$$F(x) = 1 - e^{-\int_0^x h(t)dt},$$

or

$$f(x) = h(x)e^{-\int_0^x h(t)dt}.$$

Proof

The result follows from the relationship:

$$h(x) = \frac{d}{dx} [\ln(S(x))],$$

and

$$\int_0^x h(t)dt = - \int_0^x \frac{d}{dt} [\ln(S(t))] dt$$

$$= -\ln(S(x)).$$

Property 2 - The hazard function must be non-negative (because it is a rate) and its integral over $[0, \infty]$ must be infinite:

1. $h(x) \geq 0 \forall x$
2. $\int_0^{\infty} h(t)dt = \infty$

Proof

The above properties are necessary because:

$$\frac{f(x)}{1 - F(x)} \geq 0,$$

and

$$\int_0^{\infty} h(t)dt = \int_0^{\infty} \frac{d}{dt} [\ln(S(t))] dt$$

$$= \ln(S(t)) \Big|_0^{\infty}$$

$$= \infty.$$

These properties are sufficient because the resulting CDF will be a valid CDF in terms of $h(x)$.

$$F(-\infty) = F(0) = 1 - e^{-\int_0^0 h(t)dt} = 0,$$

$$F(\infty) = 1 - e^{-\int_0^{\infty} h(t)dt} = 1,$$

and $F(x)$ is an increasing function of x because $\int_0^x h(t)dt$ is an increasing function of x .

2.1.6. Censoring

2.1.6.1. Background on Censoring

One aspect of survival analysis is that of censoring or censored sampling. Censoring happens when some of the data is missing and the complete survival time cannot be recorded. There are different types of censoring depending on the nature of the experiment and the ultimate aim of the study.

Sometimes censoring can be intentional e.g. when testing a specific component, it could take years for all the units in the study to break / expire. In this case, the study may not be economically feasible in terms of funding and time and it can be agreed at the onset to terminate the experiment either after a certain number of components have failed, or after a specific amount of time has passed.

2.1.6.2. Types of Censoring

It is practical to also look at censoring in terms of order statistics - Bain and Engelhardt, (1992). If a random sample of n units is used for a survival analysis, then the first observed failure time is automatically the first order statistic $x_{1:n}$. Similarly, the second observed failure time is the second order statistic and so one can continue until all failure times have been observed. Should this be the case, it is known as *complete sampling*.

Type II censored sampling on the right occurs when the experiment is stopped after the first r observations have been obtained.

Type II censored sampling on the left happens if for some reason the first s observations are not available.

Type I censored sample or truncated sampling occurs when the experiment is terminated after a fixed time x_0 .

As all observations are naturally ordered, not all the information is lost for the censored observations. In the case of Type II censoring, it is known that the survival time is at least x_0 .

Some cases may also result due to calibration of equipment e.g. a scale that can only measure up to 100kg. For a unit that weighs more than 100kgs in the study, it will only be known that its weight is at least 100 kg.

2.1.6.2.1. Examples of Survival and Hazard Calculation with Censoring

When looking at customer data in a marketing example, there are three possible outcomes – ACTIVE, CENSORED and STOPPED.

ACTIVE indicates that the customer relationship is still ongoing whereas STOPPED indicates that the customer relationship has ceased in that time interval because the customer has churned or cancelled the deal. CENSORED means that the customer is not included in the calculation.

The below is a hypothetical example from Berry and Linoff (2004).

Table 4 Tenure Data for Several Customers

CUSTOMER	CENSORED	TENURE
2	N	4
3	N	3
4	Y	3
5	N	2
6	Y	1
7	N	1

The above summary of customer tenure can be represented as follows:

Table 5 Tracking Customers over Several Time Periods

CUSTOMER	CENSORED	LIFETIME	TIME0	TIME1	TIME2	TIME3	TIME4	TIME5
1	Y	5	ACTIVE	ACTIVE	ACTIVE	ACTIVE	ACTIVE	ACTIVE
2	N	4	ACTIVE	ACTIVE	ACTIVE	ACTIVE	STOPPED	CENSORED
3	N	3	ACTIVE	ACTIVE	ACTIVE	STOPPED	CENSORED	CENSORED
4	Y	3	ACTIVE	ACTIVE	ACTIVE	ACTIVE	CENSORED	CENSORED
5	N	2	ACTIVE	ACTIVE	STOPPED	CENSORED	CENSORED	CENSORED
6	Y	1	ACTIVE	ACTIVE	CENSORED	CENSORED	CENSORED	CENSORED
7	N	1	ACTIVE	STOPPED	CENSORED	CENSORED	CENSORED	CENSORED

To compute the hazards for the example, one would proceed as follows:

Table 6 from Times to Hazards

CUSTOMER	TIME0	TIME1	TIME2	TIME3	TIME4	TIME5
ACTIVE	7	6	4	3	1	1
STOPPED	0	1	1	1	1	0
CENSORED	0	0	2	3	5	5
HAZARD	0%	14%	20%	25%	50%	0%

Notice that the censoring always takes place one unit after the lifetime, as intuitively expected. The first customer survived until time 5, what happened after this time period is unknown.

The Hazard at any given point in time is the number of customers that have stopped, divided by the sum of number of customers that are active and that have stopped in that time period.

2.1.6.3. Other types of Censoring

There can be other reasons why it might be necessary to exclude or censor observations in a lifetime study. To best illustrate this, a medical example is considered.

Suppose cancer patients in a survival study are treated with a new medicine that eliminates a type of cancer cell and allegedly increases the cure rate. Say these patients are observed for 10 years after having undergone treatment. Some of these patients die in a ski accident within this ten year period. Technically, these patients died not because of the cancer but another cause.

This is an example of competing risks. Competing risks arise when there are multiple causes for failure unrelated to the study – a patient can die of cancer, a patient can also die in a ski accident.

Also, patients can move without notifying the researcher of the change in address. In this case, these patients have been “lost to follow up.”

In both of the cases above, it is known that the patients were healthy up to the point where they either moved away or died in a ski accident. This information is useful and should be included in the study to calculate hazards. Afterward, one cannot know what happened and the observations are censored at the point when they exit the study. If a patient dies of some other cause (e.g. a ski accident) then the patient will be censored at time of death and this death will not be included in the hazard calculation.

Competing risks often happen in the business environment – especially in customer retention and churn. There are two main types of churn; voluntary and involuntary churn. The prior happens when a customer decides of their own accord to leave, and the latter when a customer is forced to leave e.g. if not paying their bills.

Voluntary and Involuntary churn should be analysed separately and each should have its own set of hazard curves.

3. CHAPTER 3 – The Cox Proportional Hazards Model for Survival Data

3.1. Introduction

In survival models, the aim is to relate the “time to event” dependent variable to one or more independent covariates which are relevant to the study. The values of these covariates or explanatory variables have been recorded at the onset of the study for each individual.

There are two main reasons why one would be interested in modelling survival data –

1. To answer the question of which set or subset of explanatory variables affect the hazard function, and in particular, what the effect of this is on the hazard of death
2. To obtain an estimate of the hazard function itself for an individual and use this for forecasting or preventative measures (e.g. to inform a strategy for a retentions campaign for individuals who are likely to move their business away)

In the analysis of survival data, the interest is mainly on the risk or the hazard of death at any time after the onset of the study. It is for this reason that the focus is more on modelling the hazard function rather than the survival function.

An important class of models in survival analysis is the proportional hazard models in which the unit increase in a covariate is multiplicative with respect to the hazard rate. For example, customers who take out a certain type of contract may have a higher hazard rate than customers on another type of contract.

Other types of survival models such as the accelerated failure time’s model are available and may be applicable where the assumption of proportional hazards does not hold. For example, a situation where a drug reduces a subject’s immediate risk of going into remission from cancer, but where there is no reduction in the hazard rate after one year for subjects who do not go into remission in the first year of the analysis.

3.2. Proportional Hazard Models

3.2.1. Background

The proportional hazards model was first introduced by Sir David Cox (Cox 1972). He studied the effects of initial factors (time zero covariates) on hazards extensively. By making the assumption that these factors have a uniform proportional effect on the hazards themselves, he was able to understand how to measure these effects for different factors.

The Cox model has the form:

$$\lambda(t|X) = \lambda_0(t)e^{(X\beta)}.$$

where $\lambda(t|X)$ is the hazard at time t for an individual with explanatory variable (risk profile) X , $\lambda_0(t)$ is the baseline hazard and $e^{(X\beta)}$ is a weighted function of the individual variables of the individual's risk profile.

Proportional hazards are often quoted in public, e.g. *the risk of leukemia for smokers is 1.53 times higher than for nonsmokers* etc. In this example, at the beginning of the study the researchers knew whether an individual was a smoker or not. This is an example of an initial condition.

Initial conditions can also be analysed by looking at the hazard rates for each level e.g. hazard rates of customers who signed up over an internet channel vs. customers who signed up using a mobile device.

In both of the above examples, the initial conditions are described by categorical variables i.e. smoker vs non smoker, mobile vs internet channel etc. It is also possible to have an initial condition as a continuous variable. Consider another statement about the dangers of smoking: *the risk of colon cancer increases 6.7% per pack year smoked*. With proportional hazards it is possible to determine the input of both continuous and categorical variables. In the case of continuous variables, the hazard responds logarithmically.

Survival models can be viewed as consisting of two parts namely:

- $h_0(t)$ - the underlying hazard function. This function describes how the hazards change over time at baseline levels of the covariates
- The effect variables, describing how the hazard varies in response to explanatory variables

In a marketing example, covariates could be account type as well as demographic characteristics such as age, gender and income.

As mentioned earlier, the *proportional hazards condition* states that the covariates are multiplicatively related to the hazard. As a basic example, a treatment may halve a subject's hazard at time t , while the baseline hazard may vary.

Sir David Cox found that if the proportional hazard assumption is assumed to hold, then it is possible to estimate the effect of the parameters without any consideration of the hazard function. This approach to survival data is called the application of the ***Cox proportional hazards model***, also known as ***Cox model*** or ***proportional hazards model***.

3.2.2. The General Proportional Hazards Model

Suppose that two treatments are being investigated and each has its own set of hazards associated with it:

$h_1(t)$ for treatment one, and $h_2(t)$ for treatment two.

Assume further that the hazard at time t for a patient on treatment two is proportional to the hazards at the same time for a patient on treatment one, i.e.

$h_1(t) = \varphi h_2(t)$ where φ is a constant.

The implication of this assumption is that the corresponding survival curves will not cross.

The value φ is the ratio of the hazards of death at any time for an individual on treatment two relative to treatment one i.e.

$$\varphi = \frac{h_1(t)}{h_2(t)}$$

This quantity is the relative hazard or the hazard ratio.

Since a hazard ratio cannot be negative, it is convenient to set $\varphi = e^\beta$

Let X be an indicator variable with values 0 if the patient is on treatment one, and 1 if the patient is following treatment two.

If X_i is the value of X for the i^{th} individual in the study, the hazard function for this individual can then be written as

$$h_i(t) = e^{\beta x_i} h_0(t).$$

To set the scene for the general proportional hazards model, assume that the hazard of death or failure at a particular point in time of a sample of n individuals depends on the explanatory variables X_1, X_2, \dots, X_p represented by the vector \mathbf{X} .

Let $h_0(t)$ be the hazard function for an individual for whom the values of all the explanatory variables in \mathbf{X} are zero. This quantity is called the baseline hazard function.

The Hazard function for the i^{th} individual in the sample is then given by:

$$h_i(t) = \varphi(\mathbf{X}_i) h_0(t).$$

where $\varphi(\mathbf{X}_i)$ is a function of the values of the vector \mathbf{X} . Since the relative hazard, $\varphi(\mathbf{X}_i)$ cannot be negative, it is convenient to assume e^{η_i} where η_i is a linear combination of the p explanatory or independent variables in \mathbf{X} . Therefore

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} \text{ or}$$

$$\eta_i = \sum_{j=1}^p \beta_j X_{ij}.$$

This quantity is also known as the linear component, the risk score or the prognostic index for the i^{th} individual.

The general proportional hazard model then becomes:

$$h_i(t) = e^{\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}} h_0(t).$$

Because this model can also be expressed in the form below, the proportional hazards model can be regarded as a linear model for the logarithm of the hazard ratio.

$$\ln \left(\frac{h_i(t)}{h_0(t)} \right) = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}.$$

Notice that there are no constant terms in the proportional hazards model – this is because the addition of a constant will simply scale the base line function by dividing $h_0(t)$ by the exponent of the constant and the constant term will simply be cancelled out. Notice also that no assumptions have been made regarding the actual form of the baseline function.

3.2.3. Fitting the Cox Proportional Hazards Model

Fitting the proportional hazards model involves estimating the coefficients, $\beta_1, \beta_2, \dots, \beta_p$ as well as the baseline function. These two components can be estimated separately by first estimating the coefficients, $\beta_1, \beta_2, \dots, \beta_p$ and then using the estimated coefficients to arrive at an estimate for the baseline hazard function.

To estimate the coefficients, $\beta_1, \beta_2, \dots, \beta_p$ the method of maximum likelihood is used, using sample data and maximising the likelihood function.

Suppose that data has been observed for n individuals, within which there are r distinct death times and $n - r$ right censored survival times. For the moment, it is assumed that there are no ties in the data.

The r death times are denoted as $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ so that $t_{(j)}$ is the j^{th} ordered death time.

The set of individuals who are at risk at time $t_{(j)}$ is denoted by $R(t_{(j)})$ - i.e. $R(t_{(j)})$ is the set of individuals who are alive and uncensored just before time $t_{(j)}$.

Cox (1972) showed that the likelihood for the proportional hazards model is then:

$$L(\beta) = \prod_{j=1}^r \frac{e^{(X_j^T \beta)}}{\sum_{l \in R(t_{(j)})} e^{(X_l^T \beta)'}}$$

where X_j is the vector of covariates of the individual who dies at the j^{th} ordered death time $t_{(j)}$. The summation in the denominator of the likelihood is the sum of all individuals who are at risk at time $t_{(j)}$. Notice that the product is taken for those individuals where a death time has been recorded – individuals whose survival times have been censored are excluded from the numerator. Notice also that the likelihood depends on the ranking of the death times as this determines the risk set at each death time.

Suppose now that the data consist of n observed survival times, denoted by t_1, t_2, \dots, t_n .

Define δ_i as an indicator variable that equals zero if the i^{th} survival time t_i $i = 1, 2, \dots, n$ is right censored and one otherwise.

The likelihood can then be expressed as follows:

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{e^{(X_i^T \beta)}}{\sum_{l \in R(t_i)} e^{(X_l^T \beta)}} \right\}^{\delta_i}.$$

The corresponding log-likelihood is then given by:

$$\log L(\beta) = \sum_{i=1}^n \delta_i \{ X_i^T \beta - \log \sum_{l \in R(t_i)} e^{(X_l^T \beta)} \}.$$

The above function is maximised to obtain the maximum likelihood estimates of the parameters.

David Collett (2003) gives the following rationale for the likelihood above - consider the probability that the i^{th} individual dies at some time $t_{(j)}$, conditional on $t_{(j)}$ being one of the observed ordered set of r death times, $t_{(1)}, t_{(2)}, \dots, t_{(r)}$. If the vector of explanatory variables or risk profile for the individual who dies at $t_{(j)}$ is denoted by x_j , this probability is

$$P(\text{individual with explanatory variables } x_j \text{ dies at } t_{(j)} | \text{one death at } t_{(j)})$$

Using the definition of conditional probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}.$$

One can write:

$$P(\text{individual with explanatory variables } x_j \text{ dies at } t_{(j)} | \text{one death at } t_{(j)}) =$$

$$\frac{P(\text{individual with explanatory variables } x_j \text{ dies at } t_{(j)})}{P(\text{one death at } t_{(j)})}.$$

Since the death times are assumed to be independent of one another, the denominator in this quantity can be replaced by the sum of the probabilities of death at time $t_{(j)}$ over all individuals who are at risk of death at that time (i.e. the risk set). If these individuals are indexed by l , and the corresponding risk set at time $t_{(j)}$ by $R(t_{(j)})$, the expression becomes:

$$\frac{P(\text{individual with explanatory variables } \mathbf{x}_j \text{ dies at } t_{(j)})}{\sum_{l \in R(t_{(i)})} P(\text{individual } l \text{ dies at } t_{(j)})}$$

The probabilities of death at time $t_{(j)}$ in the above expression are now replaced by probabilities of death in the interval $(t_{(j)}, t_{(j)} + \delta t)$ and dividing both the numerator and the denominator by δt the following result is obtained:

$$\frac{P(\text{individual with variables } \mathbf{x}_j \text{ dies in } (t_{(j)}, t_{(j)} + \delta t)) / \delta t}{\sum_{l \in R(t_{(i)})} P(\text{individual } l \text{ dies in } (t_{(j)}, t_{(j)} + \delta t)) / \delta t}$$

Letting $\delta t \rightarrow 0$, the ratio of the corresponding hazards of death at time $t_{(j)}$ is obtained:

$$\frac{\text{Hazard of death for individual with variables } \mathbf{x}_j \text{ at } t_{(j)}}{\sum_{l \in R(t_{(i)})} \{\text{Hazard of death for individual } l \text{ at } t_{(j)}\}}$$

If the i^{th} individual dies at time $t_{(j)}$, then the hazard function in the numerator of this expression can be written as $h_i(t_{(j)})$ and similarly the denominator is the sum of the hazards of death at time $t_{(j)}$ over all individuals who are at risk of death at this time i.e. the sum of the values $h_i(t_{(j)})$ over all individuals in the risk set $R(t_{(j)})$ at time $t_{(j)}$. Consequently the conditional probability becomes:

$$\frac{h_i(t_{(j)})}{\sum_{l \in R(t_{(i)})} \{h_l(t_{(j)})\}}$$

Substituting the Proportional Hazards Model into the above:

$$\begin{aligned} \frac{h_i(t_{(j)})}{\sum_{l \in R(t_{(i)})} \{h_l(t_{(j)})\}} &= \frac{e^{\mathbf{x}_j^T \beta} h_0(t)}{\sum_{l \in R(t_{(i)})} \{e^{\mathbf{x}_l^T \beta} h_0(t)\}} \\ &= \frac{e^{\mathbf{x}_j^T \beta}}{\sum_{l \in R(t_{(i)})} \{e^{\mathbf{x}_l^T \beta}\}} \end{aligned}$$

Taking the product over these conditional probabilities over the r death times gives the likelihood function.

To illustrate the structure of the partial likelihood, an example from *Modelling of Survival Data* will be used.

Consider a sample consisting of 5 individuals, numbered 1 to 5 with survival data as illustrated in the figure below:

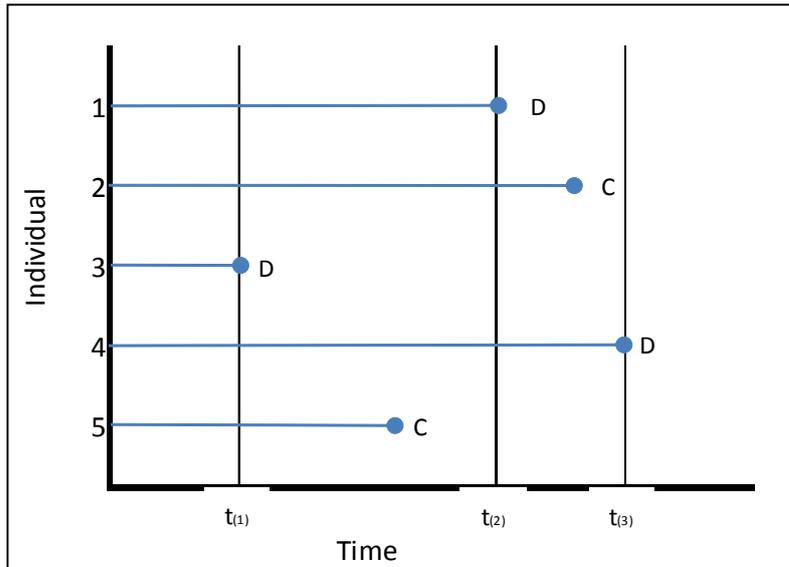


Figure 4 Survival Times of Five Individuals

The observed survival times of individuals 2 and 5 will be taken to be right censored, and the three ordered death times are denoted by $t_{(1)} < t_{(2)} < t_{(3)}$, so that $t_{(1)}$ is the death time of individual 3, $t_{(2)}$ is that of individual 1 and $t_{(3)}$ that of individual 4.

The risk set at each of the three ordered death times then consists of the individuals who are alive and uncensored just before each death time. Therefore, $R(t_{(1)})$ contains all 5 individuals, $R(t_{(2)})$ contains individuals 1,2 and 4 and $R(t_{(3)})$ has only one individual.

Let $\varphi(i) = e^{x_j^T \beta}$ for $i = 1, 2, \dots, 5$ for the risk score of the i^{th} individual.

The partial likelihood function over the three death times is then:

$$\frac{\varphi(3)}{\varphi(1) + \varphi(2) + \varphi(3) + \varphi(4) + \varphi(5)} \times \frac{\varphi(1)}{\varphi(1) + \varphi(2) + \varphi(4)} \times \frac{\varphi(4)}{\varphi(4)}$$

This likelihood function is not a true likelihood function in the traditional definition of likelihood as it does not make direct use of the actual censored and uncensored survival times. For this reason, it is referred to as a partial likelihood.

3.2.4. Dealing with Ties in the Data

It is assumed that the data in a survival analysis is continuous, however in practice survival times are often rounded to the nearest day or month and censoring and death can occur at the same time. Because of this it is possible to encounter ties in the data. The example on Multiple Myeloma for instance has tied survival times.

The likelihood function developed in the previous section assumes no ties in the survival times. Several approaches has been put forward to deal with tied times. The appropriate likelihood function in the presence of tied observations was developed by Kalbfleisch and Prentice (2002) but has a very complicated form and is computationally very taxing, especially when there are many tied times.

Several approximations to the partial likelihood have been developed that are easier to compute and to understand. Two of these will be discussed in this section.

Let \mathbf{S}_j be the vector of sums of each of the p covariates for those individuals who die at the j^{th} death time $t_{(j)}$ $j = 1, 2, \dots, r$

If there are d_j deaths at $t_{(j)}$, the h^{th} element of \mathbf{S}_j is $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$ where x_{hjk} is the value of the h^{th} explanatory variable $h = 1, 2, \dots, p$ for the k^{th} of the d_j individuals, $k = 1, 2, \dots, d_j$ who die at the j^{th} death time $j = 1, 2, \dots, r$.

The simplest approximation was given by Breslow (1974) who proposed the approximate likelihood:

$$\prod_{j=1}^r \frac{e^{(\mathbf{S}_j^T \boldsymbol{\beta})}}{\{\sum_{l \in R(t_{(j)})} e^{(\mathbf{X}_l^T \boldsymbol{\beta})}\}^{d_j}}$$

This likelihood function is easy to compute and reasonable accurate when the number of tied observations at a given death time is not too large. For this reason, this method is usually the default procedure for handling tied survival times in statistical software procedures. When there are no ties in the survival times, the above approximation reduces to the original partial likelihood function.

3.2.5. Confidence Intervals and Hypothesis testing for $\boldsymbol{\beta}$

In order to assess the contribution of the parameters estimated for a proportional hazards model, it is necessary to develop tests of hypotheses for the parameters and their associated confidence intervals.

It is important to revisit some of the quantities derived and used in the calculation of the maximum likelihood estimates. Consider the setting where n observations have been observed and are used to estimate the values of p unknown parameters: $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]$ with corresponding $\boldsymbol{\beta}$ function as described in the previous section, $L(\boldsymbol{\beta})$.

