Modelling Informative Censoring in Relative Survival

National and Kapodistrian University of Athens

This thesis was submitted in order to fulfill the requirements for the MSc diploma in Statistics and Operational research

Author:Stylianos Tzortzakis

Supervisor:Assistant Professor Fotios Siannis

3/11/22

Contents

Introduction

Survival analysis is a set of statistical methods that are used for studying the occurrence of events over a time period.
Survival analysis originally designed for studies where deaths happen.
Several applications come from sociology,medical research,engineering,economics usually with different names but the meaning is the same.
An example in survival analysis is the marriage, where the divorce can be considered as an event , meaning something that change the state of an individual from married to unmarried.
In fact , we need to know more than just who is married and who is not, meaning that we need to know when the change occurred.
As a result we observe events in the time horizon.

Groups of data in survival analysis contain survival times with positively skewed tendency meaning a histogram with longer tail to the right of the interval that contains the largest number of observations.
As we see so far, we focus on observing an individual in a study regarding the actual event of interest but in many cases the event never happen.
These survival times are called censored survival times.

There are many kinds of censoring.
In this thesis we mostly deal with right censoring and is the most common case.
Right censoring occurs after the individual has been entered into the study or when the study ends and the individual hasn't experience the event yet(is refferred to also as administrative censoring).
For instance after some period of time we lose track of a patient for an unknown

reason.
This means that the patient might has left the study because doesn't want to participate anymore or for any other reason that we don't know.

Other types of censoring are the left and the interval censoring .
Left censoring is encountered when the actual survival time of an individual is less than the observed time.
For example let's consider a study in which interest centers on the time to recurrence of a particular cancer following surgical removal of the primary tumour.
Three months after their operation, patients are examined to determine if the cancer has recurred.
At this time, some of the patients may be found to have a recurrence and for such patients, the actual time to recurrence is less than three months.
The recurrence times of these patients are considered as left censoring.
Moreover in another example, an epidemiologist wishes to know the age at diagnosis in a follow-up study of diabetic retinopathy.
At the time of the examination, a 50-year-old participant was found to have already developed retinopathy, but there is no record of the exact time at which initial evidence was found.
Thus the age at examination is a left-censored observation.
Interval censoring occurs when the event of interest is known that has been occurred between two times.
For example let's say that a patient is examined 3 months after the surgical removal and is found to be free of the disease but an examination 6 months later shows that tumour appeared again.
The time period between 3 and 6 months is considered the interval of the true recurrence of tumour and so the observed recurrence time is considered as interval censoring time .

All of these cases of censoring can be grouped into 2 categories independent and informative(dependent) censoring.
The most usual case is the independent censoring.
Independent censoring means that censoring mechanism is independet from the

event mechanism.
As a result individuals with non-informative censored times are representative of all others who are at risk at that time and have the same values of explanatory variables.

On the other hand dependent censoring occurs when there is dependence between the time to an event such as death and the time of the occurrence of censoring.
Also it is not possible to use the observed data to determine whether a data set has dependent censoring.
However, the context of the study can often provide some indication of whether there is a possibility for existence of dependent censoring.
For example such an indication may be a life threatening experience that force an individual to exit the study.
Unfortunately, there is no statistical test for informative censoring and the best thing we can do is a sensitivity analysis.

In Chapter 1 we present general techniques relative to survival analysis , in chapter 2 we discuss cases with multiple events,in chapter 3 we present some techniques for informative censoring , in Chapter 4 we focus on methods for 2 groups of data and in chapter 5 we discuss a special field in survival analysis that is refferred to as relative survival field ,but in this thesis except from this special chapter ,the term 'relative survival' is understood as something that concerns 2 groups of data(the title of this thesis also).
Finally in Chapter 6 we present a simulation regarding the problem of informative censoring in our data.

At this point i would like also to thank my supervisor , Mr. Fotios Siannis ,for his valuable guidance through this thesis.

Εισαγωγή

Η ανάλυση επιβίωσης είναι ένα σύνολο στατιστικών μεθόδων που χρησιμοποιείται στην μελέτη γεγονότων σε έναν χρονικό ορίζοντα.
Η ανάλυση επιβίωσης αρχικά σχεδιάστηκε για μελέτες που το γεγονός ενδιαφέροντος είναι ο θάνατος.
Εφαρμογές εντοπίζονται στην κοινωνιολογία,ιατρική έρευνα,μηχανική,οικονομικά όπως και σε άλλους κλάδους.
Ένα παράδειγμα είναι ο γάμος όπου το διαζύγιο θεωρείται σαν γεγονός και αλλάζει την κατάσταση του ατόμου από παντρεμένο σε διαζευγμένο. Όμως εκτός από το ποιος είναι παντρεμένος και ποιος όχι θέλουμε να μάθουμε και πότε έγινε αυτή η αλλαγή,οπότε παρατηρούμε γεγονότα πάνω στον χρονικό ορίζοντα.
Τα δεδομένα στην ανάλυση επιβίωσης είναι οι χρόνοι επιβίωσης και έχουν συνήθως θετικά ασύμμετρη συμπεριφορά.
Μέχρι στιγμής λοιπόν έχουμε σχολιάσει ότι σε μια έρευνα μας ενδιαφέρει να δούμε πότε συνέβει το πραγματικό γεγονός ενδιαφέροντος ,αλλά πολλές φορές αυτό δεν συμβαίνει ποτέ.
Τέτοιοι χρόνοι καλούνται λογοκριμένοι.
Σε αυτή την διπλωματική εργασία κυρίως ασχολούμαστε με δεξιά λογοκριμένους χρόνους που είναι η πιο συνηθισμένη περίπτωση ,παρ΄όλα αυτά υπάρχουν και άλλα είδη.
Δεξιά λογοκρισία συμβαίνει όταν ήδη το άτομο έχει εισέλθει σε μία έρευνα και για κάποιον άγνωστο λόγο φεύγει ή όταν η έρευνα τελειώσει και δεν έχει παρατηρηθεί ακόμα το γεγονός(αυτό λέγεται και λογοκρισία του διαχειριστή).

Αλλά είδη λογοκρισίας είναι η αριστερή και διαστηματική λογοκρισία.
Η αριστερή λογοκρισία συμβαίνει όταν ο πραγματικός χρόνος επιβίωσης είναι μι-

κρότερος του παρατηρηθέντος χρόνου.
Για παράδειγμα ας θεωρήσουμε μία μελέτη όπου το ενδιαφέρον επικεντρώνεται στην επανεμφάνιση ενός συγκεκριμένου τύπου καρκίνου μετά από χειρουργική αφαίρεση του κύριου όγκου.
Τρεις μήνες μετά την επέμβαση οι ασθενείς εξετάζονται ξανά για να παρατηρήθει αν ο καρκίνος εμφανίστηκε πάλι και βρίσκουμε ότι σε κάποιους ασθενείς όντως εμφανίστηκε ξανά ,οπότε συμπεραίνουμε ότι ο πραγματικός χρόνος επιβίωσης(εμφάνιση του γεγονότος ενδιαφέροντος) συνέβει σε λιγότερο από τρεις μήνες,άρα είναι αριστερή λογοκρισία.
Η διαστηματική λογοκρισία συμβαίνει σε κάποιο χρονικό διάστημα.
Για παράδειγμα αν θεωρήσουμε ότι ο ασθενής εξετάζεται μετά από 3 μήνες από την επέμβαση και βρέθηκε ότι δεν είχε εμφανιστεί πάλι ο καρκίνος ,αλλά μετά από 6 μήνες σημειώθηκε επανεμφάνιση ,οπότε στο διάστημα μεταξύ τριών και έξι μήνων λέμε ότι είχαμε διαστηματική λογοκρισία.

Όλες αυτές αυτές οι περιπτώσεις μπορούν να χωριστούν σε 2 κατηγορίες, ανεξάρτητη και εξαρτημένη λογοκρισία.
Η πιο συνηθισμένη περίπτωση είναι η ανεξάρτητη λογοκρισία.
Ανεξάρτητη λογοκρισία σημαίνει ότι ο μηχανισμός λογοκρισίας είναι ανεξάρτητος του μηχανισμού των γεγονότων.
Από την άλλη μεριά η εξαρτημένη λογοκρισία σημαίνει εξάρτηση μεταξύ των δύο μηχανισμών.
Επίσης δεν είναι δυνατό να παρατηρήσουμε εξαρτημένη λογοκρισία ,άλλα παρ΄όλα αυτά το περιεχόμενο μιας μελέτης μπορεί να δώσει κάποιο στοιχείο.

Στο πρώτο κεφάλαιο παρουσιάζουμε γενικές τεχνικές σχετικά με την ανάλυση επιβίωσης,στο δεύτερο κεφάλαιο αναλύονται περιπτώσεις πολλαπλών γεγονότων ,στο τρίτο κεφάλαιο παρουσιάζουμε τεχνικές για πληροφοριακή-εξαρτημένη λογοκρισία,στο τέταρτο κεφάλαιο εστιάζουμε σε μεθόδους για δύο ομάδες δεδομένων και στο πέπτο κεφάλαιο συζητάμε για ένα ειδικό πεδίο που αναφέρεται ως σχετική επιβιώση,αλλά με εξαίρεση αυτό το κεφάλαιο ο όρος ΄σχετική΄ θα αφορά δύο ομάδες δεδομένων.
Τέλος στο κεφάλαιο έξι γίνεται μια προσομοίωση που αφορά δύο ομάδες δεδομένων

με πληροφοριακή λογοκρισία.

Σε αυτό το σημείο θα ήθελα επίσης να ευχαριστήσω τον κύριο Φώτιο Σιάννη για την πολύτιμη καθοδήγηση καθ'όλη την διάρκεια αυτής της διπλωματικής εργασίας.

Chapter 1

Survival methods

Survival analysis is a field of statistics which describes the analysis of data from a time origin until an event happens (e.g death for humans or failure in machines) or until the end point (end of the study)
In this thesis we mention methods about individuals (e.g patients,animals,etc) from the entry until an event happens such as death or exit from the study due to unknown reasons(right censoring, see the introduction).

1.1 Functions of survival times

First of all we consider the random variable $T \geq 0$ which is absolutely continuous and represents the survival time of an individual from the time origin.
The most common function in the survival field is called survival or survivor function and is defined as

$$S(t) = P(T \geq t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)du \,, \tag{1.1}$$

meaning the probability of an individual survive longer than t , where F is the cumulative function of time T and f is the probability density function of time T.
Also is a monotonically decreasing function of time t , continuous for all values of T, $S(0) = P(T \geq 0) = P(E) = 1$, where E is the space of all possible outcomes and $S(\infty) = \lim_{t\to\infty} S(t) = \lim_{t\to\infty}(1 - F(t)) = 1 - 1 = 0$

Another useful function is called hazard function and represents the approximate probability of an individual who is in danger of an event at some time t to experience that event in an infinitesimal interval after time t , divided by the length of this interval.
In mathematical notation that is

$$h(t) = \lim_{h \to 0} \frac{P(t \leq T < t+h | T \geq t)}{h} . \tag{1.2}$$

Furthermore it is easy to show that

$$h(t) = \frac{f(t)}{S(t)} , \tag{1.3}$$

where f(t) is the density of t.
Indeed the conditional probability $P(t \leq T < t+h | T \geq t)$ equals to $\frac{P(t \leq T < t+h)}{P(T \geq t)}$.
So the expression (1.2) takes the form

$$h(t) = \lim_{h \to 0} \frac{P(t \leq T < t+h)}{hP(T \geq t)} = \frac{1}{S(t)} \lim_{h \to 0} \frac{P(t \leq T < t+h)}{h} = \frac{f(t)}{S(t)} .$$

The so called cumulative hazard function works in an analogous way for the hazard function as F (cumulative function) works for f (probability density function) meaning that is right continuous and nondecreasing function of T(in this case is continuous on all values of T and is a monotonically increasing function).
Specifically the cumulative hazard function has the following expression

$$H(t) = \int_0^t h(u) du . \tag{1.4}$$

A very useful relation between S(t) and H(t) is

$$H(t) = -\log(S(t)) . \tag{1.5}$$

This relation can be proved using (1.1) , (1.3) and (1.4) expressions.
Specifically $H(t) = \int_0^t h(u) du = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t \frac{f(u)}{1-F(u)} du$.

Now we notice that $\Big(log(1-F(t))\Big)' = \frac{-f(t)}{1-F(t)}$, since $f(t) = F'(t)$.
So $H(t) = -logS(t)|_0^t = -logS(t) + logS(0) = -logS(t)$ since $S(0) = P(T \geq 0) = 1$.
From equation (1.5) it follows immediately by multiplying with -1 and taking exp to both sides we take that

$$S(t) = exp(-H(t)) = exp\Big(-\int_0^t h(u)du\Big) . \tag{1.6}$$

As we see the expression (1.6) provides a relation between S(t),H(t) and h(t).
Also we mention that in general time can be discrete (e.g we measure in days) or continuous.
In this thesis is considered mostly the second case because is the most common one.

At this point it is worth to also mention briefly the discrete case and a unified form for the 2 cases.
Let Z be a discrete random variable with $Z \geq 0$, also suppose that Z takes the values $z_1 < \ldots z_n < \ldots,$.
We define the probability function of Z as $f(z_j) = P(Z = z_j)$ for j=1,2,....
Now the survivor function takes the form $S(z) = P(Z \geq z) = \sum_{j:z_j \geq z} f(z_j)$.
In this case the survivor function is a non-increasing step function and left continuous.
The discontinuity points are exactly the points $z_1, \ldots z_n, \ldots$, where the survivor function takes a step down which is exactly $f(z_j) = S(z_j) - S(z_{j+1}) > 0$ for $j = 1, 2 \ldots$.
For the discrete case the hazard function is defined as

$$h(z_j) = P(Z = z_j | Z \geq z_j) = \frac{f(z_j)}{S(z_j)} = \frac{S(z_j) - S(z_{j+1})}{S(z_j)} = 1 - \frac{S(z_{j+1})}{S(z_j)} ,$$

for $j = 1, 2 \ldots$.
From this expression we can take a relation between h(z) and S(z).
Indeed if we notice that

$$S(z) = P(Z \geq z) = \frac{S(z)}{S(z_n)} \frac{S(z_n)}{S(z_{n-1})} \ldots \frac{S(z_2)}{S(z_1)} S(z_1) .$$

So

$$S(z) = \prod_{j:z_j<z} \Big(1 - h(z_j)\Big) . \tag{1.7}$$

There are two ways to define the cumulative hazard function for the discrete case.

A first approach is defined as $H(z) = \sum_{z:z_j<z} h_j(z_j)$.

The second approach is defined as the (1.5) form but S(t) is replaced by the (1.7) expression.

After this presentation we are in a place where we can present a unified form for the discrete and continuous case.

First of all we introduce the Riemann-Stieltjes integral of a function K(u) .

This function is non-decreasing,right continuous and has a finite number of discontinuities in any finite interval.

Also let's say that the derivative of K(u) exists except at points of discontinuity e_j , $j = 1, 2, \ldots$ where at these points we say that K(u) has a $k_j = K(e_j) - K(e_j-)$ jump ,$K(e_j-) = \lim_{h\to\ 0} K(e_j - h)$.

Moreover we define the quantity dK(u) as $dK(u) = K'(u)du + K(u) - K(u-)$.

So if K is continuous at point u we have $dK(u) = K'(u)du$ and otherwise $dK(u) = K(u) - K(u-)$.

The Riemann-Stieltjes integral of K(u) over the interval (a,b] is defined as

$$\int_{(a,b]} dK(u) = \int_a^b K'(u)du + \sum_{j:a<e_j\leq b} k_j . \tag{1.8}$$

The cumulative function of T , F(t) where T can be either discrete or continuous has the properties of the K(u) function so the Riemann-Stieltjes integral of F over (a,b] gives

$$P(a < T \leq b) = \int_{(a,b]} dF(t) = \int_a^b F'(t)dt + \sum_{j:a<e_j\leq b} f_j .$$

Another useful tool that helps in the unified concept is the product integral.

Let's consider a partition of (a,b] $a = t_0 < t_1 < \cdots < t_n = b$ with $\Delta t_i = t_i - t_{i-1}$

and $max(\Delta t_i) \to 0$ when $n \to \infty$.

Then the product integral of a function K(t) who is continuous over (a,b] where dK(t) has the same meaning as dK(u) that was defined earlier

$$\prod_{(a,b]}(1+dK(t)) = \prod_{(a,b]}(1+K'(t)dt) = \lim_{n\to\infty}\prod_{i=1}^{n}(1+K'(t_i)\Delta t_i + o(\Delta t_i)) ,$$

where $o(\Delta t_i))$ is a function such that $\frac{o(t)}{t} \to 0, t \to 0$ and can be ignored for small values of t.

So the product integral for the continuous case over (a,b] can be written as

$$\lim_{n\to\infty}\prod_{i=1}^{n}(1+K'(t_i)\Delta t_i) .$$

By taking logarithms the relation of the product integral for the continuous case becomes

$$\log\Big(\prod_{(a,b]}(1+K'(t)dt)\Big) = \lim_{n\to\infty}\sum_{i=1}^{n} log\Big(1+K'(t_i)\Delta t_i\Big) ,$$

so as Δt_i is considered small we can take the approach

$$log\Big(1+K'(t_i)\Delta t_i\Big) \approx K'(t_i)\Delta t_i ,$$

and we can take that

$$\lim_{n\to\infty}\sum_{i=1}^{n} log\Big(1+K'(t_i)\Delta t_i\Big) = \lim_{n\to\infty}\sum_{i=1}^{n} K'(t_i)\Delta t_i ,$$

but this is exactly the Riemman integral of K'(t) over the interval (a,b] , $\int_a^b K'(t)dt$

In conclusion it follows that

$$\prod_{(a,b]}(1+K'(t)dt) = exp\Big(\int_a^b K'(t)dt\Big) ,$$

and if we take $K(t) = -H(t)$,$K'(t) = -h(t)$ where h(t) is the hazard function and H(t) is the cumulative hazard for the continuous case and (a,b]=(0,t] we

get the (1.6) expression.
So the product integral of $K(t) = -H(t)$ over the interval (0,t) or (0,t] for the continuous case is exacty the survivor function S(t) when T is continuous

The product integral for the discrete case with t_i $i = 1, \dots$ are the points of discontinuity is defined as

$$\prod_{(a,b]} (1 + dK(t)) = \lim_{n \to \infty} \prod_{i=1}^{n} (1 + K(t_i) - K(t_{i-1})) \,,$$

and for (a,b)=(0,t) and K(t)=-H(t) , where h(t) is the hazard function for the discrete case , we conclude that the product integral for the discrete case takes the (1.7) expression because

$$dH(t_j) = H(t_j) - H(t_j-) = H(t_j) - H(t_{j-1}) = h(t_j) \,,$$

where t_j is the points of discontinuity $j = 1, \dots$ and so is the survivor function when T is discrete .

The unified form for a function K(t) with e_j points of discontinuity with k_j size of jump ,$j = 1, \dots$ is defined as

$$\prod_{(a,b]} (1 + dK(t)) = \prod_{(a,b]} (1 + K'(t)dt) \prod_{j: a < e_j \leq b} (1 + k_j) \,. \tag{1.9}$$

So for K(t)=-H(t) we take a unified form for the survivor function which is

$$S(t) = P(T \geq t) = \exp\Big(- \int_0^t h(t)dt \Big) \prod_{j: t_j < t} \Big(1 - h(t_j) \Big) \,. \tag{1.10}$$

1.2 Non - parametric methods

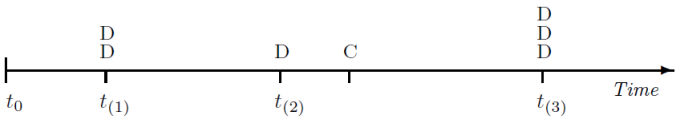
Here we focus on some non - parametric methods .

1.2.1 Kaplan-Meier estimator

The most important and widely used non-parametric estimator for the survivor function is the Kaplan Meier estimator.

We begin with an example The figure (1.2.1) provides an example of the Kaplan-

Figure 1.2.1: Kaplan-Meier example
(Image taken from Modelling survival data in medical research (Collett) third edition)



Meier estimator.Specifically in this example we have 3 distinct deaths $t_1 < t_2 < t_3$ and 7 individuals in the study.

Let's say that we are interested to compute the Kaplan-Meier estimator $\widehat{S(t)}$ when $t_2 < t < t_3$.

Over the interval $[t_0, t_1)$ no deaths happen so the individuals who are alive just before t_1 are $n_1 = 7$,at time t_1 , 2 deaths happen so $d_1 = 2$ .Over the inteval $[t_1, t_2)$ no deaths happen so the individuals who are alive just before t_2 are $n_2 = 5$,at time t_2 there is one death ,so $d_2 = 1$

As a result the probability for an individual to survive at least time t is the product of the survival probablities over the intervals $[t_0, t_1)$ and $[t_1, t_2)$, $\widehat{S(t)} = \frac{n_1-d_1}{n_1}\frac{n_2-d_2}{n_2} = \frac{5}{7}\frac{4}{5}$.

We mention also that the individuals who are alive just before t_3 are $n_3 = 3$,because there is one censored time after t_2,so this individual left the study for reasons that we don't know .Moreover $\widehat{S(t)} = 0$ for $t \geq t_3$,because no one is alive after t_3.

In a more general manner in order to construct this estimator we suppose that n individuals are in the study with $t_1, ..., t_n$ observed survival times.

Some of these observations might be right censored , meaning that after the entry of an individual in the study we lost track at some time t or the study eventually ends and the event of interest(e.g death) didn't happen yet.

For simplicity we will assume that the event of interest is the death of an individual, so suppose that there are r distinct deaths among n individuals,then we order those deaths $t_1 < ... < t_r$,$r \leq n$.

After we define $j = 1, \dots r$, n_j as the number of individuals who are alive and uncencored just before t_j , where $j = 1, \dots r$ and d_j is defined as the number of

deaths at t_j .

We consider also that all deaths happen at the beginning of the j-th $[t_j, t_{j+1})$ interval and no more deaths happen after that. An estimation of the probability of death in the interval from $t_j - h$ to t_j when h is close to zero is $\frac{d_j}{n_j}$,so the estimated survival probability is $\frac{n_j - d_j}{n_j}$.

Moreover we the probability of survival beyond t_j and just before t_{j+1} is unity ,because no deaths happen after t_j .So due to independency among the intervals $(t_j - h, t_j]$ and (t_j, t_{j+1}) we can say as $h \to 0$ that the survival probability over the interval $[t_j, t_{j+1})$ is $\frac{n_j - d_j}{n_j}$

Furthermore if deaths and cencoring happen at the same time ,deaths go first . Finally if we assume that deaths of the individuals in the sample occur independently of one another then the survivor function at time t is the product of all estimated survivor probabilities till the interval that t lies in.

More formally

$$\widehat{S(t)} = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{1.11}$$

for $t_k \leq t < t_{k+1}$ and $k = 1, ...r$. Of course if $t < t_1, \widehat{S(t)} = 1$ and if $t \geq t_r$ the survivor funcion is zero,esle if the last observation is a censored time the survivor function is not defined beyond that time.

Overall the Kaplan-Meier provides a general form for estimating the survivor function

Before this paragraph ends it is worth to mention the case where no censored observations exist .In this case $n_{j+1} + d_j = n_j$ because individuals who survive before t_j are exactly the individuals who survive before t_{j+1} plus the deaths at t_j .

The Kaplan-Meier in (1.11) takes the form $\prod_{j=1}^{k} \frac{n_{j+1}}{n_j} = \frac{n_2}{n_1} \cdots \frac{n_{k+1}}{n_k} = \frac{n_{k+1}}{n_1}$ for $k = 1, \ldots r - 1$.

We notice that n_1 represents the individuals who survive before time t_1 but in fact in every case n_1 represents all individuals in the study because no deaths happen before t_1 . Also n_{k+1} represents the individuals who survive just before t_{k+1} so at least they survive time $\geq t_{k+1}$.

Considering the above discussion for n individuals in the study

$\widehat{S(t)} = \frac{\textbf{number of individiuals who survive} \geq t}{\textbf{all individuals in the study}} = \frac{\sum_{i=1}^{n} 1_{T_i \geq t}}{n}$,

This special case of Kaplan-Meier estimator is called empririal estimator and has mean $E(\widehat{S(t)}) = E\Big(\frac{\sum_{i=1}^{n} 1_{T_i \geq t}}{n}\Big) = \frac{1}{n}\sum_{i=1}^{n} E\Big(1_{T_i \geq t}\Big) = \frac{1}{n}\sum_{i=1}^{n} P\Big(T_i \geq t\Big)$ and variance $V(\widehat{S(t)}) = V\Big(\frac{\sum_{i=1}^{n} 1_{T_i \geq t}}{n}\Big) = \frac{1}{n^2} V\Big(\sum_{i=1}^{n} 1_{T_i \geq t}\Big)$

$1_{T_i \geq t}$ takes the value 1 if $T_i \geq t$ and the value zero otherwise.

So if we consider that T_i are independent and identically distributed random variables (i.i.d) then $1_{T_i \geq t}$ follows the bernouli distribution with parameter S(t)(probability of success) in other words the survivor function .So in this case $E(\widehat{S(t)}) = S(t)$ and $V(\widehat{S(t)}) = \frac{S(t)(1-S(t))}{n}$ meaning that $V(\widehat{S(t)}) \to 0$ as $n \to \infty$.

In conclusion the empirical estimator is a consistent estimator for the survivor fuction(this is also true for the general form of Kaplan-Meier,for further details see [2]) .

We remind to the reader that a consistent estimator T_n of a parameter θ is an estimator that converges in probability to θ .

1.2.2 Confidence intervals

The Kaplan-Meier estimator provides estimation for the survivor function S(t) at each point t.Here we discuss estimation for the survivor function in time intervals.

A confidence interval is an interval that estimates the survivor function and in general every function of interest .

Let's say that the function of interest is the survivor function then a $100(1-a)\%$ confidence inteval,$(0 < a < 1)$ is defined as an interval [a,b] where a,b are real numbers with a=A(T) and b=B(T) where T is the sample of the observations $T_1, \dots T_n$(random vector) and A,B are real functions .Then [a,b] tells us that there is a prescribed probability 1-a that the value of the true survivor function is included within it .Using mathematical notation that is $P(a \leq S(t) \leq b) = 1-a$ .This relation means that if we repeat many times the collection of survival times(observations) of individuals in the study under the same circumstances and every time we calculate the inteval [a,b] , then we expect that the $100(1-a)\%$ of those intervals to contain the function S(t) .

For a specific observasion $T = (t_1,\dots t_n)$ we hope that the interval [A(T),B(T)] is one of the many intervals ($100(1-a)\%$ of them) that contains S(t)
This is a result from the definition of the confidence interval and the empirical definition of probability(Richard von Mises).
The most common and widely used confidence interval is

$$[\widehat{S(t)} - z_{a/2}se(\widehat{S(t)}), \widehat{S(t)} + z_{a/2}se(\widehat{S(t)})] \,, \tag{1.12}$$

where $z_{a/2}$,is the upper a/2 quantile point of the standar normal distribution.This confidence interval for a large sample apply in a similar manner to many functions of interest with great success and it is based on the assumption that the quantity $\frac{g(\widehat{X_n})-g(\theta)}{se(g(\widehat{X_n}))}$ where g is a real function,θ a parameter of interest and X_n , $n = 1\dots$, are the observations ,follows the standar normal distribution as $n \to \infty$.
This interval though may cause trouble because usually the observations in clinical trials follow a positively skewed distribution (mode<median<mean) .So when the estimated survivor function is close to zero or unity, symmetric intervals are inappropriate, since they can lead to confidence limits for the survivor function outside the interval (0,1).
A pragmatic solution to this problem is to replace any limit that is greater than unity by 1, and any limit that is less than zero by 0.
Another solution is to transform $S(t)$, to take values in the range of all real numbers, and obtain a confidence interval for the transformed value.Then confidence intervals are back transformed to give a confidence interval for the survivor fuction $S(t)$.
A possible transformation is the log-log transformation $\log\Big(-\log S(t)\Big)$
$P\Bigg(\log\Big(-\log\widehat{S(t)}\Big) - z_{a/2}se\Big(\log\Big(-\log\widehat{S(t)}\Big)\Big) \le \log\Big(-\log S(t)\Big) \le \log\Big(-\log\widehat{S(t)}\Big) + z_{a/2}se\Big(\log\Big(-\log\widehat{S(t)}\Big)\Big)\Bigg) = 1-a$
This leads to an $100(1-a)\%$ confidence interval for S(t) which is

$$\left[\widehat{S(t)}^{\exp\left(z_{a/2}se\left(\log\left(-\log\widehat{S(t)}\right)\right)\right)}, \widehat{S(t)}^{\exp\left(-z_{a/2}se\left(\log\left(-\log\widehat{S(t)}\right)\right)\right)}\right]. \tag{1.13}$$

$\widehat{S(t)}$,is the Kaplan-Meier estimator but also other estimators can be used (1.2.3) and (1.2.4) paragraph .Also (1.2.5) paragraph discuss methods that compute the standar deviation $se(\widehat{S(t)})$ and in general the standar error-standar deviation $se(g(X))$ for a random variable X and for any given real function g.
Another choice to transfrom $S(t)$ is the logistic transformation $\log \frac{S(t)}{1-S(t)}$.

1.2.3 Nelson-Aalen estimator

An alternative estimator for the survivor function is the Nelson-Aalen estimator. This estimator works better than Kaplan Meier in small samples ,but in large samples Kaplan Meier maybe is better as is a generalization of the empirical survivor function which is a consistent estimator.In any case the two estimators are very similar especially at the earlier survival times.In order to find Nelson-Aalen estimator for the survivor function we notice that if we plug-in the Kaplan-Meier estimator in (1.5) expression and use the same notation as in (1.2.1) paragraph we get $\widehat{H(t)} = -log\widehat{S(t)} = -log\prod_{j=1}^{k}\frac{n_j-d_j}{n_j} = -\sum_{i=1}^{k}\log\left(\frac{n_j-d_j}{n_j}\right)$ with $t_k \leq t < t_{k+1}$ and $k = 1, ...r$. This is the Kaplan-Meier estimator for the cumulative hazard function .
At this point we observe that $\log\left(\frac{n_j-d_j}{n_j}\right) = \log\left(1+\frac{-d_j}{n_j}\right) \approx \frac{-d_j}{n_j}$ for small $\frac{d_j}{n_j}$,so it follows that $\widehat{H(t)} \approx \sum_{i=1}^{k} -\frac{-d_j}{n_j}$ for small $\frac{d_j}{n_j}$.
The last sum is the Nelso-Aalen estimator for the cumulative hazard function

$$\widetilde{H(t)} = \sum_{i=1}^{k}\frac{d_j}{n_j}\,, \tag{1.14}$$

where $t_k \leq t < t_{k+1}$, $k = 1, ..., r$, $t_{r+1} = \infty$. Now it follows immediately from the expression (1.6) if we plug-in the (1.14) that the Nelson-Aalen estimator for the survivor function is

$$\widetilde{S(t)} = \prod_{j=1}^{k}\exp(\frac{-d_j}{n_j})\,, \tag{1.15}$$

where $t_k \leq t < t_{k+1}$ and $k = 1, ...r$.
The Kaplan-Meier estimator for the survivor function can be seen also as the approximation of Nelson- Aalen (first order approximation) via the taylor expansion of $\exp(-x)$ for small x .Also because $\exp(-x) \geq 1-x$ the Nelson Aalen

estimator is always above the Kaplan Meier.

Now we present an example to provide a visual comparison of Kaplan-Meier and Nelson-Aalen estimators.
Consider the chronic data hepatitis data set taken from modelling survival data in medical research(Collett)third edition(additional data sets,appendix B).
In a clinical trial described by Kirk et al. (1980), 44 patients with chronic active hepatitis were randomised to the drug prednisolone, or an untreated control group. The survival time of patients, in months, following admission to the trial, was the response variable of interest.
The table (1.2.2) shows the first group which includes the patients who took the drug.Status is the event indicator (equals to 1 if death occured and 0 otherwise) and treatment 1 is reffered to the drug group.

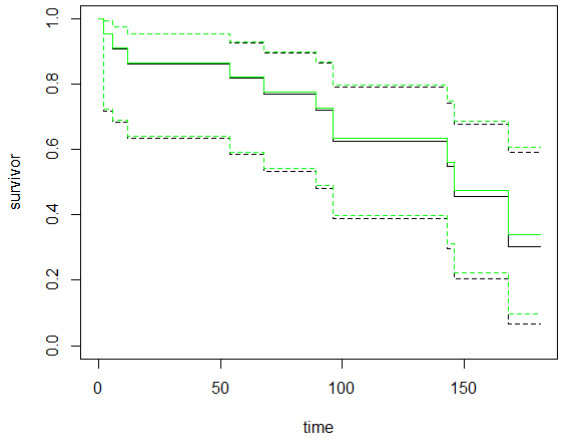
Figure 1.2.2: Patients who took the drug

	treatment	time	status
2	1	2	1
3	1	6	1
4	1	12	1
5	1	54	1
6	1	56	0
7	1	68	1
8	1	89	1
9	1	96	1
10	1	96	1
11	1	125	0
12	1	128	0
13	1	131	0
14	1	140	0
15	1	141	0
16	1	143	1
17	1	145	0
18	1	146	1
19	1	148	0
20	1	162	0
21	1	168	1
22	1	173	0
23	1	181	0

A plot for both estimators Kaplan-Meier and Nelson Aalen for the survivor function is in figure (1.2.3)

Dush lines are the boundaries(using log-log confidence intervals) (1.13) , the green color represents the Nelson-Aalen estimator and the black color the Kaplan-Meier.
This shows the previous discussion that Nelson-Aalen is always above the Kaplan-Meier and for early survival times the difference is ignorable.
The outputs in R software in figures (1.2.4),(1.2.5) estimate the survivor func-

Figure 1.2.3: Kaplan Meier vs Nelson-Aalen



tion with Kaplan-Meier and Nelson-Aalen estimators .We observe that the estimations are very close .The estimations of the survivor fuction are in the 4-th column .The 5-th and 6-th columns give the confidence intervals with the log-log menthod (1.13).The second column gives the patients who are still alive and as a result at risk of experience death.The second column gives the number of deaths at each time.