The values of $\boldsymbol{\beta}$ that maximise the maximum $L(\boldsymbol{\beta})$ are found by solving the system of p equations simultaneously:

$$\frac{d \log L(\boldsymbol{\beta})}{d \beta_j} = 0, \text{ for } j = 1, 2, \dots, p$$

Let the values of $\boldsymbol{\beta}$ that maximise $L(\boldsymbol{\beta})$ be denoted by $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p]$ so that

$$\max(L(\boldsymbol{\beta})) = L(\hat{\boldsymbol{\beta}}).$$

The *efficient score* for $\beta_j, j = 1, 2, \dots, p$ is:

$$u(\beta_j) = \frac{d \log L(\boldsymbol{\beta})}{d \beta_j}.$$

When the efficient scores are stacked to form the p -component vector, this is:

$\mathbf{u}(\boldsymbol{\beta})$, and the vector of maximum likelihood estimates can then be expressed as:

$$\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Denote the *Hessian Matrix*, or $p \times p$ matrix of second order partial derivatives of the log-likelihood function by $\mathbf{H}(\boldsymbol{\beta})$ with the $(i, j)^{th}$ element:

$$\mathbf{H}(\boldsymbol{\beta})_{ij} = \frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}.$$

The information matrix is obtained from the relationship:

$$\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta}).$$

The $(i, j)^{th}$ element of the *corresponding expected information* matrix is given by:

$$-E \left[\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right].$$

To construct tests of hypothesis for $\hat{\boldsymbol{\beta}}$, it is first necessary to derive the variance of these estimates, $var(\hat{\boldsymbol{\beta}})$. The inverse of the observed information matrix, evaluated at $\hat{\boldsymbol{\beta}}$ can be used for a reasonable approximation of the $p \times p$ variance-covariance matrix:

$$var(\hat{\boldsymbol{\beta}}) \approx \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}).$$

It follows that the square roots of the diagonal elements will yield the standard errors for $\hat{\boldsymbol{\beta}}$:

$$SE(\beta_j) = \sqrt{I^{-1}(\hat{\boldsymbol{\beta}})_{jj}}.$$

The **Likelihood Ratio Test** to test the null hypothesis of $H_0(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ has test statistic:

$$2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\mathbf{0})\}.$$

The **Wald Test** has associated test statistic:

$$\hat{\boldsymbol{\beta}}' \mathbf{I}(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\beta}}.$$

The **Score Test** statistic is based on:

$$\mathbf{u}'(\mathbf{0}) \mathbf{I}^{-1}(\mathbf{0}) \mathbf{u}(\mathbf{0}).$$

Under the null hypothesis: $H_0(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, each of the above test statistics has a chi-squared distribution with p degrees of freedom.

3.2.6. Example of Fitting the Proportional Hazards Model

The Cox Proportional Hazards model is fitted to the multiple Myeloma data using the *phreg* procedure in SAS. The Breslow method for handling ties is employed.

The independent variables in the examples are given below:

- Age – the patient's age in years
- Sex – 0 = Male, 1 = Female
- Bun – the levels of blood urea nitrogen

- Ca – serum calcium
- HB – Hemoglobin
- PCells – Percentage of plasma cells in the bone marrow
- Protein – an indicator variable to denote whether or not Bence-Jones protein was present in the urine. (0 = present, 1 = absent)

The proportional hazards model for the i^{th} individual is then:

$$h_i(t) = e^{\beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Bun}_i + \beta_4 \text{Ca}_i + \beta_5 \text{HB}_i + \beta_6 \text{PCells}_i + \beta_7 \text{Protein}_i} h_0(t).$$

The Baseline hazard function is the hazard function for an individual for whom the values of all seven of these variables are zero i.e. a male ages zero, with zero values of **Bun**, **Ca**, **HB** and **Pcells** and no **Bence-Jones protein**. Because this is difficult to interpret, it would have been possible to recode age as age – 60, which would correspond then to a male who is aged 60 and has zero values for the other independent variables. Though this facilitates interpretation, it has no effect on the explanatory variables and will not affect inference.

The estimates in the below tables are obtained after fitting the model:

Table 7 SAS Output on Fitting the Proportional Hazards Model to the Example Data

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	-0.0195	0.0280	0.4838	0.4867	0.9810	Age
sex	1	-0.2399	0.4045	0.3518	0.5531	0.7870	Sex
bun	1	0.0208	0.0059	12.2690	0.0005	1.0210	Bun
ca	1	0.0207	0.1325	0.0243	0.8761	1.0210	Ca
hb	1	-0.1329	0.0691	3.6971	0.0545	0.8760	HB
pcells	1	-0.0010	0.0066	0.0238	0.8774	0.9990	Pcells
protein	1	-0.6276	0.4269	2.1608	0.1416	0.5340	Protein

From the above, it can be seen that many of the estimates are close to zero and have high p-values. In fact, only bun and hb have a p-value smaller than 0.1 (or 10%).

However, from this output, one cannot conclude that these two variables will be the only two predictors. Like normal linear regression, proportional hazards are sensitive to correlation in the set of independent variables. To arrive at the best fit, several models need to be fitted and the output compared and evaluated to determine whether the models make intuitive and statistical sense.

3.2.7. Model Selection

As in the case of linear regression, the set of available independent variables needs to be evaluated to find the optimal set that best fits the data and makes the most sense in the application area. Fortunately, many of the selection procedures available to regression can also be applied to survival modelling and will be discussed in this section.

It is important to have a summary statistic to evaluate and compare different models fitted to the data. Since the likelihood function summarises the information contained in the data about the unknown parameters, it makes sense to start with this function and develop a metric from here. The likelihood function, when substituting the values of the estimating parameters gives a

description of the model fit and since the objective of the likelihood is to obtain the parameter set that maximises the likelihood function: the larger the value, the better. In this fashion, different models can be compared.

However, in practice it is convenient to use $-2\log\hat{L}$, where \hat{L} is the value of the likelihood function when substituting a particular set of parameter estimates. When using $-2\log\hat{L}$, the lower the value, the better. $-2\log\hat{L}$ will also always be positive. This value can be used to compare the different outputs of the selection procedures discussed below.

In addition to $-2\log\hat{L}$, one can also look at a closely related quantity, $2k - 2\log\hat{L}$ where k is the number of parameters in the model. This quantity is known as the Akaike Information Criterion or AIC. The AIC deals with the trade-off between goodness-of-fit and number of parameters (complexity) of the model i.e. a model with too many parameters can overfit the data. As with $-2\log\hat{L}$, the preferred model is the model with the lowest value of AIC.

Other criteria such as the Bayesian Information Criterion can also be used to compare different models, but will not form part of the scope of this essay.

- Forward Selection

In forward selection the algorithm starts with building one variable regressions for each variable in the set of independent variables. The first model selected is the one with the smallest value of $-2\log\hat{L}$. This variable is then kept and a two variable model is constructed for each of the remaining $p - 1$ independent variables. The two variable model with the smallest value $-2\log\hat{L}$ is then again selected. The algorithm then carries on adding variables in this fashion until the difference in $-2\log\hat{L}$ does not change below a pre-specified value. This is called a stopping value.

- Backward Elimination

In the case of backward elimination, the algorithm starts with all independent variables included in the model. It will then remove each of the p independent variables and exclude the variable that decreases $-2\log\hat{L}$ by the smallest amount. The algorithm will proceed in this fashion again until no significant change occurs anymore in $-2\log\hat{L}$.

- Stepwise Selection

The Stepwise selection algorithm is very similar to Forward Selection except that a variable that was included in a previous step can be excluded in a future step. In this fashion, stepwise selection will identify a subset of the independent variables that gives the smallest value of $-2\log\hat{L}$.

The disadvantage of these selection procedures is that it will typically lead to a specific subset in the data which may or may not make sense to the application area. These procedures also are not necessarily immune to correlation in the data and the analyst will have to evaluate and assess the output obtained.

3.2.8. Example using Selection Procedures

The previous output of the Proportional Hazards model fitted to the Multiple Myeloma data suggested that not all seven independent variables are necessary in the modelling of the hazard of death. The stepwise selection procedure was applied and the results below obtained.

Table 8 SAS Output on Fitting the Proportional Hazards Model to the Example Data using Stepwise Selection

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
bun	1	0.0185	0.0057	10.6209	0.0011	1.0190	Bun
hb	1	-0.1316	0.0621	4.4921	0.0341	0.8770	HB

From the above output it can be seen that only **Bun** and **HB** were selected in the stepwise selection procedure. **Bun** was entered in step one followed by **HB** in step two. The process then terminated meaning that the inclusion of any of the remaining independent variables could not yield a better model fit.

Both these variables are significant at the 5% level of significance. This threshold value could have been set to another value if required.

3.2.9. Interpretation of the Parameter Estimates

When a proportional hazards model is fitted, the coefficients of the explanatory variables in the model can be interpreted as the logarithms of the ratio of the hazard of death to the baseline hazard.

Consider the univariate model:

$$h_i(t) = e^{\beta X_i} h_0(t).$$

which gives the hazard of death for individual i in the dataset. Suppose that another individual in the dataset has the observed value $X_i + 1$ for the independent variable X . The ratio of these two observations to one another will yield the following:

$$\frac{e^{\beta(X+1)}}{e^{\beta X}} = e^{\beta}.$$

Therefore $\hat{\beta}$ is the estimated change in the logarithm of the hazard ratio when the value of X is increased by one unit.

Using a similar argument, the estimated change in log-hazard ratios when a value of X is increased by r units will be $r\hat{\beta}$, and the corresponding estimate of the hazard ratio is $e^{r\hat{\beta}}$.

Returning to the example of Multiple Myeloma, the parameter estimate for **Bun** is 0.0185 with the corresponding hazard ratio 1.0190. This means that a unit increase in the value of **Bun** will result in a 0.0185 change in the log of the hazard ratio, or 1.0190 on the hazard ratio itself.

3.2.10. Estimating the Hazard and Survival Functions

So far, only the estimation of the β coefficients in the linear component of the proportional hazards model has been discussed. This is all that is necessary to draw inferences about the set of independent variables in the modelling of the hazard function. Once a suitable model has been identified, it is possible to estimate the hazard function, the baseline hazard and the associated survival function.

After estimation of the β coefficients, the fitted model is.

$$\hat{h}_i(t) = e^{x_i \hat{\beta}} \hat{h}_0(t).$$

The next step is to determine $\hat{h}_0(t)$, the baseline hazard.

3.2.10.1. Estimating the Baseline Hazard

Kalbfleisch and Prentice (1973) provided an estimate of the baseline hazard, using maximum likelihood which is discussed in this section. The derivation of this estimate is quite complex and will not be included here.

Suppose there are r distinct ordered death times denoted by $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ and there are d_j deaths and n_j individuals at risk at time $t_{(j)}$.

The estimate of the baseline hazard function at time $t_{(j)}$ is then given by:

$$\hat{h}_0(t_{(j)}) = 1 - \xi_j.$$

where ξ_j is the solution to the equation:

$$\sum_{l \in D(t_{(j)})} \frac{e^{x_l^T \hat{\beta}}}{1 - \xi_j e^{x_l^T \hat{\beta}}} = \sum_{l \in R(t_{(j)})} e^{x_l^T \hat{\beta}},$$

for $j = 1, 2, \dots, r$

Where

- $D(t_{(j)})$ is the set of all individuals who die at the j^{th} ordered death time.
- $R(t_{(j)})$ is the set of all n_j individuals who are at risk at time $t_{(j)}$.

In the particular case where there are no tied death times, the left hand side of the equation will be a single term and the equation solved to give:

$$\xi_j = \left[1 - \frac{e^{x_{(j)}^T \hat{\beta}}}{\sum_{l \in R(t_{(j)})} e^{x_l^T \hat{\beta}}} \right] e^{x_{(j)}^T \hat{\beta}}.$$

Using the results arrived at so far, the hazard function can now be estimated given the vector of explanatory variables X .

$$\hat{h}_i(t) = e^{x_i \hat{\beta}} \hat{h}_0(t).$$

Integrating on both sides yields the following:

$$\int_0^t \hat{h}_i(u) du = e^{x_i \hat{\beta}} \int_0^t \hat{h}_0(u) du.$$

So that the cumulative hazard function for the i^{th} individual is given by:

$$\hat{H}_i(t) = e^{x_i \hat{\beta}} \hat{H}_0(t).$$

and therefore the estimated survival function:

$$\hat{S}_i(t) = \hat{S}_0(t) e^{-x_i \hat{\beta}},$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \dots, r$.

3.2.10.2. Model Validation

3.2.10.2.1. Introduction

Like in the case of normal linear regression, the fitted proportional hazards models need to be validated to check whether the model fit is good. The use of diagnostic procedures is strongly advised in a survival analysis. Because methods used to assess the quality of a model fit have to cope with the occurrence of censored survival times, this adds an additional layer of complexity to the available diagnostic procedure. The following questions can typically be asked:

- Does the fitted model accurately describe the hazard of death?
- Do the parameter estimates make sense and can they be explained?
- Is the model missing possible trends in the data and if so, what are these trends and what can be done to account for them?

In this section, several types of residuals and residual analysis will be explained as well as their unique advantages and disadvantages. A method for checking whether trends exist in the

independent variable set and possible transformation of the independent variables will also be discussed.

For the whole section, the assumption will be made that the survival times of n individuals are available, that there are r deaths of these n individuals and that $n - r$ individuals are right censored. Assume further that a Cox Regression has been fitted to the survival times and that the linear component of the model contains p explanatory variables X_1, X_2, \dots, X_p . The fitted model for the i^{th} individual is then, $i = 1, 2, \dots, n$:

$$\hat{h}_i(t) = e^{\mathbf{x}_i \hat{\boldsymbol{\beta}}} \hat{h}_0(t).$$

where

$$\mathbf{x}_i \hat{\boldsymbol{\beta}} = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}.$$

is the fitted linear component of the model, and $\hat{h}_0(t)$ is the estimated baseline hazard function.

3.2.10.2.2. Cox-Snell Residuals

The most widely used residual in the analysis of survival data is the Cox-Snell residual. These residuals have the name ‘‘Cox-Snell’’ as it is a particular definition of residuals given by Cox and Snell (1968).

The Cox Snell residual for the i^{th} observation is given by

$r_{ci} = e^{\mathbf{x}_i \hat{\boldsymbol{\beta}}} \hat{H}_0(t_i)$, where $\hat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i - the observed survival time of that individual.

Note that the Cox-Snell residual is the value of

$$\begin{aligned} \hat{H}_i(t_i) &= e^{\mathbf{x}_i \hat{\boldsymbol{\beta}}} \hat{H}_0(t_i) \\ &= -\log \hat{S}_i(t_i). \end{aligned}$$

where $\hat{H}_i(t_i)$ and $\hat{S}_i(t_i)$ are the estimated values of the cumulative hazard and survival functions for the i^{th} individual at time t_i

The Cox-Snell residuals have the interesting property that they follow an Exponential distribution with a unit mean. The proof of this property follows from Theorem 6.3.2 in (Bain and Engelhardt 1992) on page 198 and is given below:

Let X be a continuous random variable with pdf f_X .

Then the density of the random variable $Y = g(X)$ is given by:

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{\left| \frac{dy}{dx} \right|}.$$

Using this result, let T be the random variable associated with the survival time of an individual, and $S(t)$ the associated survival function. Consider the transformation $Y = -\log(S(T))$ where

$$g(Y) = -\log(S(T))$$

$$\therefore e^{-Y} = S(T)$$

$$\therefore g^{-1}(Y) = T = S^{-1}(e^{-Y}),$$

and

$$\frac{dy}{dt} = \frac{d\{-\log(S(T))\}}{dt} = \frac{f_T(t)}{S(t)}.$$

Therefore

$$f_Y(y) = f_X(g^{-1}(y)) / \left| \frac{dy}{dx} \right| \text{ becomes}$$

$$f_Y(y) = \frac{f_T\{S^{-1}(e^{-y})\}}{\frac{f_T(S^{-1}(e^{-y}))}{S(S^{-1}(e^{-y}))}} = e^{-y}.$$

This is the density function of an Exponential distribution with unit mean or $EXP(1)$.

From the results above, one can assume that the Cox-Snell residuals r_{ci} will behave as observations from a $EXP(1)$ distribution, if the fitted model is correct.

If the observed survival time is right censored, then the corresponding residual is also right censored. The residuals will therefore be a censored sample from a $EXP(1)$ distribution and a test of this assumption provides insight into model adequacy.

Because of the above, the Cox-Snell residuals will not exhibit the same properties as normal residuals from a linear regression i.e. have mean zero and a normal distribution. They will not be symmetrically distributed and will never be negative. They will also have a skew distribution with a mean and variance of one.

4. CHAPTER 4 – A Background on Smoothers and Additive Models

4.1. Smoothers and Smoothing

4.1.1. Introduction

A smoother is a tool used for describing and summarising data. The name smoother derives from the fact that a smoother produces an estimate of the variable that is less variable than the variable itself, hence the name smoother.

The moving average is an example of a widely used smoother, and works by averaging points in a neighbourhood of a given size. It is easy to understand and helps the analyst spot trends in the data.

An important property of a smoother is that it is not parametric – there are no rigid assumptions of the form of the dependence of the mean of Y on $\mathbf{X} = (X_1, X_2, \dots, X_p)$

Where one is only applying a smoother to the data points (x_i, y_i) , the term scatterplot smoother is used.

As mentioned earlier, smoothers are mainly used for description. Smoothers aid the analyst by enhancing the visual appearance of the scatter plot of X on Y , hence making it easier to spot trends. Secondly, smoothers are used to estimate the dependence of the mean of Y on the predictors. This fact leads to their use in additive and generalised additive models.

A smoother can be easily written in the form:

$\hat{\mathbf{y}} = S\mathbf{y}$ where $\hat{\mathbf{y}}$ is the produced smooth of the observed vector \mathbf{y} and S is a $n \times n$ smoother matrix.

Two broad categories of smoothers exist – linear and non-linear smoothers. In simple terms for a linear smoother the above smoothing matrix S does not depend on \mathbf{y} . A more mathematical definition will be given in a subsequent section.

Examples of linear smoothers are running means, cubic splines, locally weighted running lines and even the least squares line. The running median is an example of a non-linear smoother for which a smoothing matrix cannot be constructed.

The level of smoothing / amount of smoothing that a smoother does is controlled by its smoothing parameter. Usually these are the degrees of freedom (in the case of cubic and regression splines) or span. Span is denoted by w and can be interpreted as a proportion of points of the data sample used to compute the smooth i.e. for w between 0 and 1 so that $[wn]$ is an integer. When $w = 1$, every smooth will include all of the available points and with $w = 0$ the opposite.

The question of how to select the smoothing parameter has no straight forward answer and many techniques have been put forward to deal with the issue. Most common is cross-validation, however in many cases the analyst will have to make a decision based on their knowledge of the subject and the nature of the data. Also if some data driven technique is used to select a smoothing parameter or generate a test, then the smoother also becomes non-linear.

4.1.2. Neighbourhoods

As a simplistic definition, a smoother works by averaging Y -values of observations having predictor values close to a target value. The points to use for the calculated average or smoother depend on the neighbourhood, a subset of x_i 's, close to x_i for which the corresponding y_i 's are used in the smooth.

Intuitively, neighbourhoods can be selected in two ways – by taking the r nearest points to x (regardless of which side they are on), or by taking k points to the left and k points to the right of x . This latter is called a symmetric nearest neighbourhood whereas the prior is called the nearest neighbour. In the case of a symmetric neighbourhood, when there are not exactly k -points on a side available, as many points as possible are selected.

The size of the neighbourhood is typically expressed in terms of an adjustable smoothing parameter. As expected, choosing neighbourhoods of small size will result in estimates with low bias but high variance, whereas a larger neighbourhood will yield estimates with high bias but low variance. This is known as the bias-variance trade-off, and must always be kept in mind when choosing the smoothing parameter.

The indices for points belonging to x in a symmetric nearest neighbourhood are denoted by: $N^s(x_i)$ with a formal definition given by:

$$N^s(x_i) = \left\{ \max\left(i - \frac{wn - 1}{2}, 1\right), \dots, i - 1, i, 1 + 1, \dots, \min\left(i + \frac{wn - 1}{2}, n\right) \right\}.$$

4.1.3. Moving Average / Running-mean and Running Line Smoothers

The moving average or running mean is a widely used smoother, and often used in time-series analysis, especially when the data is evenly spaced. It has several drawbacks – just as averages are very sensitive to influential data points, just so is the running mean. This can lead to smooths that look wiggly or erratic. Apart from this, the running line also tends to flatten out near the endpoints of the data, and as a direct result of this, it can be severely biased.

The running mean can be defined as:

$$S(x_i) = \text{ave}(y_j)_{j \in N^s(x_i)}.$$

Smoother matrix is given below for a running mean with smoother with $w = 0.5$ and $n=10$

$$S = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Because of the running mean's sensitivity to outlying data points and bias on the end points of the data, the running line was proposed. The running line smoother computes the least squares line instead of a mean for each neighbourhood.

The running line smoother can be defined as:

$$S(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0.$$

Where $\hat{\alpha}(x_0)$ and $\hat{\beta}(x_0)$ are the least squared estimates of the regression lines fitted locally at x_0 , (i.e. of the data points in $N^S(x_i)$).

The running line fit is dominated in the interior by the mean and at the endpoints by the slope. The parameter k (number of points to the left and right of x_0) controls the appearance of the line – for large values of k , the line is smooth, whereas small values of k produces jagged and discontinuous curves. As mentioned earlier, large values of k will also yield large bias in the estimates, but low variance, whereas small values will do the inverse.

The running line smooth computes in $o(N)$ operations and is easy to implement and to interpret.

One way of improving the appearance of the running line smooth is to compute a weighted least squares fit within each neighbourhood, by weighting points close to x_0 higher than points further away from x_0 . This idea is explored further in the next section.

The running-line smoother is also zero outside of the diagonal bands, with the non-zero elements of the i^{th} element given by:

$$s_{ij} = \frac{1}{n_i} + \frac{(x_i - \bar{x}_i)(x_j - \bar{x}_i)}{\sum_{k \in N^S(x_i)} (x_k - \bar{x}_i)^2}.$$

where n_i denotes the number of datapoints in $N^S(x_i)$, j subscripts these points and \bar{x}_i represents their mean.

4.1.4. Kernel Smoothers

A Kernel Smoother is another type of popular statistical technique for estimating the value of a real valued function, when the parameters of this model are unknown. Again the estimated function is smooth, and the level of smoothness is set by a single parameter.

The kernel defines an explicit set of weights used in the calculation of the estimate. These weights are a function of the distance of the observation point x_0 to the target point x_j and usually decrease as one moves further away from x_0 .

The weight given to the j^{th} point, used for computing the estimate at x_0 is given by:

$$S_{0j} = \frac{c_0}{\lambda} d\left(\left|\frac{x_0 - x_j}{\lambda}\right|\right),$$

where $d(t)$ is an even decreasing function in $|t|$ and λ is the window-width or bandwidth. The constant c_0 is usually chosen so that the weights sum to unity, although this may not always be the case.

The kernel smoother matrix has elements;

$s_{ij} = c_i d_\lambda(x_i, x_j)$ where c_i is chosen so that the rows sum to unity (equivalent to above)

Several popular kernels used in smoothing include:

- The Epanechnikov Kernel

$$d(t) = \begin{cases} 3/4(1 - t^2), & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

which minimises the asymptotic mean squared error.

- The Minimum Variance Kernel

$$d(t) = \begin{cases} 3/8(3 - 5t^2), & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

which minimises the asymptotic variance of the estimate.

- The Gaussian Kernel, where d is the standard Gaussian density

The idea behind using a kernel estimate is as follows. For each point in the data, x_0 , calculate the kernel weights as defined above and apply to the data to obtain the estimate at x_0 . Typically, this action can be described by the Nadaraya-Watson kernel weighted average model:

$$s(x_0) = \frac{\sum_{i=1}^n d\left(\frac{x_0-x_i}{\lambda}\right)y_i}{\sum_{i=1}^n d\left(\frac{x_0-x_i}{\lambda}\right)}. \quad (2.4.1)$$

The fitted values of the kernel smoother are continuous, and as mentioned, the smoothness of the function is determined by the bandwidth parameter, λ . For k -nearest neighbours, the neighbourhood size k is equal to λ .

From here, it can readily be seen that the running line and moving average smooths can be expressed as kernel smooths. When this is possible, the term *equivalent* kernel is used and gives one a basis for comparing different smooths on common ground.

The moving average, using nearest neighbours and a kernel terminology, can be expressed as follows:

Let :

$$d(t) = \begin{cases} 1/m & \text{for } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where m is the number of points in the neighbourhood. Substituting the above into (2.4.1) yields the moving average smoother.

As expected, the running line smoother can also be expressed in terms of its equivalent kernel.

For the running line smoother, or alternatively, Local Linear Regression, the estimate is given by:

$$s(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0.$$

Where $\hat{\alpha}(x_0)$ and $\hat{\beta}(x_0)$ are the values that minimise:

$$\sum_{i=1}^n S_{0i}(y_i - \alpha(x_0) - \beta(x_0)x_i)^2.$$

Let $b(x)^T = (1, x)$, and \mathbf{B} the $N \times 2$ regression matrix with i^{th} row equal to $b(x_i)$, and $W(x_0)$ the $N \times N$ diagonal matrix with i^{th} diagonal element S_{0i} then:

$$s(x_0) = b(x_0)^T (B^T W(x_0) B)^{-1} B^T W(x_0) y.$$

These weights combine the weighting kernel as the least squared operations.

4.1.5. Regression Splines

A spline is a smooth polynomial function that is piecewise defined, i.e. it is made up of different polynomial of degree n and the first $n - 1$ derivatives at the points where they are joined. These locations / points of where the functions meet are called knots.

The term “spline” derives from a flexible piece of metal used to draw curved lines.

The number of knots k , as well as their position are all parameters in this procedure.

Important in the arena of splines is the concept of B-Splines or Basis splines. The splines have minimal support with respect to a given degree, smoothness and domain definition – any spline function of some degree n can be written as a linear combination of B-splines of that n degree.

For a given set of knots, the B-spline is unique which is why they are called basis splines. They are important and useful in that any spline function of degree k on a given set of knots can be expressed as a linear combination of the B-splines.

The space of all cubic splines with a specified sequence of K breakpoints or knots and associated continuity conditions can be generated from a single set of basis functions i.e. the spline function is written as a linear combination of the basis elements.

To illustrate this notion, consider the space of simple cubic polynomials – a suitable basis is given by $\{1, x, x^2, x^3\}$.

Any polynomial function in x can then be written as

$$\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3.$$

If a single knot is added at $x = w_1$ and the basis function $(x - w_1)_+^3$ is included where $(\)_+$ denotes the positive part, a suitable basis will then be:

$$\{1, x, x^2, x^3, (x - w_1)_+^3\},$$

and one can verify that the polynomial function:

$$\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 (x - w_1)_+^3,$$

satisfies the conditions for a cubic spline with the given knot.

Proceeding in this fashion, terms are added for each knot and the space will become $K + 4$ dimensional when K such knots are included.

For the B-splines, each d^{th} degree B-spline basis function is nonzero between at most $d + 2$ knots. B-splines are less intuitive, but computationally attractive thanks to this local nature.

A spline function $s(x)$, can be written as a linear combination of these basis functions:

$$s(x) = \sum_{j=1}^{K+d+1} \alpha_j \beta_j(x),$$

where d is the degree of the spline, K is the number of knots, and $\beta_j(x)$, $j = 1, 2, \dots, K + d + 1$, are the d^{th} degree B-spline basis elements for the specified K knots.

The covariate X is “expanded” into a set of new variables – the evaluated B-spline basis elements for the specified K knots and the unknown coefficients $\alpha_1, \dots, \alpha_p$ can be estimated by using partial likelihood methods that are used for the standard proportional hazards model.

The standard errors of the variance-covariance matrix can be used to obtain estimates of the standard errors.

The use of regression splines is attractive because no special software is needed once the breakpoints or knots have been selected. The disadvantage is that the analyst will always be required to select these points.

From clinical trials data, Hastie and Tibshirani (1990) found that typically no more than three or four knots are required to accurately describe the underlying relationship between the independent variable and the effect thereof on the dependent variable.

If there are no prior reasons for placing knots at specific locations, a good strategy is to place them at the quartiles.

Figure 5 below shows the log hazard ratio for the independent variable age, with 3 knots placed at the quartiles. Refer to section 6.2.5.4 for the complete model. The placement of the knots allows the analyst to capture the non-linear trend of age in the data, where a normal linear fit would not have been able to.

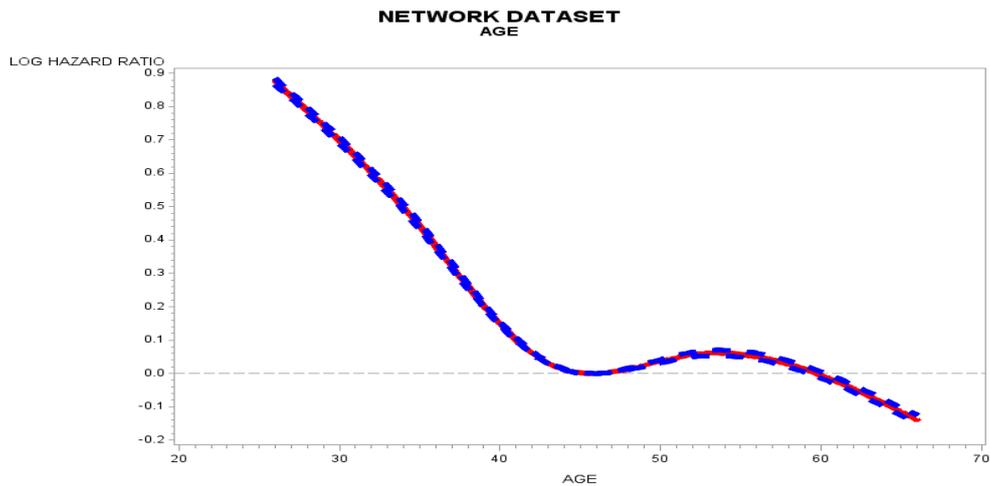


Figure 5 Log Hazard Ratio for Age

4.1.6. Cubic Smoothing Splines

The Cubic Smoothing Spline is not constructed explicitly like those considered so far – instead it arises as a solution to an optimisation problem.

Consider the following, among all functions g with two continuous derivatives, find the one that minimises the penalised residual sum of squares.

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int_{-\infty}^{\infty} [g'(z)]^2 dz. \quad (4.1.6)$$

where λ is a fixed constant.

Buja and Andreas (1989) show that the solution $\hat{g}(x_i)$ is a cubic spline with knots at each of the unique data points x_i . The two terms making up this function can be interpreted as follows:

- $\sum_{i=1}^n (y_i - g(x_i))^2$ measures the closeness to the data
- $\lambda \int_{-\infty}^{\infty} [g'(z)]^2 dz$ measures the amount of smoothing done

The constant λ controls the amount of smoothing i.e. when $\lambda = 0$ the solution is any interpolating problem whereas if $\lambda = +\infty$ the solution is the least squared line.

It has been shown that the smoothing spline is a linear smoother and that a smoother matrix can be written down. The following is taken from Green and Yandell (1985). Let

$$h_i = x_i - x_{i-1}, \text{ for } i = 1, \dots, n - 1$$

Δ is a tri-diagonal $(n - 2) \times n$ with

- $\Delta_{ii} = \frac{1}{h_i}$.
- $\Delta_{i,i+1} = -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right)$.
- $\Delta_{i,i+2} = \left(\frac{1}{h_{i+1}}\right)$.

C is a symmetric tri-diagonal matrix of order $n - 2$ with

- $C_{i-1,i} = C_{i,i-1} = \frac{h_i}{6}$.
- $C_{ii} = \frac{h_i + h_{i+1}}{3}$.

Then if $\hat{y}_i = \hat{g}(x_i)$ it can be shown that solving (4.1.6) is the same as minimising

$$\|y - \hat{y}\|^2 + \lambda \hat{y}^t K \hat{y}.$$

where $K = \Delta^t C^{-1} \Delta$ with solution to $\hat{y} = S y$ where S is given by:

$$S = (I + \lambda K)^{-1}.$$

This important result will be referred to again in section 5.1.2 in deriving the backfitting algorithm for the additive Cox proportional hazards model.

The optimisation problem can be argued and interpreted intuitively – if one isolates a region of the curve, say x then if the data density is high in that region the first part will dominate, whereas if the data is sparse, the interpolating function will be almost linear and the penalty term will dominate.

The cubic spline operator takes $O(n)$ operations to compute, thanks to the banded nature of its matrix.

4.2. Additive Models

4.2.1. Introduction

Linear regression analysis has been, and still is a very popular and powerful technique used in many industries and by many practitioners. The multiple regression model enables the analyst to describe the relationship between a dependent variable and several independent variables and is easy to compute, interpret and use. Selection criteria such as forward stepwise regression or backward elimination with the advent of faster computers has made it an even more powerful technique in that one can now identify the most predictive variables from a large dataset without much effort. The regression model given all its advantages and widespread use however has several limitations. These limitations lead one to the field of additive models and ultimately generalised additive models.

In this chapter, the linear regression model is reviewed and the additive model introduced. Further exploration of inference, parameter estimation and general theory of additive models is also explained in detail.

4.2.2. Review of Multiple Linear Regression

Given a sample of n observations of a response / dependent variable Y and k independent / design vectors \mathbf{X}_i , the objective is to model the dependence of the response variable on the independent variables or predictors. In matrix notation, this can be denoted as follows:

Response variable $\mathbf{Y} = [y_1, y_2, \dots, y_n]^T$.

Independent variables $\mathbf{X}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T$ for $i = 1, 2, \dots, n$

The multiple linear regression model describes the dependence between Y and X by means of the following relationship:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon. \quad (4.2.2.1)$$

Where $E[\varepsilon] = 0$ and $var(\varepsilon) = \sigma^2$ The errors are assumed to be independent and normally distributed.

The regression model assumes that the relationship between the dependent and independent variables is linear. This assumption makes the model extremely useful and convenient because:

- The data is described in a simple manner.
- The contribution of each independent variable towards the prediction of Y is summarised by its coefficient.
- A simple equation is obtained with which future predictions can be made.

The multiple regression model can be generalised in several ways – amongst others, surface smoothers provide a reasonable generalisation.

A surface smoother “non-parametric” regression model takes the following form:

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon, \quad (4.2.2.2)$$

where f is a pre-specified smoothing function.

A challenge with surface smoothers is choosing the shape of the kernel or neighbourhood that defines local in k dimensions. An even more serious problem for all surface smoothers has to do with the same *localness* in higher dimensions – neighbourhoods with a fixed number of points become less local as the dimensions increase (Friedman and Stuetzle 1981). Simply illustrated, if data in one dimension is considered, and span of k can be selected to capture a specific number of data points. When working in higher dimensions, the span increases as the number of dimensions increase in order to capture the same amount of data point. This leads to neighbourhoods becoming large and adversely affects the variance and reliability of obtained estimates. This problem is called the *Curse of Dimensionality*. In addition to the above, these models are also very difficult to interpret, and usefulness therefore is under question. The important characteristic of regression models which makes them easy to interpret is the fact that they are *additive* in the predictors.

Several multivariate techniques have been devised to surmount the curse of dimensionality and interpretability, one of which is to use generalise the surface smoother to be additive in its smoothing components. This leads one to consider additive models.

4.2.3. Additive Models

Like the multiple regression model, an additive model decomposes the surface smoother described in the previous sections as follows:

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon \text{ becomes}$$

$$Y = \alpha + f_1 X_1 + f_2 X_2 + \dots + f X_p + \varepsilon,$$

or

$$E[Y] = \alpha + f_1 X_1 + f_2 X_2 + \dots + f X_p, \quad (4.2.3.1)$$

where $E[\varepsilon] = 0$ and $var(\varepsilon) = \sigma^2$ as before.

The f 's are arbitrary univariate functions – one for each predictor. The above model, (4.2.3.1) retains the attractive feature of regression models in that it is additive in the predictor variables and therefore the interpretation and assessment of the model is made easier.

The additive model is a special case of the Projection Pursuit Regression, proposed by Stuetzle and Friedman (1981), as well as Alternating Conditional Expectation, by Breiman and Friedman (1985).

The additive model is able to sidestep all of the challenges listed in the previous section surrounding smoothers in higher dimensions, at the cost of approximation terms in using an additive function to model the p-dimensional surface.

Implicit in 4.2.3.1 is the assumption that $E[f_i X_i] = 0$ for all i , to prevent free constants in any of the component functions and facilitate calculation and interpretation.

A very important interpretative feature of 4.2.3.1 that is retained from the linear regression model is the fact that the variation of the fitted response surface, holding all but one predictor constant does not depend on the other predictors. This means that each of the k functions can be examined separately to assess their contribution towards the predictability of the response. This also means that the additive model provides one with a very handy data exploratory tool.

The fitted functions in additive models play the same role as the coefficients in linear regression. All of the challenges and pitfalls when fitting a regression model to data are also present in additive

models, e.g. insignificant functions must be dropped from the model and it is suggested that a model be fitted in a stepwise fashion, much like linear regression.

4.2.4. Fitting Additive Models

The backfitting algorithm, proposed by Friedman and Stuetzle (1981), is the most popular method for fitting additive models to data. It is a general algorithm, able to handle any type of smoothing functions used, however, it is iterative which may make the estimation computationally taxing.

The backfitting algorithm is derived using conditional expectation. Suppose the additive model

$$Y = \alpha + f_1X_1 + f_2X_2 + \cdots + f_pX_p + \varepsilon$$

is correct.

By taking expectations of both side of the equation, it can be seen that:

$$E[Y] = E[\alpha + f_1X_1 + f_2X_2 + \cdots + f_pX_p + \varepsilon]$$

$$E[Y] = E[\alpha + f_1X_1 + f_2X_2 + \cdots + f_pX_p] \text{ since } E[\varepsilon] = 0.$$

By conditioning on only one of the X 's, the following is obtained:

$$\begin{aligned} E[Y|X_k = x_k] &= E[\alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)|X_k = x_k] \\ &= E[\alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)|X_k = x_k] + f_k(X_k). \end{aligned}$$

Therefore:

$$f_k(X_k) = E\left[Y - \alpha - \sum_{j \neq k}^p f_j(X_j) \mid X_k = x_k\right]. \quad (4.2.4.1)$$

This suggests an iterative algorithm for computing all of the unknown functions.

Therefore, if the current estimates of the model are $\hat{f}_k, k = 1, \dots, p$ then \hat{f}_p is updated by smoothing the *partial residuals* $r_{ij} = y_i - \sum_{k \neq j} \hat{f}_k(x_{ij})$.

This is written in terms of the data as an arbitrary scatterplot smoother S below.

- The Backfitting Algorithm

1. Initialise $\alpha = \text{ave}(y_i)$, $f_j = f_j^0$, $j = 1, \dots, p$
2. Cycle , $j = 1, \dots, p, 1, \dots, p$
3. $f_j = S_j(\mathbf{y} - \alpha - \sum_{k \neq j} f_k | x_j)$.

This is continued until the individual functions do not change.

As in the case of linear regression, here it is also desirable to fit all the smoothing functions at once, and therefore the individual smoothing steps make sense. Whenever one function j is adjusted, the effects of all other functions are removed from \mathbf{y} . Therefore one can say this partial residual is smoothed against x_j .

4.3. Generalised Additive Models

4.3.1. Introduction

In the previous chapter, the additive model was introduced as a means to, amongst others; deal with the *curse of dimensionality*. Additive models can be extended to generalised additive models – very similar to the notion of extending standard linear models to generalised linear models. As in the case of generalised linear models, the predictors are assumed to be linear in the parameters, however the link between predictor and responses as well as the distribution of the responses can be quite general. A common example of this type of model which is extensively used is the logistic regression model. This model has been applied in credit scoring, propensity modelling as well as social sciences as it is easy to use and interpret and yields invaluable insight about the underlying data structures.

In logistic regression, the response variable is assumed to have a Bernoulli distribution with $\mu = P(Y = 1|X_1, \dots, X_p)$ and μ is linked to the predictors via $\ln\left(\frac{\mu}{1-\mu}\right) = \alpha + \sum_j X_j B_j$.

The family of generalised linear models provides a convenient framework for studying the common structure of such models and there is a unified convenient way for their estimation.

To illustrate the concept of a generalised additive model, consider the logistic regression model given above:

$$\ln\left(\frac{\mu}{1-\mu}\right) = \alpha + \sum_j X_j B_j.$$

The additive extension of this model will then be:

$$\ln\left(\frac{\mu}{1-\mu}\right) = \alpha + \sum_j f_j(X_j).$$

The linear form $\alpha + \sum_j X_j B_j$ is simply replaced by the additive form, $\alpha + \sum_j f_j(X_j)$.

4.3.2. Review of Generalised Linear Models

The Generalised Linear Model, introduced by John Nelder and Robert Wedderburn (Nelder, J.A. and Wedderburn R, 1972) as a means to unify several other statistical models (such as logistic, linear and Poisson regression) under one framework, in its most common form is given by:

$$g(\mu) = \eta = \alpha + \sum_j X_j B_j,$$

where

$$\mu = E[Y|X_1, \dots, X_p]. \quad (4.3.2.1)$$

and $g(\cdot)$ is called *link function*, which links the *systematic component* to the *random component* and Y (the response) is assumed to have a density function that belongs to the exponential family which can be written in the below form:

$$\rho_Y(y; \theta; \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}\right\} + c(y, \phi).$$

The parameter θ is called the *natural parameter*, and ϕ is the *dispersion parameter* and is the random component of the model.

The expectation of Y , denoted by μ is linked to the covariates by the link function:

$g(\mu) = \eta$, where $\eta = \alpha + \sum_j X_j B_j$ $g(\mu) = \eta$ and η is the systematic component which is also known as the linear predictor.

It is easy to prove that the mean is related to the natural parameter θ by $\mu = b'(\theta)$. Very often, the obvious link for any given ρ_Y is the canonical link in which $\theta = \eta$.

4.3.3. Fisher Scoring of Generalised Linear Models

Suppose that the random and systematic components, as well as the link function have been specified. Given a vector of n observations of a dependent variable Y and p corresponding independent predictor vectors (X_1, \dots, X_p) , the maximum likelihood is defined by the score equations:

$$\sum_{i=1}^n x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) V_i^{-1} (y_i - \mu_i) = 0, \quad j = 0, 2, \dots, p \quad (4.3.3.1)$$

where

$$V_i = \text{var}(Y_i).$$

In order to solve these equations, a Newton-Raphson algorithm using the expected rather than observed information matrix is used. This method is the standard practice for this setting and is called the Fisher scoring procedure.

An equivalent procedure described in their book, *Generalized Additive Models* (Hastie and Tibshinari 1990) that is easier to use for this type of problem is called *adjusted dependent variable regression* and is a form of iteratively reweighted least squared (IRLS).

Given a vector of coefficients, $\boldsymbol{\beta}^0$, with its corresponding linear predictor: $\boldsymbol{\eta}^0$, and fitted values: $\boldsymbol{\mu}^0$, the *adjusted dependent variable* is constructed.

$$z_i = \eta_i^0 + (y_i - \mu_i^0) \left(\frac{\partial \eta_i}{\partial \mu_i} \right).$$

Define weights as $w_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right) V_i^0$, where V_i^0 is the variance of \mathbf{Y} at μ_i^0

Proceed by regressing z_i on \mathbf{x}^i with weights w_i to obtain a new estimate for $\boldsymbol{\beta}$ and then compute a new $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$.

This process is repeated until the change in deviance, $\mathbf{D}(\mathbf{y}; \boldsymbol{\mu}) = 2\{l(\boldsymbol{\mu}_{max}; \mathbf{y}) - l(\boldsymbol{\mu}; \mathbf{y})\}$ is model a preset threshold. This procedure is equivalent to the Fisher scoring procedure.

5. CHAPTER 5 –Additive Modelling for Survival Times

5.1.1. Introduction

In the previous chapter, the additive model and generalised additive models were introduced conceptually. The extension of the widely known logistic regression model to the context of generalised additive models was touched on. In this section, the extension of the Cox Proportional Hazards model in this regard will be discussed.

The following notation will be used –

- Survival data of the form $(y_1, x^1, \delta_1), \dots, (y_n, x^n, \delta_n)$ where
- y_i is the survival time or censored time
- δ_i is 0 where an observation is censored and unity otherwise
- x^i is the vector consisting of p predictors for the i^{th} individual
- Distinct failure times are given by: $t_{(1)} < t_{(2)} < \dots < t_{(k)}$ with there being $d_{(i)}$ deaths or failures at time $t_{(i)}$

Recall the Proportional Hazards model is given by:

$$\lambda_i(t|\mathbf{X}) = \lambda_0(t)e^{\sum_j \beta_j X_j}.$$

where $\lambda_i(t|\mathbf{X})$ is the hazard at time t given predictor values $\mathbf{X} = (X_1, \dots, X_p)$ and $\lambda_0(t)$ an arbitrary baseline function.

The generalised additive model extension, as given by Hastie and Tibshinari (1986) of this will then be :

$$\lambda_i(t|\mathbf{X}) = \lambda_0(t)e^{\sum_j f_j(X_j)}.$$

where f is an arbitrary, unspecified smooth function.

The effect of each covariate or independent variable on the log hazard is additive and is represented by a smooth and possibly non-linear function. The transformations, f , are not chosen by the analyst before applying the model, but rather are estimated flexibly from the data.

In rare cases, all of the covariate transformations will be smooth, nonlinear functions. Typically some categorical variables are also included in the model and the levels included as dummy /

indicator binary variables. Other variables such as for example, “age” can be modeled nonlinearly although if a linear fit is sufficient it will usually be preferred for simplicity.

An important advantage of the above model is that it alleviates the need to categorise a continuous variable (e.g. exponential, quadratic etc.) in order to discover the nature of its effect.

The above generalisation comes with its own unique set of challenges and questions such as:

- How to estimate the likelihood
- Which smoother function to use
- How to validate the selected smoother function
- Model Validation

In this section, these points will be addressed and discussed in turn, starting with the original work of Hastie and Tibshirani and also looking into what other researchers have done.

5.1.2. Estimation

In their book, *Generalized Additive Models*, Hastie and Tibshirani (1990) outline the steps necessary to estimate and fit the Cox Additive Proportional Hazards model.

To estimate the original Cox Proportional Hazards model, partial likelihood is the most popular method used. Where the baseline hazard, λ_0 , assumes a parametric form, the full sampling distribution and the likelihood can be written down.

However, as in the case of using a partial likelihood in the original setting, the baseline does not feature at all and is estimated using the results of the partial likelihood.

The partial likelihood, derived in section 3.2.3, and using Peto’s approximation of ties, is given by:

$$L(\boldsymbol{\beta}) = \prod_{R \in D} \frac{e^{(\sum_{j \in D_r} \boldsymbol{\beta}^T x^j)}}{\{\sum_{j \in R_r} e^{(\boldsymbol{\beta}^T x^j)}\}^{D_r}}$$

where

- D is the set of indices of the failures
- R_r is the set of indices of the individuals at risk at time $t_r - 0$

- D_r is the set of indices of failures at time t_r

Each term in this product reflects the conditional probability of a failure at an observed failure time t_r , given that all the individuals that are still in the study and at risk at time t_r .