Figure 1.2.4: Kaplan - Meier

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	22	1	0.955	0.0444	0.7187	0.993
6	21	1	0.909	0.0613	0.6830	0.976
12	20	1	0.864	0.0732	0.6344	0.954
54	19	1	0.818	0.0822	0.5853	0.928
68	17	1	0.770	0.0904	0.5325	0.897
89	16	1	0.722	0.0967	0.4822	0.865
96	15	2	0.626	0.1051	0.3883	0.793
143	8	1	0.547	0.1175	0.2979	0.741
146	6	1	0.456	0.1285	0.2047	0.678
168	3	1	0.304	0.1509	0.0676	0.591

Figure 1.2.5: Nelson-Aalen

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	22	1	0.956	0.0434	0.7242	0.994
6	21	1	0.911	0.0600	0.6892	0.977
12	20	1	0.867	0.0716	0.6415	0.955
54	19	1	0.822	0.0806	0.5932	0.929
68	17	1	0.775	0.0886	0.5413	0.900
89	16	1	0.728	0.0949	0.4920	0.868
96	15	2	0.634	0.1033	0.3995	0.798
143	8	1	0.560	0.1149	0.3133	0.748
146	6	1	0.474	0.1253	0.2243	0.689
168	3	1	0.340	0.1445	0.0966	0.607

1.2.4 Actuarial estimator

A more practcal estimator for the survivor function is the life - table or Actuarial estimation .The origin of the name comes from an assumption we made which is that the censoring process is such that the censored survival times occur uniformly throughout the j-th interval,$j = 1, \dots n$ so that the average number of individuals who are at risk during this interval is

$$n'_j = n_j - \frac{c_j}{2} \,, \tag{1.16}$$

where we assume that the period of observations is split through some intervals(usually 5-15)but the choice depends on the person who conducts the study. We define also n_j,$j = 1, \dots n$ the number of individuals who are alive and uncensored , therefore at risk of death at the start of the j-th interval ,c_j,$j = 1, \dots n$ is the number of cencored observations over the j-th interval.
Additionally d_j,denotes the number of deaths in the j-th interval,then if we see n'_j,as n_j ,we follow a similar procedure as the Kaplan Meier's(1.2.1) to construct the estimator.
Here we observe deaths throught every interval in contrast to observe deaths at the start of each interval.The estimator then has the expression

$$S^*(t) = \prod_{j=1}^{k-1} (\frac{n'_j - d_j}{n'_j}) \,, \tag{1.17}$$

with $t_{(k-1)} \leq t < t_{(k)}$ and k=2,...n .After the n-th interval the actuarial estimator is zero and of course at the start of the study the actuarial estimator is one.

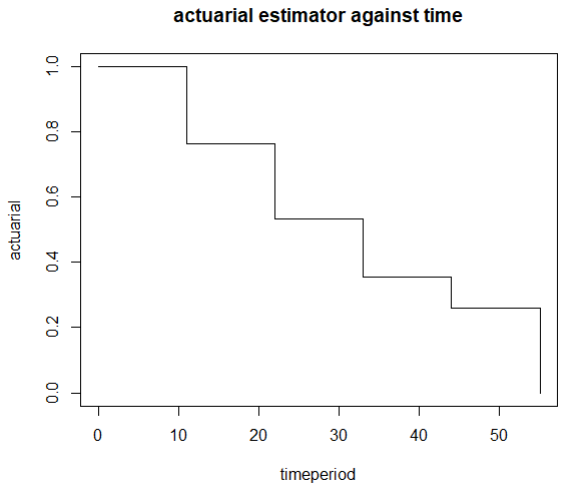
Let's consider an example with 30 individuals in a study that last 55 weeks .We will use 5 intervals with length 11,considering the theoretical analysis above we get the following table(1.2.6)
The column $n_j$new represents the quantity n'_j and the prob column represents

Figure 1.2.6: Life - table example

interval	timeperiod	d_j	c_j	n_jnew	prob	actuarialestimator
1	0	7	1	29.5	0.7627119	1.0000000
2	11	6	4	20.0	0.7000000	0.7627119
3	22	4	0	12.0	0.6666667	0.5338983
4	33	2	1	7.5	0.7333333	0.3559322
5	44	3	2	4.0	0.2500000	0.2610169

the of probability of survival in j-th interval $j = 1 \dots 5$ which is $\frac{n'_j - d_j}{n'_j}$.The intervals from figure (1.2.6) are the $[0, 11), [11, 22), \dots, [44, 55)$.So for example the probability of survival in [22,33) is $0.76 * 0.7 * 0.66 = 0.35$ and $n_2 = 30 - 7 - 1 = 22$.Moreover a plot of the life-table estimator is shown in figure (1.2.7).

Figure 1.2.7: Life - table plot



1.2.5 Standar deviation

Suppose that we have the Kaplan-Meier estimator for S(t)(same notation as in (1.2.1)), $\widehat{S(t)} = \prod_{i=1}^{k} \widehat{p_j}$ for t between the k and k+1 ordered times of death and k=1,...r where $\widehat{p_j}$,is the estimated probability that an individual survives in the time interval that begins at $t_{(j)}$,$j = 1, ..., r$, $Var(\log(\widehat{S(t)})) = \sum_{j=1}^{k} Var(\log \widehat{p_j})$
Now the key part is to assume that the number of individuals who survive in the interval that begins at t_j,follows the binomial distribution with parameters n_j, p_j,where p_j,is the true survival probability.
The random variable that defines the number of individuals who survive in the interval is $n_j - d_j$,$j = 1, ..., r$.
Then it follows that ,$Var(n_j - d_j) = n_j p_j (1 - p_j)$.
Since , $\widehat{p_j} = \frac{n_j - d_j}{n_j}$, it follows that

$$Var(\widehat{p_j}) = \frac{n_j p_j (1 - p_j)}{n_j^2} ,$$

so

$$\widehat{Var(\widehat{p_j})} = \frac{\widehat{p_j}(1-\widehat{p_j})}{n_j} = \frac{\frac{n_j-d_j}{n_j}\frac{d_j}{n_j}}{n_j} .$$

Now we have to use the taylor approximation for variance of a random variable g(X)

$$Var(g(X)) \approx \left(\frac{dg(X)}{dX}\right)^2 VarX . \tag{1.18}$$

This approximation applies to 'smooth' functions(no corners) and it is easy to prove it.

Indeed ,first we write $g(X) \approx g(\theta) + g'(\theta)(X-\theta)$,then we just take the variance of g(X) to both sides.

This is also a special case of the Delta Method. Thus if we consider that the estimated variance of the estimated probability $\widehat{p_j}$ is almost equal with the variance of the estimated probability we take $Var(\log(\widehat{p_j})) \approx \frac{1}{(\widehat{p_j})^2} Var(\widehat{p_j}) \approx \frac{1}{(\widehat{p_j})^2}\widehat{Var(\widehat{p_j})} = \frac{n_j^2}{(n_j-d_j)^2}\frac{\frac{n_j-d_j}{n_j}\frac{d_j}{n_j}}{n_j} = \frac{d_j}{n_j(n_j-dj)}$, so

$$Var(\log(\widehat{S(t)})) \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-dj)} \approx \frac{1}{[\widehat{S(t)}]^2} Var(\widehat{S(t)}) ,$$

that gives

$$\widehat{Var(\widehat{S(t)})} \approx [\widehat{S(t)}]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-dj)} .$$

Finally we have the Greenwood's formula

$$se(\widehat{S(t)}) \approx \widehat{S(t)} \left(\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-dj)}\right)^{\frac{1}{2}} , \tag{1.19}$$

for t between k and k+1 ordered times with $k = 1, \ldots, r$.

Finally we mention that with similar arguments we take similar expressions for the standar error of the Life Table and Nelson-Aalen estimator.

More specifically the form for the life-table (1.2.4) is derived easily if we think that in this framework n'_j plays the role of n_j .

$$se(S(t)^*) \approx S(t)^* \left(\sum_{j=1}^{k} \frac{d_j}{n'_j(n'_j-dj)}\right)^{\frac{1}{2}} . \tag{1.20}$$

The standar error of the Nelson-Aalen estimator(1.2.3) has the form

$$se(\widetilde{S(t)}) \approx \widetilde{S(t)} \left(\sum_{j=1}^{k} \frac{d_j}{n_j^2} \right)^{\frac{1}{2}} . \tag{1.21}$$

1.2.6 Estimating the hazard function

Here we focus on the Kaplan-Meier Estimator for the hazard function using the same notation as in (1.2.1) .
We assume that the hazard function is constant between successive death times.The probability of death in the j-th interval is $\widehat{h(t)} * l_j = \frac{d_j}{n_j}$ so

$$\widehat{h(t)} = \frac{d_j}{n_j l_j} , \tag{1.22}$$

where l_j is the length of the j-th interval and $t_j \leq t < t_{j+1}$ and j=1,...r-1, r is the total distinct deaths in the study and the hazard is zero before the first death.

The standar deviation of $\widehat{h(t)}$ can be found in a similar manner as the se$(\widehat{S(t)})$ (1.2.5).
In fact if we assume that d_j follows the binomial distribution with parameters n_j and p_j^* ,where $p_j^* = \frac{d_j}{n_j}$ is the probability of death in the j-th interval , then $Var(\widehat{h(t)}) = \frac{n_j \frac{d_j}{n_j} \frac{n_j - d_j}{n_j}}{n_j^2 l_j^2} = \frac{d_j^2}{n_j^2 l_j^2} \frac{n_j - d_j}{n_j d_j}$.
So the formula for the standar error is

$$se(\widehat{h(t)}) \approx \widehat{h(t)} \sqrt{\frac{n_j - d_j}{n_j d_j}} \tag{1.23}$$

In practice, estimates of the hazard function obtained in this way(1.22) will often tend to be difficult to use them , because plots are not very handy so a Kernel might be used to smoothen the curve .
A kernel is a non-negative integrable function $K(u)$ in which we center at each failure time.
Typically we choose a smooth-shaped kernel, with the amount of smoothing controlled by a parameter b.There are many ways to define the kernel function and selection of the appropriate amount of smoothing is one of the most difficult

problems in non-parametric hazard estimation.A Kernel for the hazard function in general has this form

$$\widetilde{h(t)} = b^{-1} \sum_{j=1}^{r} \left(K \left(\frac{t - t_j}{b} \right) \frac{d_j}{n_j} \right) , \tag{1.24}$$

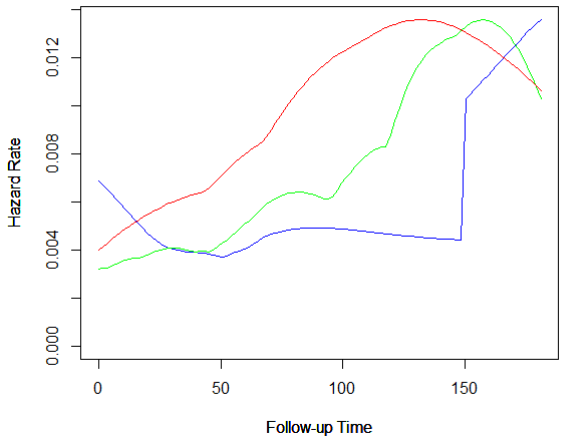
where t_j , $j = 1 \dots r$ is j-th ordered death
For example a Kernel might have the following form

$$\widetilde{h(t)} = b^{-1} \sum_{j=1}^{r} 0.75 \left(1 - \left(\frac{t - t_j}{b} \right)^2 \right) \frac{d_j}{n_j} , \tag{1.25}$$

where b is a constant and is needed to be chosen ,is also known as the bandwidth parameter and each value of this quantity controls the shape of the plot.
Moreover t belongs in the interval from b to $t_r - b$ where t_r is the last ordered death.

We use the hepatitis example(figure 1.2.2) as before to give an image of the plot of the hazard function as a kernel with different values of b . Kernels provide help to visualization. In figure(1.2.8) blue line has b=2.25 ,green is the

Figure 1.2.8: Kernels for different values of b



kernel with b=50 and red with b=100 .As b grows smoothness grows .

1.2.7 Estimating quantiles

The definition of the p-th quantile t_p is $t_p = \inf\{t : S(t) \leq \frac{100-p}{100}\}$,where $0 < p < 100$,S(t) the is survival function and inf is the infimum of this set.
In fact t_p is called generalized inverse function of p.
The t_p is well defined because the set $\{t : S(t) \leq \frac{100-p}{100}\}$ is bounded and includes at least one t,because $\lim_{t\to\infty} S(t) \to 0$, so the infimum always exists.
When the minimum of the set $\{t : S(t) \leq \frac{100-p}{100}\}$ exists we can replace the inf with min(inf=min).
In order to keep things simpler we adopt the min instead of inf in the definition of t_p.Although we keep in mind that this adoption may cause trouble .For example if we want to find that the median and the survivor function is greater than 0.5 for all t then the median doesn't exist .
When the survival function has an absolutely continuous specific form as in (1.4) section t_p can be found immidiately from the equation $S(t_p) = \frac{100-p}{100}$,but in the non-parametric framework the form of $S(t)$ is unknown.As a result we have to estimate the survivor function with the Kaplan-Meier estimator or some other estimator who is a step function of time , so $\widehat{S(t_p)} \leq \frac{100-p}{100}$,meaning that the equality may not hold.
In general the estimator of the p-th quantile is the smallest observed survival time $\widehat{t_p}$,such as ,$\widehat{S(t)} \leq \frac{100-p}{100}$,where $\widehat{S(t)}$ is the Kaplan-Meier estimator(1.2.1).Since the Kaplan-Meier change values only when deaths happen we can say that

$$\widehat{t_p} = \min\{t_i : \widehat{S(t)} \leq \frac{100-p}{100}\}\,, \tag{1.26}$$

where t_i , $i = 1\ldots, r$ is the i-th ordered death .
A formula for the standar deviation of $\widehat{t_p}$ is

$$se(\widehat{t_p}) = \frac{1}{\widehat{f}(\widehat{t_p})} se(\widehat{S}(\widehat{t_p}))\,, \tag{1.27}$$

where $\widehat{f}(\widehat{t_p}) = \frac{\widehat{S}(u_p)-\widehat{S}(l_p)}{l_p-u_p}$ is the estimated density function , u_p is the max of the set of all t_i such that $\widehat{S(t_i)} \geq 1 - \frac{p}{100} + \epsilon$ and l_p ,is the min of the set of all t_i such that ,$\widehat{S(t_i)} \leq 1 - \frac{p}{100} - \epsilon$,values of epsilon are taken to be small and $se(\widehat{S}(\widehat{t_p}))$ is found from the Greenwood formula(1.19).

Indeed first we take the approach using (1.15) , $Var(\widehat{S}(\widehat{t_p})) \approx \left(\frac{\widehat{S}(\widehat{t_p})}{d\widehat{t_p}}\right)^2 V(\widehat{t_p})$. At this point we notice that $S(t)' = -f(t)$, so we can say that $Var(\widehat{S}(\widehat{t_p})) = \widehat{f}(\widehat{t_p})^2 V(\widehat{t_p}) \iff V(\widehat{t_p}) = Var(\widehat{S}(\widehat{t_p}))\frac{1}{\widehat{f}(\widehat{t_p})}$.

A confidence interval for the p-th quantile has the form
$[\widehat{t_p} - z_{a/2}se(\widehat{t_p}), \widehat{t_p} + z_{a/2}se(\widehat{t_p})]$

1.3 A semi parametric model

1.3.1 Cox regression model

Suppose that we want to compare the survival experience of cancer patients on two different therapies.A semiparametric rather than a fully parametric hazard function might be best suited for this problem. One form of a regression model for the hazard function is

$$h_i(t) = \psi(\mathbf{x}_i)h_0(t) \,, \tag{1.28}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots x_{ip})'$ is the vector of independent variables for the i-th patient ,i=1,...,n . We assume that indepentet variables $\mathbf{x_i}$ are constants but in general their values may change over time (time varying covariates (1.5.4)),ψ,is called hazard ratio because ,$\psi = \frac{h_i(t)}{h_0(t)}$, $h_0(t)$,is called baseline hazard function. Cox (1972) was the first to propose the model and suggest using

$$\psi = \exp(\beta'\mathbf{x_i}) \,, \tag{1.29}$$

where ,β is the vector of coefficients who need estimation .We consider the model without the intercept β_0 ,because it can be included to the baseline hazard function .Also we can consider that the baseline hazard fuction is derived from the model (1.28) for the patient with $\mathbf{x} = \mathbf{0}$(null patient).

Also the model can be re-expressed in the form

$$\log \frac{h_i(t)}{h_0(t)} = \beta_1 x_{i1} + \ldots \beta_p x_{ip} \,, \tag{1.30}$$

in order to give a linear model for the logarithm of the hazard ratio . The model (1.28) is a semi-parametric model because it has the unknown parameters $\beta_1 \dots \beta_p$ that we have to estimate and the baseline function in which no assumptions are made about the actual form . Also no particular form of a probability distribution is assumed for the survival times .The model (1.28) with ψ as in (1.29) is known as the Cox regression model.
Lastly the model (1.28) is based on the proportional hazards assumption that we will discuss in (1.3.2) section .

1.3.2 Validity of the proportional assumption

If the baseline hazard $h_0(t)$ and the hazard of the i-th patient $h_i(t)$ are proportionals then ,$h_i(t) = \psi h_0(t)$,so

$$\exp\left(-\int_0^t h_i(x)dx\right) = \exp\left(-\int_0^t \psi h_0(x)dx\right) .$$

Now from (1.6) we get that

$$S_i(t) = \left(\exp\left(-\int_0^t h_0(x)dx\right)\right)^{\psi} = (S_0(t))^{\psi} . \tag{1.31}$$

The surviror functions in (1.31) are probabilities so they take values from 0 to 1 and if ,ψ, is greater than 1 , the survivor function of the patient i is greater than the baseline .
On the other hand if ψ is less than 1 the opposite is true .The point is that every time the true survivor functions do not cross .This is a necessary but not sufficient argument for the validity of the proportional hazards assumption.
Furthermore an informal graphical method is to plot the estimated survivor functions $S_i(t)$ and $S_0(t)$ over time and observe if they approximate cross or not.

A more satisfactory graphical method for assessing the validity of the proportional hazards assumption is known as the log-cumulative hazard plot

$$\log H_i(t) = \beta' \mathbf{x_i} + \log H_0(t) . \tag{1.32}$$

This equation is derived by intergrating both sides of the Cox regression model over the interval [0,t] ,so from (1.4) we take $H_i(t) = \exp(\beta' \mathbf{x_i}) H_0(t)$.Then taking the logarithm to both sides we end up in (1.32) .
We notice that differences in the log-cumulative hazard functions in (1.32) do not depend on time.In general if the proportional hazards assumption is valid the differences of the log-cumulative hazards for every pair of individuals do not depend on time.
Specifically for the i-th and j-th individual , $i \neq j$, $i, j = 1 \dots$ n
$\log H_i(t) - \log H_j(t) = \beta' \mathbf{x_i} - \beta' \mathbf{x_j}$.

This means that if we plot the log-cumulative hazard functions for each individual against time or usually the logarithm of time , the curves will be parallel.
To use this plot when individuals have an explanatory variable that is continuous,meaning that is measured in a continuous scale(e.g age) , individuals are first grouped .
For example let's say that age variable take values over the interval (10,50) ,therefore we choose for some reason to group them in 5 intervals (10,20],(20,30],(30,40],(40,50).
Then we take the Kaplan-Meier estimator of the log-cumulative hazard from the equation (1.5) for each group and plot all these estimates against $\log t$ .If the curves are approximately parallel then the proportiornal assumption is valid.

1.3.3 Fitting the model

In order to fit the Cox regression model we have to estimate the coefficients $\beta_1, \dots \beta_p$ and the baseline hazard function $h_0(t)$.A useful fact is that we can estimate those 2 separately,in particular the β's first so we can compute the hazard ratio(1.29) without computing the baseline hazard function .In order to estimate β's we will use a likelihood that is derived without direct use of the censored and event times(observations).This likelihood is called partial likelihood (Cox 1972) .
First we suppose that there are n individuals in the study , r distinct death times and n-r right cencored times(lost to follow up times).Further we assume that no ties happen on survival times (death-event or cencoring) .
The basic argument is that between successive deaths no information exists

about the effect of the explanatory variables on the hazard of death.
So we consider the probability

$$P(individual\,with\,variables\,\mathbf{x_i}\,dies\,at\,t_i|one\,death\,at\,t_i)\,,$$

where t_i is the i-th ordered death , $t_1 < \cdots < t_r$, $i = 1 \ldots r$

This probability equals to

$$\frac{P(individual\,with\,variables\,\mathbf{x_i}\,dies\,at\,t_i)}{P(one\,death\,at\,t_i)}\,,$$

then we have

$$\frac{P(individual\,with\,variables\,\mathbf{x_i}\,dies\,at\,t_i)}{\sum_{l\in R(t_i)} P(individual\,l\,dies\,at\,t_i)}\,,$$

because $P(one\,death\,at\,t_i) = \sum_{l\in R(t_i)} P(individual\,l\,dies\,at\,t_i)$, where $R(t_i)$ is called risk set and is defined as the group of individuals who are alive and uncencored just prior to t_i.
The probabilities in the numerator and denominator are replaced by the probabilities of death in the interval $[t_i, t_i + h)$, dividing by h and taking limits to zero we have

$$\frac{Hazard\,of\,death\,at\,time\,t_i\,for\,individual\,with\,variables\,\mathbf{x_i}}{\sum_{l\in R(t_i)} (Hazard\,of\,death\,at\,time\,t_i\,for\,individual\,l)}\,.$$

So the numerator is the hazard of the i-th individual .Finally taking the product from 1 to r and using the (1.28) and (1.29) we take

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta'\mathbf{x_i})}{\sum_{l\in R(t_i)} \exp(\beta'\mathbf{x_l})}\,, \tag{1.33}$$

which is the partial likelihood function for the cox model with no ties.
The (1.33) is equivalent with the expression

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta'\mathbf{x_i})}{\sum_{l\in R(t_i)} \exp(\beta'\mathbf{x_l})} \right)^{\delta_i}\,, \tag{1.34}$$

where δ_i is the indicator variable which is 0 if the i-th survival time is censored

and 1 if the i-th individual exprerience the event(e.g death) .
Taking the logarithm of this likelihood we end up with a more useful expression

$$\log(L(\beta)) = \sum_{i=1}^{n} \delta_i \left(\beta' \mathbf{x_i} - \log\Big(\sum_{l \in R(t_i)} \exp(\beta' \mathbf{x_l}) \Big) \right) . \tag{1.35}$$

The maximization process leads to no close form for β's ,so numerical methods such as Newton-Rapson procedure are used to find β's.

Consider a simple example with 6 survival times numbered from 1 to 6 .The observed survival times of individuals 2, 3 and 6 are right cencored and the others are death times $t_1 < t_2 < t_3$ for individuals 5 ,1,4 respectively .Specifically censoring for 6 is happening between the first 2 deaths and the censored times for 2 and 3 are happening between the last 2 deaths .The risk set $\mathrm{R}(t_1)$ includes all individuals ,$\mathrm{R}(t_2)$ includes 1,2,3,4 and $\mathrm{R}(t_3)$ includes 4. If $\psi(i) = exp(\beta' \mathbf{x_i})$ then partial likelihood is $\frac{\psi(5)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)}$, because the last fraction is 1.

Now we present the form of the partial likelihood when ties arise for survival times.When multiple deaths and censored times happen at some time t ,then we assume that censoring times happen just after all the deaths at t, so the risk st R(t) can be determined without further problems .
Ties can arise even in continuous case because survival times often recorded to the nearest death ,month ,year and the rounding process leads to ties.The appropriate likelihood function in the presence of tied observations has been given by Kalbfleisch and Prentice (2002) but is very complicated,so computer softwares for survival data usually use a sufficient approximation for the likelihood with ties .
An approximation has been proposed by Breslow(1974) and it's the simplest among all others

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta' \mathbf{s_i})}{\Big(\sum_{l \in R(t_i)} \exp(\beta' \mathbf{x_l}) \Big)^{d_i}} , \tag{1.36}$$

where $\mathbf{s_i}$ is the vector with of p elements , the h-th element is $s_{ih} = \sum_{k=1}^{d_i} x_{hik}$ and x_{hik} is the value of the h-th explanatory variable, $h = 1, 2, \ldots, p$, for the k-th

of d_i individuals, $k = 1, 2, \ldots, di$, who die at the i-th death time, $i = 1, 2, \ldots, r$. In this approximation, the d_i deaths at time t_i are considered to be distinct and to occur sequentially.
Other approximations proposed by Cox(1972),Efron(1977).

Next we focus on estimating the baseline hazard function which can lead to estimates for the survivor,hazard and cumulative hazard functions in the framework of the Cox regression model(1.28).
An estimate of the baseline hazard function was derived by Kalbfleisch and Prentice (1973).Suppose that there are r distinct death times $t_1 < \cdots < t_r$ and that there are d_j deaths at t_j , $j = 1 \ldots r$ and n_j individuals at risk at time t_j , $j = 1 \ldots r$.The estimated baseline hazard function at t_j is

$$\widehat{h_0}(t_j) = 1 - \hat{\xi}_j \,, \tag{1.37}$$

where $\hat{\xi}_j$ is the solution of the equation

$$\sum_{l \in D(t_j)} \frac{\exp(\hat{\beta}' \mathbf{x_l})}{1 - \hat{\xi}_j^{\exp(\hat{\beta}' \mathbf{x_l})}} = \sum_{l \in R(t_j)} \exp(\hat{\beta}' \mathbf{x_l}) \,. \tag{1.38}$$

The set $D(t_j)$ is the set of all d_j individuals who die at the j-th ordered death time and $R_{(t_j)}$ is the risk set.If no ties happen , then $d_j = 1$,$j = 1, \ldots r$ and the equation (1.38) is simplified ,meaning that the sum on the left side of (1.38) in only one term.

$$\hat{\xi}_j = \left(1 - \frac{\exp(\hat{\beta}' \mathbf{x_{(j)}})}{\sum_{l \in R(t_j)} \exp(\hat{\beta}' \mathbf{x_l})}\right)^{\exp(-\hat{\beta}' \mathbf{x_{(j)}})} \,.$$

According to the discussion above and considering that the hazard is constant between adjacent death times an appropriate estimate of the baseline hazard is given by (1.39)

$$\hat{h}_0(t) = \frac{1 - \hat{\xi}_j}{t_{j+1} - t_j} \,, \tag{1.39}$$

for $t_j \leq t < t_{j+1}$ and j=1,...r-1 ,the hazard is zero before the first death.

Cox model(1.28) according to the previous can be presented and estimated via many forms such as a form with surviror functions and with cumulative hazard functions.

More specifically if we consider that $\hat{\xi_j}$ can be regarded as an estimate of the probability that an individual survives over the interval $[t_j, t_{j+1})$ we can write

$$\widehat{S_0}(t) = \prod_{j=1}^{k} \hat{\xi}_j \ , \tag{1.40}$$

for $t_j \leq t < t_{j+1}$ and $j = 1, ...r-1$.This is also a generalization of the Kaplan-Meier estimator in (1.2.1).

From equation (1.5) we can take an estimate for baseline cumulative hazard function . Moreover estimates for the i-th patient can be derived for the hazard,survivor and cumulative hazard in the presence of covariates from the following equations.

The first equation provides an estimation for the Cox model with the (1.28) form

$$\hat{h}_i(t) = \exp(\{\hat{\beta}'\mathbf{x}_i)\hat{h}_0(t) \ .$$

Now taking the integral over [0,t] to both sides we get that

$$\hat{H}_i(t) = \exp(\hat{\beta}'\mathbf{x}_i)\hat{H}_0(t) \ .$$

Also by multiplying by -1 and taking exp to both sides from the second equation we get that(1.6)

$$\hat{S}_i(t) = \hat{S}_0(t)^{\exp(\hat{\beta}'\mathbf{x}_i)} \ . \tag{1.41}$$

Finally we mention that an explanation(see Lawless [2] for further details) for the equation (1.38) that comes from Kalbfleisch and Prentice (1973) .The main concept is to maximize the likelihood

$$\prod_{i=1}^{k} \left(\prod_{j \in D(t_i)} \left(1 - \hat{\xi}_i^{\exp(\hat{\beta}'\mathbf{x_i})}\right) \prod_{l \in R(t_{(i)}) - D(t_{(i)})} \hat{\xi}_i^{\exp(\hat{\beta}'\mathbf{x_l})} \right) \ , \tag{1.42}$$

where the second and thrird product comes from the survivor function and their over the set of deaths with ties and the censoring set respectively.

So if we maximixe with respect to $\hat{\xi}_i$ we take the desired equation .
Of course for the more complex scenario with multiple deaths at the same time we have to use numerical methods to find $\hat{\xi}_i$.
A suitable initial value is

$$1 - \hat{\xi}_0 = d_i \left(\sum_{l \in R(t_{(i)})} \exp(\hat{\beta}' \mathbf{x}_l) \right)^{-1} , \tag{1.43}$$

where d_i is the i-th ordered death time

1.3.4 Risk adjusted method

The risk adjusted method method is used to identify if the covariates of an individual give information about the estimate of a parameter of interest (such as the true survivor function).
The method uses the sample mean of all estimators for the parameter of interest and the sample mean is compared with the unadjusted estimate (for example the Kaplan-Meier). If we notice that no significant differences exist among the two estimators , then there is no reason to include the explanatory variables to etismate the survivor function .For example if the interest is focused on the estimation of the true survivor function we take the

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i(t) , \tag{1.44}$$

where $\hat{S}_i(t)$ comes from the expression (1.41).
As an illustration of this method we give an example from a data set for the survival of black ducks.
In the first year of a study on the movements and overwintering , survival of black ducks, Anas rubripes, conducted by the U.S. Fish and Wildlife Service, 50 female black ducks from two locations in New Jersey were captured and fitted with radios.
The period of the study was from 8 November 1983 to 14 December 1983 and included 31 hatch-year birds.The explanatory variables are the age(0=hatch,1=over one year),weight(in g) and length of wing(in mm).Status is 0 for censoring and 1 for death.The table (1.3.1) shows the time ,age,length ,status and weight

Figure 1.3.1: Survival table of ducks (taken from modelling survival data in medical research Collett thrid edition)(Additional data sets,appendix B)

V1	V2	V3	V4	V5	V6
1	2	1	1	1160	277
2	6	0	0	1140	266
3	6	0	1	1260	280
4	7	1	0	1160	264
5	13	1	1	1080	267
6	14	0	0	1120	262
7	16	0	1	1140	277
8	16	1	1	1200	283
9	17	0	1	1100	264
10	17	1	1	1420	270
11	20	0	1	1120	272
12	21	1	1	1110	271
13	22	1	0	1070	268
14	26	1	0	940	252
15	26	1	0	1240	271
16	27	1	0	1120	265
17	28	0	1	1340	275
18	29	1	0	1010	272
19	32	1	0	1040	270
20	32	0	1	1250	276
21	34	1	0	1200	276
22	34	1	0	1280	270
23	37	1	0	1250	272
24	40	1	0	1090	275
25	41	1	1	1050	275
26	44	1	0	1040	255
27	49	0	0	1130	268
28	54	0	1	1320	285
29	56	0	0	1180	259
30	56	0	0	1070	267
31	57	0	1	1260	269
32	57	0	0	1270	276
33	58	0	0	1080	260
34	63	0	1	1110	270
35	63	0	0	1150	271
36	63	0	0	1030	265
37	63	0	0	1160	275
38	63	0	0	1180	263
39	63	0	0	1050	271
40	63	0	1	1280	281
41	63	0	0	1050	275
42	63	0	0	1160	266
43	63	0	0	1150	263
44	63	0	1	1270	270
45	63	0	1	1370	275
46	63	0	1	1220	265
47	63	0	0	1220	268
48	63	0	0	1140	262
49	63	0	0	1140	270
50	63	0	0	1120	274

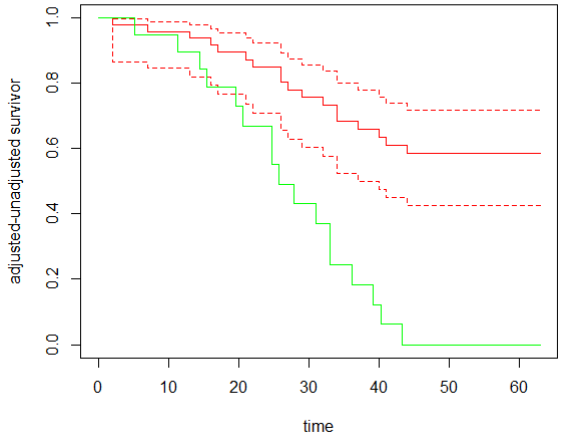
V1 is ducks as patients ,V2 time,V3 state,V4 age,V5 weight and V6 length.A plot made for adjusted(1.44) and unadjusted survivor(without the explanatory variables) (Kaplan-Meier) functions. The adjusted one is as discussed the sample mean of the survivor functions under the cox proportional hazard model.The plot (1.3.2) indicates a huge difference if we ignore the explanatory variables,but that doesn't mean that the model with the explanatory variables is the correct model .In fact in paragraph(1.3.7) we discuss methods for comparing in a more satisfactory manner two or more models.

1.3.5 Measures of explained variation

In general there are measures to explain the validity of a regression model.
Here we focus on measures who tell us if the explanatory variables used in the model , explain the actual data well ,so we can use them for further statistical inference.
In linear regression analysis this measure is

$$R^2 = \frac{SSR}{SST},$$

Figure 1.3.2: Adjusted (green)-Unadjusted(red)(dush lines are the boundaries of Kaplan-Meier) survivor functions



where SSR is the regression sum of squares and SST is the total sum of squares. This is equal to

$$R^2 = \frac{\frac{SSR}{n-1}}{\frac{SSR}{n-1} + \frac{SSE}{n-1}} = \frac{\hat{\beta}' S \hat{\beta}}{\hat{\beta}' S \hat{\beta} + \frac{Y'(I_n - P)Y}{n-1}} ,$$

where S is the variance-covariance matrix of explanatory variables(not random matrix) ,I_n is the identical $n * n$ matrix , P is the hat matirx and $\hat{\beta}$ the estimated coefficients.

In survival analysis a similar expression can be derived by Kent and O'Quigley(1988)

$$R^2 = \frac{\hat{\beta}' S \hat{\beta}}{\hat{\beta}' S \hat{\beta} + \frac{\pi^2}{6}} , \tag{1.45}$$

where $\frac{\pi^2}{6}$ is the variance of errors ϵ_i who follows the Gumbel distribution in an alternative representation of the Weibull proportional hazards model((1.4.1)and (1.5.2) sections).
Also $\hat{\beta}' S \hat{\beta}$ is an estimation of the variation in the risk score $\hat{\beta}' \mathbf{x_i}$
Other suggestions are presented by Royston and Sauerbrei (2004) and Kent and O'Quigley (1988) .

R^2 take values between 0 and 1, is largely independent of the degree of censoring, is not affected by the scale of the survival data, and increase in value as explanatory variables are added to the model.
In general big values(usually around 0.5) indicate that the model fits well with the data.

1.3.6 Residuals

Model-cheking procedures often based on quantities known as residuals.Different kind of residuals have been proposed for use under the Cox regression model.Also plots based on residuals can be helpful to identify the correct model.
Here we assume tha the cox regression model has been fitted meaning

$$\hat{h}_i(t) = \exp(\hat{\beta}' x_i)\hat{h}_0(t) \, .$$

Cox-Snell residuals

Cox-Snell residual is the most widely used residual in survival analysis.Before we explain how we can derive it, we have to mention the Nelson- Aalen or the Breslow estimator for the baseline cumulative hazard .
This estimator comes by using an approximation and overcomes the difficulty of solving the (1.38) when ties arise .So in equation (1.38) we can approximate the term $\hat{\xi}_j^{\exp(\hat{\beta}'\mathbf{x_j})}$

$$\hat{\xi}_j^{\exp(\hat{\beta}'\mathbf{x_l})} = \exp\left(\exp(\hat{\beta}'\mathbf{x_l}) \log \hat{\xi}_j\right) \approx 1 + \exp(\hat{\beta}'\mathbf{x_l}) \log \hat{\xi}_j \, .$$

Substitution of this approximation to (1.38) leads to the following equation

$$-\sum_{l \in D(t_{(j)})} \frac{1}{\log \hat{\xi}_j} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}'\mathbf{x_l}) \Longleftrightarrow -\frac{d_j}{\log \hat{\xi}_j} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}'\mathbf{x_l}) \, .$$

So

$$\hat{\xi}_j = \exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}'\mathbf{x_l})}\right) \, . \tag{1.46}$$

From equations (1.39),(1.40) and (1.5) we can take estimations for baseline hazard,survivor function and cumulative hazard function .Especially the cumulative baseline hazard is given by substituting (1.46) to (1.40) and using (1.5)

$$\hat{H}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x_l})} \,, \tag{1.47}$$

where $t_k \leq t < t_{k+1}$,$k = 1 \dots r-1$ and r is the total deaths.
This estimator is called Nelson-Aalen estimator or Breslow estimator.
Now we can explain the Cox-Snell residuals but first we will prove that $Y = -\log S(T)$ follows the exponential distribution with rate 1 ,where T is the random variable of time and S(t)the surviror function of T .
We notice that $F_Y(y) = P(Y \leq y) = P(-\log S(T) \leq y) = P(logS(T) \geq -y)$
$= P\left(S(T) \geq \exp(-y)\right) = P\left(S^{-1}(S(T)) \leq S^{-1}(exp(-y))\right)$

On the last equality we use the fact that $S^{-1}(t)$ is a decreasing function.
So $F_Y(y) = P\left(T \leq S^{-1}(exp(-y))\right)$
$= F_T(S^{-1}(exp(-y))) = 1 - S(S^{-1}(exp(-y))) = 1 - e^{-y}$

A key assumption is that if the fitted Cox model is satisfactory then the values of the survivor estimators for the individuals in the study will be close to the true values.So the cumulative hazard estimations for individuals will behave as observations from a unit exponential distibution.
The Cox-Snell residual(1968) for the i-th individual is defined

$$r_{C_i} = \exp(\hat{\beta}' \mathbf{x_i}) \hat{H}_0(t_i) = -\log(\hat{S}_i(t_i)) \,, \tag{1.48}$$

where $\hat{H}_0(t_i)$ is usually the Breslow estimator.

Many graphical procedures similar to linear regression analysis for residuals are not quite useful .Cox-Snell residuals as discussed previously have an exponential distribution with mean 1 if the fitted model is correct ,so the residuals are asymmetrically distributed.As a result index plots are not very helpful.
In general a cumulative hazard plot against r_{Ci} assess the fitness of the model .

This can be done by calculating the Cox-Snell residuals and then using the Kaplan-Meier estimator for the survivor function , but the survival times are replaced by the Cox-Snell residuals,meaning that in order to achieve the goodness of fit the survival times must follow the exponential distribution with mean 1.
If the survival time for the i-th individual is censored then the residual for that individual is censored .Finally we take that

$$\hat{H}(r_{C_i}) = -\log \hat{S}(r_{C_i}) \,,$$

and plot them against r_{C_i} .
We expect to see an approximately straight line with slope 1 and intercept 0,because if the fitted model is good then the estimated survivor function $\widehat{S(t)}$ will be approximately $exp(-t)$.This will indicate that the Cox model will fit well wth the data.

The Cox-Snell residuals have different properties from the residuals used in linear regression analysis.
In particular, they will not be symmetrically distributed about zero and they cannot be negative.

Modified Cox-Snell residuals

A problem that arises with the Cox-Snell residuals is that censored residuals are treated the same as the uncensored ones.
In general the Cox-snell residual for the i-th individual at a censored time c_i is, $r_{C_i} = \hat{H}_i(c_i)$.The actual unknown failure time of the i-th individual t_i will be greater than c_i.
If the model is correct ,then the values r_C have the exponential distribution with mean 1 ,meaning that the cumulative hazard function of an exponential distribution with mean 1 is $H(t) = -log(e^{-t}) = t$ and thus is increased linearly with time.
So bigger survival times give bigger values for the cumulative hazard ,leading us to the fact that the residuals for censored observations are smaller than the

residuals for the actual uknown survival times .
To fix this problem we consider a D positive constant known as excess residual and we use the formula for the i-th individual

$$r'_{C_i} = r_{C_i} + D\,,$$

for the censored observations.
To determine D we consider as before that since the Cox-Snell residual for the i-th individual r_{C_i} comes from an exponential distribution with mean 1 , then due to the property of lack of memory , D will also have an exponential distribution with mean 1.So we consider D = 1,because the mean of D is 1.
We mention also that the exponential distribution is the only absolute continuous distribution that has the property of lack of memory .In this case this property tell us that the probability of surviving beyond time r_{C_i} and D given that the individual already survive time beyond r_{C_i} is equal to the probability of survival beyond time D .
A formula that summarize the censored and censored case for the modified Cox-Snell residuals is given by (1.49).

$$r'_{C_i} = 1 - \delta_i + r_{C_i}\,, \tag{1.49}$$

where δ_i is the usual indicator variable and i denotes the i individual.
Another suggestion for D has been proposed by Crowley and Hu (1977) .Instead of taking the mean of exp(1) we take the median .The median can be calculated as in the (1.2.7) paragraph .More specifically $S(t_{50}) = e^{-t_{50}} = \frac{1}{2}$. So $-t_{50} = \log\frac{1}{2} \iff t_{50} = \log 2 = 0.693$ .Thus we take D=0.693 ,suggesting smaller extent on the censored residuals.

Martingale residuals

In general we can define a stochastic process M(t),$t \geq 0$ with M(t)=N(t)-L(t) ,where N(t) is the number of events(e.g deaths) on the interval [0,t) and L(t) is the cumulative intensity function which can be considered as the expected number of events over the interval [0,t) and has an analogous role in multistate models ((2.2) section).

More specifically we define the at risk process Y(t) that takes the value 1 until time t , where an event or a censored time happen and the value zero otherwise. Also we define the history of the counting process N(t), $H(t-)$ as the set of values $(N(d), Y(d))$ for all d just before t .

The intensity function $\lambda(t)$(analogous to the hazard function)(2.2 section) can be seen as the derivative over time t of the mean of dN(t)=N(t+dt)-N(t) given the history $H(t-)$.

Finally the intensity cumulative hazard function is defined as the integral over [0,t) of the intensity hazard function.

If we consider also that M(t) has zero mean ,the stochastic process M(t) is called martingale .

The martingale residuals are based on the above discussion and the formula for the i-th individual is given by (1.50)

$$r_{Mi} = \delta_i - r_{C_i} \ , \tag{1.50}$$

where δ_i is the event indicator for the i-th individual which in this case represents the number of deaths on $[0, t_i)$ and the r_{C_i} is the Cox-Snell residual that represents the average number of deaths over the $[0, t_i)$.

Also in the usual case with only one type of event we can think the intensity cumulative hazard or the intensity hazard as the ordinary cumulative hazard or hazard function,although for more details see the section (2.2).

As we see the formulas (1.49) and (1.50) has the relationship $r_{Mi} = 1 - r'_{C_i}$.Moreover because the Cox-Snell residuals are positive the values of the martingale residuals for the censored observations belong to the interval $(-\infty, 1)$ and for the uncensored observations are negative.

Another useful property that also has the usual residuals in a linear regression model is that the sum of martingale residuals sum to zero.

This fact can be proved by considering the following Riemann-Stieltjes integrals(1.8) for the i-th individual $N_i(t) = \int_0^t dN_i(u)$ and $\widehat{L_i}(t) = \int_0^t d\widehat{L}(u) = \int_0^t \widehat{\lambda(u)}du$.

The intensity function can be written using an analogous to Cox model form as $\widehat{\lambda}_i(u) = Y_i(u) \exp\left(\widehat{\beta_i}' \mathbf{x_i}(\mathbf{u})\right) d\widehat{L_0}(u)$ where $\mathbf{x_i}(\mathbf{u})$ covariate has a more general form as it depends on time u (for further details see the section (2.2),(2.11)

formula).
Also $d\widehat{L_0}(u) = \frac{\sum_{j=1}^{n} dN_j(u)}{\sum_{j=1}^{n} Y_j(u) \exp\left(\widehat{\beta_j}' \mathbf{x_j(u)}\right)}$ which is an analogous form of the Breslow estimator(1.47) for the indicator cumulative hazard .
At this point it is clear that if we consider that the Cox-Snell residuals instead of the usual cumulative hazard represent the analogous indicator cumulative hazard the sum of r_{Mi} for all n individuals is

$$\sum_{i=1}^{n} r_{Mi} = \sum_{i=1}^{n} \left(\int_0^t dN_i(u) - \int_0^t \left(Y_i(u) \exp\left(\widehat{\beta_i}' \mathbf{x_i(u)}\right) \frac{\sum_{j=1}^{n} dN_j(u)}{\sum_{j=1}^{n} Y_j(u) \exp\left(\widehat{\beta_j}' \mathbf{x_j(u)}\right)} \right) \right).$$

This sum equals to zero because is a finite sum and as result can go inside of the integrals and eliminating all the terms .

Another property that can be found also in the linear regression residuals is that the mean of the martingale residual for the i-th individual, for large sample goes to zero and can be proved by using martingale convergence theorems ,for further details see [17].

Other suggestions for residuals are the deviance residuals,Schoenfeld residuals and score residuals.For further details the reader can see [1].

Before we close the section of residuals we mention an example to demonstrate the Cox-Snell and martingale residuals .
The data set is taken from [1] .Suppose a study with 13 patients who need a kidney dialysis.That is a procedure to remove waste materials when the kidneys stop working properly.Often patiens face the danger of infaction from this procedure and the procedure must stop.
So we consider that the event time is the infection ,meaning that the event indicator is one for infection and zero for stopping the procedure for another reason(censoring).Moreover in this example we consider two explanatory variables the sex and age .The figure(1.3.3) gives all the details. As we dicuss before in order to find the explanatory variables first we must find the estimated Cox model and use the Breslow estimator .Using R we get the output in figure (1.3.4)
As result using the formulas (1.48) and (1.50) we get the following table in fig-

Figure 1.3.3: Dialysis table

patient	time	status	age	sex
1	8	1	28	1
2	15	1	44	2
3	22	1	32	1
4	24	1	16	2
5	30	1	10	1
6	54	0	42	2
7	119	1	22	2
8	141	1	34	2
9	185	1	60	2
10	292	1	43	2
11	402	1	30	2
12	447	1	31	2
13	536	1	17	2

Figure 1.3.4: Cox model

	coef	exp(coef)	se(coef)	z	p
age	0.03037	1.03084	0.02624	1.158	0.2470
sex	-2.71076	0.06649	1.09590	-2.474	0.0134

Likelihood ratio test=6.48 on 2 df, p=0.03921
n= 13, number of events= 12

ure (1.3.5). From figure(1.2.5) we see also another thing that mentioned before

Figure 1.3.5: Martingale and Cox-Snell residuals

mart_resid	resid_coxsnell
0.7199554	0.28004463
0.9276882	0.07231183
-0.2139338	1.21393382
0.9157260	0.08427404
-0.5060216	1.50602155
-0.2646313	0.26463133
0.7645308	0.23546922
0.5163173	0.48368269
-0.4379232	1.43792320
-0.2123480	1.21234799
-0.1866143	1.18661433
-0.8279315	1.82793148
-1.1948139	2.19481389

which is that the Cox-Snell residuals are positive and the martingale residuals don't exceed 1.

At this point in order to test the assumption of the goodness of fit for the Cox model we can plot the estimated cumulative hazard of Cox -Snell residuals against Cox-Snell residuals as pseudo times. From figure (1.3.6) we can say that approximately the straight line with slope one and intercept zero (red line) passes through the black dots .So even though the sample is small we can say that the Cox model fits well the data.

Furthermore we can use a martingale residual vs index(patients) plot to identify possible outliers ,meaning data from patients for whom the residuals in ablolute value are large.This plot is provided in figure (1.3.7). From figure(1.3.7) we don't observe any outlier and also we see some kind of symmetry around zero . Finally other plots can be made such as plot for martingale residuals vs age in figure (1.3.8)

Figure 1.3.6: Cumulative hazard against Cox-Snell residuals

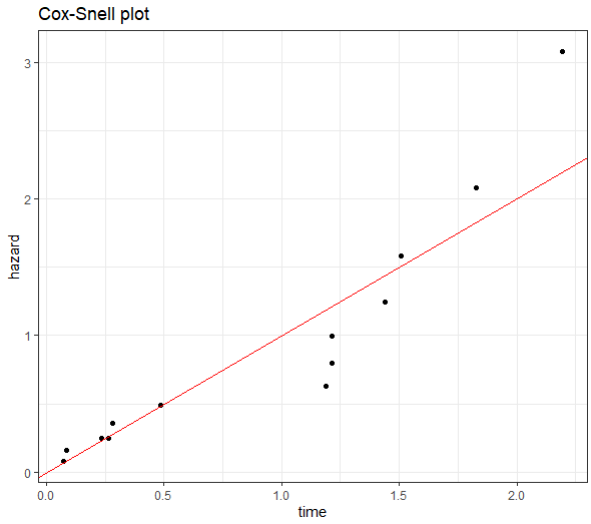


Figure 1.3.7: Martingale residuals vs index

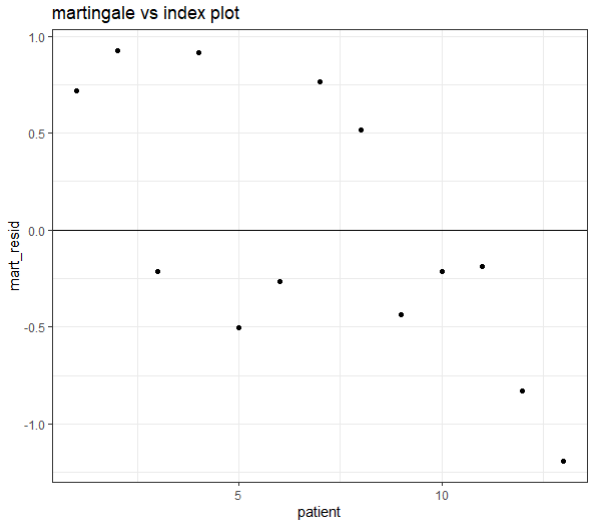
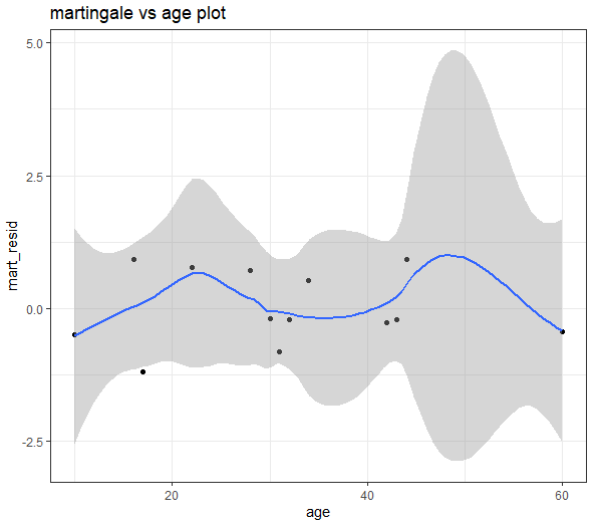


Figure 1.3.8: martingale residuals vs age



1.3.7 Hypothesis tests and model comparison

In order to identify which model fits best to the survival data many methods can be used to prove the importance of each explanatory variable.
First of all we present some classical tests that are generally used in the standar theory but are also used in survival analysis .
So we will mention the Wald test,Score test and Likelihood ratio test.

Wald and score tests

Suppose that we focus in the hypothesis $\beta_j = 0$, where $j = 1 \dots p$.
A Wald test uses the statistical function $T = \frac{\hat{\beta_j}}{se(\hat{\beta_j})}$.This quantity under the null hypothesis follows the standar normal distibution or equivalently the square of this statistic follows the chi-squared distribution with 1 degree of freedom.
The p-value or the upper $a/2$ quantile of the standar normal distribution can be used to determine the rejection or not of this hypothesis,where a is the level of significance .
More specifically if $p < a$ or $T(0) > Z_{a/2}$ then the exaplanatory variable with the coefficient β_j is significant and we keep it in the model.
However we must take notice that the hypothesis is being tested in the precense of all other β's.So a careful analysis must be done.

For example let's say that in the presence of 4 explanatory variables , two coefficients satisfy the null hypothesis in some significant level .Then we cannot conclude that these variables can be exluded from the model .This is because if we remove only X_1 for instance then X_2 might reject the null hypothesis and as result is considered important for the model .
This can definitely happen if X_1 and X_2 are highly correlated .
If only one variable satisfies the null hypothesis in the presence of all others then we may conclude that it can be rejected from the model.
This problem arises for the reason that in general coefficients are dependent on one another meaning that they correlated in some degree(positively or negatively).So alternative methods for comparing Cox models might needed.Such methods are discussed in this paragraph.
We can notice also that the standar deviation of β_j can be obtained approximately as the square root of jj-diagonal element of the inverse observed information matrix for the value $\hat{\beta}$.The observed information matrix is given by the following formula

$$\widehat{I(\hat{\beta})} = \left(\frac{\partial L(\hat{\beta})}{\partial \beta_i \partial \beta_j} \right)_{i,j} .$$

Now we consider the null hypothesis that all coefficients(β_j) are zero .If this hypothesis is true then the null model(baseline hazard) is the best model ,so there is no need to use explanatory variables for the model.
Wald test for this hypothesis uses the statistic function

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

This statistic under the null hypothesis has the chi - squared distibution with p degrees of freedom (number of β's).
On the other hand score test uses the statistic function

$$U(\mathbf{0})' I^{-1}(\mathbf{0}) U(\mathbf{0}) ,$$

where $U(\beta) = \left(\frac{d \log L(\beta)}{d\beta_j} \right)_j$ is the score vector.
This statistic also under the null hypothesis has the chi - squared distibution with p degrees of freedom.

Likelihood ratio test

Likelihood ratio test uses the statistic function

$$2\left(\log L(\hat{\beta}) - \log L(\mathbf{0})\right) ,$$

where L(0) is the likelihood for the null model.
In order to reject the null hypothesis under a level of significance α we check if p - value for the chi-squared with p degrees of freedom is less than $\alpha(p < \alpha)$.

For the general scenario let's consider two models .
The first model is $h_1(t) = \exp(\beta_1 \mathbf{x_1} + \dots \beta_p \mathbf{x_p}) h_0(t)$ and the second model has the form $h_2(t) = \exp(\beta_1 \mathbf{x_1} + \dots \beta_p \mathbf{x_p} \dots \beta_{p+q} \mathbf{x_{p+q}}) h_0(t)$.
Model 1 is nested to model 2 and the null hypothesis is $H_0 : \beta_{p+1} = \dots \beta_{p+q} = 0$
The statistic for this case is given by (1.51).

$$2\left(\log L(\hat{\beta}_B) - \log L(\hat{\beta}_A)\right) , \tag{1.51}$$

where $L(\hat{\beta}_B)$ is refered to model 2 and the other likelihood in model 1 .
This statistic has chi-squared with q degrees of freedom.(number of estimated beta's in model 2 minus number of estimated beta's in model 1)

Next it follows a discusion about some strategies that we can pursue for picking the best model.
In general a correct-best model to fit the data doesn't exist .As the famous statistician George Box(1919-2013) said 'All models are wrong but some are useful'.
Before we dive into some strategies for picking a model or models we explain the 2 types of explanatory variables , variates and factors.
Variates take numerical values that are usually on a continuous scale of measurment.
Factors on the other hand are variables that take limited values known as levels(e.g sex:Male,Female(2 levels)) and in order to fit them to the model we must

define some dummy variables.

In Cox regression model such as on all other models the first level is set to zero

.

Then if a factor has K levels we set K-1 dummy variables , each of these take the value 1 for the corresponding level and zero otherwise as in figure (1.3.9).

Figure 1.3.9: Dummy variables ,the table comes from [1]

Level of A	X_2	X_3	. . .	X_a
1	0	0	. . .	0
2	1	0	. . .	0
3	0	1	. . .	0
. . . . .				
a	0	0	. . .	1

Other subcategories of explanatory variables are the interactions and mixed terms.

Interactions come up if the model contains at least 2 factors.

For example the gender of a student and the academic performance(bad,mediocre,good,excellent). An interaction of those two factors is for instance the 'female' with 'good'.In order to put interactions into the model we consider the dummy variables X_1 which is 1 if the gender is female and zero otherwise , Y_1 which is 1 if the performance is mediocre and zero otherwise,Y_2 which is 1 if the performance is good and zero otherwise , and Y_3 which is 1 if the performance is excellent and zero otherwise. Again we remind that level 'male' and 'bad' is zero.

Overall we can make 3 interactions , for example X_1Y_2 .Also usually in statistics in order to fit an interaction to the model we include the primary factors as well,meaning that in previous example we fit X_1 and Y_2 in order to fit X_1Y_2.This is called hierarchy principal. In general though there are some cases that we may use only the interaction .

In a more general manner let's say that we have 2 factors with p and k levels respectively then interactions are (p-1)(k-1) and it is said that the interaction of those 2 factors has (p-1)(k-1) degrees of freedom.

Finally we mention the mixed term that combines a variate with a factor .This term can be used when a coefficient of a variate may differ for each level of a factor.

In figure (1.3.10) we give an example with 9 individuals ,a factor A with 3 levels and a variate X with 9 values. As we see in the table U_2 and U_3 are the dummy variables of A and for instance the 4-th indivilual adds a term x_4 in the model ,meaning one more coefficient to estimate.

Figure 1.3.10: Mixed terms ,table is from [1]

Individual	Level of A	X	U_2	U_3	U_2X	U_3X
1	1	x_1	0	0	0	0
2	1	x_2	0	0	0	0
3	1	x_3	0	0	0	0
4	2	x_4	1	0	x_4	0
5	2	x_5	1	0	x_5	0
6	2	x_6	1	0	x_6	0
7	3	x_7	0	1	0	x_7
8	3	x_8	0	1	0	x_8
9	3	x_9	0	1	0	x_9

At this point as we explained the different types of explanatory variables we continue in strategy methods for picking a good model.
A usefull criterion for comparing not necessarily nested models is the Akaike's information criterion

$$AIC = -2\log\hat{L} + 2q\,,$$

where q is the number of beta's and L the likelihood function.Smaller values lead to better models .

In general is likely that more than one models can be used to fit the data .
The usage of many variables in the model in order to get more efficiency it can be proven computationally expensive.
In this case, automatic routines for variable selection are available in many software packages.These routines are based in forward selection, backward elimination and stepwise procedure.
The forward selection adds variables one at a time .The selection of each variable based on the quantity can be based on p-values in a Wald test or on the p-values of a likelihood ratio test or even on Akaike's criterion.
The first step is to compare the null model with the models with one variable ,then the models with one variable with the models with 2 variables and so on , until a stopping rule(e.g $pvalue > 0.1$).

On the other hand the backward elimination procedure have the opposite procedure . It starts from comparing the full model with the models which have one less variable and continue in that manner until some stopping rule(e.g $pvalue < 0.1$).Finally the stewise procedure combines those two methods.

These routines usually lead to one 'best' model and this model depends on the method used .
Also models obtained by these methods usually violate the hierarchic principle.As a result these routines can be used with cautious.

Another strategy for model selection is recommended.
First we fit all the models that contain each of the variables one at a time.Then we compare them with the null model using the likelihood ratio test .
The Chi squared (1 D.F) p values determine if we include those variables to the model .We may also use the AIC criterion to give a stronger evidence.
Then the variables that are proven to be important are included in the model and we compare this model with the model that contains only one of these important variables in order to observe the significance of each variable in the presence all other variables .
If none of this comparisons (likelihood ratio tests) is significant we keep the model ,otherwise we exlude the non-significant variables from the model.
Finally we may check if some of the exluded variables from the first step are significant in the presence of all others(not likely) and also we check for possible interactions or mixed terms among the variables in the final model .

In medical research treatment effect arises .
For example suppose a study that has 2 groups.
Patients in the first group take placebo , in group 2 some medicine and we want to observe the effect of the medicine.
This effect(treatment effect) can be included in the model as a factor with 2 levels.The first level is a patient in group1(zero value) and the second level is a patient in group 2(value 1).
In general treatment effect can be seen as a variable that determines the effect

in survival time.
In the presence of a treatment effect , first we pick a model with some strategy that discussed before considering the explanatory variables without the treatment effect.
In this way the treatment effect doesn't affect the other variables.
Then we add the treatment effect to the final model and a comparison(likelihood ratio test) with the model without the treatment effect shows if the treatment effect is significant or not.
It is useful also to check if the treatment effect alone is significant.
In general a level of sifnificance a=0.15(not too strict) is recommended for general use.
Lastly we mention that there is always a need for non-statistical considerations in model building.

At this point let's consider an example for better undestanding in model selection methods(data comes from [1]).
In a placebo-controlled trial about bladder cancer, conducted by the Veterans Administration Cooperative Urological Research Group, patients with superficial bladder tumours had their tumour removed transurethrally.Then randomization to 2 groups takes place .
The first group takes the placebo and second group takes the chemotherapeutic agent, thiotepa.
The initial number of tumours in each patient, and the diameter of the largest of these, was recorded at the time of randomisation.The original data comes from D.F. Andrews and A.M. Herzberg (1985) and gives the times to up to nine tumour recurrences.
This data set , focus only on the first recurrence.
Patients who haven't experience recurrence by the end of follow up period are considered as right censored observations.
This study has 86 patients ,time is measured in months,status is zero for censoring and 1 for recurrence.The treatment variable is 1 for the first group and 2 for the second group,also init is the initial number of tumours and size is the diameter of the largest initial tumour in cm.

The figure (1.3.11) give the data for the first 6 patients in the study. At this

Figure 1.3.11: Table of first 6 patients

patient	time	status	treat	init	size
1	0	0	1	1	1
2	1	0	1	1	3
3	4	0	1	2	1
4	7	0	1	1	1
5	10	0	1	5	1
6	6	1	1	4	1

point we want to find the optimal model so we will follow the recommended strategy.

As a result in figures (1.3.12),(1.3.13) and (1.3.14) we compare the null model with the models with only one variable. As we see the figure (1.3.12) gives

Figure 1.3.12: Null model vs treat

```
Model 1: ~ 1
Model 2: ~ treat
  loglik  Chisq Df P(>|Chi|)
1 -185.14
2 -184.37 1.5356  1    0.2153
```

Figure 1.3.13: Null model vs init

```
Model 1: ~ 1
Model 2: ~ init
  loglik  Chisq Df P(>|Chi|)
1 -185.14
2 -181.79 6.6987  1  0.009649 **
```

pvalue=0.2 ,so the treatment effect is not significant on it's own.

Also the figure (1.3.13) gives pvalue=0.009 ,so the initial number of tumors is important and lastly in figure (1.3.14) gives pvalue=0.756 ,so the diameter of the largest tumour is definately not significant.

Next the figure (1.3.15) compares the model with init and treatment with the model with only the treatment variable and gives pvalue=0.004 ,indicating that the treatment effect is important in the presence of init.

Finally in figure (1.3.16) we check if the mixed term with init and treatment is significant and as it turns out it's not because pvalue=0.82.In conculsion the best model ise the model with treat and init.

Before we end this paragraph we present also the backward algorithm with the Akaike's criterion and we observe that in figure (1.3.17) we end up with the same model.

As we see in figure (1.3.17) the full model without the variable 'size' achieves the

Figure 1.3.14: Null model vs size

```
Model 1: ~ 1
Model 2: ~ size
  loglik  Chisq Df P(>|Chi|)
1 -185.14
2 -185.09 0.0965  1     0.756
```

Figure 1.3.15: Treat model vs treat and init

```
Model 1: ~ treat
Model 2: ~ init + treat
  loglik  Chisq Df P(>|Chi|)
1 -184.37
2 -180.40 7.9349  1  0.004849 **
```

Figure 1.3.16: Mixed term

```
Model 1: ~ init + treat + init:treat
Model 2: ~ init + treat
  loglik  Chisq Df P(>|Chi|)
1 -180.38
2 -180.40 0.0481  1     0.8264
```

smallest AIC among all models,so this variable is excluded from the model.Also in the second step by eliminating treat or init we end up with models with bigger AIC(> 364.8) so the process stops.

1.4 Parametric proportional hazard models

An assumption for the Cox regression model is that the baseline hazard function is defined arbitary ,so there is no need to assume a specific probability distribution for the survival times.

This means that the hazard function of a patient is hasn't a specific functional form ,so models with this assumption can be generally used for a large set of applications.

In fact in general non - parametric assumptions have a wide range of applicability,although parametric assumptions meaning a specific distibution , gives more presicion to our a study and the quantities of interest such as the median tend to have smaller deviances.

In this section we focus on parametric models .

The probability distribution that has a central role in survival analysis is the Weibull distribution , introduced by W.Weibull(1951) for industrial reliability tests and has an analogous role in survival analysis as the normal distribution in linear regression models.

1.4.1 Proportional Weibull model

First we present the Weibull distribution.

The Weibull distribution has the following density function for time t, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$.

Also the survivor function is $S(t) = \int_t^{\infty} f(u)du = \exp(-\lambda t^{\gamma})$ and the hazard

Figure 1.3.17: Null model vs size

```
Start:  AIC=366.36
Surv(timec, statusc) ~ init + size + treat

        Df    AIC
- size   1 364.80
<none>     366.36
- treat  1 367.24
- init   1 372.67

Step:  AIC=364.8
Surv(timec, statusc) ~ init + treat

        Df    AIC
<none>     364.80
- treat  1 365.58
- init   1 370.74
Call:
coxph(formula = Surv(timec, statusc) ~ init + treat)

          coef exp(coef) se(coef)      z       p
init   0.23079   1.25960  0.07542  3.060 0.00221
treat -0.51218   0.59919  0.31299 -1.636 0.10176

Likelihood ratio test=9.47  on 2 df, p=0.00878
n= 86, number of events= 47
```

function is $h(t) = \frac{f(t)}{S(t)} = \lambda\gamma t^{\gamma-1}$, where γ is the shape parameter , λ the scale parameter and they are positive parameters. IF $\gamma > 1$ then hazard increases monotonically and for $\gamma < 1$ decreases mononically.
Moreover we mention that for $\gamma = 1$ we get the exponential distribution

Now we present the Weibull proportional model.The only thing that seperates this model from the Cox model is that the baseline hazard function has a Weibull form.

$$h(t) = \exp(\beta'\mathbf{x})\lambda\gamma t^{\gamma-1} ,$$

and for each patient i that is

$$h_i(t) = \exp(\beta'\mathbf{x_i})\lambda\gamma t^{\gamma-1} . \tag{1.52}$$

In order to fit the model (1.52)to the data we have to estimate all the unknown parameters , meaning to estimate the beta's ,scale and shape parameters.
To do that we will use a general formula but first we suppose that n individuals paticipate in a study and the sample is a combination of events and right censored times.
We define the variables T_i as the event time and C_i as the censored time for the i-th individual.Also we define δ_i as the event indicator and $M_i = min(T_i, C_i)$ as the random variable that indicates what appears first(censoring or event).
The probability distribution of the pair (M_i, δ_i) is given by (1.53) and (1.54).

$$P(M_i = t_i, \delta_i = 0) = P(C_i = t_i, T_i > t_i) = P(T_i > t_i)P(C_i = t_i) = S_{T_i}(t_i)f_{C_i}(t_i). \tag{1.53}$$

$$P(M = t_i, \delta_i = 1) = P(T_i = t_i, C_i > t_i) = P(T_i = t_i)P(C_i > t_i) = S_{C_i}(t_i)f_{T_i}(t_i). \tag{1.54}$$

It is clear that in (1.53) and (1.54) we assume independence between survival and cencored times(independent censoring).Also when we referring to distributions as $P(T_i = t)$ or $P(C_i = t)$ we consider both possible scenarios meaning discrete or continuous in order to simplify the process. In fact to be more accurate for the continuous case for example we take $P(T_i \in (t, t + h)) \approx f_{T_i}(t)dt$ and we consider this as $P(T_i = t)$, meaning the probability density function.
Taking the product of (1.53) and (1.54) for all n patients and considering that C_i and T_i for $i = 1 \dots n$ are independent we end up with the following expression

$$\prod_{i=1}^{n} \left(S_{C_i}(t_i)f_{T_i}(t_i)\right)^{\delta_i} \left(S_{T_i}(t_i)f_{C_i}(t_i)\right)^{1-\delta_i} \ .$$

Rearranging this formula a little we get

$$\prod_{i=1}^{n} \left(S_{C_i}(t_i)^{\delta_i} f_{C_i}(t_i)^{1-\delta_i}\right) \prod_{i=1}^{n} \left(S_{T_i}(t_i)^{1-\delta i} f_{T_i}(t_i)^{\delta_i}\right) \ .$$

Non informative censoring indicates that the first product is considered as constant. So the likelihood is analogous to the second part. That is

$$L(\theta) = \prod_{i=1}^{n} \left(S_{T_i}(t_i)^{1-\delta i} f_{T_i}(t_i)^{\delta_i}\right) \ ,$$

where $\theta = (\mathbf{beta}, parameters\ of\ some\ distribution)$
Now by using the formula (1.3) we get an equivalent form

$$L(\theta) = \prod_{i=1}^{n} \left(S_{T_i}(t_i)h_{T_i}(t_i)^{\delta_i}\right) \ .$$

The logarithm of this likelihood is

$$\log(L(\theta)) = \sum_{i=1}^{n} \left(\delta_i \log(h_i(t_i)) + log(S_i(t_i))\right) \ . \tag{1.55}$$

At this point by using the formula (1.6) and (1.52) we get the formula for the survivor function under a Weibull baseline distribution $S(t) = \exp\left(-\exp(\beta'\mathbf{x})\lambda t^{\gamma}\right)$ and by substitution of this formula and (1.52) to (1.55) we get the (1.56) ex-

pression

$$\log(L(\theta)) = \sum_{i=1}^{n} \left(\delta_i \Big(\beta' \mathbf{x_i} + \log(\lambda\gamma) + \gamma \log(t_i) \Big) - \lambda \exp(\beta' \mathbf{x_i}) t_i^{\gamma} \right) . \qquad (1.56)$$

In order to take estimations for gamma ,lambda and beta's we take the partial derivatives of the log-likelihood under those parameters and we end up with p+2 equations (p=number of beta's).

To solve this system of equations we have to use a computer software that solves this numerically(e.g Newton-Raphson method) .

Even with small number of beta's or in the special case of the exponential distribution , computations by hand using for example the Newton - Raphson method can be difficult .