The generalised additive proportional hazards model given by:

$$\lambda_i(t|\mathbf{X}) = \lambda_0(t)e^{\sum_j f_j(X_j)}.$$

which can also be written as:

$$\lambda_i(t|\mathbf{X}) = \lambda_0(t)e^{\eta(\mathbf{X})}.$$

where

$$\eta(\mathbf{X}) = \sum_j f_j(X_j). \tag{5.1.2.1}$$

Let $\eta_i = \eta(\mathbf{x}^i)$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$.

The partial likelihood for (5.2) is then:

$$L(f_1, f_2, \dots, f_p) = \prod_{R \in D} \frac{e^{(\sum_{j \in D_r} \eta_j)}}{\{\sum_{j \in R_r} e^{(\eta_j)}\}^{d_r}}.$$

The above likelihood is not the sum of n independent terms, as is the case for a likelihood of a distribution from the exponential family of distributions and therefore the Hessian matrix (the matrix of second order partial derivatives) is not diagonal. This adds a layer of complexity to the computation of this likelihood because the “local scoring” estimation method for $f_j(\cdot)$ is not directly applicable.

Hastie and Tibshinari (1990) proposed a method to deal with this challenge. Following their train of thought, “Local Likelihood” is a means to overcome this hurdle.

In order to estimate the f_j , the challenge can be rewritten as an optimization problem:

Let:

- Q_i be the space of functions with square integrable second derivatives on Ω_i , the domain of the i^{th} predictor.
- $l(\boldsymbol{\eta}) = \log PL(\boldsymbol{\eta})$

The objective is to find $f_1 \in Q_1, f_2 \in Q_2, \dots, f_p \in Q_p$ that will maximise:

$$j(\boldsymbol{\eta}) = l(\boldsymbol{\eta}) - \frac{1}{2} \sum_1^p \lambda_i \int_{-\infty}^{\infty} f_i''(s)^2 ds. \quad (5.1.2.2)$$

In the above equation, $\lambda_i \geq 0$ ($i = 1, 2, \dots, p$) are smoothing parameters.

- $l(\boldsymbol{\eta})$ measures closeness to the data
- $\frac{1}{2} \sum_1^p \lambda_i \int_{-\infty}^{\infty} f_i''(s)^2 ds$ measures the curvature of the fitted functions

The arguments and results of (Buja and Hastie et al. 1989), establish the existence of a unique solution for the above, given that certain conditions are met. To summarise: the log partial likelihood function is concave and these results imply that if the log partial likelihood has a unique solution (up to a constant shift) the space of linear functions, then a unique maximum (up to a constant shift) will exist for 5.1.2.2.

Given that a unique solution exists, it can be seen that this solution must be a cubic spline for each of the i smoothing functions.

For any functions, $\gamma_i(x)$, let $f_i(x)$ be the cubic spline that agrees with $\gamma_i(x)$ at x_{i1}, \dots, x_{in} . Then $j(\boldsymbol{\eta})$ cannot be decreased by substitution of $f_i(x)$ for $\gamma_i(x)$, as the first term, $l(\boldsymbol{\eta})$, does not change and the second term, $-\frac{1}{2} \sum_1^p \lambda_i \int_{-\infty}^{\infty} f_i''(s)^2 ds$, is maximised by the interpolating cubic spline that goes through the points, x_{i1}, \dots, x_{in} .

This potential infinite dimensional problem can be transformed to a finite dimensional problem by selecting a suitable choice of basis for the cubic splines. Considering the evaluation of the cubic splines, $f_i(x)$, at the data points x_{i1}, \dots, x_{in} , one arrives at a convenient basis and enables one to rewrite 5.3 as:

$$\begin{aligned} j(\boldsymbol{\eta}) &= l(\boldsymbol{\eta}) - \frac{1}{2} \sum_1^p \lambda_i \int_{-\infty}^{\infty} f_i''(s)^2 ds \\ &= l(\boldsymbol{\eta}) - \frac{1}{2} \sum_1^p \lambda_i \mathbf{f}_i^T \mathbf{K}_i \mathbf{f}_i, \end{aligned}$$

where

- \mathbf{K}_i are symmetric penalty matrices
- \mathbf{f}_i represents the values of the i^{th} cubic spline evaluated at the data points x_{i1}, \dots, x_{in}

Hastie and Tibshirani (1990) continue to derive a Newton- Raphson algorithm for maximising $j(\boldsymbol{\eta})$ over $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p$.

Let $\mathbf{u} = \frac{dl}{d\boldsymbol{\eta}}$ and $\mathbf{A} = -\frac{d^2l}{d\boldsymbol{\eta}d\boldsymbol{\eta}^T}$, then it can be shown that the Newton – Raphson step to go from $\mathbf{f}_1^{old}, \mathbf{f}_2^{old}, \dots, \mathbf{f}_p^{old}$ to $\mathbf{f}_1^{new}, \mathbf{f}_2^{new}, \dots, \mathbf{f}_p^{new}$ is

$$\begin{bmatrix} \mathbf{A} + \lambda_1 \mathbf{K}_1 & \mathbf{A} & \dots & \mathbf{A} \\ \mathbf{A} & \mathbf{A} + \lambda_2 \mathbf{K}_2 & \dots & \mathbf{A} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A} & \mathbf{A} & \dots & \mathbf{A} + \lambda_p \mathbf{K}_p \end{bmatrix} \begin{bmatrix} \mathbf{f}_1^{new} - \mathbf{f}_1^{old} \\ \mathbf{f}_2^{new} - \mathbf{f}_2^{old} \\ \vdots \\ \mathbf{f}_p^{new} - \mathbf{f}_p^{old} \end{bmatrix} = \begin{bmatrix} \mathbf{u} - \lambda_1 \mathbf{K}_1 \mathbf{f}_1^{old} \\ \mathbf{u} - \lambda_2 \mathbf{K}_2 \mathbf{f}_2^{old} \\ \vdots \\ \mathbf{u} - \lambda_p \mathbf{K}_p \mathbf{f}_p^{old} \end{bmatrix}. \quad (5.1.2.3)$$

The above is an $np \times np$ system of equations, which would ordinarily require $O(\{np\}^3)$ computations. Following Hastie and Tibshirani’s arguments, this can be reduced to $O(pn)$ by leveraging of the special structure in 5.1.2.3.

Let $\mathbf{z} = \boldsymbol{\eta}^{old} + \mathbf{A}^{-1}\mathbf{u}$ and $\mathbf{S}_j = (\mathbf{A} + \lambda_j \mathbf{K}_j)^{-1}\mathbf{A}$, then 5.1.2.3 can be rewritten as:

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \dots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \dots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1^{new} \\ \mathbf{f}_2^{new} \\ \vdots \\ \mathbf{f}_p^{new} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \mathbf{z} \\ \mathbf{S}_2 \mathbf{z} \\ \vdots \\ \mathbf{S}_p \mathbf{z} \end{bmatrix}.$$

In order to solve this system of equations, one can cycle through the predictors, solving for each one in sequence all the while leaving the others fixed, and replacing the current value of a function by its newly updated value at each step. By proceeding in this fashion, the above can be rewritten as:

$$\begin{bmatrix} \mathbf{f}_1^{new} \\ \mathbf{f}_2^{new} \\ \vdots \\ \mathbf{f}_p^{new} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1(\mathbf{z} - \sum_{j \neq 1} \mathbf{f}_j^{new}) \\ \mathbf{S}_2(\mathbf{z} - \sum_{j \neq 2} \mathbf{f}_j^{new}) \\ \vdots \\ \mathbf{S}_p(\mathbf{z} - \sum_{j \neq p} \mathbf{f}_j^{new}) \end{bmatrix}. \quad (5.1.2.4)$$

This iterative procedure is known as the “Gauss-Seidel” procedure for solving linear equations. It is also know more familiarly in this context as “Backfitting”.

The overall Newton-Raphson procedure is a nested algorithm consisting of an inner loop that cycles through 5.1.2.4 updating each function in turn until convergence, and then an outer loop that recalculates $\boldsymbol{\eta}, \mathbf{z}$ and \mathbf{A} .

The matrix \mathbf{A} is not diagonal, as it would be for the exponential family and hence \mathbf{S} calculates a weighted cubic spline smooth and the algorithm needs $O(n^3)$ operations in order to apply \mathbf{S} . This latter trait makes the procedure computationally expensive to apply except for very small datasets. In order to achieve the speed of $O(n)$, the off diagonal elements of \mathbf{A} to zero and denote the resulting matrix by \mathbf{A}^* .

Hastie and Tibshirani justify this by arguing that firstly the algorithm converges using \mathbf{A}^* , and that it is clear from 5.1.2.3 that the solution it produces are solutions to the original problem.

This is because the matrix - a

$$\begin{bmatrix} \mathbf{A} + \lambda_1 \mathbf{K}_1 & \mathbf{A} & \cdots & \mathbf{A} \\ \mathbf{A} & \mathbf{A} + \lambda_2 \mathbf{K}_2 & \cdots & \mathbf{A} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A} & \mathbf{A} & \cdots & \mathbf{A} + \lambda_p \mathbf{K}_p \end{bmatrix}.$$

is nonsingular with $\mathbf{A} = \mathbf{A}^*$ and thus convergence imply that the score $\mathbf{u} - \lambda_i \mathbf{K}_i \mathbf{f}_i = 0$ for all i .

Secondly, on average, the off diagonal elements are smaller than the diagonal elements by some order of magnitude. In particular, if there is no censoring then one can show that the diagonal elements of \mathbf{A} average approximately 2, while the non-diagonal elements average approximately $\frac{1}{n}$ and therefore the Newton-Raphson procedure that makes use of \mathbf{A}^* should not be very different from the true Newton-Raphson procedure.

5.1.2.1. Further details on the computation

In this section, further details on the computation of $\frac{\partial^2 l}{\partial \eta_i^2}$ and $\frac{\partial l}{\partial \eta_i}$ required for \mathbf{A}^* and \mathbf{u} are discussed.

Let $C_i = \{k: i \in R_k\}$ (the risk set that contains individual i) and $C_{ii'} = \{k: i, i' \in R_k\}$ (the risk set containing individuals i and i'). Then,

$$\frac{\partial l}{\partial \eta_i} = \delta_i - e^{\eta_i} \sum_{k \in C_i} \frac{d_k}{\sum_{j \in R_k} e^{\eta_j}}$$

$$\frac{\partial^2 l}{\partial \eta_i^2} = -e^{\eta_i} \sum_{k \in C_i} \frac{d_k}{\sum_{j \in R_k} e^{\eta_j}} + e^{2\eta_i} \sum_{k \in C_i} \frac{d_k}{(\sum_{j \in R_k} e^{\eta_j})^2}.$$

$$\frac{\partial^2 l}{\partial \eta_i \partial \eta_{i'}} = -e^{\eta_i} e^{\eta_{i'}} \sum_{k \in C_{ii'}} \frac{d_k}{(\sum_{j \in R_k} e^{\eta_j})^2} \quad (i \neq i').$$

Hastie and Tibshirani observe that there is a close relationship between the quantity $\frac{\partial l}{\partial \eta_i}$ and the generalised residuals for the proportional hazards model that is discussed below:

The generalised residual is given by:

$$\hat{e}_i = \hat{\Lambda}(t_i) e^{\eta_i},$$

where

$$\hat{\Lambda}(t_i) = \sum_{k \in C_i} \frac{d_k}{\sum_{j \in R_k} e^{\eta_j}}.$$

Hence $\frac{\partial l}{\partial \eta_i} = \delta_i - \hat{e}_i$ and therefore if there is no censoring, \hat{e}_i and $\frac{\partial l}{\partial \eta_i}$ will be equivalent. When there is censoring present, they are not quite the same, depending on whether one makes use of the practice of adding one or $\ln 2$ to the residuals that correspond to censored observations.

The partial residual is not useful for assessing the fit of a model, but can add insight whether a covariate has been misspecified.

5.1.3. Inference and Smoothing Parameter Selection

Fitting a proportional additive model in survival analysis also has a new set of challenges – the question is asked of how to select the parameters and then also, once one arrives at a model, how can that model be validated. Development and specification of a metric to measure fit can be arrived at using intuitive arguments.

Let the deviance for model $\hat{\boldsymbol{\eta}}$ be given by:

$$dev(\mathbf{y}, \hat{\boldsymbol{\eta}}) = -2[l(\hat{\boldsymbol{\eta}}) - l(\hat{\boldsymbol{\eta}}_{max})],$$

where $\hat{\boldsymbol{\eta}}_{max}$ is the parameter value that maximises $l(\hat{\boldsymbol{\eta}})$ over all $l(\hat{\boldsymbol{\eta}})$ i.e. the saturated model.

For ease of illustration, the single predictor case is considered first. Let the true values of $\boldsymbol{\eta}$, \mathbf{u} and \mathbf{A} be given by $\boldsymbol{\eta}_0$, \mathbf{u}_0 and \mathbf{A}_0 .;

Going back to the below expression discussed earlier:

$$j(\boldsymbol{\eta}) = l(\boldsymbol{\eta}) - \frac{1}{2} \sum_1^p \lambda_i \mathbf{f}_i^T \mathbf{K}_i \mathbf{f}_i,$$

and expanding $j(\hat{\boldsymbol{\eta}})$ around $\boldsymbol{\eta}_0$ and using the fact that $\mathbf{u}_0 = \mathbf{0}$, Hastie and Tibshirani show that:

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \approx (\mathbf{A}_0 + \lambda \mathbf{K})^{-1} (\mathbf{u}_0 - \lambda \mathbf{K} \boldsymbol{\eta}_0).$$

Substituting \mathbf{A} for \mathbf{A}_0 in the above expression, one obtains:

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \approx (\mathbf{A}_0 + \lambda \mathbf{K})^{-1} (\mathbf{u}_0 - \lambda \mathbf{K} \boldsymbol{\eta}_0).$$

$\hat{\boldsymbol{\eta}} \approx \mathbf{S} \mathbf{z}_0$ where $\mathbf{z}_0 = (\mathbf{A} + \lambda \mathbf{K})^{-1} \mathbf{A} \boldsymbol{\eta}_0$. As was shown previously, and this is the smoother matrix for a cubic spline. \mathbf{z}_0 is the “true” adjusted dependent variable $\boldsymbol{\eta}_0 + \mathbf{A}_0^{-1} \mathbf{u}_0$.

The variance of \mathbf{u}_0 , $\text{Var}(\mathbf{u}_0) \approx \mathbf{A}$ and therefore $\text{Var}(\mathbf{z}_0) \approx \mathbf{A}^{-1}$ and also

$$\text{Var}(\hat{\boldsymbol{\eta}}) \approx \mathbf{S} \mathbf{A}^{-1} \mathbf{S}^T.$$

This quantity can be used to construct piecewise confidence bands for $\hat{\boldsymbol{\eta}}$.

5.1.4. Degrees of Freedom

According to Hastie and Tibshirani (1987), the effective number of parameters in the model, or degrees of freedom of a model, $\hat{\boldsymbol{\eta}}$, is defined by:

$$df(\hat{\boldsymbol{\eta}}) = n - E[\text{dev}(\boldsymbol{\eta}_{max}, \hat{\boldsymbol{\eta}})].$$

Hastie and Tibshirani (1987) further show, using a standard Taylor series argument:

$$\text{dev}(\boldsymbol{\eta}_{max}, \hat{\boldsymbol{\eta}}) \approx \mathbf{u}^T \mathbf{A}^{-1} \mathbf{u} \approx (\mathbf{z}_0 - \hat{\boldsymbol{\eta}})^T \mathbf{A} (\mathbf{z}_0 - \hat{\boldsymbol{\eta}}).$$

The expected value of this value is approximately: $n - \text{trace}(2\mathbf{S} - \mathbf{S}^T \mathbf{A} \mathbf{S} \mathbf{A}^{-1})$,

and therefore the degrees of freedom:

$$df(\hat{\boldsymbol{\eta}}) \approx \text{trace}(2\mathbf{S} - \mathbf{S}^T \mathbf{A} \mathbf{S} \mathbf{A}^{-1}).$$

This can be further simplified by using the form of \mathbf{S} for cubic splines to obtain

$$df(\hat{\boldsymbol{\eta}}) = \text{trace}(2\mathbf{S} - \mathbf{S}^2).$$

This quantity can be useful as a rough guide to assessing the significance of model terms. Empirical evidence from past research shows that this quantity has an approximate chi-squared distribution with df degrees of freedom.

For all of the calculations above dealing with the variance and the degrees of freedom, the diagonal matrix approximation \mathbf{A}^* , is used in place of \mathbf{A} to make computations easier.

5.1.5. Selecting a Smoothing Parameter

The smoothing parameter is often selected subjectively by the analyst, or in line with an intuitive or expected notion, however it is sometimes useful to look at methods for automatic selection and to validate or inform selection choices.

Global cross-validation is an approach to parameter validation that can easily be used. Here, one works through all the points in the dataset, computing the entire estimation procedure n times and each time leaving out one point in the dataset. Obviously this is not ideal for bigger datasets and the procedure will be computationally expensive.

Note however that the choice of the smoothing parameter is dependent on the smoothing parameters for the other terms in the model.

Hastie and Tibshirani in their example determined a smoothing value for the continuous variables of the model that results in a moderate amount of smoothing, (approx. 4 degrees of freedom for continuous variables). Hastie and Tibshirani then carried out a backwards stepwise selection, testing and validating at each step whether the fit for each independent variable can be simplified from a smooth fit to a linear one or omitted in its entirety from the model. To quantify these, $-2\log PL$ is used as a measure of fit.

5.1.6. Tests of Hypothesis

If a Proportional Hazards model has been specified, it is important to assess the contribution of the independent covariate effects and their significance. Should a value not be significant, it can be safely excluded from the model.

Consider the example where all effects are “linear” i.e no smoothers have been applied, except for a single covariate. The model would then have the following form for the hazard ratio.

Recall that in the proportional Hazards model, the hazard ratio is parameterised by:

$$\log[\lambda(t|\mathbf{x}, z)/\lambda_0(t)] = \mathbf{X}\boldsymbol{\beta} + h(z), \quad (5.1.6.1)$$

where $\lambda_0(t)$ is as an unspecified baseline hazard function, and the unknown function h gives the effect.

Let $B_1(z), \dots, B_{m+4}(z)$ be the cubic B-spline basis and parameterise h by

$$h(z) = \theta_0 z + \sum_{k=1}^{m+2} \theta_k B_k(z). \quad (5.1.6.2)$$

The full model with the theta parameters can be estimated using the traditional test for hypothesis – in this case the above parameterisation of the function h is treated as a transformation of the original variable.

It is also possible to add a penalised term to the likelihood and develop test of hypothesis from there. The results from “Spline-Based Tests in Survival Analysis” by R Gray (1994) are given below.

In his article, Robert Gray (1994) examines a method for testing hypothesis on covariate effects / independent variables in a proportional hazards model. The approach is to formulate a flexible parametric alternative using fixed knot splines, together with penalty functions that will penalise noisy alternatives more than smooth ones so that the power of the tests are focused towards more smooth alternatives.

The test statistics are the analogs of the ordinary likelihood based statistics, but computed from a penalised likelihood formed by subtracting the penalty function from the ordinary log-likelihood. Robert Gray explains that methods for formal inference in the setting of additive proportional hazards models using splines have not been well developed. Also, an additional complexity is the calculation of the full information matrix; however useful approximations have been given and discussed earlier in this section.

Robert Gray’s approach to developing a test of hypothesis is to model the effect of an independent variable / covariate with a moderate number of knots and the use penalty functions in the estimation as would be done for non-parametric smoothing splines. Hastie and Tibshirani (1990) refer to this approach as “generalised ridge regression”.

- Tests for the Covariate Effects

Recall that in the proportional Hazards model, the hazard ratio is parameterised by:

$$\log[\lambda(t|\mathbf{x}, z)/\lambda_0(t)] = \mathbf{X}\boldsymbol{\beta} + h(z). \quad (5.1.6.1)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, and the unknown function h gives the effect of z on the outcome.

Two hypotheses are of interest here on the unknown function h :

1. $h = 0$.

This is the hypothesis of no effect - the function h has no contribution

2. $h = \theta_0 z$, where θ_0 is some unknown parameter.

This is the hypothesis that h is linear and as such no value is gained by using a smooth function to gauge its effect.

Consider m knots' locations that are specified. Robert Gray indicates that in his experience based on numerical results, the exact number and locations of the knots are not important so long as the number is large enough ("*often as few as 10 should be sufficient*") and they should be reasonable spread out. To ensure that an equal amount of data is present between the knots is also an approach recommended by Hastie and Tibshirani where they advocate placing the knots at the three quartiles.

Let $B_1(z), \dots, B_{m+4}(z)$ be the cubic B-spline basis and parameterize h by

$$h(z) = \theta_0 z + \sum_{k=1}^{m+2} \theta_k B_k(z). \quad (5.1.6.2)$$

There are only $m + 2$ of the B-spline terms that are used in this expression because the space of the cubic b-splines includes a constant and linear functions – the constant is absorbed in the underlying / baseline hazard and the linear terms is specified on its own in the equation above.

Let

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_{m+2})^T$$

and

$$\boldsymbol{\eta} = (\theta_0, \theta_1, \theta_2, \dots, \theta_{m+2})^T = (\theta_0, \boldsymbol{\theta}^T)^T.$$

The two above hypothesis that are of interest for the unknown function become h :

Two hypotheses are of interest here on the unknown function h :

1. $\boldsymbol{\theta} = \mathbf{0}$.

This is the hypothesis of no effect- the function h has no contribution

2. $\boldsymbol{\eta} = \mathbf{0}$.

In order to define the test statistic: define the ordinary log partial likelihood $L(\boldsymbol{\beta}, \boldsymbol{\eta})$ for the model 5.1.6.1 above, where the smooth function h is parameterised by 5.1.6.2.

Robert Grey suggests that to focus more power towards the smoother alternatives, a penalty function can be subtracted from $L(\boldsymbol{\beta}, \boldsymbol{\eta})$. The standard penalty function for the use with cubic splines is:

$$\frac{1}{2}\alpha \int [h''(u)]^2 du,$$

where α is a smoothing parameter that controls the degree of smoothing used. It is also possible to use other penalty functions, however in the context of splines, the above is very popular and preferred.

Robert Grey continues by observing that only the parameters in $\boldsymbol{\theta}$ appear in this penalty and also that this penalty function is quadratic in the parameters and can therefore be rewritten as:

$$\frac{1}{2}\alpha \boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta} = \frac{1}{2}\alpha \boldsymbol{\eta}'\mathbf{P}^*\boldsymbol{\eta}.$$

In this quadratic form, the matrix \mathbf{P} is positive-definite and is only a function of the knot locations.

The matrix \mathbf{P}^* is a $(m+3) \times (m+3)$ with the first row and column populated by zero, and \mathbf{P} in the remainder of the matrix.

It is now possible to re-write the penalised likelihood as:

$$L_p(\boldsymbol{\beta}, \boldsymbol{\eta}) = L(\boldsymbol{\beta}, \boldsymbol{\eta}) - \frac{1}{2}\alpha \boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}.$$

Let the values of the parameters that maximise the penalised likelihood, $L_p(\boldsymbol{\beta}, \boldsymbol{\eta})$ be given by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$.

For the hypothesis $\boldsymbol{\eta} = \mathbf{0}$:

Let $\widehat{\boldsymbol{\beta}}_{0l}$ be the maximum partial likelihood estimator for $\boldsymbol{\beta}$ when $\boldsymbol{\eta} = \mathbf{0}$. Let the standard partial likelihood score vector be denoted by $S(\boldsymbol{\beta}, \boldsymbol{\eta}) = (S'_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\eta}), S'_{\boldsymbol{\eta}}(\boldsymbol{\beta}, \boldsymbol{\eta}))'$. Let \mathbf{I} be the information matrix from the unpenalised partial likelihood, with subscripts denoting the submatrices i.e $\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}$ for the derivatives with respect to $\boldsymbol{\eta}$.