As an illustration if we have the most simple case of the exponential case, then $h(t) = \exp(\beta' \mathbf{x})\lambda$, $S(t) = \exp\left(-\exp(\beta' \mathbf{x})\lambda t\right)$ and by taking the derivatives of the(1.56) with $\gamma = 1$ for λ and beta's $\frac{\partial log L(\theta)}{\partial \beta_j}, \frac{\partial log L(\theta)}{\partial \lambda}$, we have the following equations

$$-\sum_{i=1}^{n} \exp(\beta' \mathbf{x_i}) t_i + \frac{r}{\lambda} = 0 \,,$$

$$\sum_{i=1}^{n} \delta_i x_{ij} - \lambda \sum_{i=1}^{n} x_{ij} \exp(\beta' \mathbf{x_i}) t_i = 0 \,,$$

for $j = 1 \ldots p$ respectively and r total events(e.g deaths).

This system can be solved only numerically .As a result the Newton - Raphson method can be used.

Before we explain this method we notice that we can reduce the complexity of this problem by one if we solve the first equation for λ and substitute this to the second set of equations(p in total).

More specifically the estimated λ is

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} \exp(\hat{\beta}' \mathbf{x_i}) t_i} \,.$$

We notice that in the absence of covariates the above estimation is the number of deaths divided by the total survival times.

So we have to solve the set of p equations($j = 1 \dots p$)

$$g_j(\beta) = \sum_{i=1}^{n} \delta_i x_{ij} - \frac{r}{\sum_{i=1}^{n} \exp(\beta' \mathbf{x_i}) t_i} \sum_{i=1}^{n} x_{ij} \exp(\beta' \mathbf{x_i}) t_i = 0 \, .$$

This set of equations can't be solved numerically , so we have to use the Newton-Raphson or some other method ,in order to find the maximum points.
The Newton-Raphson method uses the partial derivatives $\frac{\partial \ g_j(\beta)}{\partial \beta_i}$ and the $g_j(\beta)$ for $i, j = 1 \dots p$.
In fact the method uses the observed information matrix $I(\beta) = \left(\frac{\partial \ g_j(\beta)}{\partial \beta_i} \right)_{ij}$ and the score function $U(\beta) = \Big(g_1(\beta), \dots, g_p(\beta) \Big)$.
First we choose a suitable initial value(e.g the value zero $\beta = 0$) to start the algorithm .The initial value can be crucial because if it has a large distance from the true value the algorithm can be stuck .
Then we follow the iterative process (1.57) until some criterion stops the algorithm (e.g small distances between 2 consecutive solutions , 10^{-2} for example).

$$\beta_{k+1} = \beta_k + I^{-1}(\beta_k) U(\beta_k) \, .$$

After we take the estimates we can calculate all the quantities of interest with the plug-in principle.

1.4.2 Assessing the Weibull assumption

In order to be sure whether a Weibull distribution is suitable there are some options that can be examined.
A first option is to estimate the hazard using non-parametric methods as in (1.2.6) section.If the hazard is reasonably constant over time we may fit an exponential distribution.
On the other hand if the hazard increases or decreases monotonically then we may choose the Weibull distribution.
This procedure can be done also by estimating the cumulative hazard (1.14) and see if the graph behaves as the graph of a cumulutaive Weibull hazard ,but for the survivor function for example the Weibull expression is more complicated ,so we may want to avoid it.

A second option is to use the log-cumulative hazard plot by using the (1.58).The survivor function of a Weibull distribution is $S(t) = \int_t^\infty f(u)du = \exp(-\lambda t^\gamma)$,so we end up with the following equation

$$\log(-\log S(t)) = \log\lambda + \gamma \log t\,. \tag{1.57}$$

Then we plug-in the Kaplan-Meier or the Nelson-Aalen estimator for the survivor function.
If the plot of the log-cumulative hazard against the $\log t$ gives an approximate straight line then we may use the Weibull distribution.
Also this plot gives approximate estimates for λ and γ .The $\log\lambda$ is the intercept of the line and γ is the slope.
This option can be used also when we may consider fitting a proportional model with some explanatory variables.

1.4.3 Gompertz model

Gompertz model is used in biological and demography sciences and can be used as an alternative approach in the framework of the proportional hazard assumption for parametric models.
Gompertz distribution introduced by Gompertz(1825) in order to model the human mortality.
The hazard of this distribution is

$$h(t) = \lambda\exp(\theta t) = exp(a + \theta t)\,,$$

where λ , α and θ are positive parameters.
For $\theta = 0$ we take the exponential distribution.As the Weibull hazard ,the Compertz hazard decreases and increases monotonically.
In order to use this distribution in the proportional hazard model we simply substitute the baseline hazard with the Compertz hazard.
The model for each i patient is

$$h_i(t) = \exp(\beta'\mathbf{x_i})\lambda\exp(\theta t)\,.$$

1.5 Non-Proportional hazards

When the proportional assumption is not tenable ,we need to use an alternative method to fit the data.
As a result we will introduce the accelerated failure time model who contains a family of distributions useful for these problems.

1.5.1 The accelerated failure time model

Suppose that each i individual has p explanatory variables ,then the general form of the accelerated hazard is

$$h_i(t) = \exp(-\omega_i) h_0\Big(\frac{t}{\exp(\omega_i)}\Big) , \tag{1.58}$$

where $\omega_i = \alpha' \mathbf{x_i}$ and α' ,x_i are the coefficient vector and the explanatory vector for the i-th individual.
Also the baseline hazard represents the zero patient.
By taking integrals to both sides of the (1.59) equation and transforming the integral $\int_0^t h_0\Big(\frac{u}{\exp(\omega_i)}\Big)du$ by using the substitution $u = x\exp(\omega_i)$ we take

$$H_i(t) = \exp(-\omega_i)\exp(\omega_i)\int_0^{\frac{t}{\exp(\omega_i)}} h_0(x)dx .$$

So $H_i(t) = H_0\Big(\frac{t}{\exp(\omega_i)}\Big)$ and by using (1.5) we get also that the accelarated survivor function is $S_i(t) = S_0\Big(\frac{t}{\exp(\omega_i)}\Big)$.
In general the accelerated failure time models determine the speed of the progression of a disease , meaning that this model can slow down or speed up the survival time of an individual according to the explanatory variables.

An alternative representation of the accelarated model is the log-linear model

$$\log T_i = \mu + \alpha_1 x_{i1} + \ldots \alpha_p x_{ip} + \sigma\epsilon_i , \tag{1.59}$$

where i is the i-th individual , μ and σ are the intercept and scale parameters respectively.
Also the ϵ_i is the error for the i-th individual ,it represents the deviation of the

values of $\log T_i$ from the linear component and is assumed that has a parametric distribution.

In addition the positive values of alphas indicate an increase in survival times with increasing values of explanatory variables and vice versa.For example , we suppose that α is negative and the only explanatory variable is the age ,then the survival times of the older individuals are smaller($logT_{younger} > logT_{older}$)

At this point we will prove that (1.60) leads to the general accelerated failure model.

First we consider the survivor function for the i-th individual by using (1.60)

$$S_i(t) = P(T_i \geq t) = P\Big(\exp(\mu + \alpha_1 x_{i1} + \dots \alpha_p x_{ip} + \sigma\epsilon_i) \geq t \Big) =,$$

$$P\Big(\exp(\mu + \alpha' \mathbf{x_i} + \sigma\epsilon_i) \geq t \Big) =,$$

$$P\Big(\exp(\mu + \sigma\epsilon_i) \geq \frac{t}{\exp(\alpha' \mathbf{x_i})} \Big) .$$

The baseline survivor function is

$$S_0(t) = P\Big(\exp(\mu + \sigma\epsilon_i) \geq t \Big) ,$$

for $\mathbf{x_i} = \mathbf{0}$

As a result $S_i(t) = S_0\Big(\frac{t}{\exp(\alpha' \mathbf{x_i})}\Big)$ is the general form of the accelerated failure time model for the survivor function.

The term $\exp(\omega_i)$ is refferred to as the acceleration factor,where $\omega_i = \alpha' \mathbf{x_i}$.

Moreover the survival function for the i-th individual under the model (1.60) can be defined from the survival function of the error ϵ_i

$$S_i(t) = P(T_i \geq t) = P(\log T_i \geq \log t) = P(\mu + \alpha_1 x_{i1} + \dots \alpha_n x_{ip} + \sigma\epsilon_i \geq \log t) =,$$

$$P\Big(\epsilon_i \geq \frac{\log t - \mu - \alpha_1 x_{i1} - \dots - \alpha_p x_{ip}}{\sigma}\Big) = S_{\epsilon_i}\Big(\frac{\log t - \mu - \alpha_1 x_{i1} - \dots - \alpha_p x_{ip}}{\sigma}\Big).$$

We mention also that most computer softwares use the log-linear model also for the case of the proportional Weibull model and in the next section we will see the reason for that.

1.5.2 The Weibull accelerated model

In this section we will present the importance of the Weibull distribution in survival analysis.

Suppose the assumption of the Weibull distribution.

The baseline hazard according to the Weibull distribution is $h_0(t) = \lambda\gamma t^{\gamma-1}$ and the accelarated failure time model for the i-th individual is

$$h_i(t) = exp(-\omega_i)\lambda\gamma\Big(\frac{t}{\exp(\omega_i)}\Big)^{\gamma-1} = \bigg(\Big(exp(-\omega_i)\Big)^{\gamma}\lambda\bigg)\gamma t^{\gamma-1}\,.$$

Also the proportional Weibull version is

$$h_i(t) = \Big(exp(\psi_i)\lambda\Big)\gamma t^{\gamma-1}\,,$$

where ω_i and ψ_i represent the risk vectors.

So in both cases the hazard of each patient preserves the Weibull distribution.

As a result it is said that the Weibull distribution possess the accelerated failure time property and the proportional hazard property.

In fact is the only distribution that has both of these properties.

Thus the proportional model is equivalent with the accelerated model and $\psi_i = -\gamma\omega_i$.

Under the (1.59) model if the ϵ_i has the Gumbel distribution then we will prove T_i has the Weibull distribution.

The survivor function of the Gumbel distribution is $S_{\epsilon_i}(t) = \exp(-\exp(t))$,where t is taking values in the set of real numbers.In (1.5.1) we have proved that

$$S_i(t) = S_{\epsilon_i}\Big(\frac{\log t - \mu - \alpha_1 x_{i1} - \cdots - \alpha_p x_{ip}}{\sigma}\Big)$$

Then under the Gambel distribution we take

$$S_i(t) = \exp\left(-\exp\left(\frac{\log t - \mu - \alpha'\mathbf{x_i}}{\sigma}\right)\right) = \exp\left(-\exp\left(\frac{-\mu - \alpha'\mathbf{x_i}}{\sigma}\right)t^{\frac{1}{\sigma}}\right) = \exp(-\lambda_i t^{\gamma}). \tag{1.60}$$

This leads to the fact that the survivor function of the i-th patient is a Weibull distribution with parameters λ_i and γ.

Next we will investigate the connection between the proportional Weibull model

and the accelarated Weibull failure time model.
By using the proportional Weibull model for the hazard function (1.52) we get by integrating on both sides over [0,t] the cumulative hazard for the i-th individual $H_i(t) = \exp(\beta' \mathbf{x_i} \lambda t^\gamma)$.
Also by using (1.6) the survivor function is

$$S_i(t) = \exp\Big(- \exp(\beta' \mathbf{x_i}) \lambda t^\gamma) \Big) . \tag{1.61}$$

The relation between the parameters of (1.60) and (1.61) is $\lambda = \exp(\frac{-\mu}{\sigma})$, $\gamma = \frac{1}{\sigma}$ and $\beta_j = -\frac{\alpha_j}{\sigma}$.
These reparametrizations are very useful when we use a computer software.

1.5.3 Fitting the accelerated Model and Model checking

In order to get estimates for μ , σ and α's the log-likelihood (1.56) can be used. Under the (1.59) model we prove that for the i-th individual

$$S_i(t_i) = S_{\epsilon_i}\Big(\frac{\log t_i - \mu - \alpha' \mathbf{x_i}}{\sigma}\Big) = S_\epsilon(\kappa_i) .$$

Also the derivative of $1 - S_i(t_i)$ gives the density function for the i-th individual

$$f_i(t_i) = \frac{1}{\sigma t_i} f_\epsilon(\kappa_i) .$$

By subtitution of these quantities to (1.56) we get that

$$\log Ł(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \Big(- \delta_i \log(\sigma t_i) + \delta_i \log f_{\epsilon_i}(\kappa_i) + (1 - \delta_i) \log S_{\epsilon_i}(\kappa_i) \Big) . \tag{1.62}$$

So the Newton-Rapson method (1.4.1) can be used to find the estimators of α's ,μ and σ.

At this point we present the chronic active hepatitis example in (1.2.2) figure but now we will include also the patients in the second group don't took the drug(figure(1.5.1)).

The accelerated Weibull model using R gives the following results in figure(1.5.2).

Figure 1.5.1: Hepatitis ,full data set

	treatment	time	status
2	1	2	1
3	1	6	1
4	1	12	1
5	1	54	1
6	1	56	0
7	1	68	1
8	1	89	1
9	1	96	1
10	1	96	1
11	1	125	0
12	1	128	0
13	1	131	0
14	1	140	0
15	1	141	0
16	1	143	1
17	1	145	0
18	1	146	1
19	1	148	0
20	1	162	0
21	1	168	1
22	1	173	0
23	1	181	0
24	2	2	1
25	2	3	1
26	2	4	1
27	2	7	1
28	2	10	1
29	2	22	1
30	2	28	1
31	2	29	1
32	2	32	1
33	2	37	1
34	2	40	1
35	2	41	1
36	2	54	1
37	2	61	1
38	2	63	1
39	2	71	1
40	2	127	0
41	2	140	0
42	2	146	0
43	2	158	0
44	2	167	0
45	2	182	0

Figure 1.5.2: Accelarated model

```
             Value Std. Error     z      p
(Intercept) 4.481      0.317 14.14 <2e-16
treath      1.054      0.510  2.07  0.039
Log(scale)  0.237      0.169  1.40  0.161

Scale= 1.27

Weibull distribution
Loglik(model)= -157   Loglik(intercept only)= -159.3
        Chisq= 4.54 on 1 degrees of freedom, p= 0.033
Number of Newton-Raphson Iterations: 5
n= 44
```

So the estimated values for the parameters under the (1.60) is $\hat{\mu} = 4.481$, $\hat{\sigma} = 1.27$, $\hat{\alpha} = 1.054$.

The estimated treatment effect $\hat{\alpha}$ is positive ,this means that the therapy prednisolone is trying to slow down the progression of the liver disease,because the patients who don't take the drug in contrast with the figure (1.5.1) are considered as 'zero',so the '1' values (patients who took the drug) increase the log - survival times and as a result the survival times.

Also the accelaration factor is $\exp(-\hat{\alpha}) = 0.34$.

Moreover the output (1.5.2) indicates the goodness of fit with the survival data, because p-value=0.033 by using the likelihood ratio test.

At this point we will also fit the Cox model in figure(1.5.3). The relative hazard $\psi = 0.4358 < 1$ indicates that the hazard meaning the danger of death in the new treatment(Prednisolone) is less than the old treatment(control group,the patients who don't take the drug).

So the patients in the new treatment(patients who took the drug) tend to live

Figure 1.5.3: Cox model

```
  n= 44, number of events= 27

          coef exp(coef) se(coef)      z Pr(>|z|)
treath -0.8306    0.4358   0.3973 -2.091   0.0366 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
treath    0.4358      2.295       0.2    0.9494

Concordance= 0.633  (se = 0.048 )
Likelihood ratio test= 4.47  on 1 df,   p=0.03
Wald test            = 4.37  on 1 df,   p=0.04
Score (logrank) test = 4.61  on 1 df,   p=0.03
```

longer.

Moreover the standar error using the Cox model for the treatment effect is 0.3973 and the likelihood ratio test,Wald test and Score test have small p-values ,meaning that the Cox model is also a good fit for the survival data .

Assuming also that the Cox model is more flexible in terms that does not include any assumption about the baseline function and considering that gives a smaller deviance for the treatment effect we may considering to pick this instead of the Weibull model.On the other hand papametric models have smooth curves in contrast with the curvers of Cox models , so it is easier to discern a pattern,but are highly sensitive , meaning that they have high dependancy on the distribution we use.

In any case those 2 models give similar results and we can pick any of them.

A plot shows the survivor functions for those 2 models.Red and green smooth lines are the survivor functions of the control group and treatment group(took the drug) respectively under the Weibull model and dush lines give the surviror functions under the Cox model in the same way. Before we end this section we will mention some forms of residuals that are used in the parametric case.

Suppose that we already fit the log - linear model (1.59).

A natural residual under this model is called the standardised residual.

This residual is simply comes from (1.59) if we solve the equation for the ϵ_i term ,so the residual for the i-th individual has the following form

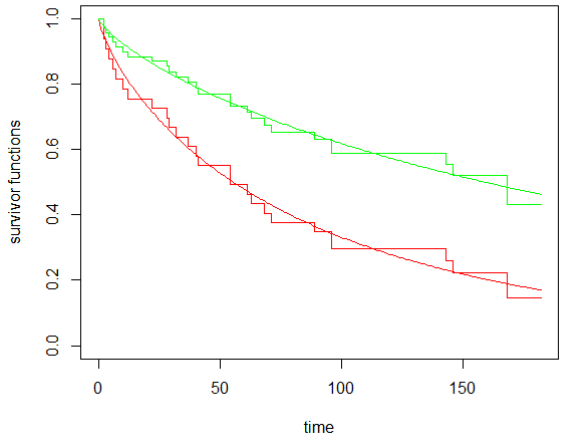
$$\hat{\epsilon_i} = \frac{\log t_i - \hat{\mu} - \hat{\alpha}'\mathbf{x_i}}{\hat{\sigma}} ,$$

where t_i is the observed survival time.

If the model is correct , the plot of the survivor function $-\log S_{\epsilon_i}(\hat{\epsilon_i})$ against the $\hat{\epsilon_i}$ will give a straight line with slope 1 and intercept zero .

This fact is explained in (1.3.6).The reason is that the $-\log S(t)$ has the expo-

Figure 1.5.4: Cox model vs Accelarated model for each group



nential distribution with unit mean and $\widehat{S}_i(t) = S_{\epsilon_i}\left(\frac{\log t_i - \hat{\mu} - \hat{\alpha}'\mathbf{x_i}}{\hat{\sigma}}\right)$.
Of course as in (1.3.6) instead of the survivor function for the errors, we use the Kaplan-Meier estimator.

Another residual is the Cox-Snell residual (1.48) , but the survivor function in this case is the $\widehat{S}_i(t) = S_{\epsilon_i}\left(\frac{\log t_i - \hat{\mu} - \hat{\alpha}'\mathbf{x_i}}{\hat{\sigma}}\right)$.
As we see the standardised residuals and Cox-Snell residuals have a close relation.

Moreover if we want to check for outliers we can use the martingale residuals (1.50) ,but the Cox-Snell residuals has the form of this section.

1.6 Time - dependent models

In the previous sections models have explanatory variables that are recorded from the origin of the study.
Although many studies contain survival data that are monitored for the duration of the study and as a result it's necessary to take into account the values of the valriables that change over time,otherwise there is a danger that the model might be misleading.

There are two types of variables that change over time.The external variables and the internal variables.

The internal variables associate to a specific individual and can't be measured if the patient is dead or in general leave the study.
These variables arise when multiple measurements of certain attributes happen over time.
Also this class of variables contains the indicator variables which indicate whether a patient suffers from a disease or not over time.

On the contrary the external variables are variables that don't necessarily taking into account the survival of a patient for their existence.
For example the age of a patient is an external variable meaning that over time we know exactly the age of a patient independently of the survival time.
Another example is the temperature of a room that a study is conducted.The temperature exists independently of any particular individual being alive or not and the changes in temperature may affect the lifetime of an individual.
In a sense external variables are deterministic variables except for cases such as the last example .

At this point we mention also that there is a chance for time-varying coefficients, meaning a term $\beta(t)$.If this term is linear it can be fitted in the corresponding explanatory variable(e.g tbX=bX(t)),but if the coefficient has a non linear form it might be difficult to fit in the model.

In conclusion we mention that there is a rule that must be taken into account when modelling survival data.
This rule says that we cannot predict the survival of an individual by using covariate values from the future and by violating this rule it may have a little or no effect on the model, but in any case we have to be cautious when we pick a model.

1.6.1 Generalized Cox Model

In the presence of time-varying variables Cox model can be generalized by simply substitute all constant covariates x with x(t) in the (1.28) model and we get the (1.63) form

$$h_i(t) = \exp\Big(\sum_{j=1}^{p} \beta_j x_{ij}(t)\Big) h_0(t)\,. \tag{1.63}$$

Moreover from the expression (1.35) and by considering covariates as functions of time we take the following expression

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \Big(\sum_{j=1}^{p} \beta_j x_{ij}(t_i) - \log \sum_{l \in R(t_i)} \exp\Big(\sum_{j=1}^{p} \beta_j x_{lj}(t_i)\Big)\Big)\,. \tag{1.64}$$

In order to find the estimated beta's from the (1.64) we must know the values of all covariates in the model at each death time in the risk set $R(t_i)$.
This can be done for the deterministic external variables(predetermined values) but for the internal and external variables which are not predetermined, it might be a problematic situation.
For example let's say that we have only one variable in the model, the 'blood pressure' and 2 individuals in the study.
This variable takes non-deterministic values over time and influence in a different way the survival times of those 2 individuals.
Further we suppose that the i-th individual died at $t_i < t_k$,where t_k is the survival time of the j-th individual, then the risk set (all individuals alive just before t_i) involves both of these individuals.
As a result we must know the values $x_k(t_i)$, $x_i(t_i)$.
More specifically the contribution in (1.64) of the i-th individual is

$$\beta x(t_i) - \log\Big(\exp(\beta x(t_i) + \exp(\beta x(t_k))\Big)\,. \tag{1.65}$$

Let's say also that the variable 'blood pressure' for each patient is measured at time l_1,l_2,l_3, where $l_1 < l_2 < t_i < l_3 < t_j$.
One option might be to take into account the measurments at l_2 ,meaning the shorter time before the time t_i in order to calculate the expression (1.65).

1.6.2 Counting process format

Many computer packages in order to fit time-dependent variables use the counting process format. In this format we use time intervals who contain constant values of all the explanatory variables.
The event indicator(0 means alive and 1 means dead) is set to zero for all intervals until the final interval in which it might be censoring(0) or event(1).
The upper limit of the last interval is the event time(event indicator is 1) or censoring time(event indicator remains zero).
Each interval associates with the values of 'start', 'stop'and 'status'.

As an example let's consider an artificial example with 8 patients with liver disease, who randomized in a placebo treatment(zero's) or in a new treatment(one's).
The Lbr is the natural logarithm of the bilirubin value (in $\mu mol/l$).
Biliburin is a substance in blood and high values of this substance cause hyperbilirubinemia that often is a sign of liver disease.
Let's say also that survival times represent months and the patients have to return to the clinic after some periods of time in order to measure the biliburin level.
The analytic representation is shown in figure (1.6.1). For instance the patient

Figure 1.6.1: Start - Stop format

individuals	start	stop	status	treatment	lbr
1	0	33	0	0	5.15
1	33	67	0	0	5.20
1	67	90	1	0	5.22
2	0	32	0	0	5.20
2	32	67	0	0	5.30
2	67	80	0	0	5.50
3	0	35	0	0	5.21
3	35	70	0	0	5.60
3	70	80	1	0	5.80
4	0	38	0	0	5.30
4	38	71	0	0	5.32
4	71	75	1	0	4.55
5	0	40	0	1	5.40
5	40	80	0	1	4.50
5	80	91	0	1	3.80
6	0	33	0	1	5.16
6	33	85	0	1	5.00
6	85	87	1	1	4.50
7	0	50	0	1	5.60
7	50	80	0	1	3.20
7	80	83	1	1	4.25
8	0	31	0	1	5.23
8	31	63	0	1	5.60
8	63	93	0	1	3.80

1 is in the placebo group and goes in clinic 3 times for measurments.
The figure (1.6.2) shows the resulting estimations of lbr and treatment for a naive analysis that don't take into account the different values of lbr.
Also in figure (1.6.3) is the results of a correct approach((1.64) model). More-

Figure 1.6.2: Naive approach

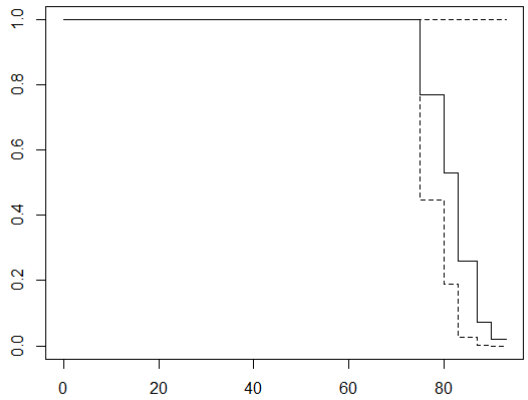
```
 n= 8, number of events= 5

              coef exp(coef) se(coef)      z Pr(>|z|)
treatment2  -2.3106    0.0992   1.6162 -1.430    0.153
lbr2         5.6441  282.6097   5.0050  1.128    0.259

           exp(coef) exp(-coef) lower .95 upper .95
treatment2    0.0992  10.080574  0.004176 2.357e+00
lbr2        282.6097   0.003538  0.015520 5.146e+06

Concordance= 0.864  (se = 0.125 )
Likelihood ratio test= 2.88  on 2 df,   p=0.2
Wald test            = 2.05  on 2 df,   p=0.4
Score (logrank) test = 2.49  on 2 df,   p=0.3
```

Figure 1.6.3: Correct approach

```
   n= 24, number of events= 5

              coef exp(coef) se(coef)      z Pr(>|z|)
treatment -0.6006    0.5485   1.6840 -0.357    0.721
lbr        0.4430    1.5574   1.2249  0.362    0.718

           exp(coef) exp(-coef) lower .95 upper .95
treatment     0.5485     1.8232   0.02022     14.88
lbr           1.5574     0.6421   0.14118     17.18

Concordance= 0.727  (se = 0.129 )
Likelihood ratio test= 1.59  on 2 df,   p=0.5
Wald test            = 1.5  on 2 df,   p=0.5
Score (logrank) test = 1.65  on 2 df,   p=0.4
```

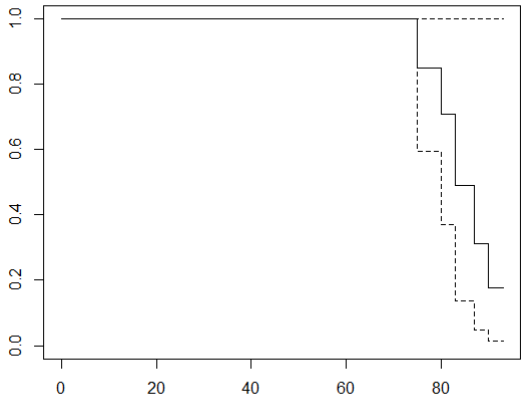
over a plot is presented for the estimated surviror function against t under the naive approach in figure (1.6.4) and for the correct approach in figure(1.6.5).

Figure 1.6.4: Survivor plot ,naive



From the figures (1.6.2) and (1.6.3) we get quite different estimations that can lead to mistakes when we try to make predictions about future values of lbr for example.

Figure 1.6.5: Survivor plot, correct



1.6.3 Parametric models

Little work has been done for parametric models in the presence of time-dependent variables.Although Petersen(1986) shows how we can fit such models. This paper considers 2 types of external variables X(t) and Z(t).The first variable has covariates discrete or continuous who remain constant for finite subperiods of time.

On the other hand Z(t) has covariates that change all the time.

The hazard fuction is

$$h\Big(t|X(t),Z(t)\Big) = \lim_{h\to\ 0} \frac{P\Big(t \le T < T+h|T \ge t, X(t), Z(t)\Big)}{h}.$$

The survivor function for survival beyond a time t_k is

$$S\Big(t_k|X(t_k),Z(t_k)\Big) = \exp\Big(-\int_0^{t_1} h(s|X(0),Z(s))ds\Big)\times\cdots\times\exp\Big(-\int_{t_{k-1}}^{t_k} h(s|X(t_{k-1}),Z(s))ds\Big),$$

where the limits on the integrals reflect the duration in a state(meaning different values of X(t) who changing over time).

An important fact is that X(t) is constant over these intervals so is independet from the time path.

For further details the reader can read [9].

Chapter 2

Multiple events

In the previous sections was presented the usual case scenario in which the indiviuals can only experience one event.
In general though,in many studies we focus on multiple events.
Multiple events are divided into two categories.
The first case is the events of the same type.For example the recurrence of headaches.
The second case involves events with different types.For example in a patient with liver disease the events might be 'death','transplant' and a specific bilirubin value.
Moreover in studies where the primary event is death individuals can experience several non-fatal events which formulate an event history.These models are reffered to as multistate models
The most simple example is the alive-dead situation in which an individual has 2 possible states, alive or dead and the transition rate from the state 'alive' to the state 'dead' is the hazard function of death at time $h(t)$.
Also in a situtation with more than 2 states,let's consider the patients with liver disease and four states.The first state is 'transplant',the second state is 'failure'(the transplanted organ fails),the third stage is the 'retransplant' and the fourth stage is 'death'.
In this example a patient can experience 3 events,death , failure of the new transplant or retranspant.

Moreover the event history starts from state 1 'transplant' and it can lead to the state 4 'death' or to the state 2 'failure'.From the 'failure' state,the patient can end up dead(death state) or may have another try for new transplant(retransplant state 3) and at some time eventually will be dead(death state). So the possible event histories are $1 \to 4$, $2 \to 4$, $3 \to 4$, $1 \to 2 \to 4$, $1 \to 2 \to 3 \to 4$.
All these possible histories have their own transition rates.

Finally we mention a special case of multiple events that is called competing risks.In this case different types of events can happen but the occurance of any of these events precludes the occurance of the other ones.
For example a patient can die by many possible causes(cancer,stroke etc) but only one cause is the true cause of death.

2.1 Competing Risks

In competing risks framework, individuals have several potential causes of death ,these causes referred to as risks.These risks compete to become the actual cause of death.
Each cause precludes the occurrence of the event from any other cause.
In contrary with the standar survival analysis who contains an event indicator δ_i for each patient with value 1 if the event happened(e.g death) ,in competing risks event indicator is replaced by a more general indicator δ_{ij} which is 1 if the i-th individual dies from the j-th cause.

2.1.1 Usual methods in competing risks

Kaplan -Meier approach

A first approach in competing risks might be to estimate the survivor functions for each cause with the Kaplan-Meier estimator.
This can be done by considering each cause one at a time as the event and the other causes as censored observations.
Such data are reffered to as cause-specific data.
So in fact we estimate the probability of death beyond time t, if cause j is the

only cause of death, meaning that all other causes are removed.

In general this approach has several problems.The obvious one is that Kaplan-Meier doesn't take into account the other causes of death,but also underestimates or overestmates the actual survival probabilities for each cause.Moreover this method in order to be valid the causes must be independent,an assumption that cannot be tested and thus for these reasons we might have to seek alternative methods.

To give some light on this method we use an example(data from [3]).

This example involves patients with prostate cancer.Time is measured(in months) from diagnosis with prostate cancer till death.

Death has 2 causes.The first one is the cause 'prostate cancer' and the second one is the death from other causes.

Also the patients in total are 62 and the detailed table(first 20 patients) is presented in figure (2.1.1). From the figure (2.1.1) we split the second column

Figure 2.1.1: Prostate cancer

survtime	status	status_other_cause	status_prostate_cancer
18	0	0	0
18	0	0	0
3	0	0	0
2	2	1	0
5	0	0	0
110	0	0	0
41	0	0	0
11	0	0	0
13	0	0	0
100	0	0	0
37	2	1	0
22	0	0	0
77	1	0	1
33	0	0	0
61	0	0	0
59	0	0	0
7	0	0	0
16	0	0	0
38	0	0	0
19	0	0	0

'status'(0=censoring,1=prostate cancer,2-death from other cause) into 2 other columns,'status for other cause' and 'status for prostate cancer '.As a result the 3 and 4 columns provide the survival data to fit the Kaplan-Meier estimators.

In figures (2.1.2) and (2.1.3) the Kaplan-Meier estimations are presented for each cause. In figure (2.1.4) is presented the plot of survivor functions for the

Figure 2.1.2: Kaplan-Meier for prostate cancer

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
26	30	1	0.967	0.0328	0.905	1
40	21	1	0.921	0.0547	0.819	1
56	12	1	0.844	0.0889	0.686	1
66	6	1	0.703	0.1483	0.465	1
77	3	1	0.469	0.2154	0.191	1

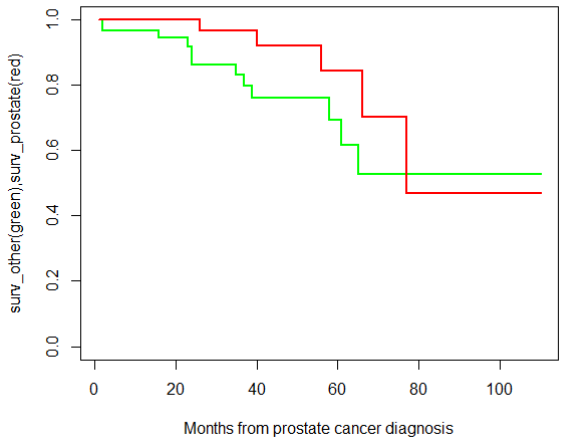
2 causes of death.

The last time recorded is 110 and is a cencored time so the survivor functions don't take values further than this time.

Figure 2.1.3: Kaplan-Meier for other causes

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	61	2	0.967	0.0228	0.924	1.000
16	45	1	0.946	0.0308	0.887	1.000
23	34	1	0.918	0.0406	0.842	1.000
24	33	2	0.862	0.0539	0.763	0.975
35	27	1	0.830	0.0606	0.720	0.958
37	26	1	0.798	0.0662	0.679	0.939
39	22	1	0.762	0.0724	0.633	0.918
58	11	1	0.693	0.0933	0.532	0.902
61	9	1	0.616	0.1102	0.434	0.875
65	7	1	0.528	0.1247	0.332	0.839

Figure 2.1.4: Kaplan-Meier plots for each cause



From figure (2.1.2) and (2.1.3) we get that the probability of dying in the time period until 110 due to prostate cancer is 1-0.469=0.531 and due to other causes 1-0.528=0.472.

The sum of these probabilities give that the value of the overall cumulative distribution at 110 is $1.003 > 1$,indicating overestimation(small in this example) in one or both probabilities .

Cause specific functions

In competing risks analogous functions as in the (1.1) section are used, but they focus particularly on each cause of death.

These functions are mentioned as cause specific functions.

The cause specific hazard for the j-cause is defined as

$$h_j(t) = \lim_{h \to 0} \frac{P\Big(t \leq T < t+h, C = j | T \geq t\Big)}{h}, \tag{2.1}$$

where C represents the cause of death and take the values $j = 1 \dots n$.
The event $\{t \leq T < T + h, C = j | T \geq t\}$,occurs when the survival time for the j-th cause is between [t,t+h) given that the survival time from all causes is equal or greater than time t.
An analogous form of the (1.3) is derived from (2.1).
Indeed,

$$P\Big(t \leq T < t + h, C = j | T \geq t\Big) = \frac{P\Big(t \leq T < t + h, C = j\Big)}{P(T \geq t)} .$$

So

$$h_j(t) = \frac{1}{P(T \geq t)} \lim_{h \to 0} \frac{P\Big(t \leq T < t + h, C = j\Big)}{hP(T \geq t)} = \lim_{h \to 0} \frac{P\Big(t \leq T < t + h, C = j\Big)}{h} = \frac{f_j(t)}{S(t)}, \tag{2.2}$$

where the $f_j(t)$ is the j cause specific density function and S(t) is the overall survival function.

The overal hazard function is the sum of the hazards from each cause.This can proved easily since the events $\{t \leq T < T + h, C = j | T \geq t\}$,$j = 1 \dots n$ are disjoint(can't occur 2 or more of these at the same time).
As a result the probability in (2.1) splits into n probabilities for each cause and we end up with the expression for the overall hazard $h(t) = \sum_{i=1}^{n} h_i(t)$.
Now by taking integrals to both sides of these formula over [0,t] we get the overall cumulative hazard $H(t) = \sum_{i=1}^{n} H_i(t)$ and from the (1.6) we get also that $S(t) = \exp(-\sum_{i=1}^{n} H_i(t)) = \prod_{i=1}^{n} \exp(-H_i(t)) = \prod_{i=1}^{n} S_i(t)$, where $S_i(t)$ is the survivor function for the i-th cause(surviving beyond t and dying from the i-th cause).
The $1 - S_i(t)$ it can be understood as the probability of dying from the i-th cause in the hypothetical world where all the other causes don't exist.

The most useful functions in competing risks are the cause specific cumulative distribution function(or cumulative incidence) and the overall cumulative distibution function.
The reason is that before a death occurs for a cause when all causes are in

present.

The j-th cause specific cumulative distribution function is defined as $F_j(t) = P(T \leq t, C = j)$ and from the previous discussion the overall cumulative distribution function is $\sum_{i=1}^{n} F_j(t)$.

The cumulative incidence function can be interpreted as the probability of dying of the j-th cause in the precense of all other causes.

As $t \to \infty$, $F_j \to P(C = j) \neq 1$ so the cause specific cumulative distribution function is not a proper distribution function and for this reason is also called subdistribution function.

In addition from (2.2) equation we have that $f_j(t) = h_j(t)S(t)$ amd by taking integrals to both sides over [0.t] we get that $F_j(t) = \int_0^t h_j(u)S(u)du$.

From this equation we can take a proper estimation for the cause specific cumulative distribution function, in contrast with the Kaplan-Meier approach in the beginning of this section.

This estimation has the following form

$$\widehat{F_j}(t) = \sum_{j:t_i \leq t} \frac{\delta_{ij}}{n_i} \widehat{S}(t_{i-1}) \,, \tag{2.3}$$

where $\frac{\delta_{ij}}{n_i}$ is the Nelson-Aalen estimator for the cause specific hazard of the j-th cause(see the (1.14) formula,also δ_{ij} for the i-th patient is one for the j-th cause and zero otherwise) and the estimated survivor function is an overall Kaplan-Meier estimate by considering all causes as one type of event.

From this formula we can take also proper estimations for the overall survival and cumulative functions since $\widehat{F(t)} = \sum_{i=1}^{n} \widehat{F_j}(t)$ and $\widehat{S(t)} = 1 - \widehat{F(t)}$.

Moreover the (2.3) formula uses the survival times up to t from all causes of death and a result it not possible to estimate the cumulative incidence for any cause by using cause-specific functions.

We remind to the reader that the cause specific functions use cause specific data(all causes are considered as censored observations except for the cause of interest).

Modelling in competing risks

The basic model in competing risks for the i-th patient and j-th cause is the following

$$h_{ij}(t) = \exp(\beta_j' \mathbf{x}_i) h_{0j}(t)\,, \tag{2.4}$$

where $\beta_j' = (\beta_{1j}, \ldots, \beta_{pj})$(coefficient vector) and $\mathbf{x}_i$ is the vector with the explanatory values.

When the baseline hazard is unspecified , we fit a Cox model and otherwise(parametric structure) we fit a parametric model.In either case the cause specific data are used as mentioned in the Kaplan Meier approach(1 for the specific cause and zero for the other causes).

The partial likelihood of (2.3) for all causes under the Cox model(compare with the (1.34)) is

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{m} \Big(\frac{\exp(\beta_{\mathbf{j}}{}' \mathbf{x_i})}{\sum_{l \in R(t_{(i)})} \exp(\beta_{\mathbf{j}}{}' \mathbf{x_l})} \Big)^{\delta_{ij}}\,.$$

We notice that this likelihood contains m usual Cox likelihoods,one for each cause and of course assumes that the causes of death are independet from one another.

At this point we have to mention also the case for a parametric baseline hazard in (2.3) equation.

The likelihood is given by the following equation(see also the ((1.4.1) paragraph)

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{m} h_j(t_i)^{\delta_{ij}} S_j(t_i)\,,$$

where $h_j(t)$ and $S_j(t)$ are the cause cause specific hazard and cause specific survival($S_j(t) = \exp(-H_j(t))$) respectively.

This likelihood also as in the Cox case contains m parametric likelihoods(compare with the equation above (1.55)),one for each cause.

Finally we have to mention that in the above parametric likelihood we can't substitute directly in the cause specific functions a known specific distribution form for the survival times of the j-th cause.

For example if the distribution of the survival times for the j-th cause has the Weibull distribution, we cannot adjust directly the Weibull distribution forms to the cause -specific functions, because the cause specific cumulative distribution function is not a proper distribution.

Lets say that the survival times for the j-th cause follows the exponential distribution with rate λ_j(the most simpe Weibull case).

Then we have that

$F_j(t) = P(T \leq t_i, C = j) = P(T \leq t_i | C = j)P(C = j)$.

As a result the cause specific cumulative distribution function has the form $F_j(t) = (1 - e^{\lambda_j t}) * p_j$.

Then it follows from (2.2) that the cause-specific hazard for the j-th cause is

$$h_j(t) = f_j(t)/S(t) = \frac{p_j \lambda_j e^{\lambda_j t}}{\sum_{j=1}^{m} \left(p_j e^{\lambda_j t} \right)} .$$

The last denominator comes from the fact that the S(t)(overall survivor function) equals to $1 - \sum_{i=1}^{m} F_j(t)$ and so

$$S(t) = 1 - \sum_{i=1}^{m} F_j(t) = 1 - \sum_{i=1}^{m} (1 - e^{\lambda_j t}) * p_j = 1 - \sum_{i=1}^{m} p_j + \sum_{j=1}^{m} \left(p_j e^{\lambda_j t} \right) = 1 - 1 + \sum_{j=1}^{m} \left(p_j e^{\lambda_j t} \right).$$

So it is not λ_j and thus the cause specific baseline hazard in (2.3) has this complicated form.

2.1.2 Fine and Gray model

In order to deal with the problem of the untestable assumption of independent censoring and to estimate the cause specific cumulative incidence $F_j(t)$ when explanatory variables arise as we mention before we can't use cause specific models.

These problems can be solved by using the Fine and Gray model(1999).

This model provides information on how the explanatory variables affect the cumulative incidence for each cause.

We will first start by explaning the key quantity for this concept which is reffered to as subhazard.

The subhazard function for the i-th cause is defined as(compare with the (1.5)

and the fact that the hazard is derivative of the cumulative hazard)

$$h_i^*(t) = -\frac{d}{dt}\log(1 - F_i(t)) \, .$$

So $h_i^*(t) = \frac{1}{1-F_i(t)}\frac{dF_i(t)}{dt}$ and the $1 - F_i(t)$ is the probality of surviving beyond time t or dying before time t from a cause different than i.

Of course in order to avoid misunderstandings we mention that $1 - F_i(t) \neq S_i(t)$ where $S_i(t)$ is the cause specific survival function for the i-th cause who previously explained.

At this point we further notice that the derivative $\frac{dF_i(t)}{dt} = \lim_{h\to\ 0}\frac{F_i(t+h)-F_i(t)}{h}$. The numerator in the limit reflects the probability of dying in an infestimal interval from cause j and survive beyond t in the precence of all causes or dying in an infestimal interval beyond t from cause j and also dying from some other cause before time t.

By combinig this interpretation with the interpretation of $1 - F_i(t)$ we get the following expression for the subhazard(i individual)

$$h_i^*(t) = \lim_{h\to\ 0}\frac{P(t \leq T \leq t+h, C = i | T \geq t \, or \{T \leq t \, and \, C \neq i\})}{h} \, .$$

As a result the (2.5) formula gives an awkward interpretation as the instantaneous death rate at time t from cause i, given that an individual has not previously died from cause i,meaning that it allows deaths from other causes before time t.

The basic step has been done.Now we consider a Cox regression model for the subhazard function that has the form

$$h_{ij}^*(t) = exp(\beta_j' \mathbf{x_i}) h_{0j}^*(t) \, ,$$

where j represents the j-th cause and i the i-th individual.Also this model is called Fine and Gray model.

The model is fitted in the usual way(1.3.5) but with a modification(we use weights).

For the j-th cause the fitted model is

$$\log(L(\beta)) = \sum_{i=1}^{n} \delta_i \left(\beta_j' \mathbf{x_i} - \log\Big(\sum_{l \in R(t_i)} w_{il} \exp(\beta_j' \mathbf{x_l}) \Big) \right) ,$$

where $R(ti)$ is the risk set of all individuals who have not experienced the j-th cause before the i-th event time t_i, for whom the survival time is greater than or equal to t_i and those who have experienced some other cause before t_i, for whom the survival time is less than or equal to t_i.
The weights have the following form

$$w_{il} = \frac{\widehat{S}_C(t_i)}{\widehat{S}_C\Big(min(t_i, t_l)\Big)} ,$$

where $\widehat{S}_C(t)$ is the Kaplan-Meier estimation for the censoring mechanism where the censoring times are considered as event times and the event times for all causes are considered as censored times.
We further notice that when $t_i < t_l$, $w_{il} = 1$.This occurs for the individuals in the risk set who haven't experience the j-th cause before t_i.
Also when $t_i > t_l$ an individual in the risk set experience some other cause and the weight in that case is $w_{il} < 1$,because the survivor function is a decreasing function of time.
As a result the main goal of using these weights is to give a minor role in deaths from other causes,The main goal of using these weights is to give a minor role in deaths from other causes,because the weights become smaller with increasing time between the occurrence of a competing risk and the event of interest , so that earlier deaths from a competing risk have a small impact on the results.

2.2 The general case, multiple events framework

In the introduction of this chapter were mentioned several cases of multiple events situations such as multistate models,repetition of an event of the same type,or occurance of multiple events with different types.
In order to use models for these situations we have to define the $\{N_i(t)\}_{t \geq 0}$ stochastic process for the i-th individual.

This process is reffered to as counting process and counts the number of occurances of some event over the time period $(0, t]$.
The graph of this process starts from zero and is a step function that increases a step of 1 unit if an event is occured.
Another useful stochastic process is the at risk process $\{Y_i(t)\}_{t\geq 0}$ that represents the process which is 1 when the i-th individual is uncensored and at risk of an event occurring at time t and zero otherwise.
Finally it is also useful the concept of history.
The history or filtration up to t $H(t-)$ is defined as the set of values $(N_i(u), Y_i(u))$ where $u < t$.
Every counting process $N_i(t)$ has an associated intesity function $\lambda_i(t)$,$t \geq 0$.
The intensity function for the i-th individual $\lambda_i(t)$ is defined as

$$\lambda_i(t) = \frac{P\Big(dN_i(t) = 1|H(t-)\Big)}{dt},$$

where $H(t-)$ is the history or filtration of the process up to but not including time t and $dN_i(t) = N_i(t + dt) - N_i(t)$.
For the reason that $dN_i(t)$ takes the value 1 when an event is occured in an infestimal interval(zero otherwise), the probability in the numerator reflects the probability of an event in an infestimal interval given the history of the process.

From the definition of the intensity function, it comes up that in the usual survival framework where only one type of event occurs(e.g death(one time) or independent headaches(multiple times)),the intensity function for the i-th individual who is at risk of an event at time t is equal to the usual hazard function(1.2)
Indeed if we imagine the usual survival framework as a counting process,the i-th individual at time zero has zero chance of experience the event of interest ,so $N_i(0) = 0$.
Moreover the counting process is increasing by one unit when the event is occured and the history up to t is $T_i \geq t$,because the event that occurs in an infestimal interval above time t only depends on the event of being at risk at time t(this is true also for independent reccurent events).

As a result we get the equation

$$\lambda_i(t) = Y_i(t)h_i(t)\,,$$

in which, as long as the i-th individual is at risk of an event or censoring at time t, the intensity function is the hazard for the i-th individual.
Also when an event or censoring occurs before time t the intensity function and the hazard are zero so they can be thought as the same thing.
We can further define the cumulative intensity function such as with the usual hazard , as the integral of the intensity function over the interval [0,t]

$$\Lambda_i(t) = \int_0^t \lambda_i(u)du\,.$$

Models with intensity functions

A general model for the intesity function of the i-th individual is the following one

$$\lambda_i(t) = Y_i(t)f\Big(t, \mathbf{x_i}(t)\Big)\,. \tag{2.5}$$

where $f\Big(t, \mathbf{x_i}(t)\Big)$ is some function of time t and of time dependent covariates for the i-th individual.Moreover the $Y_i(t)$ is the at risk process.
A model for reccurent data(repeated events) that takes the Cox approach(1.28) for the f function in (2.5) is called Anderson and Gill model(1982).
The model has the following form for the i-th individual

$$\lambda_i(t) = Y_i(t)\exp(\beta_i'\mathbf{x_i}(t))\lambda_0(t)\,.$$

In this model the within reccurent times for each individual are assumed independet,so the history of the process for the i-th individual is $T_i \geq t$,thus the hazard functions $h_i(t)$,$h_0(t)$ and the intensity functions $\lambda_i(t)$,$\lambda_0(t)$ can be thought of as the same thing.
Under this model we construct the partial likelihood for the counting process format in a similar manner as mentioned in (1.3.3) section.

The form of the partial likelihood from a realisation of the counting process is

$$L(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left(\frac{Y_i(t) \exp(\beta' \mathbf{x_i}(t))}{\sum_{l=1}^{n} Y_l(t) \exp(\beta' \mathbf{x_l}(t))} \right)^{dN_i(t)} .$$

We notice that the fraction inside the product contributes in the likelihood when an event is about to occur for the i-th individual in an infestimal interval after t.

The numerator in this case is $\exp(\beta' \mathbf{x_i}(t))$ and the denominator is the sum $\exp(\beta' \mathbf{x_l}(t))$ for all $l \neq i$ individuals who don't experience an event or cencoring before time t plus the term $\exp(\beta' \mathbf{x_l}(t))$.

In statistical packages the Anderson and Gill model is fitted by using the counting process format(1.6.2) in a more general manner(repeated events).

We construct intervals starting from zero.

Each interval is associated with a status variable which is 1 for an event and zero for censoring.

In order to take into account a possible dependance between repeated events(within subject dependence) a term r_i is used that is reffered to as random effect.

The random effect r_i is not a constant,in fact is considered as the observed value or realization of n(the number of individuals in the study) independent and identically random variables from the normal distribution with mean zero and variance σ^2.

The Anderson and Gill model with the random effect equipped takes the following form

$$\lambda_i(t) = Y_i(t) \exp(\beta_i' \mathbf{x_i}(t) + r_i) \lambda_0(t) .$$

By fitting this model we will get an estimation for the variance σ^2 that summarize the extent of differences in random effects r_i.

For further details about random effects the reader can see [1].

An extension of the Anderson and Gill model for recurrent events is the Prentice, Williams and Peterson (1981) model.

This model for the i-th patient and j-th occurance has the following form

$$\lambda_{ij}(t) = Y_{ij}(t) \exp(\beta_j' \mathbf{x_i}(t)) \lambda_{0j}(t) \ ,$$

where $Y_{ij}(t) = 1$ for the i-th individual who is uncensored and at risk of the j-th occurance at time t and zero otherwise.
This model allows the within dependence of repeated events,since the intensity function can vary between the repeated occurances for the i-th individual and it is the most preffered model for these situations.
In order to fit this model in a statistical software we use the start-stop format with stratums for each repeated occurance of the event of interest.
It is worth to mention that the same model can be used to fit different types of events for an individual,so $Y_{ij}(t) = 1$ for the i-th individual who is uncensored and at risk of the j-th type event at time t and zero otherwise.
Such models are reffered to as Wei, Lin and Weissfeld models.

We close this chapter by mentioning that in a multistate model such as the transplant example in the introduction of this chapter, we fit the following model

$$\lambda_{ijk}(t) = Y_{ijk}(t) \exp(\beta_{jk}' \mathbf{x_i}(t)) \lambda_{0jk}(t) \ ,$$

where i denotes the i-th individual,j denotes the j-th state and k the k-th state. As in the models that were mentioned above the start-stop format is used for each indvidual but now according to the transitions from one state to another.

Chapter 3

Informative censoring

The methods that were described in the previously chapters are only valid if the censoring is independent.
This means that in the precense of dependent censoring, by fitting models under independent censoring ,is resulting to biased inference .
Informative censoring occurs when there is a dependence between the time of an event such as death and the time of the occurance of censoring.
The problem with this situation is that it is not possible to use the observed data to determine the existance or absence of the dependent censoring.
In general the best thing we can do is a sensitivity analysis, meaning to examine if the casual models with independent assumptions provide biases that can be ignored.
The contex of the study usually gives some indication about the presence of informative cencoring.For example a patient who leaves the study due to some therapy that became life - threatening,so we expect early dropouts.

One way to examine the informative assumption is to plot the observed survival times against the values of each explanatory variable and if there is a greater proportion of censoring in some range of values of an explanatory variable,then maybe informative censoring exists.

One approach in a sensitivity analysis is to consider 2 supplementary analy-

sis.

First we consider that the individuals with censored survival times are dying immediately after the censoring time,so all censored times are transformed to event times.These are the 'high-risk' individuals.

Then we consider the 'low-risk' individuals, meaning all censored times are replaced by the longest survival times.

If no differences are observed by comparing the standar analysis with the other 2 analysis, then we can say that the results are not sensitive in the precense of informative censoring.

Another method in sensitivity is reffered to as the 'Siannis' method (see [5] and [7]) and is presented in the following section.

3.1 Parametric models in sensitivity analysis

In this section we will present parametric models for sensitivity analysis that are described by Siannis(2004)[5] and Siannis,Copas and Lu (2005)[7] .

Suppose that T and C are random variables that represent the survival and cencored times respectively.

Then we suppose that $\mathbf{x_i}$ is the explanatory vector of the i-th individual.

Under the proportional assumption the hazard function of T is $h_T(t, \theta, \mathbf{x_i}) = \exp(\theta' \mathbf{x_i}) h_{T0}(t)$ and the hazard function of C is $h_C(t, \gamma, \mathbf{x_i}) = exp(\gamma' \mathbf{x_i}) h_{C0}(c)$,where $\theta' \mathbf{x_i}$ and $\gamma' \mathbf{x_i}$ are the risk score and censoring score respectively.

A plot of the values of the estimated risk score against the values of the estimated censoring score can provide a helpful identification of the existance or absence of the informative censoring.

The baseline functions $h_{T0}(t)$ and $h_{C0}(t)$ have a specific parametric distributional form.

Further we define a parameter δ that represents the level of dependence between the event and cencoring processes.

In order to fit a model that takes into account the informative censoring, we must first present a previous analysis on this subject.

Also in the following discussion is presented a quantity that is called corellation bias, which is essentially the effect in the the precense of informative censoring

and provides the essentials for assessing the goodness of fit of the independent model(non-informative censoring).

Fitting the model

The 'Siannis' method considers that

$$P(C = c|T = t) = f_C\Big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t, \theta)\Big)\,, \tag{3.1}$$

,where θ and γ are uknown scalar parameters,but in general can be vectors.
Also θ corresponds with the function of T , γ with the function of C, i_γ is the information function that equals to the variance of the usual score function $\frac{\partial}{\partial\gamma}\log f_C(t,\gamma)$ and $B(t,\theta)$ is the bias function.
The bias function is depending only on the event process of T and is providing the means to insert T in the (3.1) expression.
The choice of $B(t,\theta)$ depends on the researcher.
Finally we mention that when $\delta = 0$ T and C are independent and the censoring is ignorable.
The joint function of T and C is

$$f_{T,C}(t,c) = P(T=t)P(C=c|T=t) = f_T(t) f_C\Big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\Big)\,. \tag{3.2}$$

As in (1.4.1) paragraph in order to simplify the process we consider the $P(T=t)$ as the marginal density of T.

At this point a first order approximation(multivariate aspect) around (c,γ) for the $f_C\Big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\Big)$ in (3.1) gives that

$$f_C\Big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\Big) \approx f_C(c,\gamma) + \Big(\nabla f_C(c,\gamma)\Big)'(\mathbf{x} - \mathbf{x_0})\,,$$

where $\Big(\nabla f_C(c,\gamma)\Big)' = \Big(\frac{\partial f_C(c,\gamma)}{\partial c}, \frac{\partial f_C(c,\gamma)}{\partial\gamma}\Big)$,$\mathbf{x} = \Big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\Big)$ and $\mathbf{x_0} = (c,\gamma)$.
As a result the (3.1) expression is almost equal to

$$f_C(c,\gamma) + (c-c)\frac{\partial f_C(c,\gamma)}{\partial c} + (\gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta) - \gamma)\frac{\partial f_C(c,\gamma)}{\partial\gamma}\,.$$

This leads to the quantity $f_C(c,\gamma) + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\frac{\partial f_C(c,\gamma)}{\partial \gamma}$.

Further we notice that $\frac{\partial f_C(c,\gamma)}{\partial \gamma} = f_C(c,\gamma)\frac{\partial \log f_C(c,\gamma)}{\partial \gamma}$ and because $U_C(c,\gamma) = \frac{\partial \log f_C(c,\gamma)}{\partial \gamma}$ is the score function of C, we get the following approach for the (3.1) formula

$$f_C\Big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\Big) \approx f_C(c,\gamma) + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta) f_C(c,\gamma) U_C(c,\gamma)\,.$$

This approximation is used to (3.2) and leads to the following expression

$$f_{T,C}(t,c) \approx f_T(t,\theta) f_C(c,\gamma)\Big(1 + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta) U_C(c,\gamma)\Big)\,. \tag{3.3}$$

An integration over T gives the marginal distribution of C that leads to a natural constraint of the bias function.
More specifically

$$f_C(c,\gamma) = \int_0^\infty \Big(f_T(t,\theta) f_C(c,\gamma)\Big(1 + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta) U_C(c,\gamma)\Big)\Big) dt\,.$$

So the marginal distribution of C is

$$f_C(c,\gamma) = f_C(c,\gamma)\Big(\int_0^\infty f_T(t,\theta)dt + \delta i_\gamma^{-\frac{1}{2}} U_C(c,\gamma)\int_0^\infty B(t,\theta) f_T(t,\theta) dt\Big)\,.$$

The first improper integral equals to 1 because $f_T(t,\theta)$ is the probability density function of T.
As a result the integral $\int_0^\infty B(t,\theta) f_T(t,\theta) dt$ must be zero ,meaning that the mean of $B(t,\theta)$ must be zero.
Also without loss of generality the variance of $B(t,\theta)$ is assumed to be 1.

Now with the same logic as in (1.4.1) section the log-likelihood is

$$L_\delta(\theta,\gamma) = \sum_{i=1}^n \Big(I_i \log P\Big(T = t_i, I_i = 1\Big) + (1 - I_i)\log P\Big(C = t_i, I_i = 0\Big)\Big)\,,$$

where $t_i = min(T,C)$ and I_i is 1 when an event occurs for the i-th individual and zero otherwise.

This expression is combined with the (3.3) approximation and leads to the log - likelihood in the presence of informative censoring.
More specifically,for instance the probability

$$P(C = t_i, I_i = 0) = P(C = t_i, T > t_i) = \int_{t_i}^{\infty} P(C = t_i, T = u)du \,,$$

and

$$\log\left(1 + \delta i_{\gamma}^{-\frac{1}{2}} \mu(t_i, \theta) U_C(t_i, \gamma)\right) \approx \delta i_{\gamma}^{-\frac{1}{2}} \mu(t_i, \theta) U_C(t_i, \gamma) \,.$$

Next by straightforward algebra and the usage of (3.3) the log -likelihood takes the form

$$L_{\delta}(\theta, \gamma) \approx L_0(\theta, \gamma) + \delta i_{\gamma}^{-\frac{1}{2}} \sum_{i=1}^{n} \left((1 - I_i)\mu(t_i, \theta) U_C(t_i, \gamma) - I_i B(t_i, \theta) \frac{\partial H_C(t_i, \gamma)}{\partial \gamma} \right), \tag{3.4}$$

where

$$L_0(\theta, \gamma) = \sum_{i=1}^{n} \left(I_i \log h_T(t_i, \theta) + (1 - I_i) \log h_C(t_i, \gamma) - H_T(t_i, \theta) - H_C(t_i, \gamma) \right),$$

is the log-likelihood under the independent censoring assumption,

$$\mu(t_i, \theta) = \frac{\int_{t_i}^{\infty} B(u, \theta) f_T(u, \theta) du}{S_T(t_i, \theta)} \,,$$

and H ,h represents the usual cumulative hazard and hazard for T and C.
In general δ is taken to be small with values around $(-0.3, 0.3)$.
The estimates for the location parameters θ and γ can be derived either from the model with the independent censoring L_0 or from the model with informative censoring L_{δ},so for convinience we assumme that the estimations are taken from the L_0 model which is equivelant with the (1.55) formula.

Correlation bias

By differentiating the (3.4) expression with respect to θ we can take a very useful quantity ,that is mentioned as correlation bias for θ.

$$\widehat{\theta_\delta}-\widehat{\theta_0} \approx \delta i_\gamma^{-\frac{1}{2}}(\phi(\theta))^{-1}\sum_{i=1}^{n}\left((1-I_i)\frac{\partial\mu(t_i,\theta)}{\partial\theta}U_C(t_i,\gamma)-I_i\frac{\partial B(t_i,\theta)}{\partial\theta}\frac{\partial H_C(t_i,\gamma)}{\partial\gamma}\right), \tag{3.5}$$

where $\phi(\theta)^{-1} = \left(-\frac{\partial^2 L_0(\theta,\gamma)}{\partial\theta^2}\right)^{-1} \approx Var(\widehat{\theta_\delta}) \approx Var(\widehat{\theta_0})$ is the observed information, $\widehat{\theta_\delta}$ is the maximum likelihood estimation for the model under the informative censoring and $\widehat{\theta_0}$ is the maximum likelihood estimation under the usual independent model.

More specifically in order to prove the (3.5) expression we take the following two equations by the definition of the maximum likelihood estimator

$$\frac{\partial L_\delta}{\partial\theta}\Big|_{\widehat{\theta_\delta}} = 0 \, and \, \frac{\partial L_0}{\partial\theta}\Big|_{\widehat{\theta_0}} = 0\,.$$

Then a linearization of the derivative of the likelihood at $\widehat{\theta}$ around the true value, θ gives the following equations

$$\frac{\partial L_\delta}{\partial\theta}\Big|_\theta + \frac{\partial^2 L_\delta}{\partial\theta^2}\Big|_\theta(\widehat{\theta_\delta}-\theta) = 0\,,$$

and

$$\frac{\partial L_0}{\partial\theta}\Big|_\theta + \frac{\partial^2 L_0}{\partial\theta^2}\Big|_\theta(\widehat{\theta_0}-\theta) = 0\,.$$

At this point by taking the approximation $\frac{\partial^2 L_\delta}{\partial\theta^2}\Big|_\theta \approx \frac{\partial^2 L_0}{\partial\theta^2}\Big|_\theta$ and by considering that

$$\frac{\partial L_\delta}{\partial\theta}\Big|_\theta + \frac{\partial^2 L_0}{\partial\theta^2}\Big|_\theta(\widehat{\theta_\delta}-\theta) = \frac{\partial L_0}{\partial\theta}\Big|_\theta + \frac{\partial^2 L_0}{\partial\theta^2}\Big|_\theta(\widehat{\theta_0}-\theta)\,,$$

we take the (3.5) expression.

The same logic can be used to find the corellation bias for γ.

By plotting correlation bias against the values of δ gives two possible outcomes. In the first scenario the values of bias for given values of δ might be large,so the independent model is not robust on small changes of δ.

On the other hand if the values of the corellation bias are small for small changes

of δ , the independent model is robust.

Also δ can be regarded as the maximum possible correlation between the event and censoring processes,meanining that $correlation(T, C) \leq |\delta|$.

In general a sensitivity analysis can be performed on any of the parameters θ and γ or to both of them at the same time ,but we focus only on the first case .

Proportional parametric models

At this point we focus specifically on the proportional assumption.

First we suppose that we have the model in the beginning of section (3.1) for T and C, but in the most simple case ,meaning that θ and γ are not vectors of covariates and x(explanatory vector) doen't exist.

Next we observe that under this proportional model $U_T(t, \theta) = 1 - H_T(t, \theta)$,$U_C(c, \gamma) = 1 - H_C(c, \gamma)$,$i_\theta = 1$,$i_\gamma = 1$.

Also we consider the choice of $B(t, \theta) = i_\theta^{-1/2} U_T(t, \theta)$ because this form achieves symmetry in (3.3) expression.

These observations lead to the following formulas, $B(t, \theta) = 1 - H(t, \theta)$ and $\mu(t, \theta) = -H(t, \theta)$

In order to explain these expressions let's prove the first equation

$$U_T(t, \theta) = \frac{\partial}{\partial \theta} \log f_T(t, \theta) = 1 - H_T(t, \theta) \, .$$

The other expressions are calculated in a similar manner.

Firstly according to the (1.1) the density function is written as $f_T(t, \theta) = -\frac{\partial S_T(t,\theta)}{\partial t}$.

Also from (1.6), $S_T(t, \theta) = \exp(-H_T(t, \theta)) = \exp(-e^\theta H_{T0}(t))$

The derivative of the survivor function of T with respect to t , is

$$\frac{\partial}{\partial \theta} S_T(t, \theta) = -e^\theta H_{T0}(t) \exp(-e^\theta H_{T0}(t)) \, .$$

The logarithm of $-\frac{\partial}{\partial \theta} S_T(t, \theta)$ is then $\theta + H_0(t) - e^\theta H_0(t)$.

Thus the derivative of this quantity with respect to θ , is $U_T(t, \theta) = 1 - H_{T0}(t) e^\theta = 1 - H_T(t, \theta)$.

Now by subtitution of the above formulas to (3.3) and (3.4) expression we get

the following symmetric forms

$$f_{T,C}(t,c) \approx f_T(t,\theta) f_C(c,\gamma)\Big(1+\delta\Big(1-H_C(c,\gamma)\Big)\Big(1-H_T(t,\theta)\Big)\Big)\,,$$

$$L_\delta(\theta,\gamma) \approx L_0(\theta,\gamma)+\delta\sum_{i=1}^{n}\Big(H_T(t_i,\theta)H_C(t_i,\gamma)-I_iH_C(t_i,\gamma)-(1-I_i)H_T(t_i,\theta)\Big)\,, \tag{3.6}$$

and the correlation bias for θ is

$$\widehat{\theta_\delta}-\widehat{\theta_0} \approx \delta(i(\theta))^{-1}\sum_{i=1}^{n}\Big(H_T(t_i,\theta)H_C(t_i,\gamma)-(1-I_i)H_T(t_i,\theta)\Big)\,. \tag{3.7}$$

In order to fit the proportional hazard model for T and C in the beginning of (3.1),we just consider that the (3.6) log-likelihood involves also explanatory vectors $\mathbf{x_i}$ for each individual and the parameters θ and γ are vectors.
Also we mention that inferences about δ cannot be drawn,meaning that a plot of the likelihood against the values of δ gives a fairly flat curve.

Confidence intervals

In general the correlation bias can be written as

$$\widehat{\theta_\delta}-\widehat{\theta_0} \approx \delta K\,,$$

where K is called sensitivity index and can be calculated from the independent model.
This form provides an asymptotic confidence interval of θ for small values of δ with the following form

$$[\widehat{\theta_0}-\delta K-z_{a/2}\phi(\theta)^{-\frac{1}{2}},\widehat{\theta_0}-\delta K+z_{a/2}\phi(\theta)^{-\frac{1}{2}}]\,,$$

where $z_{a/2}$ is the usual upper a/2 quantile for the standar normal distribution and $\widehat{\theta_0}$ is the maximul likelihood estimation for θ under the usual independent model.
In a more general manner we can construct confidence intervals for a function of interest G(θ).

More specifically by using a fist order approxiamtion we take

$$G(\widehat{\theta_\delta}) - G(\widehat{\theta_0}) \approx \delta K G'(\widehat{\theta_0}) \, .$$

As a result a similar confidence interval as the previous one, but for G(θ) has the following form

$$[G(\widehat{\theta_0}) - \delta K G'(\widehat{\theta_0}) - z_{a/2}\phi(\theta)^{-\frac{1}{2}}, G(\widehat{\theta_0}) - \delta K G'(\widehat{\theta_0}) + z_{a/2}\phi(\theta)^{-\frac{1}{2}}] \, .$$

Weibull model

At this point let's consider the proportional Weibull model(1.4.1) for the two processes of T and C.
Then the hazards functions have the following forms for the i-th individual $h_T(t, \theta, \mathbf{x_i}) = \exp(\theta' \mathbf{x_i}) h_{T0}(t)$ and $h_C(t, \gamma, \mathbf{x_i}) = \exp(\gamma' \mathbf{x_i}) h_{C0}(t)$.
Also $h_{T0}(t) = \lambda \alpha t^{\alpha - 1}$ and $h_{C0}(t) = \lambda_c \alpha_c t^{\alpha_c - 1}$ represent the baseline hazards for T and C respectively.

An approximation to the change in the risk score for the i-th individual(correlation bias) $\theta' \mathbf{x_i}$ in the presence of a small amount of dependence δ is given by using (3.7)

$$W(\mathbf{x_i}) = \delta \frac{\sum_{j=1}^{n} \left(\exp(\widehat{\gamma}' \mathbf{x_i}) t_j^{\hat{\alpha} + \widehat{\alpha_c}} - (1 - I_j) t_j^{\hat{\alpha}} \right)}{\sum_{j=1}^{n} t_j^{\hat{\alpha}}} \, , \tag{3.8}$$

where the t_j represents the i-th cencored or event time and the I_j the event indicator.
For a given value of δ a plot for the values of (3.8) against a range for the values of the estimated risk score can provide information about the sensitivity of the risk score ,meaning that this plot will indicate the values of the censoring score that may result to non-negligible dependent censoring impact on the risk score.

The formula (3.8) provides also the means to examine other quantities of interest such as the survivor function or the median .
Under the informative censoring the survivor function of T for the i-th individual

takes the form

$$S_i(t) = S_0(t)^{\exp(\theta' \mathbf{x_i} + W(\mathbf{x_i}))} .$$

This is the (1.41) form,but in this equation we add also the correlation bias in the risk factor.
Thus for small values of δ we can calculate the estimated values of this equation and observe the possible changes under the informative censoring.