Note that $\frac{\partial L_p(\boldsymbol{\beta}_0, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = S_{\boldsymbol{\eta}}(\widehat{\boldsymbol{\beta}}_0, \boldsymbol{\eta})$ and that the negative of the $\boldsymbol{\eta}\boldsymbol{\eta}$ portion of the information matrix of the penalised likelihood is $\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} + \alpha \mathbf{P}^*$ and the other components simply the corresponding components of \mathbf{I} .

Robert Gray gives three different test statistics:

1. A Penalised quadratic score statistic

$$Q_s = \mathbf{S}'_{\boldsymbol{\eta}}(\widehat{\boldsymbol{\beta}}_0, \mathbf{0})(\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}} + \alpha \mathbf{P}^*)^{-1} \mathbf{S}_{\boldsymbol{\eta}}(\widehat{\boldsymbol{\beta}}_0, \mathbf{0}).$$

where

$$\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}} = \mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} - \mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\beta}} \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\eta}}.$$

2. A likelihood ratio statistic

$$Q_l = 2[L_p(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}) - L_p(\widehat{\boldsymbol{\beta}}_0, \boldsymbol{\eta})].$$

3. A Wald-type test statistic

$$Q_w = \widehat{\boldsymbol{\eta}}' (\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}} + \alpha \mathbf{P}^*) \widehat{\boldsymbol{\eta}}.$$

For each of the above, the test statistic will reject for large values of the statistic. The Likelihood ratio statistic is similar to the deviance statistics discussed in an earlier section and in Hastie and Tibshirani (1990).

To construct an hypothesis that the effect is linear, in other words: $\boldsymbol{\theta} = \mathbf{0}$, is done in exactly the same way, except with θ_0 included with $\boldsymbol{\beta}$ instead of $\boldsymbol{\eta}$. It is possible to consider more general forms of h , but that will not be considered here.

To obtain the approximate distributions of the test statistics, Robert Grey only considers the case where the number of knots as well as the number of parameters is held fixed as the sample size increases. Where the knots are assumed to remain fixed, it is further assumed that the usual

conditions are satisfied so that standard asymptotic expansions will hold for the unpenalised partial likelihood. – Anderson and Gill (1982).

Then under the null hypothesis, the statistics Q_s , Q_l and Q_w all have the same asymptotic distribution which is that of

$$\sum \lambda_j Z_j^2,$$

where the Z_j are independent standard normal random variables and the λ_j are the eigenvalues of the matrix

$$\lim \mathbf{I}(\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}) \left(\mathbf{I}(\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}) + \alpha \mathbf{P}^* \right)^{-1}.$$

Robert Gray references the works of Imhof (1961) and Davies (1973, 1980) for their development of methods for the distribution of a linear combination of chi-squares based on inverting the characteristic function. In his own simulations and examples, Robert Gray used eigenvalues derived from the data for the power of the tests.

Asymptotically the expected value under the null hypothesis of any of the three statistics is:

$$\sum \lambda_j = \text{trace}(\lim \mathbf{I}(\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}) (\mathbf{I}(\boldsymbol{\eta}\boldsymbol{\eta}|\boldsymbol{\beta}) + \alpha \mathbf{P}^*)^{-1}).$$

For unpenalised likelihood, this would be the degrees of freedom of the test. This definition for degrees of freedom also correlated to that of Hastie and Tibshirani.

In the above approach, the value of α is not specified directly. Rather, a value for the degree of freedom is specified and the corresponding value of α is used. This approach relies heavily on the data analyst specifying an appropriate number for the degrees of freedom (how much smoothness is required) and then being able to translate that into a value for α , given the suitability of the data.

Where the sample size is small, a few degrees of freedom can be specified. Other methods such as cross-validation can be useful but it is not clear if such a method would lead to best tests. Also insightful is Robert Gray's recommendation to have the number of knots specified for the spline must be twice as large as the degrees of freedom.

5.1.7. A Specific Application of a Mixed Model

In their article: *A semiparametric multilevel survival model*, Zang and Steele (2004), proposed a semiparametric multilevel survival for clustered duration data. The investigation concerned the first birth intervals (time from marriage to first child) of women in Bangladesh.

The model that was proposed is an example of a mixed model – where some of the predictors are not transformed and the remainder is modelled using B-splines.

The proposed model takes the form:

$$h(t|Z_i, X_i) = h_0(t)e^{Z_i\beta + f(X_i)}, \quad (5.1.7.1)$$

where $Z_i = [X_1, X_2, \dots, X_p]$ - a matrix of observed variables for p independent covariates, with β a p -valued vector of associated parameter estimates, and $f(X_i)$ is an unknown smooth function, applied to an independent variable not contained in Z_i .

The above model is a special case of the Generalised Additive Model.

Common candidates for $f(\cdot)$ are splines and local polynomials. Note that when a B-spline is used where the variable is effectively transformed into its b-spline, no special considerations need to be taken to compute the likelihood and the method in section 3.2.3 can be used.

However, working from first principles to estimate the likelihood for the model 5.1.7.1 starting with the method of partial likelihood (Cox, 1972) and with Peto's (1972) approximation to ties, the partial likelihood can be explicitly written as:

$$\sum_{l=1}^L \left\{ \sum_{j \in D_l} \left(Z_j \beta + f(X_j) - \log \left[\sum_{k \in R_l} e^{Z_k \beta + f(X_k)} \right] \right) \right\}.$$

where D_l is the number of events at time $t_{(l)}$ and the associated set of indices for individuals at risk at time $t_{(l)}$ is R_l .

Zang and Steele explain that there are many ways of dealing with the non-linear component of the above partial likelihood and they took the local linear approach because of its “design-adaptive and automatic boundary correction”, quoting Fan and Gijbels (1996). Their approach to the smooth function is to use a Kernel function.

From Taylor's expansion, for any x one can write:

$$f(X) \approx f(x) + f'(x)(X - x) = a + b(X - x)$$

when X is close to x .

When combining the above Taylor expansion with the above partial likelihood function, one obtains:

$$\sum_{l=1}^L \left\{ \sum_{j \in D_l} K_h(X_j - x) (Z_j \beta + a + b(X_j - x)) - \log \left[\sum_{k \in D_l} e^{Z_k \beta + a + b(X_k - x) K_h(X_k - x)} \right] \right\},$$

where

$$K_h = K(\cdot/h)/h,$$

and $K(\cdot)$ is a Kernel function.

In the above expression, a cancels and an estimator for $f(x)$ cannot be obtained directly. However, the derivative of $f(x)$ can be estimated.

Let $\hat{\beta}(x)$ and $\hat{b}(x)$ maximise the above expression. Let the estimator for $f'(x)$ be taken as $\hat{b}(x)$, which in turn can be used to develop an estimator for $f(x)$ by:

$$\hat{f}(x) = \int_c^x f'(u) du = \int_c^x \hat{b}(u) du.$$

Set $c = 0$ for convenience. An estimator for β can be obtained as follows:

Let $X_{(1)} < X_{(2)} < \dots < X_{(m)}$ be the distinct values of X_1, X_2, \dots, X_n .

On the basis of $X_{(k)}$, the estimator for $\hat{\beta}(X_{(k)})$, $k = 1, \dots, m$ is obtained.

The estimator for β is taken as the average of $\hat{\beta}(X_{(k)})$:

$$\hat{\beta} = \frac{1}{m} \sum_{k=1}^m \hat{\beta}(X_{(k)}).$$

Zang and Steele state that this estimator is based entirely on the information that is provided by the semi-parametric structure which makes the estimate more efficient.

They also indicate that though the base-line hazard is viewed as a nuisance parameter, it can be estimated using a method such as Breslow's estimator (Breslow, 1972, 1974) after an estimation of β and f have been obtained.

Using Breslow's estimator, the estimate of the base-line hazard function $h_0(\cdot)$ at time $t_{(l)}$, $l = 1, \dots, L$ is given by:

$$\hat{h}_0(t_{(l)}) = \left[\sum_{k \in R_l} e^{Z_k \hat{\beta} + \hat{f}(X_k)} \right]^{-1}.$$

By smoothing $\hat{h}_0(t_{(l)})$ against $t_{(l)}$ i.e. by viewing $(t_{(l)}, \hat{h}_0(t_{(l)}))$, $l = 1, \dots, L$, as a sample from the model:

$$y = h_0(t) + \varepsilon,$$

and making use of the local linear technique, it follows that the estimator for $\hat{h}_0(t)$ can be obtained by:

$$\hat{h}_0(t) = (1, 0)(\mathbf{T}^T \mathbf{W} \mathbf{T})^{-1} \mathbf{T}^T \mathbf{W} \hat{\mathbf{h}}_0,$$

where

$$\mathbf{T} = \begin{pmatrix} 1 & t_{(1)} - t \\ \vdots & \vdots \\ 1 & t_{(L)} - t \end{pmatrix}$$

$$\hat{\mathbf{h}}_0 = \begin{pmatrix} \hat{h}_0(t_{(1)}) \\ \vdots \\ \hat{h}_0(t_{(L)}) \end{pmatrix},$$

and

$$\mathbf{w} = \text{diag}\{K_h(t_{(1)} - t), \dots, K_h(t_{(L)} - t)\}.$$

The natural estimator for the cumulative base-line hazard

$$\Lambda_0(t) = \int_0^t h_0(u) du,$$

is

$$\hat{\Lambda}_0(t) = \int_0^t \hat{h}_0(u) du.$$

In practice, the approximation

$$\hat{\Lambda}_0(t) = \sum_{t_{(l)} \leq t} \hat{h}_0(t_{(l)}),$$

is used.

6. CHAPTER 6 –An Application of Additive Modelling for Survival Times

6.1. Introduction

In this section, the classic and additive Proportional Hazards models will be applied to Telecommunications / Network data. The results will be interpreted and analysed for insights and to determine if tools can be developed to aid in customer lifetime modelling.

The aim of the analysis is to study the survival of postpaid customers – which consist of Top Up and Contract packages. The relationship with the customer ends when the customer disconnects from the network.

Looking only at data for 2013, the following figures have been observed for a major Telecommunications network in South Africa (referred to as “The Network”).

Every month sees on average 30 000 subscribers who disconnect from and effectively end their relationship with the Network. The average value of a postpaid subscriber is R279 per month. This translates to a financial loss of approximately R8.5 million per month, not taking into consideration the future value of the subscriber.

Table 9 Number of Disconnected Subscribers per Month

Month	Number of Disconnects	Value
Jan-13	30429	R 8 459 262.00
Feb-13	24803	R 6 895 234.00
Mar-13	32876	R 9 139 528.00
Apr-13	32151	R 8 937 978.00
May-13	35147	R 9 770 866.00
Jun-13	32088	R 8 920 464.00
Jul-13	36529	R 10 155 062.00
Aug-13	36313	R 10 095 014.00
Sep-13	39926	R 11 099 428.00
Oct-13	38008	R 10 566 224.00
Nov-13	23848	R 6 629 744.00
Dec-13	19115	R 5 313 970.00
Total	381233	R 105 982 774.00

The aim of the network is to mine the subscriber usage data and develop insights, recommendations and potentially predictive tools that will aid the network in securing a future relationship with the subscriber.

The value is also evident from the above table in the financial losses that the network could possibly prevent.

It is expected that the Proportional Hazards models, when applied to usage data will yield insights in behaviour that will be of value to the Network.

6.2. Data preparation

6.2.1. Data received

The original dataset received for the analysis contained 6 865 877 records. The unique identifier is subs_id000 and there were 3 701 754 unique values of this field.

The dataset can be divided into two parts namely a *performance* component and a *movement* component. The *performance* component contains behavioral historic performance of the subscribers, specifically:

- Val – the monthly value
- Voice – the monthly value for voice
- SMS – the monthly value for SMS
- Data – the monthly value for data
- Duration – the monthly call duration
- SMS_events – the total number of SMS events in the month
- MB – the total megabytes used for data in the month

The above characteristics are available for October 2012 up to and including September 2013.

The dataset also contains static fields such as:

- price plan
- payment method

- service provider
- business consumer class
- sales region
- sales channel
- sales town

The rows of the *performance* component are sometimes duplicated if a subscriber had more than one movement in the time window October 2012 – September 2013.

Movements include –

- Connections
- Port-ins
- Conversions
- Migrations
- Disconnections
- Port-outs
- SP-changes

For every movement, there is detail on the price plan, payment method, service provider to and from as well as the movement date.

Since it is possible for a subscriber to have multiple movements, the joining of the *movements* component to the *performance* component caused rows in *performance* to be duplicated – this is the reason that there are fewer unique sub_id000 than total rows in the table.

6.2.2. Data Manipulation Steps

6.2.2.1. Introduction

The below data manipulation steps describe how the original data received was analysed and manipulated to arrive at a dataset that:

- Displays Postpaid Customers who were active on the network as at 1 January 2013 and that were,

- At risk of disconnecting from the network to a Top Up package in the analysis window of 1 January 2013 – December 2013

6.2.2.2. Steps

The *performance* component was isolated from the original dataset. Rows were deduplicated on subs_id000 – in total 3 164 123 rows deleted from 6 865 877 so that the new table displays performance with 113 variables for 3 701 754 subscribers.

The *movements* component was isolated from the original dataset and since the aim of the analysis has only subscribers who disconnected, the movement table was not required.

The physical date of disconnection was available and was used to create the fields:

- BGI – which is binary with 0 when a subscriber has disconnected and 1 when a subscriber is still active on the network
- The customer lifetime

The following characteristics were constructed for analysis and inclusion in the model:

- Tenure in months – the total tenure of the customer as at 1 January 2013
- Time to upgrade – the number of months until the subscriber can upgrade from 1 January 2013
- Usage Characteristics:
 - Age – then subscriber age in years as at 1 January 2013
 - ave_val as the average value for October 2012 to December 2012
 - ave_voice as the voice value for October 2012 to December 2012
 - ave_sms as the average SMS value for October 2012 to December 2012
 - ave_data as the average data for October 2012 to December 2012
 - ave_dur as the average duration for October 2012 to December 2012
 - ave_smsevents as the average SMS events for October 2012 to December 2012
 - ave_mb as the average megabytes for October 2012 to December 2012

For the above usage characteristics, missing values were recoded to zero.

- Lifetime:
 - Lifetime in days from 1 January 2013 to disconnection date or 15 December 2013 if no movement occurred

Two records with shifted values were removed.

Subscribers who connected onto the network after 1 January 2013 were also removed – there were 46 757 such subscribers.

Subscribers who disconnected from the network before 1 January 2013 were excluded – there were 394 475 such subscribers.

The final working dataset had 3 260 520 rows.

6.2.3. Exploratory Analysis

6.2.3.1. Introduction

Crowley, Leblanc, Gentleman and Salmon (1995) in their article discuss several exploratory methods that can be used when conducting a survival analysis. This is because despite a lot of research in the field of survival analytics, a lot of the practice and application still retains a “black box” flavour.

Crowley, Leblanc, Gentleman and Salmon introduce several tools that can be used as exploratory tools to get started with a survival analysis. These include Box plots, running median plots and non-parametric estimation of the Cox regression function.

Two aspects of survival analysis are responsible for the lack of use of exploratory methods in a survival analysis:

1. Censoring – for some individuals only partial information will be available
2. The Cox Proportional Hazards Model does not lend itself easily to visual representations on displays of data i.e. in a linear regression, one can easily plot the dependent variable against the independent variable to gain an understanding of the nature of the relationship between the two variables.

The purpose of the methods introduced is to describe the relationship between response and covariate – in effect using the data to derive some idea of the relationship that exists as opposed to test hypothesis that certain relationships hold.

For this practical, some of these methods will be used and is described in this section.

6.2.3.2. Exploratory data analysis

The following independent variables are considered in the survival analysis:

- Tenure in months
- Time to upgrade
- Age ave_val
- ave_voice
- ave_sms
- ave_data
- ave_dur
- ave_smsevents
- ave_mb

Dependent variables:

- Survival time – the length in days from 1 January 2013
- BGI – a binary variable to indicate whether the disconnected or remained active in the time analysis time window.

Disconnect distribution

The distribution of disconnects is displayed below:

Table 10 Number of Disconnected Subscribers per Month

Month	Number of Disconnects
Jan-13	30429
Feb-13	24803
Mar-13	32876
Apr-13	32151
May-13	35147
Jun-13	32088
Jul-13	36529
Aug-13	36313
Sep-13	39926
Oct-13	38008
Nov-13	23848
Dec-13	19115
Total	381233

Distribution Analysis

Summary statistics were created for the independent variables to gain insight into their distributions:

Table 11 Summary Statistics of Independent Variables

Variable	Mean	Std Dev	N Miss	Lower Quartile	Median	Upper Quartile
ave_val	278.4399	649.7297	0	8.0104	123.6	314.5565
ave_voice	217.9485	474.356	0	0	87.25333	256.555
ave_sms	26.75826	136.6045	0	0	5.0502	26.4898
ave_data	33.73314	354.7175	0	0	0	0.038
ave_dur	13655.12	27808.06	0	1058.33	5822.33	14817.33
ave_smsevents	67.0828	190.5567	0	0.666667	17.33333	68.66667
ave_mb	58984.54	541807.7	0	0	0	127.498
AGE	44.49936	12.0853	77	35	43	52
time_to_upgrade_months	18.88122	16.71181	370157	15	22	29
Tenure_months	77.89524	53.59374	0	31	67	113

The average value of a subscriber is R278, with voice R217 for voice, R26 for SMS and R33 for data. Average duration is 13655 seconds with 67 SMS sent and 58 984 megabytes of data consumed.

The postpaid subscriber has an average tenure of 78 months and is 44 years old. The 77 missing values for AGE were replaced with the average value of this characteristic so that their records will not be excluded from the regression.

The positive values of time to upgrade in months were binarised to also be included in the model as it does not make intuitive sense to impute the missing values with the average value.

Correlation:

The Pearson Correlation was calculated for the independent and dependent variables. Notable pairs of variables to be mindful of include:

- Ave Val and Ave Data
- Ave Val and Ave Dur
- Ave Voice and Ave Duration
- Ave SMS and Ave SMS Events
- Ave data and Ave MB

Table 12 Correlation of Independent Variables

	Tenure months	time to upgrade months	ave val	ave voice	ave sms	ave data	ave dur	ave smsevents	ave mb	AGE	BGI	lifetime
Tenure months	1	0.11508	0.09258	0.08855	0.05748	0.02903	0.18266	0.14136	0.04288	0.23051	0.17889	0.15294
time to upgrade months	0.11508	1	0.10133	0.10081	0.05989	0.02979	0.12476	0.10352	0.03198	-0.02111	0.1489	0.14407
ave val	0.09258	0.10133	1	0.80821	0.36211	0.61143	0.65961	0.36256	0.33851	-0.00158	0.01954	-0.0009
ave voice	0.08855	0.10081	0.80821	1	0.17335	0.07635	0.74353	0.27748	0.05643	0.00066	0.01666	-0.00465
ave sms	0.05748	0.05989	0.36211	0.17335	1	0.04635	0.1678	0.5037	0.0331	0.00656	0.02476	0.01743
ave data	0.02903	0.02979	0.61143	0.07635	0.04635	1	0.14928	0.09905	0.53183	-0.0063	0.00398	-0.00213
ave dur	0.18266	0.12476	0.65961	0.74353	0.1678	0.14928	1	0.39028	0.15836	0.01842	0.04355	0.0222
ave smsevents	0.14136	0.10352	0.36256	0.27748	0.5037	0.09905	0.39028	1	0.10666	0.01332	0.0572	0.04361
ave mb	0.04288	0.03198	0.33851	0.05643	0.0331	0.53183	0.15836	0.10666	1	-0.01529	0.01097	0.00835
AGE	0.23051	-0.02111	-0.00158	0.00066	0.00656	-0.0063	0.01842	0.01332	-0.01529	1	0.10405	0.09565
BGI	0.17889	0.1489	0.01954	0.01666	0.02476	0.00398	0.04355	0.0572	0.01097	0.10405	1	0.85739
lifetime	0.15294	0.14407	-0.0009	-0.00465	0.01743	-0.00213	0.0222	0.04361	0.00835	0.09565	0.85739	1

The pairs highlighted above make intuitive sense as for example, average data usage in revenue and physical megabytes consumed will be related. The same holds for Voice and Duration. Since the packages are mostly for telephony and data, value and average duration and data will have a relationship.

Dependence analysis:

The Independent variables were analysed to determine whether there was an upfront difference in the distributions of the independent variables with regards to subscriber who had and who did not have a movement in the time window January 2013 – December 2013.

The independent variables were reformatted into intervals based on the quartiles and the disconnect percentage in each interval calculated. This is done to understand how disconnection may be related to the independent variables. The results are displayed graphically below.