Moreover the median of the proportional Weibull model for T under the independent model can be found by the equation(see 1.2.7) $S_T(t_{50}) = \frac{100-50}{100}$.
More specifically(see 1.4.1),for the i-th individual we get

$$S_T(t_{50}) = \exp\Big(- \lambda t_{50}^{\alpha} \exp(\theta' \mathbf{x_i})\Big) = \frac{1}{2}$$

As a result the median has the following form

$$t_{50} = \Big(\frac{\log 2}{\lambda \exp(\theta' \mathbf{x_i})}\Big)^{\frac{1}{\alpha}} .$$

So under the dependent cencoring we add in the risk score the correlation bias and take

$$t'_{50} = \Big(\frac{\log 2}{\lambda \exp(\theta' \mathbf{x_i} + W(\mathbf{x_i}))}\Big)^{\frac{1}{\alpha}} .$$

The estimated values of the quantity

$$\frac{t_{50} - t'_{50}}{t_{50}} = 1 - \exp\Big(- \frac{W(\mathbf{x_i})}{\alpha}\Big) ,$$

gives the relative reduction of the median for the i-th individual and for some value of δ.
Also a plot against the censoring scores $\gamma' \mathbf{x_i}$ provide the possible changes in the median for each individual under the precense of the informative censoring.

An example

As it mentioned before, by using an independent model when informative censoring exists is resulting to biases.
So the estimates under the independent model overestimate or underestimate

the survivor function.
A positive association between C and T, means that an individual with censored event time is expected to live shorter than those who remain at risk(e.g patients who experience life -threatening therapy).
On the other hand negative association, means that the individuals with censored event times may be those who would otherwise had a longer time before the occurrence of the event of interest.(e.g after a specific intervention,a patient is cured and decides that is not necessary to be in the study anymore).

At this point let's consider a study that interests to determine the mortality rate for patients registered for a liver transplant.
The data were obtained from the UK Transplant Registry on the time from registration to death on the list.
The study contains 281 patients with primary biliary cholangitis (often referred to as primary biliary cirrhosis).
This is a type of a liver disease that can get gradually worse over time and without treatment, it may eventually lead to liver failure.
The patients were first registered for a liver transplant in the five-year period from 1 January 2006.
Status variable is unity for a patient who has died while waiting for a transplant and zero when their time from listing has been censored(the patients who receive a transplant also have censored times).
Furthermore, BMI is the body mass index of each patient(kg/m^2) , UKLED is a disease score, where bigger values indicate greater need for a transplant.Finally the gender factor is 1 for males and zero for females.
The table for the first 34 patients is shown in figure (3.1.1). In order to present a sensitivity analysis first we must find the estimations under the independent assumption.
The estimations for the parameters of the distribution of T can be found from (1.55).
Similarly if we focus on the distribution of C,meaning the time to censor-

Figure 3.1.1: Transplant table

patient	time	status	age	gender	BMI	UKELD
1	1	0	60	0	24.24	60
2	2	1	66	0	30.53	67
3	3	0	71	0	26.56	61
4	3	0	65	1	23.15	63
5	3	0	62	0	22.55	64
6	4	1	56	0	36.39	73
7	5	0	52	0	24.77	57
8	5	0	65	0	33.87	49
9	5	1	58	0	27.55	75
10	5	1	57	0	22.10	64
11	6	0	62	0	21.60	55
12	7	0	56	0	25.69	66
13	7	0	52	0	32.39	59
14	8	1	45	0	28.98	66
15	9	0	50	0	31.67	60
16	9	0	65	0	24.67	57
17	9	0	44	0	24.34	64
18	10	0	67	0	22.65	61
19	12	0	67	0	26.18	57
20	13	0	57	0	22.23	53
21	14	0	38	0	17.03	60
22	14	0	66	0	21.51	55
23	15	1	63	0	24.32	61
24	16	1	52	0	34.13	62
25	16	0	57	0	44.14	54
26	16	0	58	1	27.06	56
27	17	1	56	0	31.63	60
28	17	0	67	0	30.36	58
29	18	1	63	0	22.77	62
30	18	1	62	0	24.91	67
31	20	0	67	0	22.21	54
32	21	0	45	1	24.30	49
33	21	0	61	0	22.79	60
34	23	0	61	1	21.30	66

ing(censoring as event and vice versa) , we get that

$$L(\theta) = \prod_{i=1}^{n} \left(S_{C_i}(t_i)^{\delta i} f_{C_i}(t_i)^{1-\delta_i} \right)$$

A log-cumulative hazard plot(without prognostic variables) in figure (3.1.2) as mentioned in (1.4.2) with thr log(-log) Kaplan-Meier against logt gives approximately a straight line As a result the Weibull distribution seems to fit well with the data and we will use the Weibull proportional model and the following correlation bias(3.8) from the previous paragraph.

The two extreme values in figures (3.1.2) with logH 10 and -10, in fact are negative and positive infinity for the values one and zero of the survivor function respectively.

Estimates for T and C mechanisms under the independened assumption can be provided by the next 2 outputs using R in figures (3.1.3) and (3.1.4). These estimations consider the log-linear model(1.59) ,so in order to take the θ's γ's and the baseline parameters we have to reparametrize the estimations for the prognostic effects as in section (1.5.2).

Moreover the p-values(Wald tests) for gender and BMI in figure (3.1.3) indicate that may be exluded from the model.In fact the backward algorithm in figure (3.1.5) suggest exactly that. Despite that fact we keep all the prognostict variables ,so the model may be overfitted.

The reason for keeping all explanatory variables is that with this example we

Figure 3.1.2: Log-cumulative hazard plot

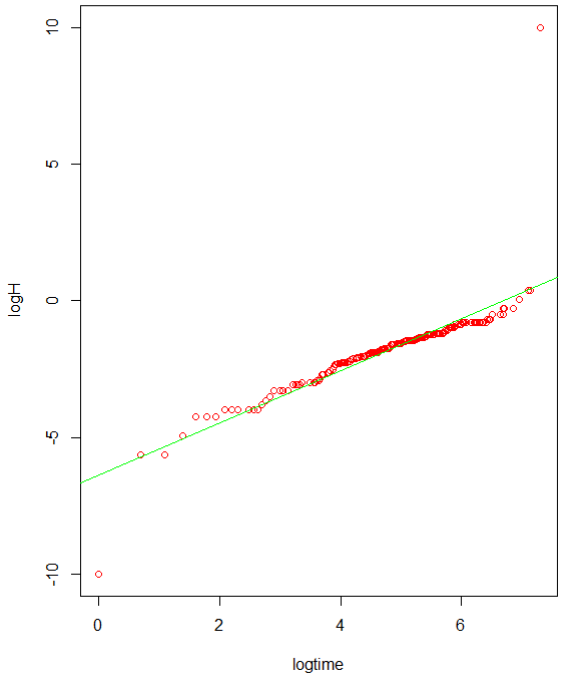


Figure 3.1.3: Estimations for the model under T

```
                Value Std. Error     z       p
(Intercept) 24.1275     2.2907 10.53 < 2e-16
age         -0.0811     0.0206 -3.94 8.3e-05
gender       0.1750     0.4748  0.37    0.71
BMI         -0.0222     0.0230 -0.97    0.33
UKELD       -0.2157     0.0271 -7.95 1.8e-15
Log(scale)  -0.0154     0.0900 -0.17    0.86

Scale= 0.985

Weibull distribution
Loglik(model)= -451.6   Loglik(intercept only)= -495.6
        Chisq= 87.93 on 4 degrees of freedom, p= 3.6e-18
Number of Newton-Raphson Iterations: 7
n= 281
```

focus mainly on a sensitivity analysis and not on a full analysis for picking the 'best' model.

A plot of the estimated risk scores against the estimated censoring scores in figure (3.1.6) shows that there is a positive corellation between the failure and cencoring mechanisms. The figure (3.1.6) is an indication of informative censoring and tells us that the patients who are in danger of death,are those with higher censoring.

In this study censoring times involves the patiens who received a transplant and as a result the patients who tend to die sooner than others receive the transplant

Figure 3.1.4: Estimations for the model under C

```
              value Std. Error     z     p
(Intercept)  7.62606    1.00576  7.58 3.4e-14
age          0.00699    0.00788  0.89 0.37477
gender      -0.18392    0.19501 -0.94 0.34561
BMI          0.01926    0.01407  1.37 0.17095
UKELD       -0.05348    0.01515 -3.53 0.00042
Log(scale)  -0.01885    0.05166 -0.36 0.71523

Scale= 0.981

Weibull distribution
Loglik(model)= -1414.6   Loglik(intercept only)= -1422
        Chisq= 14.93 on 4 degrees of freedom, p= 0.0049
Number of Newton-Raphson Iterations: 5
n= 281
```

Figure 3.1.5: Backward algorithm for T

```
Start:  AIC=915.21
Surv(time, status) ~ age + gender + BMI + UKELD

         Df    AIC
- gender  1 913.36
- BMI     1 914.14
<none>      915.21
- age     1 933.26
- UKELD   1 975.72

Step:  AIC=913.36
Surv(time, status) ~ age + BMI + UKELD

        Df    AIC
- BMI    1 912.32
<none>      913.36
- age    1 931.26
- UKELD  1 976.96

Step:  AIC=912.32
Surv(time, status) ~ age + UKELD

        Df    AIC
<none>      912.32
- age    1 930.32
- UKELD  1 977.46
Call:
survreg(formula = Surv(time, status) ~ age + UKELD, data = transplant0,
    dist = "weibull")

Coefficients:
(Intercept)         age       UKELD
23.68560577 -0.07887966 -0.21994495

Scale= 0.9756679

Loglik(model)= -452.2   Loglik(intercept only)= -495.6
        Chisq= 86.82 on 2 degrees of freedom, p= <2e-16
n= 281
```

first.

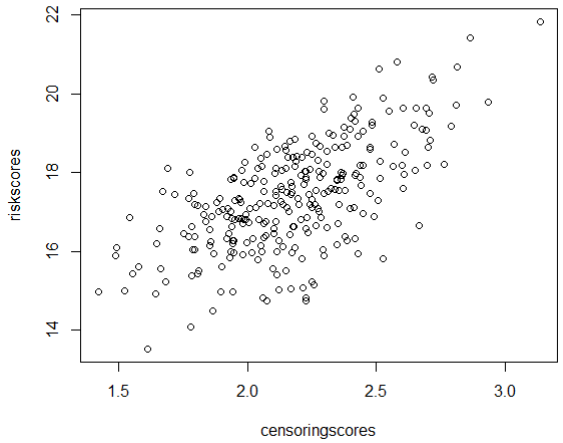
3.2 Semi-parametric model for dependent censoring

In this section we will discuss about a Cox type model that takes into account the informative cencoring.

First of all let's consider the Anderson and Gill model(just as an illustration for the method),as discussed in the paragraph(models with intensity functions),but also we can consider any kind of Cox type model that is anticipated for usage (e.g the usual Cox model, the (1.63) model etc).

In order to take into account the dependent censoring under the Anderson and Gill model we can use some special weights for each individual that we will

Figure 3.1.6: risk scores vs censoring scores



explain,but first we present the likelihood under this model.

$$L(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left(\frac{w_i(t) Y_i(t) \exp(\beta' \mathbf{x_i}(t))}{\sum_{l=1}^{n} w_l(t) Y_l(t) \exp(\beta' \mathbf{x_l}(t))} \right)^{dN_i(t)} . \tag{3.9}$$

As we see the only extra factor that seperates this likelihood from the Anderson and Gill likelihood is the weight $w_i(t)$ for each individual.

We notice also that the weight is considered as time-varying variable and as a result it may change over the follow up time period,so the start-stop format must be used in order to fit the model in a statistical package.

Moreover the weights are calculated from the estimated survivor functions of the censoring process(event as cencored times and vice-versa) under the independent assumption and these estimates can be derived from a model of our choice,but the log-linear Weibull model(1.60) is suggested for general use.

More specifically the survivor function of this model(1.60) for the censoring process and for the i-th individual is

$$S_{Ci}(t) = exp\Big(- exp\Big(logt - log\mu_C - \omega_C' \mathbf{x_i} \Big) / \sigma_C \Big) . \tag{3.10}$$

Then the weights are calculated as the inverse of (3.10) expression for each individual.

This method is called Inverse Probability of Censoring Weighted(IPCW).
The logic behind this idea is explained with an example.
Suppose that an individual has 1/4 probabibility of censoring at or after some event time t,then on average three other individuals with the same explanatory variables(identical individuals) will have censored survival times before t.
Consequently if these three individuals had not been censored before time t ,then their contribution to the partial likelihood would be the same as the first individual.
Thus we give a contibution(weight) 4 to the individual in the start of this conversation at time t (event time).
We notice also that greater probabilities of censoring for an individual before time t,give greater weights and vice-versa.
As a result by using this technique we try to include in the chosen model the dependance between the survival and censored times.

As an illustration let's consider the transplant example from previous paragraph.
By using R in figure(3.2.1) the Weibull model for the censoring mechanism gives the estimated coefficients to calculate the (3.10) expression Also we provide the

Figure 3.2.1: Weibull model for C

```
               Value Std. Error     z     p
(Intercept)  7.62606    1.00576  7.58 3.4e-14
age          0.00699    0.00788  0.89 0.37477
gender      -0.18392    0.19501 -0.94 0.34561
BMI          0.01926    0.01407  1.37 0.17095
UKELD       -0.05348    0.01515 -3.53 0.00042
Log(scale)  -0.01885    0.05166 -0.36 0.71523

Scale= 0.981

Weibull distribution
Loglik(model)= -1414.6   Loglik(intercept only)= -1422
        Chisq= 14.93 on 4 degrees of freedom, p= 0.0049
Number of Newton-Raphson Iterations: 5
n= 281
```

start - stop format for the first 6 patients in figure (3.2.2) in order to expain the calculations for the weights. The 'k' column in figure (3.2.2) has elements with

Figure 3.2.2: Star-stop format for the first 6 patients

id	time	status	age	gender	BMI	UKELD	cens.score	tstart	tstop	state	k	estim.S.cen	weights
1	1	0	60	0	24.24	60	-2.322148	0	1	0	1	0.9955153	1.004505
2	2	1	66	0	30.53	67	-2.533361	0	2	1	1	0.9887670	1.011361
3	3	0	71	0	26.56	61	-2.254029	0	2	0	0	0.9915379	1.008534
3	3	0	71	0	26.56	61	-2.254029	2	3	0	1	0.9872362	1.012929
4	3	0	65	1	23.15	63	-2.652541	0	2	0	0	0.9873258	1.012837
4	3	0	65	1	23.15	63	-2.652541	2	3	0	1	0.9809036	1.019468
5	3	0	62	0	22.55	64	-2.554629	0	2	0	0	0.9885224	1.011611
5	3	0	62	0	22.55	64	-2.554629	2	3	0	1	0.9827012	1.017603
6	4	1	56	0	36.39	73	-2.811239	0	2	0	0	0.9851177	1.015107
6	4	1	56	0	36.39	73	-2.811239	2	4	1	1	0.9700706	1.030853

1 or 0.The 1 represents an event or censoring and the 0 that the individuals are still at risk of an event or censoring.
Furthermore we notice that the weights for each individual change when an event(death) is occured while are still alive.
After the construction of the table we fit the usual Cox model or any model of our preference according to the start-stop format.

Finally is mentioned that the weights can get quite large values(very large probabilities of censoring before time t).
In this situation it is more efficient to stabilize the weights.
The stabilised weights have the following form

$$w_{istab} = \frac{\widehat{S(t)}}{\widehat{S_{Ci}(t)}} .$$

The numerator in this expression is the usual Kaplan-Meier estimator of the survivor function.
These weights are not causing any trouble in expression the (3.9) because the likelihood doesn't change and they provide stability to the model.
Another useful thing is to take into account a robust estimate for the variance-covariance matrix by using the sandwich estimate(Lin and Wei (1989)).
This estimate involves the observed information matrix and the efficient scores and is written as

$$I^{-1}(\hat{\beta})U'(\hat{\beta})U(\hat{\beta})I^{-1}(\hat{\beta}) .$$

This quantity is helpful because it corrects the possible overestimation of the standar errors ,meaning standar errors which are smaller than they should be.

Chapter 4

Survival methods for comparison of 2 groups of data

This chapter presents methods regarding to 2 groups of survival data such as non-parametric tests for comparing 2 groups ,some parametric and semi-parametric models and a sensitivity analysis for 2 groups of data.

4.1 Non-parametric tests

4.1.1 Log-rank test

Suppose that 2 groups of data are labelled as group 1 and group 2,also d_{1j} and d_{2j} denote the number of deaths at t_j which is the j-th distinct death among r such deaths.
Moreover $n_{1j} - d_{1j}$ and $n_{2j} - d_{2j}$ are the individuals who survive beyond t_j and n_{1j} , n_{2j} are the individuals at risk and uncensored just before t_j.
Finally d_j ,n_j are the total deaths and total individuals at risk respectively, at t_j .

The null hypothesis in a log-rank test, states that there is no difference in

the survival experience among the 2 groups.

Under the null hypothesis and the assumption of constant total individuals and deaths in the study, all quantities can be detemirned by the d_{1j}.

As a result the d_{1j} is considered to have the hypergeometric distribution.

Consider a box with 2 different type of balls ,the green ones represent the deaths d_{1j} and the red ones the individuals $n_{1j} - d_{1j}$ who survive beyond t_j, then pick 1 ball at a time without replacement until the n_{1j} stage.

Furthermore we define k as the number of deaths picked.

The probability that d_{1j} takes the value k is

$$\frac{\binom{d_j}{k}\binom{n_j-d_j}{n_{1j}-k}}{\binom{n_j}{n_{1j}}} .$$

The statistical function $L = \sum_{j=1}^{r}(d_{1j} - \mu_{1j})$ has zero mean and measures the deviation of the observed values of deaths from their expected values μ_{1j} in group 1.

So in order to take a better measure for the deviance between the d_{1j} and μ_{1j} we have to square this statistic and if we consider a relatively large number of deaths , we get that $\frac{L}{\sqrt{V}}$ approaches the standar normal distribution, where V represents the variance of L.

These facts are resulting to the following statistic $T = \frac{L^2}{V}$.

More specifically the mean of d_{1j} is

$$\mu_{1j} = n_{1j}\frac{d_j}{n_j} ,$$

and the variance of L,because the deaths for each individual are independent is

$$V = Var\Big(\sum_{j=1}^{r}(d_{1j} - \mu_{1j})\Big) = \sum_{i=1}^{r} Var(d_{1j} - \mu_{1j}) = \sum_{i=1}^{r} Var(d_{1j}) = ,$$

$$\frac{n_{1j}d_j(n_j - d_j)n_{2j}}{n_j^2(n_j - 1)} .$$

Large values of this statistic provide evidence against the null hypothesis.

Also under the null hypothesis this statistic follows approximately the chi-squared distribution with one degree of freedom.

Other names for this test are reffered to as Mantel-Cox and Peto-Mantel-Haenszel.

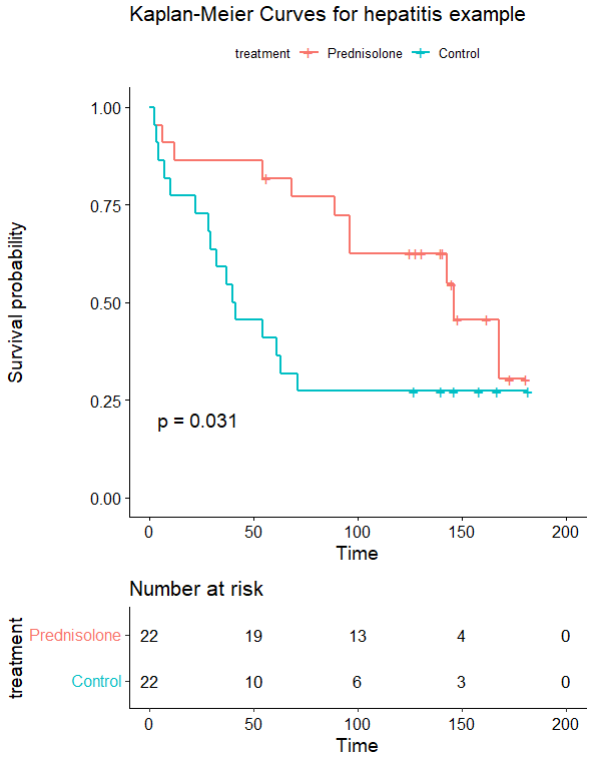
Furthermore if the alternative hypothesis is based on the proportional hazard assumption ,the log-rank test is recommended for general use.

For example such tests consider that the null hypothesis of no survival differences is equivalent with the assumption that the hazard ratio(ψ)(1.29) equals to 1 and the alternative hypothesis may be that the hazard ratio is less than 1 (one sided hypothesis) or different from 1(two sided hypothesis)

Lets consider again the hepatitis example as in (1.5.3) section(it doesn't matter the choice of zero's and one's or one's and two's in order to define the treatment group traditionally though the first choice is picked).

The figure(4.1.1) shows that the survivor functions of the 2 groups do not cross ,so it is safe to say that the proportional assumption holds. Also from the figure

Figure 4.1.1: Proportional assumption test



(4.1.1) it is clear that the usage of the prednisolone increases the survival for the

individuals in the study and the p-value for the log-rank test is p=0.031 which provides a strong evidence against the null hypothesis of no differences in the groups.

For instance if the significant level is $\alpha = 0.05$ then $p < \alpha$.

Moreover the detailed results from the log - rank test are given in figure(4.1.2)

Figure 4.1.2: Log-rank test

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
treatment=1 22       11     16.4      1.77      4.66
treatment=2 22       16     10.6      2.73      4.66

 Chisq= 4.7  on 1 degrees of freedom, p= 0.03
```

As we see the log-rank statistic is 4.66(rounded=4.7) and each group due to symmetry gives the same value .

4.1.2 Wilcoxon-Breslow test

Another test that examines the difference in survival experience between 2 groups of data is the Wilcoxon test and is recommended for other types of alternative hypothesis(different from the proportional assumption hypothesis). The statistic is similar with the log-rank statistic. More specifically the statistic is

$$\frac{W^2}{V'} ,$$

where $W = \sum_{i=1}^{r} n_j(d_{1j} - \mu_{1j})$ and V' is the variarnce of W and is equal to

$$V' = n_j^2 V ,$$

where V is the variance from the log-rank test.

Again this statistic follows approximately the chi-squared distribution with 1 degree of freedom.

As we notice,the statistic W weights the differences $d_{1j} - \mu_{1j}$ by using the total number of individuals at risk at time t_j and is trying in that way to give less weight in the longest survival times.

Thus so is less sensitive than the log-rank statistic in the longest survival times.

4.1.3 Stratified tests

When we consider that a variable such as age,gender,etc ,has an impact on the survival experience at each group ,it is usefull to calculate the log-rank or the Wilcoxon statistic at each strata(e.g male,female) and then summarize those calculations in order to take a more informative assesment about the differences in survival experiences at each group.
For example a stratified log-rank test uses the statistic

$$\frac{\left(\sum_{i=1}^{n} L_i\right)^2}{\sum_{i=1}^{n} V_i},$$

where L_i and V_i are the statistic as in (4.1.1) and the variance of L_i respectively for the i-th strata.
Again this statistic follows approximately the chi-squared distribution with one degree of freedom.

As an illustration let's say that in hepatitis example(1.5.3), we want to examine if the age affects the survival at each group.
Specifically we consider 2 strata.
The fist one contains the patients below the age of 45 and the other one the patients above 45.
The first 24 patiens are presented in figure (4.1.3) The log-rank test gives the

Figure 4.1.3: First 24 patients in hepatitis example

	treatment	time	status	age	agegroup
1	1	2	1	76	45+
2	1	6	1	25	45-
3	1	12	1	65	45+
4	1	54	1	28	45-
5	1	56	0	30	45-
6	1	68	1	37	45-
7	1	89	1	85	45+
8	1	96	1	90	45+
9	1	96	1	47	45+
10	1	125	0	36	45-
11	1	128	0	37	45-
12	1	131	0	38	45-
13	1	140	0	39	45-
14	1	141	0	40	45-
15	1	143	1	41	45-
16	1	145	0	42	45-
17	1	146	1	56	45+
18	1	148	0	44	45-
19	1	162	0	44	45-
20	1	168	1	46	45+
21	1	173	0	47	45+
22	1	181	0	48	45+
23	2	2	1	49	45+
24	2	3	1	50	45+

following results in figure (4.1.4) We see that the results are similar with the unajusted one's in figure (4.1.2),so we didn't have to use the stratified statistic and as result the age doesn't seem to affect the survival of the individuals in

Figure 4.1.4: Stratified log-rank test

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
treatment=1 22       11     16.8      2.02      6.14
treatment=2 22       16     10.2      3.35      6.14

 Chisq= 6.1  on 1 degrees of freedom, p= 0.01
```

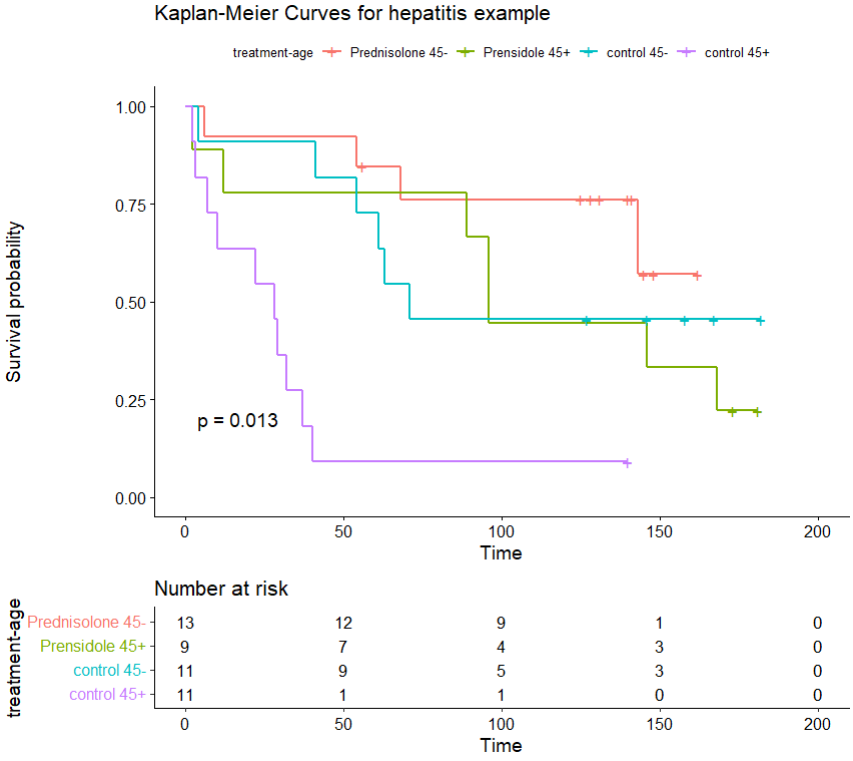
the study.

Moreover a plot in figure (4.1.5) of the survival curves for each strata in every group shows that each strata in the 2 groups have the proportional hazard assumption(alternative hypothesis).

This plot is usefull ,because combined with the results from log-rank test can help to identify which treatment is the superior one.

In this case as with the unadjusted log-rank test(4.1.2) , we see that the prednisolone treatment is better.

Figure 4.1.5: Proportional assumption test for the stratified case



4.2 The Cox model for 2 groups of data

In chapter (1.3) was investigated thoroughly the usual Cox model(1.28).
In this paragraph we will mention an important special case for 2 groups of data
.
Suppose that we want to compare 2 groups of data.
The fist group is called the standar group and involves the patients who receive the standar treatment and the second group is reffered to as the new group which involves the patients who receive the new treatment.
As it is mentioned before the Cox model is valid under the proportional assumption and the model takes the form

$$h_N(t) = \psi h_S(t) \, ,$$

where ψ is called hazard ratio and it is positive.
If $\psi < 1$ then $h_N(t) < h_S(t)$, meaning that the new treatment is better than the old one , because the danger of death for an individual is smaller ,otherwise the old treatment is better if $(\psi > 1)$.
Also if $\psi = 1$, then the new treatment is the same as the old one(null model) and then the estimation of the hazard is based on non-parametric methods as in (1.2.1)-(1.2.7).
Moreover in order to put this model under the family of Cox models ,we use an indicator function that takes the value 1 if the patient is on the new treatment and zero otherwise.
Further we substitute ψ we the quantity $\exp(\beta x_i)$ where β is a scalar coefficient and the model takes the form

$$h_i(t) = \exp(\beta x_i) h_0(t) \, .$$

The β coefficient is the treatment effect.
In (1.5.3) section were presented an example (hepatitis example) that use this model.
A more general model for 2 groups of data is a model that contains treatment effect and also some other variables such as some demographic variables(e.g

age,sex etc).

4.3 A parametric model for 2 groups of data

At this point we will mention a special case of the (1.58) model that is used to compare 2 groups of data.
Such models are valid under the accelerated failure property.
The model in this case is the following one

$$h_N(t) = \frac{1}{\psi} h_S\Big(\frac{t}{\psi}\Big) .$$

This model tell us that the danger of death in the new treatment is ψ times the danger of death in the old treatment and as a result the values of ψ shows how fast or slow the death will come.
If $\psi < 1$, then the patients in the new treatment are more likely to die faster than those in the old treatment and vice-versa when $\psi > 1$.
The model for the i-th patient in order to fit in the family of the accelarated model takes the following form

$$h_i(t) = \exp(-ax_i) h_0\Big(\frac{t}{\exp(ax_i)}\Big) , \tag{4.1}$$

where $\exp(-ax_i)$ represents the ψ^{-1} factor which is reffered to as the accelaration factor and x_i is the indicator variable that shows the group of the i-th patient.Also α is a scalar coefficient.
Of course this model can be generalized with more explanatory variables as in formula (1.58).

Q-Q plot

A plot for assesing the validity of the accelarated failure time model (4.1) can be provided by a quantile-quantile plot.
The p-th quantile(percentile in this case) for a continuous survivor function is defined as $t_p = S^{-1}\Big(\frac{100-p}{100}\Big)$.
So under the accelerated model the survivor functions for each group have the following relation $S_2(t_{p2}) = S_1\Big(\frac{t_{p2}}{\psi}\Big)$(1.5.1 paragraph) which leads to the fol-

lowing relation for the quantiles for each group $t_{1p} = \psi^{-1} t_{2p}$.

This relation suggests that if we plot the estimated quantiles of group1 against the estimated quantiles of group 2(Kaplan-Meier estimates),then the result would be an approximate straight line with slope equal to the accelaration factor ψ^{-1} and zero intrercept.

At this point let's use an example to explain this procedure.

In a study that is occured at the university of Oklahoma Health Sciences Center, data were obtained on the survival times of patients with kidney cancer.

Group 1 involes the patients who hadn't received a nephrectomy(0) and group2 involves the patients who had received a nephrectomy(1).

Nephrectomy is the surgical removal of kidney.

In figure (4.3.1) are presented the data for the first 10 patients The Kaplan-

Figure 4.3.1: Nephrectomy table,data is used comes from [1]

	nephrectomy	time	status
1	0	9	1
2	0	6	1
3	0	21	1
4	0	15	1
5	0	8	1
6	0	17	1
7	0	12	1
8	1	104	0
9	1	9	1
10	1	56	1

meier estimates for the survivor functions of each group are presented in figure (4.3.1)

Figure 4.3.2: Kaplan Meier estimates for each group

nephrectomy=0

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	7	1	0.857	0.132	0.6334	1.000
8	6	1	0.714	0.171	0.4471	1.000
9	5	1	0.571	0.187	0.3008	1.000
12	4	1	0.429	0.187	0.1822	1.000
15	3	1	0.286	0.171	0.0886	0.922
17	2	1	0.143	0.132	0.0233	0.877
21	1	1	0.000	NaN	NA	NA

nephrectomy=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	29	1	0.966	0.0339	0.9013	1.000
6	27	1	0.930	0.0479	0.8404	1.000
8	26	1	0.894	0.0579	0.7874	1.000
9	25	3	0.787	0.0773	0.6489	0.954
10	22	1	0.751	0.0816	0.6069	0.929
14	21	1	0.715	0.0852	0.5663	0.903
18	20	2	0.644	0.0905	0.4887	0.848
26	18	2	0.572	0.0935	0.4154	0.788
35	16	1	0.536	0.0942	0.3801	0.757
36	15	2	0.465	0.0943	0.3124	0.692
38	13	1	0.429	0.0936	0.2799	0.658
48	12	1	0.393	0.0923	0.2483	0.623
52	11	2	0.322	0.0883	0.1880	0.551
56	9	1	0.286	0.0854	0.1593	0.514
68	8	1	0.250	0.0819	0.1318	0.475
72	7	1	0.215	0.0776	0.1056	0.436
84	5	1	0.172	0.0730	0.0746	0.395
108	3	1	0.114	0.0675	0.0360	0.363
115	1	1	0.000	NaN	NA	NA

Also a table for the 10-th till 90-th percentile is presented in figure (4.3.3)

These percentiles can be taken as the smallest time such as the survivor function is below $\frac{100-p}{100}$ (paragraph (1.2.7)).

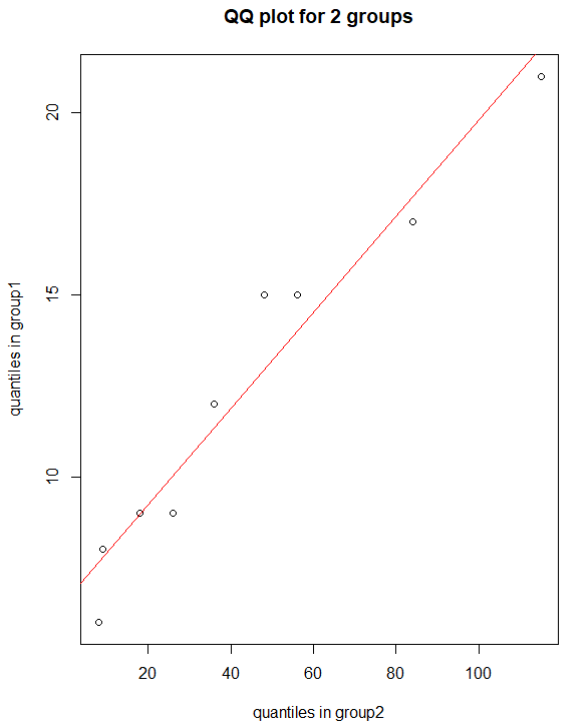
Finally a plot of the second column against the third column in figure (4.3.3)

Figure 4.3.3: Percentile table

percentiles	group1	group2
10	6	8
20	8	9
30	9	18
40	9	26
50	12	36
60	15	48
70	15	56
80	17	84
90	21	115

shows an approximate straight line,so the accelarated failure time model can be used.

Figure 4.3.4: QQ plot



Moreover the straight line gives an approximate accelaration factor $0.13 < 1$.As a result the nephrectomy procedure tries to slow done the process of death because is less than unity.

The adoption of an accelarated weibull model gives the following estimations in figure(4.3.5)

This output confirms the neccesity of the nephrectomy procedure on patients and also gives that the accelaration factor is $exp(-1.413) \approx 0.24$ which is close to the estimation of the Q-Q plot.

Figure 4.3.5: Accelarated estimations

```
             value Std. Error     z       p
(Intercept)  2.554      0.293  8.71 < 2e-16
nephrectomy  1.413      0.332  4.26 2.1e-05
Log(scale)  -0.255      0.144 -1.77   0.077

Scale= 0.775

Weibull distribution
Loglik(model)= -146.2   Loglik(intercept only)= -151.4
        Chisq= 10.5 on 1 degrees of freedom, p= 0.0012
Number of Newton-Raphson Iterations: 6
n= 36
```

Thus we see that with only one plot we can take valuable assumptions .
Finally we mention that if the end point is not death but recovery from a disease, then the acceleration factor has an opposite meaning in relation with the previous analysis on the (4.1) model.

4.4 Sensitivity analysis under informative censoring on 2 groups of data

Suppose that we have 2 groups of data A and B and that we are interested in a sensitivity analysis according to (3.1) section.
So for the 2 groups of data under the proportional hazard assumption we consider the following 2 models for the T mechanism(event process), that are presented also in [5].

$$h_A(t) = e^{\theta} h_0(t) = e^{u' x_A} ,$$

and

$$h_B(t) = e^{k} h_A(t) = e^{k+\theta} h_0(t) = e^{u' x_B} h_0(t) ,$$

where θ is the parameter of the failure distribution in group A , k is the hazard ratio parameter , $u = (\theta, k)'$, $x_A = (1, 0)'$ and $x_B = (1, 1)'$.
Also $h_0(t)$ is the baseline hazard of a known distribution(e.g exponential,weibull etc).
The main interest is to perform a sensitivity analysis on vector u.
The log - likelihood for each group has the form of the (3.6) expression, but we take different $\delta' s$ in order to allow some flexibility in the levels of dependence within each group.
Moreover the log - likelihood that we will use to draw inferences about u is the sum of the log - likelihoods(3.6) for each group(independent log-likelihoods),

$L_\delta = L_\delta^A + L_\delta^B$.

The observed information now takes the form of a matrix

$$\phi(u) = -\frac{\partial L_0}{\partial u \partial u'} = -\begin{pmatrix} \frac{\partial^2 L_0}{\partial \theta^2} & \frac{\partial^2 L_0}{\partial k \partial \theta} \\ \frac{\partial^2 L_0}{\partial \theta \partial k} & \frac{\partial^2 L_0}{\partial k^2} \end{pmatrix},$$

where $L_0 = L_0^A + L_0^B$ and after some calculations under the models $h_A(t)$ and $h_B(t)$ the correlation bias for u is then

$$\widehat{u_\delta} - \widehat{u_0} = (\phi(u))^{-1} \begin{pmatrix} \delta_A \sum_{i=1}^{n_A} \left(H_T^A(t_i) H_C^A(t_i) - (1 - I_i^A) H_T^A(t_i) \right) + \delta_B \sum_{i=1}^{n_B} \left(H_T^B(t_i) H_C^B(t_i) - (1 - I_i^B) H_T^B(t_i) \right) \\ \delta_B \sum_{i=1}^{n_B} \left(H_T^B(t_i) H_C^B(t_i) - (1 - I_i^B) H_T^B(t_i) \right) \end{pmatrix},$$

where

$$\phi(u) = \begin{pmatrix} \sum_{i=1}^{n_A} H_T^A(t_i) + \sum_{i=1}^{n_B} H_T^B(t_i) & \sum_{i=1}^{n_B} H_T^B(t_i) \\ \sum_{i=1}^{n_B} H_T^B(t_i) & \sum_{i=1}^{n_B} H_T^B(t_i) \end{pmatrix}.$$

As a result after some calculations the corellation bias for u takes the following form

$$\begin{pmatrix} -\delta_A \frac{\sum_{i=1}^{n_A} \left(H_T^A(t_i) H_C^A(t_i) - (1 - I_i^A) H_T^A(t_i) \right)}{\sum_{i=1}^{n_A} H_T^A(t_i)} \\ \delta_A \frac{\sum_{i=1}^{n_A} \left(H_T^A(t_i) H_C^A(t_i) - (1 - I_i^A) H_T^A(t_i) \right)}{\sum_{i=1}^{n_A} H_T^A(t_i)} - \delta_B \frac{\sum_{i=1}^{n_B} \left(H_T^B(t_i) H_C^B(t_i) - (1 - I_i^B) H_T^B(t_i) \right)}{\sum_{i=1}^{n_B} H_T^B(t_i)} \end{pmatrix}. \tag{4.2}$$

The first row in (4.2) is the bias of θ and the second raw is the bias of k.

Also we notice that the first row equals to the corellation bias of θ for the group A alone as in the expression $(\widehat{\theta} - \widehat{\theta_0})$(3.7),so the model for group B in the above analysis does not affect the estimated correlation bias of θ.

The second raw in (4.2) equals to the opposite of the correlation bias of θ for group A$(\widehat{\theta} - \widehat{\theta_0})$ plus the correlation bias of $\theta + k$, $\widehat{\theta} + \widehat{k} - (\widehat{\theta_0} + \widehat{k_0})$ considering the analysis with only group B.

Thus is the correlation bias $\widehat{k} - \widehat{k_0}$.

Similar levels of dependence with the same sign (δ_A¿0 and $\delta_B > 0$ or $\delta_A < 0$ and $\delta_B < 0$) implies that the bias of the risk parameter k becomes smaller.

On the other hand if δ_A and δ_B are opposite then the bias of k becomes larger.

The values of δ_A and δ_B represent the maximum correlation between the T and C mechanisms, so they are naturally restricted to take values between -1 and 1.

Finally we mention that if a Weibull model is used,each row of (4.2) can be calculated by using the (3.8) expression and for the special case of an exponential distribution we take $\hat{\alpha} = 1$ and $\hat{\alpha_c} = 1$.
More specifically in the Weibull case let's say that the basiline hazard(group A) has the following form for the T mechanism

$$h_{T0}(t) = \lambda \alpha t^{a-1} .$$

Also the form for the C mechanism is

$$h_{C0}(t) = \lambda_c \alpha_c t^{a_c - 1} .$$

As a result the first row in (4.2) is

$$\widehat{\theta} - \widehat{\theta_0} = \delta_A \frac{\sum_{j=1}^{n} \left(\exp(\widehat{\gamma}) t_j^{\hat{\alpha} + \widehat{\alpha_c}} - (1 - I_j) t_j^{\hat{\alpha}} \right)}{\sum_{j=1}^{n} t_j^{\hat{\alpha}}} ,$$

where α_c is the shape parameter of the baseline hazard under the cencoring model and γ is the distribution parameter under the C mechanism for the group A(first model at the start of this paragraph).
Moreover the second raw has the following form

$$\widehat{k} - \widehat{k_0} = -\delta_A \frac{\sum_{j=1}^{n} \left(\exp(\widehat{\gamma}) t_j^{\hat{\alpha} + \widehat{\alpha_c}} - (1 - I_j) t_j^{\hat{\alpha}} \right)}{\sum_{j=1}^{n} t_j^{\hat{\alpha}}} + \delta_B \frac{\sum_{j=1}^{n} \left(\exp(\widehat{\gamma} + \widehat{k_c}) t_j^{\hat{\alpha} + \widehat{\alpha_c}} - (1 - I_j) t_j^{\hat{\alpha}} \right)}{\sum_{j=1}^{n} t_j^{\hat{\alpha}}},$$

where k_c is the relative risk under the censoring model for the group B(second model at the start of this paragraph).

4.5 Inverse probability weights (IPW)

Again we consider 2 groups of data A and B who represent 2 different treatments.
The focus of our interest is to compare the 2 treatments and on how each treatment affects an outcome of interest.
The ideal study design would be to use a randomized trial , meaning each patient is randomly assingned to group A or group B.

This method aims to balance any differences in baseline characteristics so that any differences in the outcome may be attributed to the treatment.
The randomization though , usually is costly and sometimes impractical or unethical.
On the other hand observational studies are less costly and more practical for researchers.
However, even with a well-designed observational study, subjects in different treatment groups are not likely to be comparable with respect to their baseline characteristics.
For example the group A may has lower percentage of men ,meaning that the issue of imbalance arise,so our conclusions about the treatment effect may be misleading due to potential selection bias(is the bias that arises by the selection of patients for analysis in such way that a proper randomization is not achieved).
Moreover possible confounding variables(a type of external variable which has correlation with the independent variables and affect the dependent variable(response variable).
For example hot temperatures can cause people to both eat more ice cream and spend more time outdoors under the sun, resulting in more sunburns.
In this example temperature is the confounding variable ,ice cream consumption is the independent variable and sun burn is the outcome.
A common method to face the imbalance is to first consider an appropriate model for the data (e.g multivariate regression model,Cox model,accelarated-failure time model etc).
After the choice is considered the model with only the treatment effect a factor and is observed the significance of the treatment effect.
If the treatment effect is significant then we adjust the other explanatory variables inside the model and observe if the treatment effect is still significant.
In this case we have a strong evidence to conclude that the treatment effect is significant.
Another strategy is mentioned in (1.3.7) section,that first picks the appropriate without considering the treatment effect and then adds the treatment effect into the model in order to examine the significance of this factor.
In any case the usage of the model which contains only the treatment effect it

can be generally provide useful results.
However, in studies where sample sizes may be small in relative with the number of unbalanced variables, this method may not work or may be invalid.

The usage of propensity scores(specific probabilities) can provide an alternative method of addressing the issue of imbalance.
The propensity score is defined as the probability of an individual being assigned to one of the two treatments given all information(explanatory variables) ,before assignment.
Let's say that the response variable(treatment assingment that takes the values 1 and zero) is the Y_i for the i-th individual and X_i is the explanatory vector, then these probabilities(propensity scores) have the following form

$$P(Y_i = 1|X_i) = p_i \, and \, P(Y_i = 0|X_i) = 1 - p_i \, .$$

The estimation of the propensity scores is based on the data collected, such that the patients with similar scores(probabilities) would be the patients with similar explanatory values.
Thus the patients with similar propensity scores are comparable.

A method that uses the inverse of the propensity scores in order to balance the expanatory variables , is called inverse probabilty weights (IPW).
We notice that this inverse technique is mentioned for a different purpose in (3.2) section and we named it as IPCW(see the paragraph below of the (3.10) formula).
In this section the idea of (IPW) is to weigh the patients by the inverse of their propensity scores.
Thus the patients with higher propensity scores(probabilitties) or equivalently overrepresented patients, will take a lower weight and vice - versa for the underrepresented patients.
The result will be a pseudo sample which contains balanced expanatory variables among the 2 groups .
For instance let's consider that group A has 5 males and group B has 10 males ,so the gender variable is unbalanced between the 2 groups and the other pos-

sible expanatory variables are considered balanced.
The propensity score or probability of group A is 5/15 and the propensity score or probability of group B is 10/15.
Then the weight of the group A is 15/5=3 and the weight of the group B is 15/10 =3/2.
So the pseudo sample for the A group has $3 * 5 = 15$ females and the B group has $15/10 * 10 = 15$ females.As a result we end up with a balanced gender factor.

In the general case, where more variables are unbalanced ,the approach is to use a multivariate logistic regression model with the treatment assingment as the response variable(binary response) and the other variables (measured before assingment) as independent variables(balanced and unbalanced variables are included).
Thus the logistic model will estimate the propensity scores $P(Y_i = 1|X_i) = p_i$, $P(Y_i = 0|X_i) = 1 - p_i$ and then we will take the inverse of these probabilities and use them as weigths in the chosen model.
The original (IPW) method often artificially increase the total sample size as in the previous example.
Also small propensity scores provide huge weights and this fact may lead to an increase in the variances of the estimated coefficients.
A solution to this problem is to stabilize the weights.
The stabilized weights are defined as $\frac{\pi}{p_i}$ for $Y_i = 1$ and $\frac{1-\pi}{1-p_i}$ for $Y_i = 0$, where π is the probability of treatment without considering covariates.
Moreover as it is mentioned in (3.2) section(last paragraph) we may consider also a robust estimate for the coefficients.
For example if 30 out of total 50 individuals are in the treatment group then $\pi = 0.6$ and is the same for all 30 individuals regardless of their baseline characteristics.
After the stabilization, the weights are used to fit the model of interest.(e.g weighted Cox regression model, weighted logistic regression model, weighted linear regression model etc).

Chapter 5

Relative survival

The relative survival concept was introduced by Berkson(1942),although the term was first introduced in [16].
Berkson proposed the relative survival suvival ratio (all cause survival of the patients divided by all-cause survival that would be expected in the absence of the specific disease under study), as an estimator for the net survival(the survival probability in a hypothetical world where patients could only die of the specific disease).
In general the relative survival field involves 2 data sets ,the observed data on patients and the general population mortality data of a country or a region.
We observe the patients in a period of time and the interest focuses on a specific disease.
In the presence of the specific disease of interest a patient may also die from other causes.
The relative survival field is based on the assumption that the hazard function and all the other functions of interest that involve deaths from other causes, are represented by the population data set which can be obtained from the general-population mortality tables.

One of the main goals in relative survival field is to compare the disease of interest for different populations (e.g populations of different countries).
This can be done by computing the net survival ,which is defined as the hy-

pothetical survival probability that takes into account only the deaths from the disease of interest and thus is independent from the general population trends,because is assumed that the disease of interest is rare and by removing it from the population life tables has negligible effect on the hazard of the general population data set.

Moreover the main difference from a usual competing risks approach(Kaplan-Meier approach (2.1.1)), is that the relative survival field, as we will see in the following text, provide estimators for the net survival, that don't need the knowledge of the causes of death.

Finally an important assumption is that given a known set of covariates the time to death due to the disease of interest and the time to death due to other causes are conditionally independent.

As a result the hazard due to other causes(general population hazard) and the excess hazard(hazard for the disease of interest) gives the following expression(see also [12])

$$\textbf{observed hazard} = \textbf{excess hazard} + \textbf{population hazard}\ . \tag{5.1}$$

We notice from this formula that the population hazard is smaller than the observed hazard.

At this point we define for the i-th individual the time to death due to the disease of interest T_{E_i} ,also T_{P_i} is the time to death due to other causes(general population) and $T_i = \min(T_{E_i}, T_{P_i})$, where ,$i = 1 \dots n$ is the observed time(the occurance of one cause of death precludes the other causes).

Furthermore we assume independent and identical cencored times C_i and also we define the follow up time $F_i = min(T_i, C_i)$,$i = 1 \dots n$ and the usual event indicator δ_i.

The i-th individual in the observed data set has a vector X_i of covariates , D_i represents the vector of the demographic variables(sex,age etc) from the general population mortality tables and is considered as a subset of X_i,$i = 1 \dots n$.Usually we take $X_i = D_i$.

So we assume that the i-th individual in the general population set has identical demographic variables with the i-th individual in the study.

Moreover the survival fuctions $S_{E_i}(t) = P(T_{Ei} > t|X_i)$, $S_{P_i}(t) = P(T_{Pi} > t|D_i)$ and $S_{O_i}(t) = P(T_{Oi} > t|X_i)$ are the excess, population and observed survival respectively for the i-th individual, $i = 1, \ldots, n$.
Finally we mention that the observed data(deaths for all causes) for the i-th patient ,$i = 1, \ldots n$, is defined by the (F_i, δ_i, X_i).

Relative survival ratio

Suppose that $S_O(t)$ is defined as the total observed survival function of the patients in study and $S_P(t)$ is the total population survival function ,$S_P(t)$ is also reffered to as the expected survival function.
The population function $S_P(t)$ can be found from the general population tables of the individuals who have the same demographic characteristics as the individuals in the study group ,but are free of the specific disease under study.
The relative survival ratio $S_R(t)$ is defined as(see also [12])

$$S_R(t) = \frac{S_O(t)}{S_P(t)}. \tag{5.2}$$

This is a measure that compares the observed survival to the survival of the disease-free group, who have the same demographic characteristics as the study group.
If $S_R(t) < 1$ then the mortality of the patient group, exceeds the mortality of the general population group (free of the specific disease under study),so a patient in the study group live less than a similar-identical person (same demographic variables) without the disease.
Of course if $S_R(t) = 1$ then the mortality of the 2 groups is the same.
The maintenance of the relative survival ratio to 1 over a reasonable number of years in the follow - up period,indicates that some ratio of patients in the study escape from the specific disease.
Thus the general population group can be regarded as a control group(see also [16]).

Net survival

Net survival is defined as the survival due to the excess hazard alone ,meaning that is defined as the survival probability in a hypothetical world where patients can only die of cancer

Net survival has the following form(see also [12] and [19])

$$S_E(t) \approx \frac{1}{n}\sum_{i=1}^{n} S_{Ei}(t) = \frac{1}{n}\sum_{i=1}^{n} \exp\Big(-\int_0^t h_{Ei}(u)du\Big) = \frac{1}{n}\sum_{i=1}^{n} \frac{S_{Oi}(t)}{S_{Pi}(t)}\,, \quad (5.3)$$

where the first approximation is the definition of the net survival and it can be seen as an estimation of the theoretical probability of death due the cause of interest in the hypothetical world,where only one cause of death exists.

The second equality is derived from the (1.6) formula and the last equality it comes from the assumption, that the events of the i-th patient, $\{T_{Ei} > t|X_i\}$ and $\{T_{Pi} > t|D_i\}$ are independent and also $X_i = D_i$,meaning that the individuals in the study are fully determined by the demographic covariates.

As a result, because the event $\{T_{Oi} > t|X_i\}$ is written as

$$\{T_{Oi} > t|X_i\} = \{T_{Ei>t}|X_i, T_{Pi>t}|D_i\}\,,$$

we get that $S_{Oi}(t) = S_{Ei}(t)S_{Pi}(t)$.

Moreover from the equation (1.6)

$$S_E(t) = \exp\Big(-\int_0^t h_E(u)du\Big)\,,$$

where $h_E(t)$ is the usual hazard of the cause of interest(1.2).

The net survival is a key measure in the relative survival field, since it is the only quantity independent of the population mortality and thus directly comparable. The relative survival ratio has the advantage of a clear interpretation in the real world ,but it is usually less desirable than net survival due to its strong dependence on the population mortality trends.

On the other hand the fact that net survival is not a real world measure must be kept in mind at all times.

As it is mentioned in the beginning of this section one of the main goals in the relative survival field is to compare a disease of interest between 2 different

populations through the net survival , thus in the following text we will present estimations for the net survival.

Observable net survival

As it is pointed out in [12], in practice the net survival is not observable when at least 2 causes of death are involved.
For this reason we focus on the estimation of the observable net survival(it is also reffered to as net survival) which is defined as

$$S_E^*(t) = \exp\Big(- \int_0^t h_E^*(u)du \Big) , \tag{5.4}$$

where $h_E^*(u)$ is the cause specific hazard of the cause of interest in (2.1) formula. We notice that is involved the cause specific framework ,but in the relative survival field we will estimate the cause specific survival, in the presence of the assumption of the conditional indepentent events $\{T_{Ei>t}|X_i\}$ and $\{T_{Pi>t}|D_i\}$ with $X_i = D_i$, $h_E(t) = h_E^*(t)$ and thus under this assumption, the observable net survival and net survival, give the same quantity.
Of course this assumption may not hold,but in any case the estimations for the net survival,can also be regarded as estimations for the net survival,with a small or huge bias according to each case.
Finally we mention, that a method in order to estimate the observed net survival is the Kaplan-Meier approach in (2.1.1).

5.1 Non-parametric estimators for the net survival

Non-parametric estimations of the net survival (5.3) can be taken by estimating the relative survival ratio(5.2) with different approaches.
Thus we will present all current estimators of the relative survival ratio(5.2) ,Ederer I,Ederer II,Hakulinen.
Also we will present the newest suggested estimator of the net survival which is called Pohar-Perme estimator[12].
The main difference of the first 3 estimators, is that the estimator of the overall

population survival in the denominator of (5.2), $S_P(t)$,has a different form. These estimations have the following general form

$$S'_R(t) = \frac{\widehat{S_O}(t)}{\widehat{S_P}(t)} ,$$

where $\widehat{S_O}(t)$ is a usual non-parametric estimator of the observed data(e.g (1.11) or (1.15) or (1.17)) and the denominator $\widehat{S_P}(t)$ determines each of these 3 estimators.

Ederer I estimator

Originally Ederer I estimator proposed in [13] in order to estimate the population survival $S_P(t)$ and is defined as

$$\widehat{S_P}(t) = \frac{1}{n}\sum_{i=1}^{n} S_{P_i}(t) ,$$

where i denotes the i-th person in the disease free group from the general population who has the same demographic variables as the patient i in the study group.
We notice that this estimation of the population survival does not take into account the time at which a patient dies or is censored and thus the patients are considered to be at risk indefinitely.

Ederer II estimator

The Ederer II estimator originally was proposed in [14] in order to estimate the population survival of a group of patients at time t and is defined as

$$\widetilde{S_P}(t) = \frac{\sum_{i=1}^{n} Y_i(t) S_{Pi}(t)}{\sum_{i=1}^{n} Y_i(t)} , \tag{5.5}$$

where $Y_i(t)$ is the at risk indicator for the i-th patient, as defined in (2.2),but we remind to the reader that this variable takes the value 1,if an event or censoring is occured after time t and also takes the value 0 otherwise.
Thus the Ederer II estimator takes into account the times of patients until an event or cencoring was occured and provides a better estimation of the net

survival.

In fact the ederer II estimator,estimates the observed net survival(for further details see [12]).

Pohar-Perme estimator

According to the (5.1) formula and the competing risks framework(2.1)(see also [12]) we can take that

$$h_O(t) = h_E^*(t) + h_P^*(t) \,,$$

where $h_E^*(t) = \lim_{h\to\ 0} \frac{P(t\leq T_E\leq t+h|T\geq t)}{h}$ and $h_P^*(t) = \lim_{h\to\ 0} \frac{P(t\leq T_P\leq t+h|T\geq t)}{h}$ are the cause-specific hazards for E and P respectively.

From this equation it follows that

$$H_E^*(t) = H_O(t) - H_P^*(t) \,,$$

where H_E^* and $H_P^*(t)$ are the cause-specific cumulative hazards(2.1.1).

From this equation Andersen and Vaeth (1989) in [15], take a natural estimation for the cumulative excess hazard of the Ederer II estimator that has the following form

$$\widetilde{H_E^*}(t) = \int_0^t \frac{d\Big(\sum_{i=1}^n N_i(u)\Big)}{\sum_{i=1}^n Y_i(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i(u) dH_{Pi}(u)}{\sum_{i=1}^n Y_i(u)} \,, \tag{5.6}$$

where $Y_i(t) = I(T_i \geq t, C_i \geq t$ and $dN_i(t) = I(T_i \leq t, C_i \leq t)$.

The $dN_i(t)$ records the observed event or censoring for the i-th individual before or at time t and equals to 1 if and only if the i-th individual experience an event or censoring before or at t.

Therefore for n individuals the sum $Y^*(t) = \sum_{i=1}^n Y_i(t)$ is the number of individuals alive and uncensored just before t and the sum $dN^*(t) = \sum_{i=1}^n dN_i(t)$ record the number of events(e.g deaths) and censoring before or at time t.

This notation provides a different form of what we called Nelson-Aalen estimator in (1.14).

Indeed we can define the estimation of the cumulative hazard H(t),as $d\hat{H}(t) = \hat{h}(t)dt$, where $d\hat{H}(t) = \frac{dN^*(t)}{Y^*(t)}$(this can be derived via the maximum likelihood method ,for further details see [2]).

Also by using the Riemann-Stieltjes integral (1.8) we take the alternative form

of the Nelson-Aalen estimator $\hat{H}(t) = \int_0^t d\hat{H}(u) = \int_0^t \frac{dN^*(u)}{Y^*(u)}$,which is exactly the first term on the right hand side of (5.6).

At this point in order to eliminate tha bias present in the risk set $\{Y_i(t), i = 1, \dots n\}$,Pohar-Perme(2012) in [12] used a weighted version of the (5.6) formula. More specifically the idea is to divide $Y_i(t)$ and $N_i(t)$ with the $S_{Pi}(t)$ for the i-th individual, $i = 1 \dots n$.
With this way each $Y_i(t)$ and $N_i(t)$ take increased values for smaller $S_{Pi}(t)$.
As a result the Pohar-Perme estimator for the net survival is defined through the followng cumulative excess hazard

$$\widetilde{H_E'}(t) = \int_0^t \frac{d\Big(\sum_{i=1}^n N_i'(u)\Big)}{\sum_{i=1}^n Y_i'(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i'(u) dH_{Pi}(u)}{\sum_{i=1}^N Y_i'(u)} ,$$

where $Y_i'(u) = \frac{Y_i(u)}{S_{Pi}(u)}$ and $N_i'(u) = \frac{N_i(u)}{S_{Pi}(u)}$ for the i-th individual.
Also in [12] is pointed out that this estimator can't be expressed as the relative survival ratio $\frac{\widehat{S_O}(t)}{\widehat{S_P}(t)}$ that is mentioned in the beginning of (5.1) section.
Finally in [12] it is also proved that the Pohar-Perme(PP) estimator is a consistent estimator for the net survival and is the current recommended non-parametric estimator for the net survival.

5.1.1 Estimations under informative sencoring

As it is noticed in chapter 3 there is always the possibility for informative cencoring in the study.
The above estimators provide biases in the presence of informative censoring,so a solution to these situations must be provided.

Hakulinen estimator

The Hakulinen estimator was proposed in [21] in order to take into account the presence of the informative censoring.
Under the informative cencoring, $\widehat{S_O}(t)$(e.g Kaplan-Meier) in the numerator provide biases in the estimation of the net survival through the relative survival $S_R'(t) = \frac{\widehat{S_O}(t)}{\widehat{S_P}(t)}$.

Thus the estimator proposed by Hakulinen, is trying to put a similar bias in the denominator.

This can be done by considering the following estimation for the population survival

$$S'_P(t) = \frac{\sum_{i=1}^{n} C_i(t) S_{Pi}(t)}{\sum_{i=1}^{n} C_i(t)}\,, \tag{5.7}$$

where $C_i(t)$ takes the values 1 and 0.

More specifically for the i-th individual $C_i(t) = 1$ if t is less or equal to the potential follow-up time and takes the value 0 otherwise .

The potential follow up time for a patient is considered as the period time of diagnosis till the last observed time(usually the end of the study).

In other words when an interim censoring or an event is occured(end of the potential follow up time) ,the matched individual in the population group ,also experience censoring or death and no longer contributes in the estimation(5.7).

The Hakulinen estimator was originally desinged for cancer registry data in which usually the patients remain in study until the end and thus only administrative censoring may occur,but in other studies cencoring may happen during the study.

Also as it is mentioned in [8] the Hakulinen estimator does not perform well under non-informative censoring and may be considered to be used only in the presence of informative censoring.

Further in [8] is given an example, where the data set includes a group of older patients and a group of younger patients in a period of 20 years with the same size.

The half of the younger patients was diagnosed in the beginning of the study and the other half after 10 years,but all the older patients was diagnosed in beginning of the study.

In this example the informative censoring arise.

Of course the Kaplan-Meier estimation in the numerator of the $S'_R(t)$ underestimates the actual survival of the observed data,because of the inbalance in the first 10 years(more old people) and the denominator which is the population survival, also underestimates the actual survival in the population.

As a result the biases are similar and the Hakulinen estimator provide a good estimation for the net survival.

Finally it is mentioned that if the potential follow up period is infinite for each individual,then $C_i(t) = 1$ for each i individual and the Hakulinen estimator coincide with the Ederer I estimator.

Weighted estimations

In [8] was proposed a solution in the presence of informative censoring for the Ederer I and PP estimator through the inverse probabilty of censoring weighted((3.2) section under the (3.10) expression).
First let's consider the Ederer I case.
From the (5.1) expression under the assumption that the events $\{T_{Ei>t}|X_i\}$ and $\{T_{Pi>t}|D_i\}$ with $X_i = D_i$ are independet , we take that

$$h_O(t) = h_E(t) + h_P(t)\,,$$

where h denotes the usual hazard in (1.2).
The population survival can be estimated as the average of the population survival for each i individual(Ederer I estimator).
Thus the derivative of t of the expression

$$\widehat{S_P}(t) = \frac{1}{n}\sum_{i=1}^{n} S_{P_i}(t)\,,$$

by using (1.1) leads to the following formula for the density function of P

$$\widehat{f_P}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{P_i}(t)\,,$$

but by using the (1.3) expression we get that

$$\widehat{h_P}(t) = \frac{\sum_{i=1}^{n} S_{P_i}(t)h_{P_i}(t)}{\sum_{i=1}^{n} S_{P_i}(t)}\,.$$

Moreover , the corresponding cumulative hazard relation for O,P and E is given by the following equation

$$H_O(t) = H_E(t) + H_P(t)\,,$$

As a result the corresponding Ederer I cumulative hazard estimation , $\widehat{H_E}(t)$ is given by the following equation

$$\widehat{H_E}(t) = \int_0^t \frac{d\Big(\sum_{i=1}^N N_i(u)\Big)}{\sum_{i=1}^N Y_i(u)} - \int_0^t \frac{\sum_{i=1}^N S_{Pi}(u) dH_{Pi}(u)}{\sum_{i=1}^n S_{Pi}(u)} \,. \tag{5.8}$$

The first integral in (5.8) as it was mentioned before in this chapter,that is the Nelson-Aalen estimator of the observed cumulative hazard.
In the presence of informative censoring the Nelson-Aalen estimator is affected,but on the other hand the population survival version of the Ederer I estimator is not affected,thus the (ICPW) gives the following form in the the Nelson-Aalen estimator in the (5.8) formula

$$\int_0^t \frac{d\Big(\sum_{i=1}^n \frac{N_i(u)}{S_{Ci}(u)}\Big)}{\sum_{i=1}^n \frac{Y_i(u)}{S_{Ci}(u)}} \,, \tag{5.9}$$

where $S_{Ci}(u)$ is the survivor function for the i-th patient under the independent censoring model (3.10).

Lastly we mention that the informative censoring version of the (PP) estimator is given by the following formula (for further details see [8])

$$\widetilde{H_E'}(t) = \int_0^t \frac{d\Big(\sum_{i=1}^n \frac{N_i(u)}{S_{Pi}(u)S_{Ci}(u)}\Big)}{\sum_{i=1}^n \frac{Y_i(u)}{S_{Pi}(u)S_{Ci}(u)}} - \int_0^t \frac{\sum_{i=1}^n \frac{Y_i(u)}{S_{Pi}(u)S_{Ci}(u)} dH_{Pi}(u)}{\sum_{i=1}^n \frac{Y_i(u)}{S_{Pi}(u)S_{Ci}(u)}} \,.$$

5.2 Parametric models for estimating the net survival

In this section is presented a parametric model(see also [19]) that estimates the net survival.
First we take into account the (5.1) expression and we consider that the excess hazard has the following form

$$h_E(t) = \exp(\beta' \mathbf{x}) \,,$$

where β is the coefficient vector and $\mathbf{x}$ is the expanatory vector.
Then according to (5.1) we get the following model

$$h_O(t) = h_P(t) + \exp(\beta'\mathbf{x}) \,. \tag{5.10}$$

Of course the log-likelihood for the observed survival of this model by using (1.55) is

$$\log(L(\beta)) = \sum_{i=1}^{n} \left(\delta_i \log(h_{Oi}(t_i)) + log(S_{Oi}(t_i))\right) \,.$$

Moreover by using the formula (1.6) gives the following result

$$\log(L(\beta)) = \sum_{i=1}^{n} \left(\delta_i \log\left(h_{Pi}(t_i) + \exp(\beta'\mathbf{x_i})\right)\right) - \sum_{i=1}^{n} \int_0^{t_i} h_{Pi}(u)du - \sum_{i=1}^{n} t_i \beta'\mathbf{x_i} \,.$$

The x_i vector for each $i = 1 \ldots n$ individual may depend on time.
In general , we split the follow up time into bands.
In each band it is assumed that the hazard is constant which implies that the number of deaths in each band-interval, follows the poisson distribution with mean $\mu_j = \lambda_j y_j$,where y_j is the length of each interval and λ_j the rate of death in the j-th interval.
In this case according the model (5.10) has the following form

$$\frac{\mu_j}{y_j} = \frac{d'_j}{y_j} + \exp(\beta'\mathbf{x}) \,,$$

where d'_j is the expected number of deaths from general population mortality table.
By taking logarithms we get a poisson general linear model(GLM) with a link that is not the usual.
More specifically the model has the following form

$$\log(\mu_j - d'_j) = log(y_j) + \beta'\mathbf{x} \,,$$

where the link is $\log(\mu_j - d'_j)$ and the offset $log(y_j)$.
For more details and examples the reader can see [19].

Chapter 6

Simulation

In this chapter we will present 3 simulated examples regarding the possible presence of the informative censoring in our sample.
The first example associates with 1 sample , is examined according to chapter 3 and it aims to explain the calculation of the correlation bias under the presence of one explanatory variable(not a proper simulation).
The logic here is that someone gave us the data.
In the second example we calculate the bias by generating a certain amount of censoring each time for dependent T and C mechanisms(informative censorng) in order to obsevre if there is any serious difference by ignoring the informative presence.
Finally in the last example we examine the case of 2 groups of data with possible informative censoring seperately and then together(4.4) in order to see if the possible bias of the treatment effect can be ignorable.

So for the first example we consider one group of 100 patients with survival times from the exponential distribution with rate 1.
Next we take the indicator variable for each patient from the bernouli distribution with 1/2 probability of event and we also consider the age as the explanatory variable from the uniform distribution in the interval (30,60).
In order to be able to find the correlation bias of theta in the (3.8) expression and in general to calculate the survivor function under the possible informative

censoring , as discussed before we have to estimate the coefficient of the age parameter under the censoring mechanism(see also the (3.2) section and the (3.10) expression).

The figure (6.0.1) gives exactly that. More specifically this is an estimation under the log-linear model (1.59) that gives $\alpha = 0.007$.