From the below graphs, it appears that subscribers who disconnect will have:

- a shorter tenure on the network
- a shorter time to upgrade
- Lower value
- The same voice, data, SMS and data usage
- a younger age

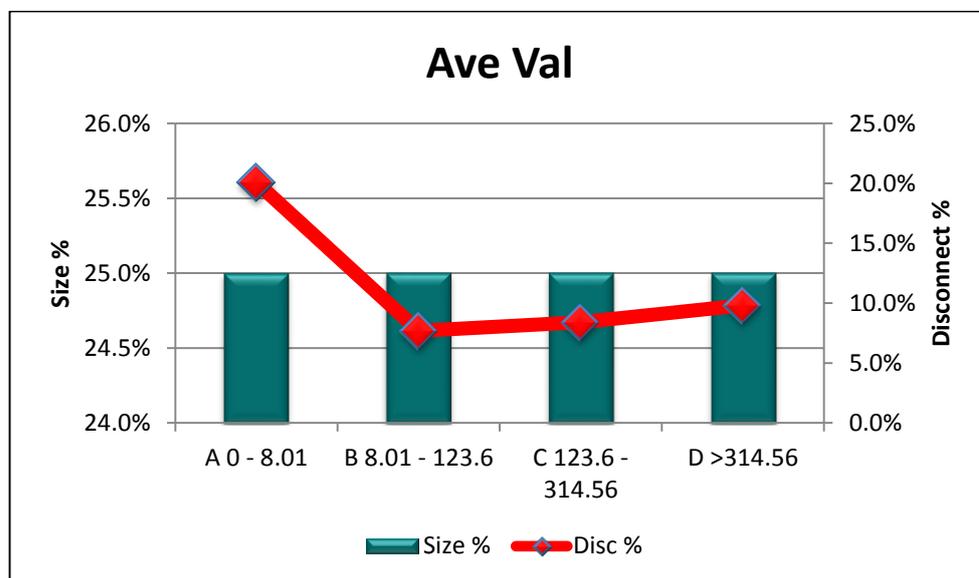


Figure 6 Disconnection and Size Distribution for Average Value

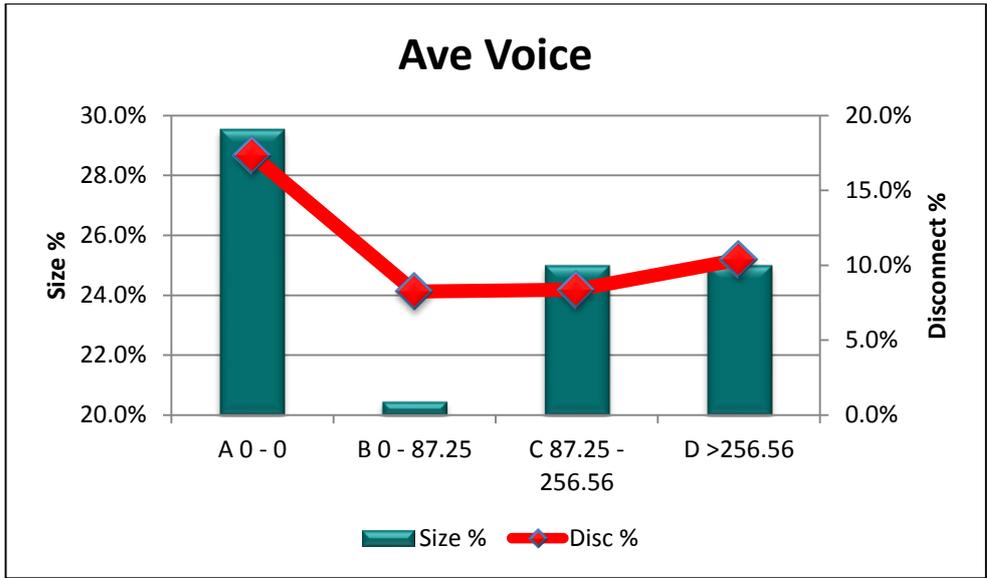


Figure 7 Disconnection and Size Distribution for Average Voice

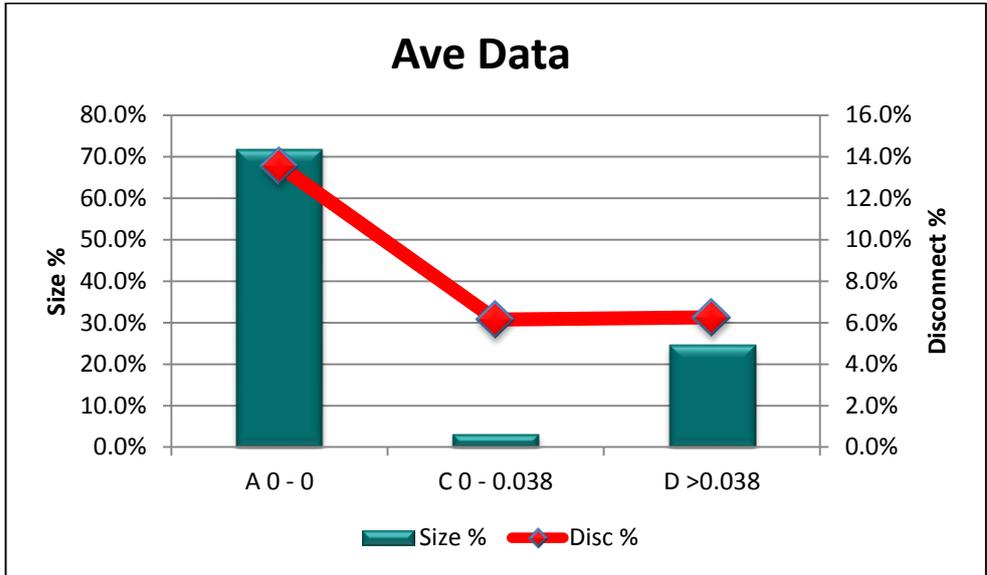


Figure 8 Disconnection and Size Distribution for Average Data

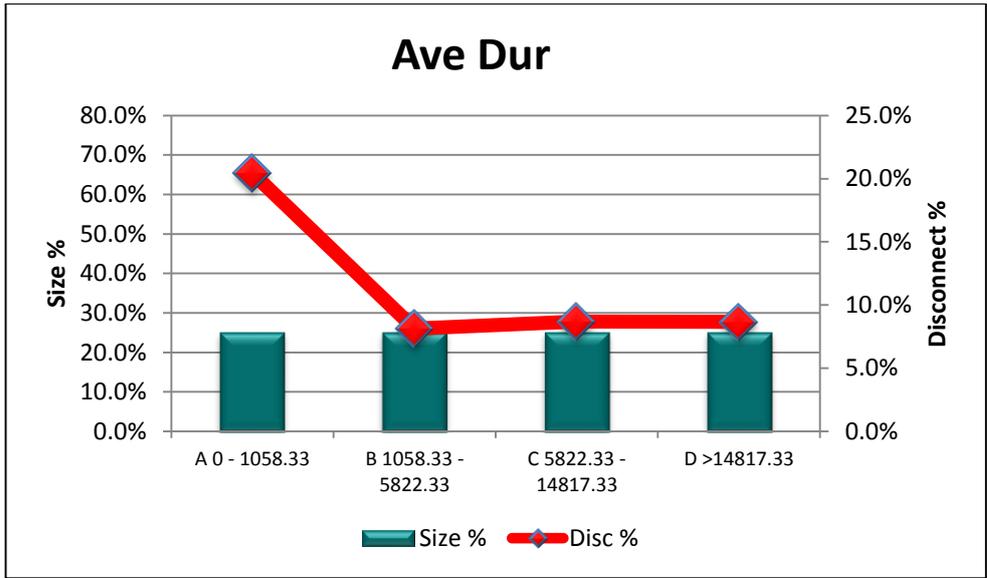


Figure 9 Disconnection and Size Distribution for Average Duration

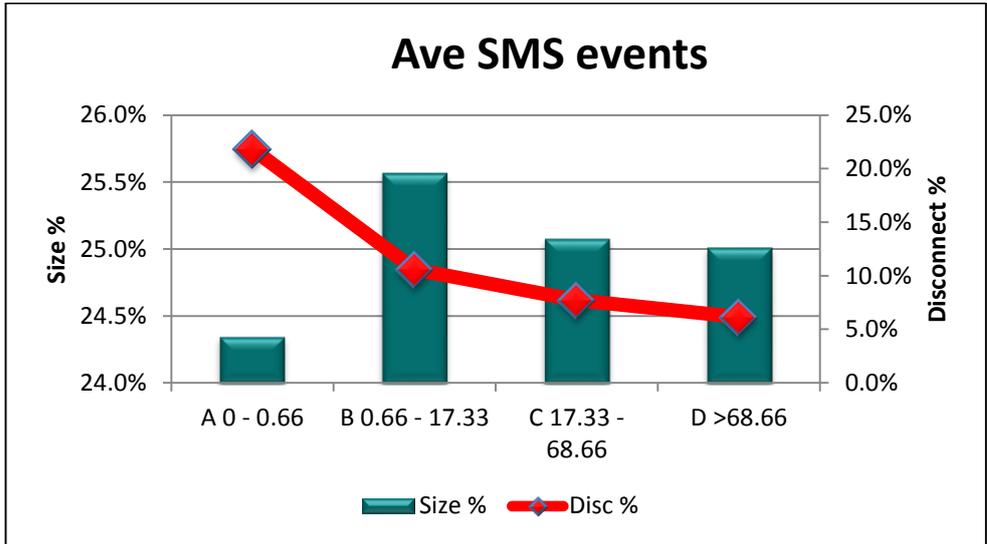


Figure 10 Disconnection and Size Distribution for Average SMS Events

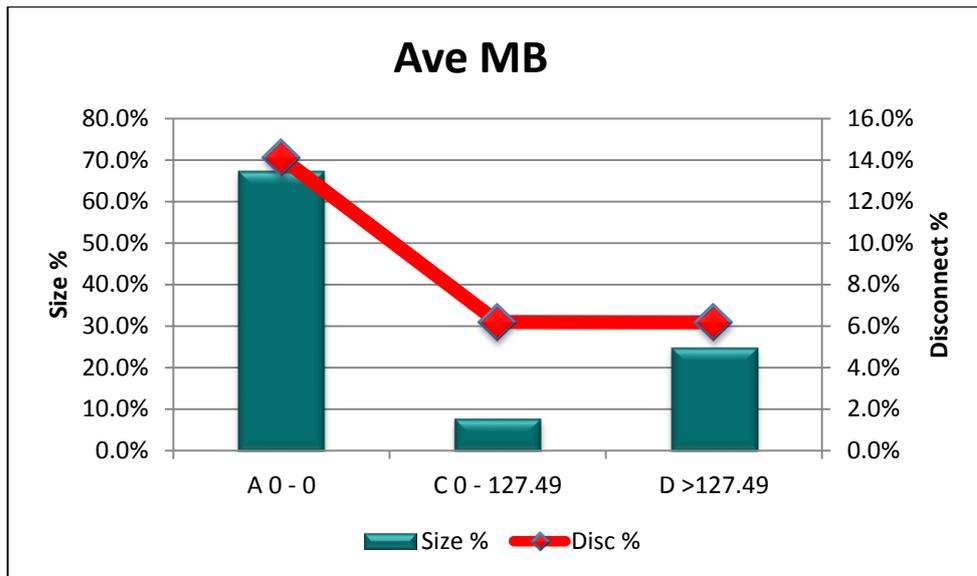


Figure 11 Disconnection and Size Distribution for Average Megabytes

For the usage characteristics; average value, average voice, average data, average duration, average SMS events and average megabytes, low values are associated with a higher disconnection rate that generally decreases as the value of the characteristic increases. This makes intuitive sense as one expects subscribers who are more ‘active’ to have a lower likelihood of disconnecting.

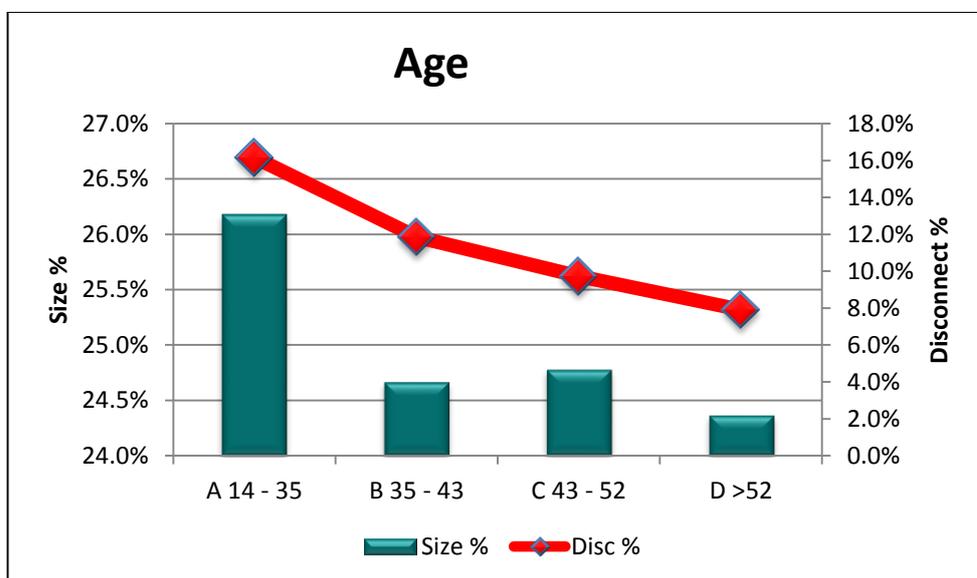


Figure 12 Disconnection and Size Distribution for Age

Looking at subscriber age, younger subscribers have a higher disconnection percentage than older subscribers – this also makes intuitive sense as one expects younger subscribers to be more ‘volatile’ and they may be less loyal towards the brand.

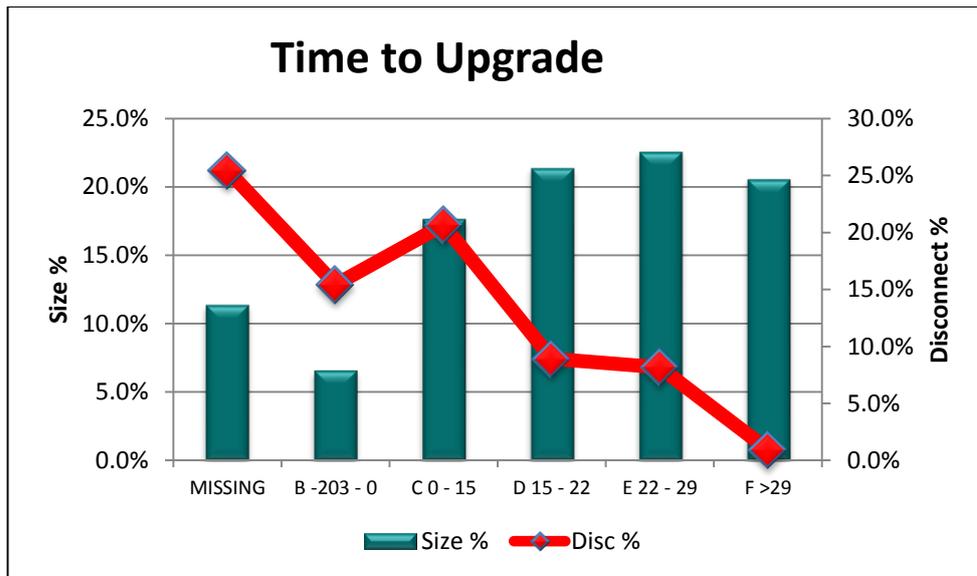


Figure 13 Disconnection and Size Distribution for Time to Upgrade

When analysing the time to upgrade a negative value (upgrade date is in the past and overdue) has a low disconnection percentage – this can be the case of subscribers who are loyal to the brand and will upgrade when they need to.

Upgrading is also an opportunity for a subscriber to terminate their relationship with the network and therefore a decreasing rate of disconnection percentage as time to upgrade increases makes sense – the closer the upgrade date to today, the higher the chance that subscribers are looking to other deals and may not stay with the network.

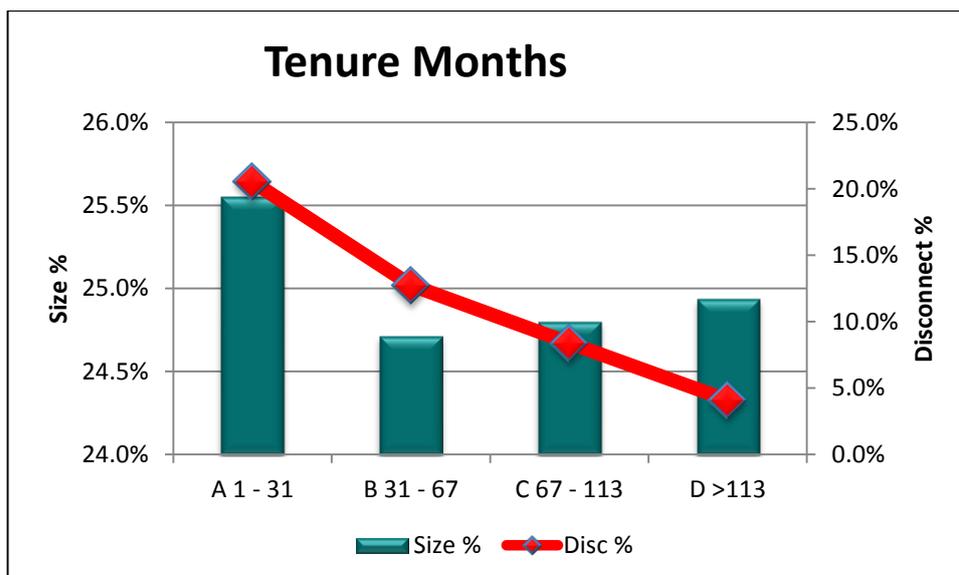


Figure 14 Disconnection and Size Distribution for Tenure

Lastly, the customer tenure shows that the longer a subscriber has been with the Network, the higher the chance that they will remain with the network.

6.2.4. Classic Proportional Hazards Modelling

6.2.4.1. Introduction

The Cox proportional Hazards model was fitted to the data and the results discussed in this section.

A Stepwise regression was run to determine which variables would be predictive in determining the hazard of disconnecting from the network.

Thereafter, results were analysed to determine whether they make intuitive and business sense, are interpretable and valid.

6.2.4.2. Modelling Strategy and Results

A Stepwise regression on the independent variable set yielded the following results – all variables are available to be selected and entered into the model. As mentioned earlier, the *time to upgrade* characteristics had been transformed into dummy variables and another indicator variable, *payment_method_dummy* to indicate whether a subscriber is a “Top Up” or a “Contract” subscriber was created. At onset, it is believed that the behavioural characteristics between these two levels of payment method will be different.

The selection results from the stepwise are given in the table below:

Table 13 Stepwise Selection

Summary of Stepwise Selection								
Step	Effect		DF	Number	Score	Wald	Pr > ChiSq	Effect
	Entered	Removed						
1	upgrade_months_02		1	1	240739		<.0001	[0,8] (-5.42R,6.25%)
2	upgrade_months_05		1	2	41350.4		<.0001	(28,33] (7.14R,15.67%)
3	AGE		1	3	41304.8		<.0001	
4	upgrade_months_06		1	4	31248.7		<.0001	(33, HIGH] (58.8R,7.82%)
5	upgrade_months_03		1	5	14742.6		<.0001	(8,22] (1.3R,32.74%)
6	upgrade_months_04		1	6	36119.3		<.0001	(22,28] (1.33R,19.57%)
7	upgrade_months_01		1	7	3546.9		<.0001	[-998,-1] (-1.43R,6.49%)
8	ave_smsevents		1	8	2762.85		<.0001	
9	ave_val		1	9	3942.25		<.0001	
10	payment_method_dummy		1	10	1173.24		<.0001	
11	ave_data		1	11	1042.61		<.0001	
12	ave_sms		1	12	148.977		<.0001	

The *time to upgrade* dummy variables were selected in the first two steps. After selecting Age, all of the remaining *time to upgrade* levels were selected. Of the usage characteristics, only *average sms events, value, data* and *SMS* made it into the model. Finally, the *payment method indicator* is also featured.

The *time until a subscriber upgrades* makes intuitive sense as many subscribers have the opportunity to change to another deal very easily.

Age also makes intuitive sense as the longer a older subscriber exhibit lower disconnection rates than younger, more 'volatile' subscribers.

The Maximum Likelihood Estimates for the above model are given below:

Table 14 MLE for Stepwise Selection

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter	Standard	Chi-Square	Pr > ChiSq	Hazard	Label
		Estimate	Error			Ratio	
AGE	1	-0.02703	0.000153	31117.5	<.0001	0.973	
ave_val	1	0.0000832	1.18E-06	4943.2	<.0001	1	
ave_sms	1	0.0000343	4.79E-06	51.3272	<.0001	1	
ave_data	1	-0.0000429	1.94E-06	486.556	<.0001	1	
ave_smsevents	1	-0.0015	2.21E-05	4645.29	<.0001	0.998	
payment_method_dummy	1	-0.11803	0.00352	1126.76	<.0001	0.889	
upgrade_months_01	1	-0.32467	0.00676	2305.98	<.0001	0.723	[-998,-1] (-1.43R,6.49%)
upgrade_months_02	1	0.80643	0.00482	28013.1	<.0001	2.24	[0,8] (-5.42R,6.25%)
upgrade_months_03	1	-0.92395	0.00466	39276.7	<.0001	0.397	(8,22] (1.3R,32.74%)
upgrade_months_04	1	-0.96517	0.00541	31848.9	<.0001	0.381	(22,28] (1.33R,19.57%)
upgrade_months_05	1	-2.63705	0.01098	57703.7	<.0001	0.072	(28,33] (7.14R,15.67%)
upgrade_months_06	1	-4.63383	0.04197	12191.3	<.0001	0.01	(33, HIGH] (58.8R,7.82%)

The parameter estimates are all very small with small associated standard errors for the usage characteristics. Their Hazard ratios are also close to one.

Looking at the time to upgrade, an interesting trend in parameter estimates is noted:

For a negative time to upgrade the estimate is negative – and then negative again and decreasing for values of 8 and higher.

The -2LogL of the model is 10841247. 10792750

The negative estimates of the parameter indicate that this parameter will decrease the hazard, whereas a positive will increase the hazard. Looking at the *time to upgrade* for the interval 0 to 8 months, the estimate is 0.80643, meaning that the hazard will be much higher for these values of *time to upgrade* than for the remainder of the interval.

This is visually represented below by the survival function which is nearly flat for the twelve month analysis window.

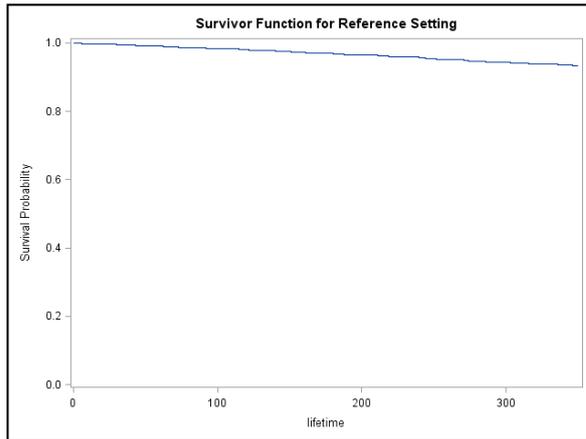


Figure 15 Survival Function for Disconnections

The reference values of the independent variables are given below – these are the average values of each variable and are used in the baseline hazard to compare against.

Table 15 Values for Reference Setting

AGE	44.4994
ave_val	278.44
ave_sms	26.7583
ave_data	33.7331
payment_method_dummy	0.46896
upgrade_months_recoded_01	0.06602
upgrade_months_recoded_02	0.06249
upgrade_months_recoded_03	0.32726
upgrade_months_recoded_04	0.19568
upgrade_months_recoded_05	0.15679
upgrade_months_recoded_06	0.07824

6.2.5. Proportional Hazards Modelling using Cubic Spline functions

6.2.5.1. Introduction

In the previous sections, the Classic / Popular Cox proportional hazards model was fitted to the data and yielded useful results.

In this section, some transformations of the independent variable set using cubic spline functions will be explored and discussed.

6.2.5.2. Background

Heinzl and Kaider (1997) in their article “Gaining more flexibility in the Cox proportional hazards regression models with cubic spline functions” discuss the use of cubic spline functions in the Cox regression model as well as how to apply such transformations using SAS.

The use of cubic splines makes the Cox regression model more flexible and is easy to apply and interpret. With the use of other smoothers such as running medians, the usefulness of developing a predictive tool would have been restricted, however in this case a predictive tool can be developed using procedures and software that are available to most organisations.

Cubic spline functions were used by Durrleman and Simon (1989) to investigate and detect possible non-linear independent variable and lifetime relationships in the Cox model. They can also be used to detect possible time dependence in the covariates.

Despite their appeal and interesting features in the analysis and exploration of data, the cubic splines have not enjoyed a lot of attention because of their rather bulky formulas.

Heinzl and Kaider provide a useful SAS programming macro that:

- Uses “put” statements to generate SAS code to apply cubic splines in the Cox model (PROC PHREG)
- Prepares the results for plotting (PROC IML)

- Plots the results (PROC GPLOT)

The user obtains an executable SAS program.

6.2.5.3. The RCS Macro

The purpose of the RCS macro is to overcome the hurdle of having to code the large mathematical expressions associated with cubic splines and to produce a program that will contain the formulas and that can be edited by the user thereafter in line with their requirements.

The Macro can be used to test or analyse:

1. The non-linear functional relationships of continuous independent variables
2. Interactions of covariates with time
3. Interactions of binary time-dependent independent variables over time

The Macro requires the following inputs to execute:

- TITLE – The title of the analysis
- DATA- The name of the SAS data set that will be used in the analysis
- DIRDATA – The location of the input directory where the dataset can be found
- PROGRAM – The name of the SAS file that will be generated from the macro that will contain the PHREG, IML and GPLOT statements to run the analysis
- TIME – The name of the variable associated with survival time
- STATUS – The name of the variable that indicates whether the observation is censored or not. Censored observations are coded as zero and failures otherwise
- COV1 – COV20 – The names of up to 20 independent variables to may be included in the analysis. These variables need to be specified in consecutive order
- WHAT1 – WHAT20 – These statements tell the RCS macro what to do with the corresponding covariate i.e.
 - WHATn = 0 – to model a non-linear effect with cubic splines
 - WHATn = 1 – to assess an interaction of a covariate with time
 - WHATn = 2 – to model the interaction of a binary time dependent variable over time
 - Otherwise – to model the independent variable as a normal variable in the model

- KNOTS1 – KNOTS20 – the location of knots if the n^{th} covariate is to be modelled using splines. The number of knots will look for the associated action to execute in the corresponding WHATn statement
- Graph – to produce graphics for the relative hazard ratio function or the log relative hazard function or both
- TIMEUNIT – the time unit label of the X-axis

6.2.5.4. Modelling Strategy and Results

From the previous section, the following variables were selected from the stepwise regression for inclusion in the model:

Continuous variables:

- Age
- Ave_val
- Ave_sms
- Ave_data
- Ave_smsevents

Dummy / indicator variables:

- payment_method dummy
- Upgrade_months_recoded_01 (for time to upgrade -998 – 0)
- Upgrade_months_recoded_02 (for time to upgrade 0 - 8)
- Upgrade_months_recoded_03 (for time to upgrade 9 - 22)
- Upgrade_months_recoded_04 (for time to upgrade 23 - 28)
- Upgrade_months_recoded_05 (for time to upgrade 28 - 33)
- Upgrade_months_recoded_06 (for time to upgrade 34 or more)

Using the RCS macro, the continuous variables can be tested for non-linearity i.e. testing the null hypothesis that the contribution of the variable into the model can be modeled using a linear

effect, and the variables can be transformed and added to the model as splines. The contribution and significance of the splines can also be tested.

The linear variable of time to upgrade will also be tested with splines and the dummy variables removed to see if the model fit can be improved.

Heinzl and Kaider advise that knots can be placed using the percentiles of the distribution of the independent variable. Most common strategies include placing knots at these percentiles:

- {5,50,95}
- {5,25,75,95}
- {5,25,50,75,95}

for 3,4, or 5 knots. Empirical evidence suggests that 3 – 5 knots normally suffice and that knots must be placed at the quantiles, near but not at the extremes and roughly uniform over the quantiles.

The percentiles for the independent variables were calculated and are displayed in the table below:

Table 16 Percentiles of Continuous Variables

Variable	Percentile				
	5th	25th	50th	75th	95th
AGE	26	35	43	52	66
ave_val	0	8.01	123.6	314.56	1025.22
ave_sms	0	0	5.05	26.49	114.65
ave_data	0	0	0	0.038	92.95
ave_smsevents	0	0.67	17.33	68.67	281.33

From the above table, it appears that AGE can be modeled using 5 knots, *ave_val*, *ave_sms*, *ave_sms_events* using 3 knots and *ave_data* using 2 knots.

For *ave_data*, because there are only two knots due to the skewness of the distribution, the linear effect of the model will be used instead.

The results of applying this strategy are displayed and presented below.

The value of -2logL is 10792750 for this model – lower than what is observed for the classic model.

Table 17 MLE for Cox Model with Cubic Splines

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter	Standard	Chi-Square	Pr > ChiSq	Hazard	95% Hazard Ratio Confidence	Label
AGE	1	-0.0433	0.000813	2837.6974	<.0001	0.958	0.956	0.96
Age_Spline__1_1	1	-0.0000602	3.93E-06	235.0776	<.0001	1	1	1
Age_Spline__1_2	1	0.0003673	1.28E-05	828.705	<.0001	1	1	1
Age_Spline__1_3	1	-0.0006016	1.52E-05	1566.0381	<.0001	0.999	0.999	1
ave_val	1	0.00147	3.95E-05	1385.6991	<.0001	1.001	1.001	1
Ave_val_spline__2_1	1	-1.30E-08	3.73E-10	1225.0306	<.0001	1	1	1
ave_sms	1	-0.00247	0.00128	3.7301	0.0534	0.998	0.995	1
Ave_sms_spline__3_1	1	0.0000217	4.78E-06	20.6804	<.0001	1	1	1
ave_data	1	-0.0000448	1.78E-06	632.8868	<.0001	1	1	1
Ave_data_spline__5_1	1	-1.86E-06	2.26E-07	67.8014	<.0001	1	1	1
ave_smsevents	1	-0.16503	0.00183	8122.7437	<.0001	0.848	0.845	0.85
ave_smsevents_spline__6_1	1	0.00485	6.24E-05	6046.0397	<.0001	1.005	1.005	1.01
ave_smsevents_spline__6_2	1	-0.00505	0.000065	6028.9947	<.0001	0.995	0.995	1
ave_smsevents_spline__6_3	1	0.0001977	2.67E-06	5483.5057	<.0001	1	1	1
payment_method_dummy	1	-0.15496	0.00369	1765.7674	<.0001	0.856	0.85	0.86
upgrade_months_01	1	-0.46158	0.00685	4540.8756	<.0001	0.63	0.622	0.64 [-998,-1] (-1.43R,6.45%)
upgrade_months_02	1	0.80872	0.00485	27812.773	<.0001	2.245	2.224	2.27 [0,8] (-5.42R,6.25%)
upgrade_months_03	1	-0.76186	0.00478	25400.361	<.0001	0.467	0.462	0.47 (8,22] (1.3R,32.74%)
upgrade_months_04	1	-0.92024	0.00544	28649.135	<.0001	0.398	0.394	0.4 (22,28] (1.33R,19.57%)
upgrade_months_05	1	-2.5972	0.01099	55843.218	<.0001	0.074	0.073	0.08 (28,33] (7.14R,15.67%)
upgrade_months_06	1	-4.44377	0.04202	11182.546	<.0001	0.012	0.011	0.01 (33, HIGH] (58.8R,7.82%)

The Maximum Likelihood estimates are given in the above table. For the characteristic AGE, the variables Age_spline_1_1 to Age_spline_1_3 are the associated cubic splines, as are ave_val_spline_2_1 for ave_val, ave_sms_spline_3_1 for ave_sms and ave_data_spline_5_1 for average data and ave_smsevents_6_1 to ave_smsevents_6_3 for average_smsevents.

The linear effect for ave_sms is not significant.

The log hazard ratio plotted against each of the continuous variables that have been transformed are given below.

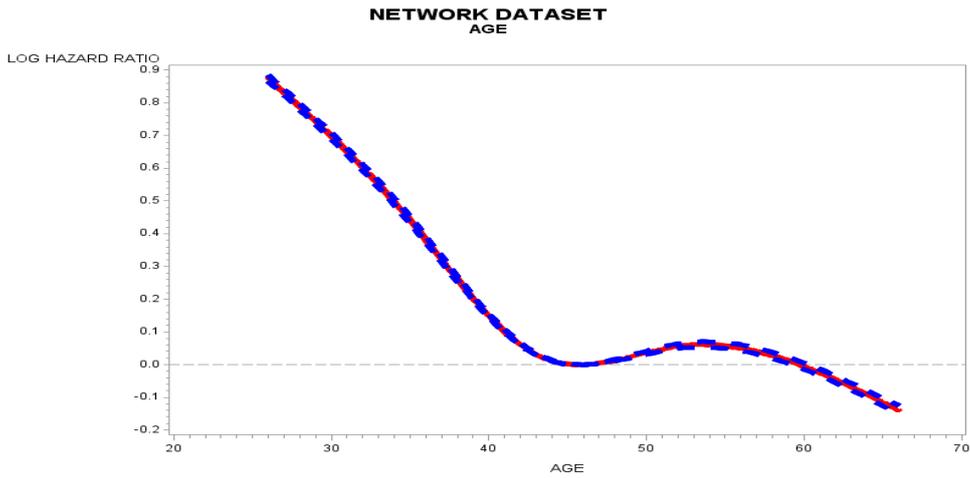


Figure 16 Log Hazard Ratio for Age

Looking at the log hazard ratio for Age, a distinct non-linear trend can be observed. The log hazard ratios are decreasing up to Age = 50 where after it is flat and increases again and then again decreases after about 53 and dipping below zero at 58. This indicates that though the hazard decreases with Age, it only really decreases the hazard function after age 58.

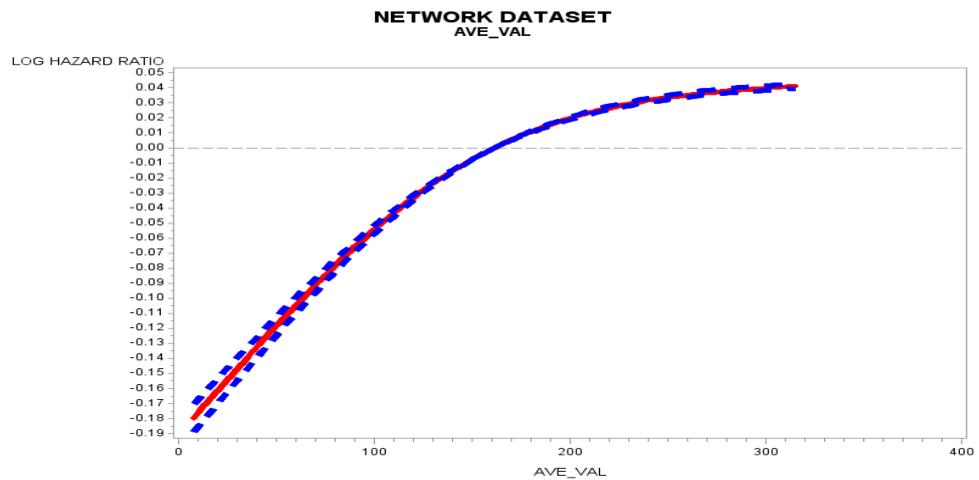


Figure 17 Log Hazard Ratio for Average Value

When analysing the average value of a consumer against the log hazard ratio show a weak linear trend that can be better approximated by the non-linear functions. This is not a trend one expects to see as higher values are associated with a higher log hazard rates. However, this is balanced by the other effects in the model which show a more intuitive trend.

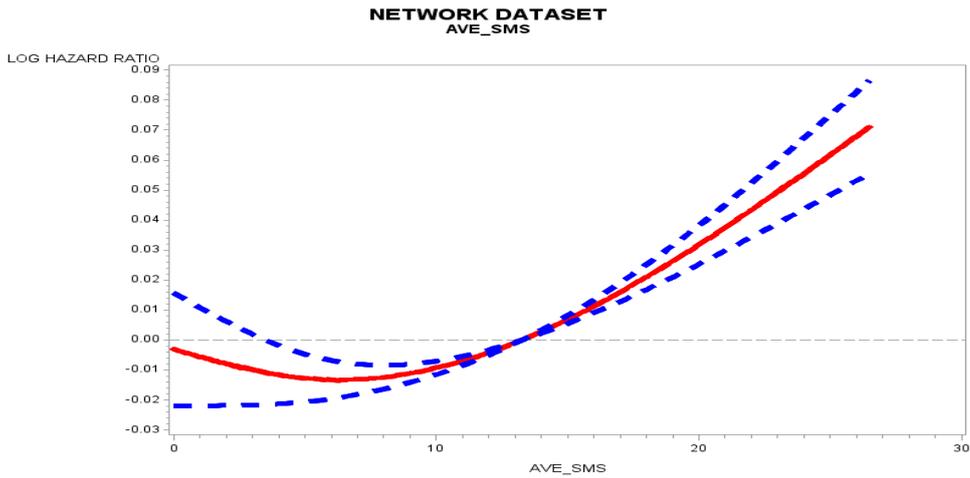


Figure 18 Log Hazard Ratio for Average SMS

The higher the value of average SMS, the higher the log hazard ratio - the standard error bands also fan out close to zero and after 13 for this characteristic. Again, as with average value – this trend in the average value generated from SMS is not what one expects, however since this characteristic is correlated to average value (is one of the constituting factors of average value), the trend is expected.

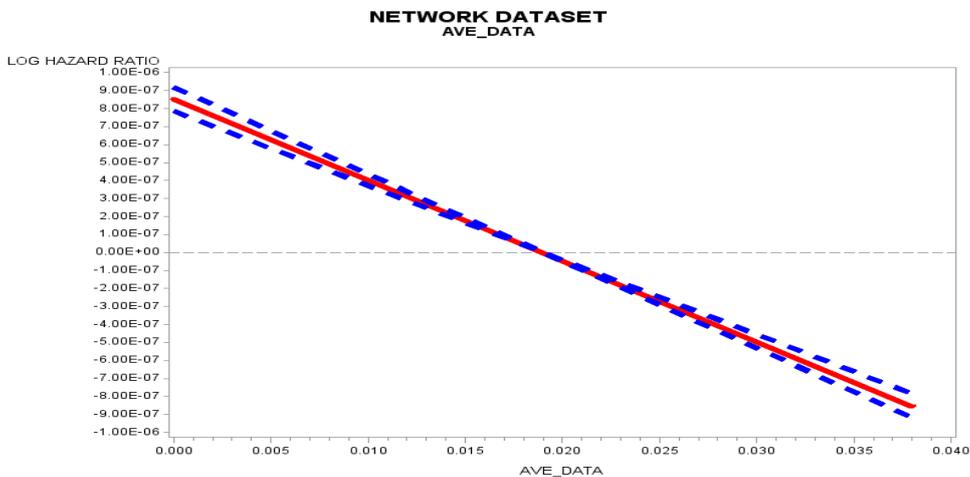


Figure 19 Log Hazard Ratio for Average Data

The log-hazard ratio for the average value from data for a subscriber shows a distinct (almost linear) decreasing trend – the higher one’s data value, the lower the log hazard ratio. In general the log hazard ratios of this variable are very small. The observed trend makes intuitive sense, as it shows that the higher data values bring down the hazard of a subscriber disconnecting from the network. As a result, average data is modelled linearly.

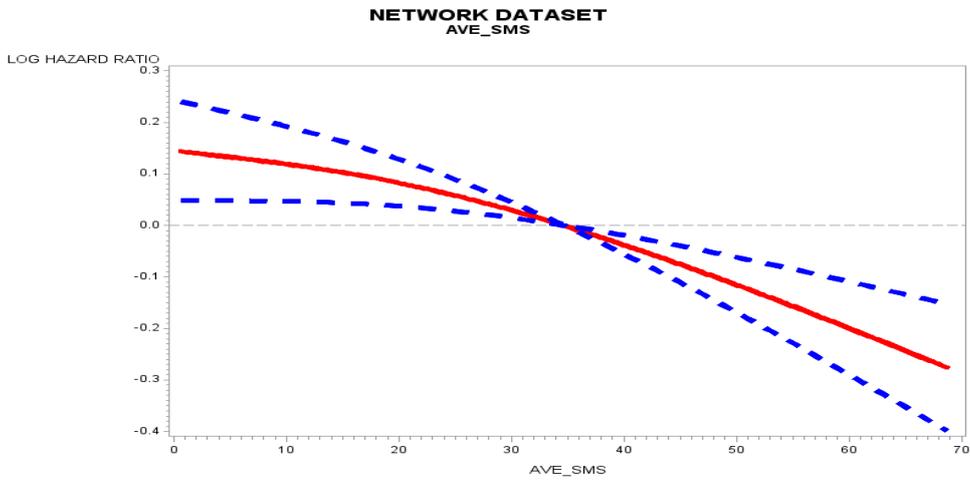


Figure 20 Log Hazard Ratio for Average SMS

As in the case of average data, the average sms value for a subscriber also shows that higher sms values will be associated with lower log hazard ratios.

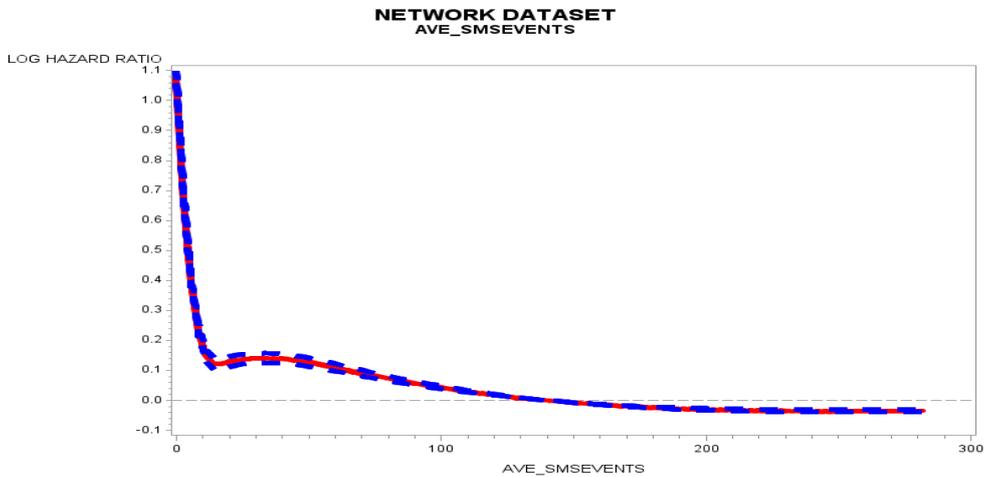


Figure 21 Log Hazard Ratio for Average SMS Events

The physical number of SMS's that are sent – the average sms events shows a clear non-linear trend that is sharply decreasing for low values, then increases and decreases gradually again as the values of the characteristic grows.

The above trend indicates that higher values of usage will decrease the hazard of leaving the network, but only gradually after a certain level. The breakpoint at about 16 average events is a clear differentiator of behavior.

6.2.5.5. Final Model

Taking the results from the previous sections into consideration, a stepwise procedure is performed on the variables as well as their transformed counterparts to arrive at a final model to describe the data and that can be used as a predictive tool. All effects were reselected, except for *_6_2* – the second spline function of *average sms events*. The maximum likelihood estimates of the model are given in the table below:

Table 18 MLE of Stepwise Regression

Analysis of Maximum Likelihood Estimates									
Parameter	DF	Parameter	Standard	Chi-Square	Pr > ChiSq	Hazard	95% Hazard Ratio Confidence	Label	
AGE	1	-0.04318	0.000813	2822.6488	<.0001	0.958	0.956	0.96	
Age_Spline__1_1	1	-0.000064	3.93E-06	265.0235	<.0001	1	1	1	
Age_Spline__1_2	1	0.0003815	1.28E-05	894.1581	<.0001	1	1	1	
Age_Spline__1_3	1	-0.0006199	1.52E-05	1661.3997	<.0001	0.999	0.999	1	
ave_val	1	0.0004333	3.74E-05	134.0574	<.0001	1	1	1	
Ave_val_spline__2_1	1	-3.26E-09	3.53E-10	85.3796	<.0001	1	1	1	
ave_sms	1	-0.05286	0.00109	2359.1949	<.0001	0.949	0.946	0.95	
Ave_sms_spline__3_1	1	0.0002286	4.00E-06	3271.3998	<.0001	1	1	1	
ave_data	1	-0.0000463	1.73E-06	716.9735	<.0001	1	1	1	
Ave_data_spline__5_1	1	-0.0000115	1.95E-07	3460.9395	<.0001	1	1	1	
ave_smsevents	1	-0.02655	0.000378	4943.4386	<.0001	0.974	0.973	0.98	
ave_smsevents_spline__6_1	1	7.03E-06	1.31E-07	2891.4279	<.0001	1	1	1	
ave_smsevents_spline__6_3	1	-9.37E-06	1.77E-07	2787.8717	<.0001	1	1	1	
payment_method_dummy	1	-0.16299	0.00369	1946.7637	<.0001	0.85	0.843	0.86	
upgrade_months_01	1	-0.42283	0.00683	3828.5374	<.0001	0.655	0.646	0.66	[-998,-1] (-1.43R,6.49%)
upgrade_months_02	1	0.82051	0.00485	28675.163	<.0001	2.272	2.25	2.29	[0,8] (-5.42R,6.25%)
upgrade_months_03	1	-0.77168	0.00477	26132.265	<.0001	0.462	0.458	0.47	(8,22] (1.3R,32.74%)
upgrade_months_04	1	-0.91372	0.00544	28219.64	<.0001	0.401	0.397	0.41	(22,28] (1.33R,19.57%)
upgrade_months_05	1	-2.58985	0.01099	55518.294	<.0001	0.075	0.073	0.08	(28,33] (7.14R,15.67%)

upgrade_months_06	1	-4.45168	0.04205	11205.537	<.0001	0.012	0.011	0.01	(33, HIGH] (58.8R,7.82%)
-------------------	---	----------	---------	-----------	--------	-------	-------	------	-----------------------------

The reference set of covariates for is given in the following table.

Table 19 Reference set of Covariates for Plotting

Reference Set of Covariates for Plotting	
AGE	44.49936
ave_val	278.4399
ave_sms	26.75826
ave_data	33.73314
ave_smsevents	67.0828
payment_method_dummy	0.468962
upgrade_months_01	0.066021
upgrade_months_02	0.062489
upgrade_months_03	0.327256
upgrade_months_04	0.19568
upgrade_months_05	0.156788
upgrade_months_06	0.07824
Age_Spline__1_1	13659
Age_Spline__1_2	4669.247
Age_Spline__1_3	1253.45
Ave_val_spline__2_1	20451931
Ave_sms_spline__3_1	8667.562
Ave_data_spline__5_1	55544
ave_smsevents_spline__6_1	1840393
ave_smsevents_spline__6_3	1236709

The SAS code used to transform the cubic splines and to estimate the final parameters is given in the appendix. Using the code and the parameter estimates above, the risk score can be calculated for each individual and then the baseline hazard over each time period that has been observed in the dataset.

6.2.5.6. Discussion

The results from the previous section show that a better model fit can be obtained when using non-linear effects in the Cox-proportional hazards model.

Though the model was fitted at a global level, the results displayed show interesting patterns of customer behavior that can help the Network to better understand its customers in terms of their behavior.

The cubic splines that are utilized in the modelling strategy can easily be implemented using an automated decisioning system and the hazard ratios, and survival probabilities estimated at a subscriber level.

The classic Cox Proportional Hazards model also shows a lot of value in terms of ease of implementation and use for the Network. Because of the ease of use of the cubic splines and the lift the model displayed when using these characteristics, it is advised that the cubic splines model be used when considering implementation.

7. Chapter 7 – Conclusion

The classic Cox Proportional Hazards model can be used a tool to gain insight into customer behavior with regards to disconnecting from the Network.

Both the classic model and well as the more generalised model using splines, are easy to use, to interpret and can be used to set a survival probability and a hazard ratio at subscriber level. These latter quantities are useful metrics that can be used to segment customers into a retention strategy i.e. customers that show a short expected lifetime / high hazard from disconnecting from the network can be campaigned and an appropriate incentive offered to convince the subscribers to remain with the Network.

Given these quantities, the expected future lifetime per subscriber and per payment method can be calculated and used for budgetary purposes. As alluded to earlier, the targets of CVM departments are often related to disconnection percentages, and when combined with a likely to disconnect score, reasonable and stretch targets can be set and the executive strategies of the network applied accordingly. This makes these models and strategies incredibly powerful tools.

Though both of the models are easy to use, especially given tools such as the RCS macro, the classic model remains the most widely known and easiest to understand. In terms of a modelling strategy, using indicator variables where a different hazard rate is expected will make the classic model easier to implement and to explain than the cubic splines model e.g. the time to upgrade for a subscriber. Where trends are known, dummy variables can be constructed to model these intervals and these will be easier to implement and to present to the business than a cubic spline function.

However, since the proportional hazards model does not lend itself easily to visual representations when exploring the data, the cubic splines model does a better job of uncovering trends that might be hidden in the data. The splines can be specified without a lot of subjective input from the analyst and executed using most available software packages.

The trends that are uncovered from the splines modelling can be used to suggest possible transformations of the independent variable set, just as in the case of GAM's.

More importantly is the ease of use, and though many sophisticated techniques have been put forward to automatically detect trends and fit models, these models are in most cases very difficult to implement. Both the classic and Cox model using splines can be repeated and represents a fair blend of automatic detection and ease of use.

The final models have been made available to the Network and a survival probability, baseline hazard and hazard rate supplied per subscriber. These metrics will be used to set budgets, and set strategies for retaining postpaid customers. These metrics were also calculated at store and regional level to understand if there are perhaps more prominent demographic factors that demand the Network's attention e.g. infrastructure problems that may cause inadequate support of postpaid subscribers.

Using these tools, the Network is armed with insight to input into better CVM strategies and in general a better understanding of its customers and how to ensure they remain with the network.

8. REFERENCES

- Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *Annals of statistics, Volume 10, number 4*, 1100-1120.
- Bain, L.J. and Engelhardt, M. (1992) *Introduction to probability and mathematical statistics*, Duxbury Press Belmont, CA.
- Berry, M.J. and Linoff, G.S. (2004) *Data mining techniques*, * Wiley Computer Publishing.
- Breiman, L. and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391), 580-598.
- Breslow, N. (1974) Covariance analysis of censored survival data. *Biometrics*, 89-99.
- Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear Smoothers and Additive Models. *The Annals of Statistics*, 17(2), 453-510.
- Collett, D. (2003) *Modelling survival data in medical research*, CRC press.
- Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187-220.
- Cox, D.R. and Snell, E.J. (1968) A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 248-275.
- Crowley, J., LeBlanc, M., Gentleman, R. and Salmon, S. (1995) Exploratory methods in survival analysis. *Lecture Notes-Monograph Series*, 55-77.
- Davies, R.B. (1980) Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3), 323-333.
- Davies, R.B. (1973) Numerical inversion of a characteristic function. *Biometrika*, 60(2), 415-417.
- De Boor, C. (1978) *A practical guide to splines*, Springer-Verlag New York.
- Durrleman, S. and Simon, R. (1989) Flexible regression models with cubic splines. *Statistics in medicine*, 8(5), 551-561.
- Fan, J. and Gijbels, I. Local Polynomial Modelling and its Applications. 1996. *Chapman, Hall, London*, .
- Friedman, J.H. and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American statistical Association*, 76(376), 817-823.
- Grambsch, P.M. and Therneau, T.M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3), 515-526.
- Gray, R.J. (1994) Spline-Based Tests in Survival Analysis. *Biometrics*, 50(3), 640-652.
- Green, P.J. and Yandell, B.S. (1985) *Semi-parametric generalized linear models*, Springer.

- Hastie, T. and Tibshirani, R. (1990) Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 1005-1016.
- Hastie, T. and Tibshirani, R. (1990) *Generalized additive models*, Chapman and Hall/CRC.
- Hastie, T., Sleeper, L. and Tibshirani, R. (1992) Flexible covariate effects in the proportional hazards model. *Breast cancer research and treatment*, 22(3), 241-250.
- Hastie, T. and Tibshirani, R. (1986) Generalized Additive Models. *Statistical Science*, 1(3), 297-310.
- Heinzl, H. and Kaider, A. (1997) Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer methods and programs in biomedicine*, 54(3), 201-208.
- Hutton, J. 2011. *Mobile Phones Dominate in South Africa*. [Online] Available from: <http://www.nielsen.com/us/en/insights/news/2011/mobile-phones-dominate-in-south-africa.html>
- Imhof, J. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4), 419-426.
- Kalbfleisch, J.D. and Prentice, R.L. (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2), 267-278.
- Kalbfleisch, J.D. and Ross, L. Prentice (2002), *The Statistical Analysis of Failure Time Data*.
- Kaplan, E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282), 457-481.
- Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384.
- Smith, G. (2013). *The Reality of Mobile Usage and Social Media Growth in South Africa*. [Online]. Available from: <http://imagi-nation.co.za/reality-mobile-usage-social-media-growth-south-africa/>
- Tibshirani, R. and Hastie, T. (1987) Local likelihood estimation. *Journal of the American Statistical Association*, 82(398), 559-567.
- Zhang, W. and Steele, F. (2004) A semiparametric multilevel survival model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(2), 387-404.

9. APPENDIX – Multiple Myeloma Data Set

1. Multiple Myeloma

The data in the table below, obtained from Krall, Uthoff and Hartley (1975) , relate to 48 patients aged between 50 and 80 years. Not all of the patients had died by the time the study was completed and so their survival time will be right censored.

Censored times are coded with 0 being censored, and 1 not censored and the patient having died from multiple myeloma.

At the time of diagnosis, the values of a number of explanatory variables were recorded for each patient. These are

- Age – the patient’s age in years
- Gender – 1 = Male, 2 = Female
- Bun – the levels of blood urea nitrogen
- Ca – serum calcium
- HB – Hemoglobin
- PCells – Percentage of plasma cells in the bone marrow
- Protein – an indicator variable to denote whether or not Bence-jones protein was present in the urine.

Patient number	Survival Time	Status	Age	Sex	Bun	Ca	HB	Pcells	Protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0

11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11	16	1
34	5	1	77	1	23	8	9	29	0
35	10	1	61	1	13	10	14	19	0
36	91	1	58	2	27	11	11	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	19	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

10. APPENDIX - SAS CODE

```
LIBNAME __DATA 'C:\Users\xxx\Documents\xx\Studies\MSC\Network Survival Model\SAVE\';

TITLE ' NETWORK DATASET ';

PROC PHREG DATA=__DATA.Network_survival_model_rcs COVOUT OUTEST=__RCS;

MODEL LIFETIME*CENSOR1(0) = AGE Age_Spline__1_1 Age_Spline__1_2 Age_Spline__1_3 AVE_VAL
Ave_val_spline__2_1 AVE_SMS Ave_sms_spline__3_1 AVE_DATA AVE_SMS Ave_dataAve_data__5_1
AVE_SMSEVENTS
ave_smsevents__6_1 ave_smsevents__6_2 ave_smsevents__6_3 payment_method_dummy
upgrade_months_recoded_01
upgrade_months_recoded_02 upgrade_months_recoded_03
upgrade_months_recoded_04 upgrade_months_recoded_05
upgrade_months_recoded_06 /RL;

***** spline modelling of fixed covariate AGE;
***** with 5 knots located at;
***** 26 35 43 52 66;
Age_Spline__1_1=((AGE-26)**3)*(AGE>26)
-((AGE-52)**3)*(AGE>52)
*(66-26)/(66-52)
+((AGE-66)**3)*(AGE>66)
*(52-26)/(66-52);
Age_Spline__1_2=((AGE-35)**3)*(AGE>35)
-((AGE-52)**3)*(AGE>52)
*(66-35)/(66-52)
+((AGE-66)**3)*(AGE>66)
*(52-35)/(66-52);
Age_Spline__1_3=((AGE-43)**3)*(AGE>43)
-((AGE-52)**3)*(AGE>52)
*(66-43)/(66-52)
+((AGE-66)**3)*(AGE>66)
*(52-43)/(66-52);

***** spline modelling of fixed covariate AVE_VAL;
***** with 3 knots located at;
***** 8.01 123.6 314.56;
Ave_val_spline__2_1=((AVE_VAL-8.01)**3)*(AVE_VAL>8.01)
-((AVE_VAL-123.6)**3)*(AVE_VAL>123.6)
*(314.56-8.01)/(314.56-123.6)
+((AVE_VAL-314.56)**3)*(AVE_VAL>314.56)
*(123.6-8.01)/(314.56-123.6);

***** spline modelling of fixed covariate AVE_SMS;
***** with 3 knots located at;
***** 0 5.05 26.49;
Ave_sms_spline__3_1=((AVE_SMS-0)**3)*(AVE_SMS>0)
-((AVE_SMS-5.05)**3)*(AVE_SMS>5.05)
*(26.49-0)/(26.49-5.05)
+((AVE_SMS-26.49)**3)*(AVE_SMS>26.49)
*(5.05-0)/(26.49-5.05);

***** linear modelling of fixed covariate AVE_DATA;
```

```

***** spline modelling of fixed covariate AVE_SMS;
***** with 3 knots located at;
***** 0.67 17.33 68.67;
Ave_data_spline__5_1=((AVE_SMS-0.67)**3)*(AVE_SMS>0.67)
-((AVE_SMS-17.33)**3)*(AVE_SMS>17.33)
*(68.67-0.67)/(68.67-17.33)
+((AVE_SMS-68.67)**3)*(AVE_SMS>68.67)
*(17.33-0.67)/(68.67-17.33);

***** spline modelling of fixed covariate AVE_SMSEVENTS;
***** with 5 knots located at;
***** 0 0.67 17.33 68.67 281.33;
ave_smsevents_spline__6_1=((AVE_SMSEVENTS-0)**3)*(AVE_SMSEVENTS>0)
-((AVE_SMSEVENTS-68.67)**3)*(AVE_SMSEVENTS>68.67)
*(281.33-0)/(281.33-68.67)
+((AVE_SMSEVENTS-281.33)**3)*(AVE_SMSEVENTS>281.33)
*(68.67-0)/(281.33-68.67);
ave_smsevents_spline__6_2=((AVE_SMSEVENTS-0.67)**3)*(AVE_SMSEVENTS>0.67)
-((AVE_SMSEVENTS-68.67)**3)*(AVE_SMSEVENTS>68.67)
*(281.33-0.67)/(281.33-68.67)
+((AVE_SMSEVENTS-281.33)**3)*(AVE_SMSEVENTS>281.33)
*(68.67-0.67)/(281.33-68.67);
ave_smsevents_spline__6_3=((AVE_SMSEVENTS-17.33)**3)*(AVE_SMSEVENTS>17.33)
-((AVE_SMSEVENTS-68.67)**3)*(AVE_SMSEVENTS>68.67)
*(281.33-17.33)/(281.33-68.67)
+((AVE_SMSEVENTS-281.33)**3)*(AVE_SMSEVENTS>281.33)
*(68.67-17.33)/(281.33-68.67);

*----- Testing variable: AGE -----;
EFFECT1: TEST AGE, Age_Spline__1_1, Age_Spline__1_2, Age_Spline__1_3;
NONLIN1: TEST Age_Spline__1_1, Age_Spline__1_2, Age_Spline__1_3;

*----- Testing variable: AVE_VAL -----;
EFFECT2: TEST AVE_VAL, Ave_val_spline__2_1;
NONLIN2: TEST Ave_val_spline__2_1;

*----- Testing variable: AVE_SMS -----;
EFFECT3: TEST AVE_SMS, Ave_sms_spline__3_1;
NONLIN3: TEST Ave_sms_spline__3_1;

*----- Testing variable: AVE_SMS -----;
EFFECT5: TEST AVE_SMS, Ave_data_spline__5_1;
NONLIN5: TEST Ave_data Ave_data_spline__5_1;

*----- Testing variable: AVE_SMSEVENTS -----;
EFFECT6: TEST AVE_SMSEVENTS, ave_smsevents__spline__6_1, ave_smsevents__spline__6_2,
ave_smsevents__spline__6_3;
NONLIN6: TEST ave_smsevents__spline__6_1, ave_smsevents__spline__6_2,
ave_smsevents__spline__6_3;
RUN;
===== End of PROC PHREG =====;

*----- Graph for AGE -----;
PROC IML;
NPOINTS=101; * Number of points to build the graphic;
LOWEREND=26; *Smallest value for X-axis;

```

```

UPPEREND=66; *Largest value for X-axis;
REF=(26+66)/2; *Reference value for X-axis;
X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
S1=((X-26)##3)#(X>26)
  -((X-52)##3)#(X>52)
  #((66-26)/(66-52)
  +((X-66)##3)#(X>66)
  #((52-26)/(66-52)
  -((REF-26)##3)#(REF>26)
  +((REF-52)##3)#(REF>52)#((66-26)/(66-52)
  -((REF-66)##3)#(REF>66)#((52-26)/(66-52));
S2=((X-35)##3)#(X>35)
  -((X-52)##3)#(X>52)
  #((66-35)/(66-52)
  +((X-66)##3)#(X>66)
  #((52-35)/(66-52)
  -((REF-35)##3)#(REF>35)
  +((REF-52)##3)#(REF>52)#((66-35)/(66-52)
  -((REF-66)##3)#(REF>66)#((52-35)/(66-52));
S3=((X-43)##3)#(X>43)
  -((X-52)##3)#(X>52)
  #((66-43)/(66-52)
  +((X-66)##3)#(X>66)
  #((52-43)/(66-52)
  -((REF-43)##3)#(REF>43)
  +((REF-52)##3)#(REF>52)#((66-43)/(66-52)
  -((REF-66)##3)#(REF>66)#((52-43)/(66-52));
XMAT=(X-REF)| |S1| |S2| |S3;
HV={ AGE Age_Spline__1_1 Age_Spline__1_2 Age_Spline__1_3 };
USE __RCS; READ ALL VAR HV INTO C;
READ ALL VAR { _NAME_ } INTO HC; CLOSE __RCS;
B=C[1,]` ; HC=REPEAT(HC,1,NCOL(HV));
HV=REPEAT(HV,NROW(HC),1);
HV=(upcase(HC)=upcase(HV))[,+];
HV=LOC(HV#(1:NROW(C))`); C=C[HV,];
F=XMAT*B; FU=XMAT*C*XMAT`; FREE XMAT;
FU=SQRT(VECDIAG(FU)); FO=F+1.96*FU; FU=F-1.96*FU;
Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
F=F//FO//FU; FE=EXP(F); X=REPEAT(X,3,1);
CREATE __RCS1 VAR { F FE Z X }; APPEND; CLOSE __RCS1;
QUIT;

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

PROC Gplot DATA=__RCS1;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 ' AGE ';
LABEL X=AGE;
LABEL F=LOG HAZARD RATIO;
RUN;

*----- Graph for AVE_VAL -----;
PROC IML;
NPOINTS=101; * Number of points to build the graphic;
LOWEREND=8.01; *Smallest value for X-axis;

```

```

UPPEREND=314.56; *Largest value for X-axis;
REF=(8.01+314.56)/2; *Reference value for X-axis;
X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
S1=((X-8.01)##3)#(X>8.01)
  -((X-123.6)##3)#(X>123.6)
  #(314.56-8.01)/(314.56-123.6)
  +((X-314.56)##3)#(X>314.56)
  #(123.6-8.01)/(314.56-123.6)
  -((REF-8.01)##3)#(REF>8.01)
  +((REF-123.6)##3)#(REF>123.6)#(314.56-8.01)/(314.56-123.6)
  -((REF-314.56)##3)#(REF>314.56)#(123.6-8.01)/(314.56-123.6);
XMAT=(X-REF) | | S1;
HV={ AVE_VAL Ave_val_spline__2_1 };
USE __RCS; READ ALL VAR HV INTO C;
READ ALL VAR { _NAME_ } INTO HC; CLOSE __RCS;
B=C[1,]; HC=REPEAT(HC,1,NCOL(HV));
HV=REPEAT(HV,NROW(HC),1);
HV=(upcase(HC)=upcase(HV))[,+];
HV=LOC(HV#(1:NROW(C))`); C=C[HV,];
F=XMAT*B; FU=XMAT*C*XMAT; FREE XMAT;
FU=SQRT(VECDIAG(FU)); FO=F+1.96*FU; FU=F-1.96*FU;
Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
F=F//FO//FU; FE=EXP(F); X=REPEAT(X,3,1);
CREATE __RCS2 VAR { F FE Z X }; APPEND; CLOSE __RCS2;
QUIT;

```

```

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

```

```

PROC GLOT DATA=__RCS2;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 ' AVE_VAL ';
LABEL X=AVE_VAL;
LABEL F=LOG HAZARD RATIO;
RUN;

```

```

*----- Graph for AVE_SMS -----;
PROC IML;
NPOINTS=101; * Number of points to build the graphic;
LOWEREND=0; *Smallest value for X-axis;
UPPEREND=26.49; *Largest value for X-axis;
REF=(0+26.49)/2; *Reference value for X-axis;
X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
S1=((X-0)##3)#(X>0)
  -((X-5.05)##3)#(X>5.05)
  #(26.49-0)/(26.49-5.05)
  +((X-26.49)##3)#(X>26.49)
  #(5.05-0)/(26.49-5.05)
  -((REF-0)##3)#(REF>0)
  +((REF-5.05)##3)#(REF>5.05)#(26.49-0)/(26.49-5.05)
  -((REF-26.49)##3)#(REF>26.49)#(5.05-0)/(26.49-5.05);
XMAT=(X-REF) | | S1;
HV={ AVE_SMS Ave_sms_spline__3_1 };
USE __RCS; READ ALL VAR HV INTO C;
READ ALL VAR { _NAME_ } INTO HC; CLOSE __RCS;
B=C[1,]; HC=REPEAT(HC,1,NCOL(HV));

```

```

HV=REPEAT(HV,NROW(HC),1);
HV=(upcase(HC)=upcase(HV))[,+];
HV=LOC(HV#(1:NROW(C))`); C=C[HV,];
F=XMAT*B; FU=XMAT*C*XMAT`; FREE XMAT;
FU=SQRT(VECDIAG(FU)); FO=F+1.96*FU; FU=F-1.96*FU;
Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
F=F//FO//FU; FE=EXP(F); X=REPEAT(X,3,1);
CREATE __RCS3 VAR { F FE Z X }; APPEND; CLOSE __RCS3;
QUIT;

```

```

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

```

```

PROC GPLOT DATA=__RCS3;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 ' AVE_SMS ';
LABEL X=AVE_SMS;
LABEL F=LOG HAZARD RATIO;
RUN;

```

*----- Graph for AVE_DATA -----;

```

PROC IML;
NPOINTS=101; * Number of points to build the graphic;
LOWEREND=0; *Smallest value for X-axis;
UPPEREND=0.038; *Largest value for X-axis;
REF=(0+0.038)/2; *Reference value for X-axis;
X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
XMAT=(X-REF);
HV={ AVE_DATA };
USE __RCS; READ ALL VAR HV INTO C;
READ ALL VAR { _NAME_ } INTO HC; CLOSE __RCS;
B=C[1,] ; HC=REPEAT(HC,1,NCOL(HV));
HV=REPEAT(HV,NROW(HC),1);
HV=(upcase(HC)=upcase(HV))[,+];
HV=LOC(HV#(1:NROW(C))`); C=C[HV,];
F=XMAT*B; FU=XMAT*C*XMAT`; FREE XMAT;
FU=SQRT(VECDIAG(FU)); FO=F+1.96*FU; FU=F-1.96*FU;
Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
F=F//FO//FU; FE=EXP(F); X=REPEAT(X,3,1);
CREATE __RCS4 VAR { F FE Z X }; APPEND; CLOSE __RCS4;
QUIT;

```

```

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

```

```

PROC GPLOT DATA=__RCS4;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 ' AVE_DATA ';
LABEL X=AVE_DATA;
LABEL F=LOG HAZARD RATIO;
RUN;

```

*----- Graph for AVE_SMS -----;

```

PROC IML;
NPOINTS=101; * Number of points to build the graphic;

```

```

LOWEREND=0.67; *Smallest value for X-axis;
UPPEREND=68.67; *Largest value for X-axis;
REF=(0.67+68.67)/2; *Reference value for X-axis;
X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
S1=((X-0.67)##3)#(X>0.67)
  -((X-17.33)##3)#(X>17.33)
  #((68.67-0.67)/(68.67-17.33)
  +((X-68.67)##3)#(X>68.67)
  #((17.33-0.67)/(68.67-17.33)
  -((REF-0.67)##3)#(REF>0.67)
  +((REF-17.33)##3)#(REF>17.33)#((68.67-0.67)/(68.67-17.33)
  -((REF-68.67)##3)#(REF>68.67)#((17.33-0.67)/(68.67-17.33));
XMAT=(X-REF) | S1;
HV={ AVE_SMS Ave_dataAve_data__5_1 };
USE __RCS; READ ALL VAR HV INTO C;
READ ALL VAR { _NAME_ } INTO HC; CLOSE __RCS;
B=C[1,] ; HC=REPEAT(HC,1,NCOL(HV));
HV=REPEAT(HV,NROW(HC),1);
HV=(upcase(HC)=upcase(HV))[,+];
HV=LOC(HV#(1:NROW(C))) ; C=C[HV,];
F=XMAT*B; FU=XMAT*C*XMAT; FREE XMAT;
FU=SQRT(VECDIAG(FU)); FO=F+1.96*FU; FU=F-1.96*FU;
Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
F=F//FO//FU; FE=EXP(F); X=REPEAT(X,3,1);
CREATE __RCS5 VAR { F FE Z X }; APPEND; CLOSE __RCS5;
QUIT;

```

```

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

```

```

PROC Gplot DATA=__RCS5;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 ' AVE_SMS ';
LABEL X=AVE_SMS;
LABEL F=LOG HAZARD RATIO;
RUN;

```

```

*----- Graph for AVE_SMSEVENTS -----;

```

```

PROC IML;
NPOINTS=101; * Number of points to build the graphic;
LOWEREND=0; *Smallest value for X-axis;
UPPEREND=281.33; *Largest value for X-axis;
REF=(0+281.33)/2; *Reference value for X-axis;
X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
S1=((X-0)##3)#(X>0)
  -((X-68.67)##3)#(X>68.67)
  #((281.33-0)/(281.33-68.67)
  +((X-281.33)##3)#(X>281.33)
  #((68.67-0)/(281.33-68.67)
  -((REF-0)##3)#(REF>0)
  +((REF-68.67)##3)#(REF>68.67)#((281.33-0)/(281.33-68.67)
  -((REF-281.33)##3)#(REF>281.33)#((68.67-0)/(281.33-68.67));
S2=((X-0.67)##3)#(X>0.67)
  -((X-68.67)##3)#(X>68.67)
  #((281.33-0.67)/(281.33-68.67)
  +((X-281.33)##3)#(X>281.33)

```

```

#(68.67-0.67)/(281.33-68.67)
-((REF-0.67)##3)#(REF>0.67)
+((REF-68.67)##3)#(REF>68.67)#(281.33-0.67)/(281.33-68.67)
-((REF-281.33)##3)#(REF>281.33)#(68.67-0.67)/(281.33-68.67);
S3=((X-17.33)##3)#(X>17.33)
-((X-68.67)##3)#(X>68.67)
#(281.33-17.33)/(281.33-68.67)
+((X-281.33)##3)#(X>281.33)
#(68.67-17.33)/(281.33-68.67)
-((REF-17.33)##3)#(REF>17.33)
+((REF-68.67)##3)#(REF>68.67)#(281.33-17.33)/(281.33-68.67)
-((REF-281.33)##3)#(REF>281.33)#(68.67-17.33)/(281.33-68.67);
XMAT=(X-REF)|S1|S2|S3;
HV={ AVE_SMSEVENTS ave_smsevents__6_1 ave_smsevents__6_2 ave_smsevents__6_3 };
USE __RCS; READ ALL VAR HV INTO C;
READ ALL VAR { _NAME_ } INTO HC; CLOSE __RCS;
B=C[1,] ; HC=REPEAT(HC,1,NCOL(HV));
HV=REPEAT(HV,NROW(HC),1);
HV=(upcase(HC)=upcase(HV))[,+];
HV=LOC(HV#(1:NROW(C))) ; C=C[HV,];
F=XMAT*B; FU=XMAT*C*XMAT ; FREE XMAT;
FU=SQRT(VECDIAG(FU)); FO=F+1.96*FU; FU=F-1.96*FU;
Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
F=F//FO//FU; FE=EXP(F); X=REPEAT(X,3,1);
CREATE __RCS6 VAR { F FE Z X }; APPEND; CLOSE __RCS6;
QUIT;

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

PROC Gplot DATA=__RCS6;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 ' AVE_SMSEVENTS ';
LABEL X=AVE_SMSEVENTS;
LABEL F=LOG HAZARD RATIO;
RUN;

```