So $\beta_{age} = -0.007$(relations under the (1.61) expression). So the expression (3.8)

Figure 6.0.1: Estimation under the censoring mechanism

```
call:
survreg(formula = Surv(timeA, 1 - indi) ~ age, dist = "exponential")
              value Std. Error     z    p
(Intercept)  1.18231    0.82942  1.43 0.15
age         -0.00702    0.01764 -0.40 0.69

Scale fixed at 1

Exponential distribution
Loglik(model)= -79.9   Loglik(intercept only)= -80
        Chisq= 0.16 on 1 degrees of freedom, p= 0.69
Number of Newton-Raphson Iterations: 5
n= 100
```

for the i-th patient is written as

$$W(age_i) = \delta \frac{\sum_{j=1}^{n} \Big(\exp(\beta_{age} age_i) t_j^2 - (1 - I_j) t_j \Big)}{\sum_{j=1}^{n} t_j} \tag{6.1}$$

The table with the data for the first 6 patients is given in figure (6.0.2). Biases

Figure 6.0.2: Table,example 1

idA	timeA	indi	age
1	0.01057119	1	52
2	0.01495641	0	51
3	0.03280299	1	60
4	0.05147864	1	44
5	0.05707920	1	34
6	0.06885357	1	37

for $\delta = 0.2$ are given in figure (6.0.3) and as we notice they are quite large ,so we get a first glance that even for small correlation between the censoring and failure mechanisms we cannot take the independent model as a choice to fit the data. Moreover , it is useful to calculate the censoring scores $\beta_{age} age_i$

Figure 6.0.3: Correlation biases for each patient

```
17.81372 17.46945 20.56786 15.05958 11.61690 12.64970 15.05958 14.02678 13.68251 11.61690 14.71531 15.05958 15.74811 16.78092 19.19079 17.12518
17.81372 11.96117 10.92837 10.92837 10.92837 15.74811 18.50225 14.71531 16.09238 12.99397 19.53506 17.46945 12.64970 15.74811 17.12518 18.84652
11.61690 17.12518 18.15799 20.56786 12.99397 19.53506 17.46945 17.12518 17.81372 14.37104 10.58410 18.15799 18.84652 15.74811 19.19079 11.61690
11.27263 17.46945 19.19079 18.15799 13.68251 12.64970 14.02678 18.50225 18.84652 14.37104 15.74811 15.05958 19.19079 12.64970 18.15799 12.99397
19.53506 15.74811 18.84652 17.81372 13.33824 15.74811 16.09238 10.92837 15.40385 17.46945 11.27263 19.53506 17.81372 17.81372 11.96117 19.19079
16.43665 12.64970 11.61690 18.84652 14.02678 14.71531 11.27263 18.84652 19.87933 13.33824 17.81372 14.37104 10.58410 14.37104 12.64970 15.05958
13.68251 19.19079 11.96117 20.56786
```

and the risk scores $\beta^*_{age} age_i$ for each i individual in the study ,where β^*_{age} is the estimation of the coefficient of age under the event-failure mechanism.

The figure (6.0.4) provides this estimation. So $\beta^*_{age} = -0.005$.

The risk scores and censoring scores are provided in figures (6.0.5) and (6.0.6).

Figure 6.0.4: Estimation under the T mechanism

```
survreg(formula = Surv(timeA, indi) ~ age, dist = "exponential")
             Value Std. Error    z    p
(Intercept) 0.32316    0.70477 0.46 0.65
age         0.00561    0.01529 0.37 0.71

Scale fixed at 1

Exponential distribution
Loglik(model)= -89.9   Loglik(intercept only)= -90
        Chisq= 0.13 on 1 degrees of freedom, p= 0.71
Number of Newton-Raphson Iterations: 5
n= 100
```

Figure 6.0.5: Censoring scores

```
0.3648229 0.3578071 0.4209495 0.3086963 0.2385381 0.2595855 0.3086963 0.2876488 0.2806330 0.2385381 0.3016805 0.3086963 0.3227280 0.3437754
0.3928862 0.3507913 0.3648229 0.2455539 0.2245064 0.2245064 0.2245064 0.3227280 0.3788546 0.3016805 0.3297438 0.2666014 0.3999020 0.3578071
0.2595855 0.3227280 0.3507913 0.3858704 0.2385381 0.3507913 0.3718387 0.4209495 0.2666014 0.3999020 0.3578071 0.3507913 0.3648229 0.2946647
0.2174906 0.3718387 0.3858704 0.3227280 0.3928862 0.2385381 0.2315222 0.3578071 0.3928862 0.3718387 0.2806330 0.2595855 0.2876488 0.3788546
0.3858704 0.2946647 0.3227280 0.3086963 0.3928862 0.2595855 0.3718387 0.2666014 0.3999020 0.3227280 0.3858704 0.3648229 0.2736172 0.3227280
0.3297438 0.2245064 0.3157121 0.3578071 0.2315222 0.3999020 0.3648229 0.3648229 0.2455539 0.3928862 0.3367596 0.2595855 0.2385381 0.3858704
0.2876488 0.3016805 0.2315222 0.3858704 0.4069179 0.2736172 0.3648229 0.2946647 0.2174906 0.2946647 0.2595855 0.3086963 0.2806330 0.3928862
0.2455539 0.4209495
```

By rounding to 2 decimal points these scores , we get the following plot in figure

Figure 6.0.6: Risk scores

```
1] -0.2915366 -0.2859302 -0.3363884 -0.2466849 -0.1906201 -0.2074395 -0.2466849 -0.2298654 -0.2242590 -0.1906201 -0.2410784 -0.2466849 -0.2578978
4] -0.2747172 -0.3139625 -0.2803237 -0.2915366 -0.1962266 -0.1794072 -0.1794072 -0.1794072 -0.2578978 -0.3027496 -0.2410784 -0.2635043 -0.2130460
7] -0.3195690 -0.2859302 -0.2074395 -0.2578978 -0.2803237 -0.3083561 -0.1906201 -0.2803237 -0.2971431 -0.3363884 -0.2130460 -0.3195690 -0.2859302
3] -0.2803237 -0.2915366 -0.2354719 -0.1738007 -0.2971431 -0.3083561 -0.2578978 -0.3139625 -0.1906201 -0.1850136 -0.2859302 -0.3139625 -0.2971431
3] -0.2242590 -0.2074395 -0.2298654 -0.3027496 -0.3083561 -0.2354719 -0.2578978 -0.2466849 -0.3139625 -0.2074395 -0.2971431 -0.2130460 -0.3195690
8] -0.2578978 -0.3083561 -0.2915366 -0.2186525 -0.2578978 -0.2635043 -0.1794072 -0.2522913 -0.2859302 -0.1850136 -0.3195690 -0.2915366 -0.2915366
9] -0.1962266 -0.3139625 -0.2691108 -0.2074395 -0.1906201 -0.3083561 -0.2298654 -0.2410784 -0.1850136 -0.3083561 -0.3251755 -0.2186525 -0.2915366
2] -0.2354719 -0.1738007 -0.2354719 -0.2074395 -0.2466849 -0.2242590 -0.3139625 -0.1962266 -0.3363884
```

(6.0.7) which indicates negative correlation among the 2 mechanisms , meaning larger risk scores corresponding to smaller censoring scores ,so the patients who tend to live longer are censored. At this point let's also test ,something that we already know ,meaning that the exponential choice fits well with the data.

In order to do that we use the log-cumulative hazard plot in (1.57) equation.The plot is in the figure (6.0.8). Finally , we end this example by showing in figure (6.0.9) the difference among the survivor functions under the independent and informative censoring respectively for the first patient with age=52 and delta=0.01

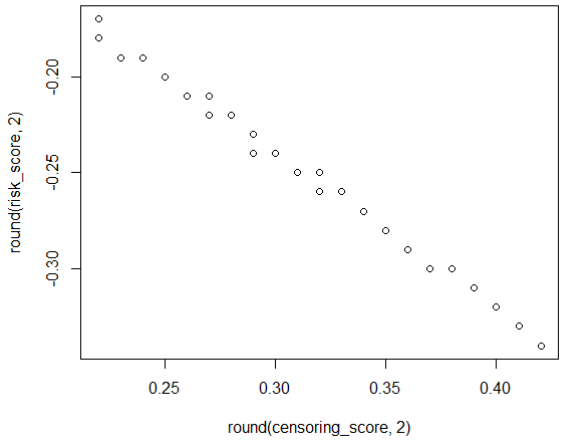
The survivor function under the informative censoring is calculated from the equation under the (3.8) expression and even a very small correlation(0.01) is resulting to non-negligible informative censoring.

Now for the second example we consider 1000 simulations of 100 patients and the failure mechanism T follows the exponential distribution with known λ.

We want to produce k percent of censoring under the assumption of dependence between the T and C mechanisms.

The censored random variable C is also considered following the exponential distribution with parameter μ but we have to find the specific parameter of the

Figure 6.0.7: Risk scores vs Censoring scores



censoring distribution in order to achieve the amount of k percent censoring.
That leads to the calculation of the probability $P(C < T)$.
This probability can be calculated with two ways.
The first one is the usual

$$P(C < T) = \int_0^\infty \int_0^t f_{T,C}(t, c)dcdt \,,$$

and the second way is based on the generalized law of total probability(continuous case)

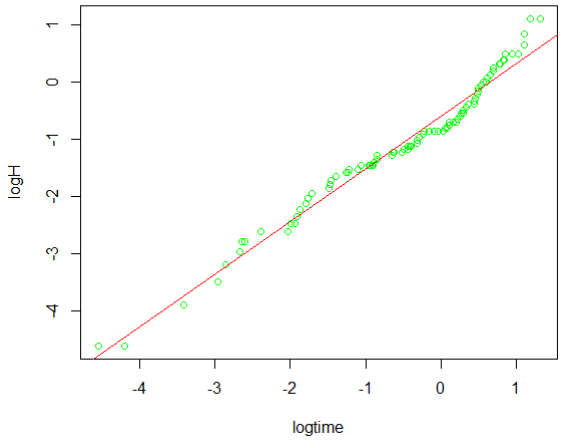
$$P(C < T) = \int_0^\infty P(C < T|T = t)f_T(t)dt \,,$$

For example if we assume that T and C are independent (non-informative censoring) ,

$$P(C < T|T = t) = P(C < t|T = t) = P(C < t) = 1 - exp(-\mu t) \,.$$

So we substitute this together with the density $f_T(t) = \lambda exp(-\lambda t)$ with the second way and after some calculations we conclude that $P(C < T) = \frac{\mu}{\mu+\lambda}$.
Further we set this probability equal to the desired amount of censoring k and we get the relation $\mu = \frac{\lambda k}{1-k}$.
Now for T and C dependent exponential random variables(informative censor-

Figure 6.0.8: Risk scores vs Censoring scores



ing) we will use the expression

$$P(C < T) = \int_0^\infty \int_0^t f_{T,C}(t,c)dcdt\,,$$

in order to find the appropriate event indicator under k percent of censoring when λ and δ are known.

Further we will examine if the informative censoring is negligible.

So in order to calculate this probability we will use the joint density of T and C (3.3).

As we discussed in chapter 3 we choose $B(t,\lambda) = i_\lambda^{-1/2} U_T(t,\lambda)$.

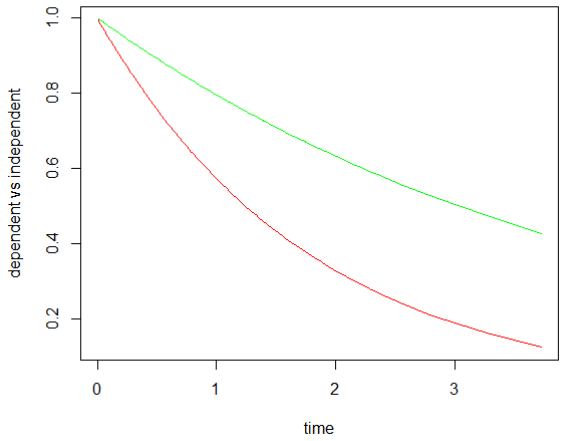
Further we calculate the quantities ,$i_\lambda^{-1/2} = Var(U_T(t,\lambda))^{-1/2}$ and $U_T(t,\lambda) = \frac{\partial}{\partial\lambda}\log f_T(t,\lambda)$.

So we get $U_T(t,\lambda) = (\log\lambda - \lambda t)' = 1/\lambda - t$.

As a result $Var(U_T(t,\lambda) = Var(1/\lambda - T) = 1/\lambda^2$ and the joint density (3.3) is written as

$$f_{T,C}(t,c) = \lambda exp(-\lambda t)\mu exp(-\mu c)\Big(1 + \delta(1-\lambda t)(1-\mu t)\Big)\,.$$

Figure 6.0.9: Dependent(red) vs independent censoring



By using this form to calculate the double integral we take after some calculations that

$$P(C < T) = 1 - \frac{\lambda}{\lambda+\mu} + \frac{\lambda\delta\mu}{(\mu+\lambda)^2} - \frac{2\lambda^2\mu\delta}{(\mu+\lambda)^3} = k\,.$$

This leads to the following polynomial equation

$$(1-k)\mu^3+\Big(3(1-k)\lambda-\lambda+\delta\lambda\Big)\mu^2+\Big(3(1-k)\lambda^2-2\lambda^2-\lambda^2\delta\Big)\mu+(1-k)\lambda^3-\lambda^3=0.$$

For $\delta = -0.2$, k=0.3,0.4,0.5,0.6 and $\lambda = 1$ we get a similar bias(for lambda) around 1 ,which tell us that we cannot ignore the presence of the informative censoring even with small portion of censoring.

In the third example we consider 2 groups of data , T_A has the $exp(\lambda)$,T_B has the $exp(2\lambda)$,C_A has the $exp(\mu)$ and C_B has the $exp(2\mu)$.
In this way the patients in group B tend to die and leave from the study due to uknown reasons faster ,so the uknown reasons might be that they experience a life threatening therapy.
For $\delta = -0.2$,$\lambda = 1$ and $k = 0.5$ gives the results in figure (6.0.10). So although that the biases of A and B alone are big , the bias of the treatment effect is quite small and it may not cause trouble if we estimate the treatment effect

Figure 6.0.10: Biases of lambda for A,B ,of mu for A,B and for the treatment with that order

```
$meanA
[1] 1.004077

$meanB
[1] 1.312841

$meanA2
[1] 0.9992949

$meanB2
[1] 2.611889

$meantreat
[1] 0.3046862
```

under the independent version.

We remind that according to (4.4) the bias of treatment is the bias of B minus the bias of A alone(for the λ) .

Also other values were tested and the results were quite similar .

R code

```
#chronic active hepatitis
#first group
library(survival)
library(muhaz)#package for non parametric hazards
data1=c(2,6,12,54,56,68,89,96,96,125,128,131,140,141,143,145,146,148,162,168,173,181)
    delta1=c(1,1,1,1,0,1,1,1,1,0,0,0,0,0,1,0,1,0,0,1,0,0)

group1KM=survfit(Surv(data1,delta1 ) ~ 1, conf.type="log-log")
summary(group1KM )
plot(group1KM)
group1NA=survfit(Surv(data1,delta1)~1,conf.type="log-log",type='fh')
 summary(group1NA)
 plot(group1NA)
 hazgroup1=muhaz(data1,delta1,max.time = 181)

 hazgroup1smooth1=muhaz(data1,delta1,max.time = 181,bw.grid=50,bw.method = 'global',b.cor='none')
 hazgroup1smooth=muhaz(data1,delta1,max.time = 181,bw.grid=100,bw.method = 'global',b.cor='none')
#larger b's give greater smoothings Kernel without the square
 plot(hazgroup1smooth)
 plot(group1KM)
 par(new=TRUE)
 plot(group1NA,xaxt='n',yaxt='n',col='green',xlab='time',ylab='survivor')
#figure 1.2.8
plot(hazgroup1,col='blue')
par(new=TRUE)
plot(hazgroup1smooth1,xaxt='n',yaxt='n',col='green')
par(new=TRUE)
plot(hazgroup1smooth,xaxt='n',yaxt='n',col='red')

##############################
#black ducks
blackducks=read.table('C:/Users/steli/Desktop/Survival_of_black_ducks.dat',skip=1)
time=blackducks[[2]]
status=blackducks[[3]]
age=blackducks[[4]]
weight=blackducks[[5]]
length=blackducks[[6]]
 library(survival)
Surv(time,status)
cox1=coxph(Surv(time,status)~age+weight+length)
coxc=coxph(Surv(time,status)~age+weight+length,ties='exact')#same result

summary(cox1)
anova(cox1) #weight has small contribute
coxw=coxph(Surv(time,status)~weight)#only with weight
summary(coxw)
anova(coxw)
cumhaz=basehaz(cox1,centered=FALSE)
cumhaz[c(1,2)]=cumhaz[c(2,1)]
```

```
plot(cumhaz,type='s',xlab='time',ylab='baseline cumulative hazard')
basesurv=matrix(0,ncol=2,nrow=length(cumhaz[,1]))
basesurv[,2]=exp(-as.matrix(cumhaz[,2]))
basesurv[,1]=cumhaz[,1]
plot(basesurv,type='s',xlab='time',ylab='baseline survivior function')
indv=matrix(c(blackducks[[4]],blackducks[[5]],blackducks[[6]]),ncol =3)
riskscore=NULL
for(i in 1:length(indv[,1])){
riskscore[i]=exp(indv[i,]%*%matrix(as.vector(cox1$coefficients),ncol=1))}
surv=array(0,dim=c(length(cumhaz[,1]),2,length(riskscore)))
for(i in 1:length(riskscore)){
surv[,2,i]=basesurv[,2]^riskscore[i]
surv[,1,i]=basesurv[,1]}
#risk adjusted method
survivoradjusted=matrix(0,nrow=length(cumhaz[,1]),ncol=2)
survivoradjusted[,2]=apply(surv[,2,],c(1),mean)
survivoradjusted[,1]=cumhaz[,1]
plot(survivoradjusted,type='s',xlab='time',ylab='adjusted survivor')

#Kaplan-Meier(unadjasted survivor)
Surv(time,status)
KM=survfit(Surv(time,status)~1,conf.type='log-log')
summary(KM)
plot(KM,col='red')
par(new=TRUE)
plot(survivoradjusted,type='s',xlab='time',ylab='adjusted-unadjusted survivor',
xaxt='n',yaxt='n',col='green')
#THIS GIVES HUGE DIFFERENCE
##############################################################################################
#actuarial estimator example
interval=c(1,2,3,4,5)
timeperiod=c(0,11,22,33,44,55)
d_j=c(7,6,4,2,3)
c_j=c(1,4,0,1,2)
n_j=c(30,22,12,8,5)
n_jnew=n_j-c_j/2
prob=(n_jnew-d_j)/n_jnew

actuarial=NULL
actuarial[1]=1
actuarial[2]=prob[1]
actuarial[3]=prob[1]*prob[2]
actuarial[4]=prob[1]*prob[2]*prob[3]
actuarial[5]=prob[1]*prob[2]*prob[3]*prob[4]
actuarial[6]=0
 actuarialestimator=actuarial[-6]
data.frame(interval,timeperiod,d_j,c_j,n_jnew,prob,actuarialestimator)
plot(timeperiod,actuarial,type='s',main='actuarial estimator against time')
#############
#cancer recurrence
cancer=read.table('C:/Users/steli/Desktop/Recurrence of bladder cancer.dat',skip=1)
colnames(cancer)=c('patient','time','status','treat','init','size')
head(cancer)

library (survival)
#First we take each variable seperately

cox0=coxph(Surv(time,status)~1,data=cancer)#null model
cox11=coxph(Surv(time,status)~treat,data=cancer)#treat model
cox12=coxph(Surv(time,status)~init,data=cancer)#init model
cox13=coxph(Surv(time,status)~size,data=cancer)#size model
anova(cox0,cox11)
anova(cox0,cox12)
anova(cox0,cox13)

#we choose init
```

```
#Then we include the treatment variable
cox21=coxph(Surv(time,status)~init+treat,data=cancer)
anova(cox11,cox21)



#Finally we consider the mixed term of index with the treatment in the final model
cox3=coxph(Surv(time,status)~init+treat+init:treat,data=cancer)
cox31=coxph(Surv(time,status)~init+treat,data=cancer)

anova(cox3,cox31)
step(coxph(Surv(time,status)~init+size+treat,data=cancer))#backward algorithm  under the AIC criterion



#############################################################################################
#hepatitis accelarated vs Cox
hepatitis2=read.table('C:/Users/steli/Desktop/Chronic_active_hepatitis.dat',skip=1)
hhh=read.table('C:/Users/steli/Desktop/Chronic_active_hepatitis.dat')
timeh=hepatitis2[[2]]
statush=hepatitis2[[3]]
treath=NULL
length(hepatitis2[[1]])
sum(hepatitis2[[1]]<2)
treath[(1:22)]=rep(1,22)
# the patients who took the drug are 1 and the others are zero
treath[(23:length(hepatitis2[[1]]))]=rep(0,22)
weib=survreg(Surv(timeh,statush)~treath,dist='weibull')
coxprop=coxph(Surv(timeh,statush)~treath)
summary(weib)
summary(coxprop)
surv=survfit(coxprop,newdata=data.frame(list(treath=c('control','Prednisolone'))))
muhat=weib$coefficients[1]
sigmahat=weib$scale
lambdahat=exp(-muhat)
time0=0:182
basesurv=1 - pweibull(time0, shape=1/sigmahat,scale=1/lambdahat)
alphahat=weib$coefficients[2]
survnew=basesurv^(exp(-alphahat/sigmahat))

plot(surv, col=c("red", "green"),xlab = 'time',ylab='survivor_functions')
lines(basesurv~time0,col='red')
lines(survnew~time0,col='green')
exp(-weib$coefficients[[2]])#acccelaration factor
#########################################################################
#TIME DEPENDANCY
individuals=c(rep(1,3),rep(2,3),rep(3,3),rep(4,3),rep(5,3),rep(6,3),rep(7,3),rep(8,3))
start=c(0,33,67,0,32,67,0,35,70,0,38,71,0,40,80,0,33,85,0,50,80,0,31,63)
stop=c(33,67,90,32,67,80,35,70,80,38,71,75,40,80,91,33,85,87,50,80,83,31,63,93)
treatment=c(rep(0,12),rep(1,12))
lbr=c(5.15,5.20,5.22,5.20,5.30,5.50,5.21,5.60,5.80,5.30,5.32,4.55,5.40,4.50,3.80,5.16,5.00,4.50,5.60,
3.20,4.25,5.23,5.60,3.80)
status=c(rep(0,2),1,rep(0,2),0,rep(0,2),1,rep(0,2),1,rep(0,2),0,rep(0,2),1,rep(0,2),1,rep(0,2),0)
liver=data.frame(individuals,start,stop,status,treatment,lbr)
summary(coxph(Surv(start, stop, status) ~ treatment+lbr),data=liver)
time=c(90,80,80,75,91,87,83,93)
status2=c(1,0,1,1,0,1,1,0)
treatment2=c(rep(0,4),rep(1,4))
lbr2=c(5.15,5.20,5.21,5.3,5.4,5.16,5.6,5.23)
summary(coxph(Surv(time, status2) ~ treatment2+lbr2))
```

```
timedep=survfit(coxph(Surv(start, stop, status) ~ treatment+lbr),data=liver)
plot(timedep)
cox=survfit(coxph(Surv(time, status2) ~ treatment2+lbr2))
plot(cox)


#########################################################
####transplant
transplant0=read.table('C:/Users/steli/Desktop/Time_to_death_while_waiting_for_a_liver_transplant.dat',
skip=1)
colnames(transplant0)=c('patient','time','status','age','gender','BMI','UKELD')

survreg(Surv(time,status)~1,data=transplant0,dist='weibull')#weibull model without prognostic variables
length(transplant0$status)
#transormations from zero's to one's and vice versa
kk=function(x){
  for(i in 1:length(x)){
  if(x[i]==0){ x[i]=1
  }else x[i]=0
  }
  return(x)
}
kk(transplant0$status)
#alternatively just put 1-delta for the censoring model
survreg(Surv(time,kk(transplant0$status))~1,data=transplant0,dist='weibull')
cens=survreg(Surv(time,kk(transplant0$status))~age+gender+BMI+UKELD,data=transplant0,dist='weibull')
 #full censoring model
risk=survreg(Surv(time,status)~age+gender+BMI+UKELD,data=transplant0,dist='weibull') #full
model
step(risk)  #backward algorithm
#assesing weibull
riskfit=survfit(Surv(time,status)~1,data=transplant0)
riskfit$surv
summary(riskfit)
#loghazard
logH=log(-log(riskfit$surv))
logH[1]=-10;logH[204]=10
#logtime
logtime=log(riskfit$time)
plot(logH~logtime,col='red')
lm(logH~logtime)
abline(lm(logH~logtime),col='green')
#plot risk scores against censoring scores
riskcoef=-risk$coefficients[2:5]/risk$scale
censcoef=-cens$coefficients[2:5]/cens$scale
age0=transplant0$age
gender0=transplant0$gender
trans0=transplant0$BMI
UKELD0=transplant0$UKELD
prognostic=matrix(c(age0,gender0,trans0,UKELD0),nrow=length(age0),ncol=4)
riskscores=NULL
censoringscores=NULL
for(i in 1:length(age0)){
  riskscores[i]=riskcoef%*%matrix(prognostic[i,],ncol=1)
  censoringscores[i]=censcoef%*%matrix(prognostic[i,],ncol=1)
}
plot(riskscores~censoringscores)




#############################
```

```
####weighted cox model with dependent censoring(previous transplant example)
cens#censoring weibull estimation
risk0=coxph(Surv(time,status)~age+gender+BMI+UKELD,data=transplant0)#time cox model (independent)
accel.cens.sore=NULL
#comptute censoring accelarated score
censcoef0=cens$coefficients[2:5]
for(i in 1:length(age0)){
  accel.cens.sore[i]=censcoef0%*%matrix(prognostic[i,],ncol=1)
}
#define stop-start data frame
transplant1=data.frame(transplant0,accel.cens.sore)
colnames(transplant1)=c('id','time','status','age','gender','BMI','UKELD','cens.score')


start.stop=tmerge(transplant1[1:6,],transplant1[1:6,],id=id,state=event(time,status),k=tdc(c(0,0,2,2,2,2)))
estim.S.cen=exp(-exp((log(start.stop$tstop)-cens$coefficients[1]-start.stop$cens.score)/cens$scale))
#survivor function of censoring for stopping times
weights=estim.S.cen^-1#weights for  stopping times

stat.stop.t=data.frame(start.stop,estim.S.cen,weights)
#cox model with weights only for 6 patients
coxph(Surv(time,status)~age+UKELD+BMI+gender+cluster(id),data=stat.stop.t)
risk0=coxph(Surv(time,status)~age+gender+BMI+UKELD,data=transplant0[1:6,])
#independent for 6 patients

##############################################################################
####hepatitis tests
test=read.table('C:/Users/steli/Desktop/Chronic_active_hepatitis.dat',skip=1)
colnames(test)=c('treatment','time','status')

#install.packages('survminer')
library(survminer)
library(Rcpp)
library(survival)
#plot ensures the proportional hazard assumption
ggsurvplot(survfit(Surv(time,status)~treatment,data=test),pval=TRUE,risk.table=TRUE,
legend.labs=c('Prednisolone', 'Control'),
legend.title='treatment',title='Kaplan-Meier_Curves_for_hepatitis_example')
survdiff(Surv(time,status)~treatment,data=test)#log-rank test
age=c(76, 25, 65, 28, 30, 37, 85, 90, 47, 36, 37, 38, 39, 40, 41, 42, 56, 44, 44, 46, 47,
 48, 49, 50, 44, 52,
 53, 54, 55, 56, 57,
 58, 68, 20, 39, 22, 23, 20, 37,48, 40, 33, 18, 25)
agegroup=as.factor(c('45+','45-','45+','45-','45-','45-','45+','45+','45+','45-','45-','45-',
'45-','45-','45-','45-','45+','45-','45-','45+','45+','45+','45+','45+',
'45-','45+','45+',
'45+','45+','45+',
'45+','45+','45+','45-',
'45-','45-','45-','45-','45-','45+','45-','45-','45-','45-'))


test1=data.frame(test,age,agegroup)
survdiff(Surv(time,status)~treatment+strata(agegroup),data=test1)#strata test
#plot for each strata for each group
ggsurvplot(survfit(Surv(time,status)~treatment+strata(agegroup),data=test1),pval=TRUE,risk.table=TRUE,
legend.labs=c('Prednisolone_45-',
 'Prensidole_45+','control_45-','control_45+'),legend.title='treatment-age',
title='Kaplan-Meier_Curves_for_hepatitis_example')



#QQ PLOT NEFRECTOMY DATA
nefr=read.table('C:/Users/steli/Desktop/Treatment_of_hypernephroma.dat',skip=1)
head(nefr)
library(dplyr)
nefr1=nefr
```

```
nefr2=select(nefr1,-2)

colnames(nefr2)=c('nephrectomy','time','status')
nefr2[1:10,,]
kaplan_meier2=survfit(Surv(time,status)~nephrectomy,data=nefr2)#kaplan meier for each group
summary(kaplan_meier2)
estquantile1=c(6,8,9,9,12,15,15,17,21);estquantile2=c(8,9,18,26,36,48,56,84,115)
plot(estquantile2,estquantile1,main='QQ_plot_for_2_groups',xlab='quantiles_in_group2',
ylab='quantiles_in_group1')
abline(lm(estquantile1~estquantile2),col='red')
percentile=c(10,20,30,40,50,60,70,80,90)
mat_QQ=matrix(c(percentile,estquantile1,estquantile2),nrow=9,ncol=3)
colnames(mat_QQ)=c('percentiles','group1','group2')

lm(estquantile1~estquantile2)#take the approximate accelaration factor
summary(survreg(Surv(time,status)~nephrectomy,data=nefr2))
  exp(- summary(survreg(Surv(time,status)~nephrectomy,data=nefr2))$coef[2])
##############################################################################3
#dialysis example with residuals
dialysis=read.table('C:/Users/steli/Desktop/Infection_in_patients_on_dialysis.dat',skip=1)
colnames(dialysis)=c('patient','time','status','age','sex')
cox_model=coxph(formula=Surv(time,status)~age+sex,data=dialysis)
mart_resid=residuals(cox_model,type='martingale')
marti_index_plot=ggplot(data = dialysis, mapping = aes(x = patient, y = mart_resid)) +
  geom_point() +

  labs(title = "martingale_vs_index_plot") +
  theme_bw() + theme(legend.key = element_blank())
 marti_index_plot+geom_hline(yintercept = 0)
 ggplot(data = dialysis, mapping = aes(x = age, y = mart_resid)) +
   geom_point() +
   geom_smooth()+
   labs(title = "martingale_vs_age_plot") +
   theme_bw() + theme(legend.key = element_blank())
 ggplot(data = dialysis, mapping = aes(x = sex, y = mart_resid)) +
   geom_point() +

   labs(title = "martingale_vs_sex_plot") +
   theme_bw() + theme(legend.key = element_blank())

resid_coxsnell =-mart_resid +dialysis$status
fit_coxsnell <- coxph(formula = Surv(resid_coxsnell, status) ~ 1,
                      data    = dialysis)
 cumulative_haz=basehaz(fit_coxsnell,centered=FALSE)
resplot=ggplot(data = cumulative_haz, mapping = aes(x = time, y = hazard)) +
  geom_point() +

  labs(title = "Cox-Snell_plot") +
  theme_bw() + theme(legend.key = element_blank())

resplot+geom_abline(intercept=0, slope=1,col='red')
data.frame(mart_resid,resid_coxsnell)
#############################################################
#competing risks KM
###asaur package
prostate=prostateSurvival[c(1,100,200,300.400,1000,2000,3000,4000,4500,5000,6000,7000:7050),4:5]
prostate1=within(prostate,{status_prostate_cancer=as.numeric({status == 1})
status_other_cause=as.numeric({status==2})})
#####kaplan meier for the event of interest
KM_prostate= survfit(Surv(survTime,status_prostate_cancer)~1,data=prostate1)
#KM for other causes
KM_other= survfit(Surv(survTime,status_other_cause)~1,data=prostate1)
time=KM_prostate$time
prob_other=1-KM_other$surv
 plot(KM_other$surv ~ time, type="s", ylim=c(0,1), lwd=2,
        xlab="Months_from_prostate_cancer_diagnosis",ylab='surv_other(green),surv_prostate(red)',
```

```
col="green")
 lines(KM_prostate$surv ~ time, type="s", col="red", lwd=2)
 length(prostate1$survTime)
 summary(KM_other);summary(KM_prostate)
################################## simulation
library(survival)

##################################
#one group analysis based on an exponential with rate 1
#(just an example not a proper simulation)
groupA=rexp(100,1)# survival times
timeA=sort(groupA)
idA=1:100
age=round(runif(100,30,60))#explanatory variable
indi=rbinom(100,1,1/2)#indicator variable
table=data.frame(idA,timeA,indi,age)
#bias for an individual
biasA=function(delta,variable){
   biasA=delta*(sum(term2*variable)-apply(group_A,2,sum)[4])/apply(group_A,2,sum)[2]
   return(biasA)}
term1=(1-indi)*timeA
term2=(timeA)^2
modelA=survreg(Surv(timeA,1-indi)~age,dist='exponential')#censoring mechanism
modelB=survreg(Surv(timeA,indi)~age,dist='exponential')#event mechanism
term3=NULL
 for(i in 1:100){ term3[i]=exp(-modelA$coefficients[2])*age[i]}
biases=NULL
for(i in 1:100){
   biases[i]=biasA(0.2,term3[i])}
group_A=data.frame(idA,timeA,indi,term1,term2,term3)
100-sum(indi)# number of  censored patients
#assesing exponential via Kaplan-Meier
fit=survfit(Surv(timeA,indi)~1)
logH=log(-log(fit$surv))
logtime=log(fit$time)
plot(logH~logtime,col='green')
lm(logH~logtime)
abline(lm(logH~logtime),col='red')
#risk scores vs censoring scores
censoring_score=NULL
for(i in 1:100){
   censoring_score[i]=-modelA$coefficients[2]*age[i]}
risk_score=NULL
for(i in 1:100){
   risk_score[i]=-modelB$coefficients[2]*age[i]}
plot(round(risk_score,2)~round(censoring_score,2))
plot(biases)
plot(censoring_score)
lambda=exp(-modelA$coefficients[1])
###plot the survival curve
basesurv=1-pexp(timeA,rate=lambda)
indsurv=basesurv^(exp(-modelB$coefficients[2]*age[1]))
depsurv=basesurv^(exp(-modelB$coefficients[2]*age[1]+biasA(0.01,term3[1])))
plot(depsurv~timeA,type='l',col='red',xlab='time',ylab='dependent_vs_independent')
lines(indsurv~timeA,type='l',col='green')
##################################
##1000 simulations of 100 patients
#newton rapson to calculate the rate of the censoring process
#for a specific amount of censoring under the informative model
newtonsim=function(delta,lambda,k){
   ep=10
   xold=0
   while(ep!=3){
      xnew=xold-((1-k)*xold^3+(3*(1-k)-1+delta)*lambda*xold^2+(3*(1-k)-2-delta)*lambda^2*xold+
(-k)*lambda^3)/(3*(1-k)*xold^2+2*(3*(1-k)-1+delta)*lambda*xold+(3*(1-k)-2-delta)*lambda^2)
      if(abs(xnew-xold)<10^-4){
```

```
      ep=3} else{

        xold=xnew}
  }
  return(xnew)
}

sim=function(delta,lambda,k){
  biaslambdaA=NULL#bias lambda for A
  biaslambdaB=NULL#bias lambda forB
  biasmuA=NULL#bias mu for A
  biasmuB=NULL#bias mu for B
  biastreat=NULL#bias of treat under the together situation
  for(i in 1:1000){
  TA=rexp(100,lambda) #events for A
  TB=rexp(100,2*lambda)#events for B
  CA=rexp(100,newtonsim(delta,lambda,k))#censoring for A
  CB=rexp(100,2*newtonsim(delta,2*lambda,k))#censoring for B
  A=data.frame(TA,CA)
  B=data.frame(TB,CB)
  min_timeA=apply(A,1,min) #min of TA,CA
  min_timeB=apply(B,1,min) #min of TB,CB
  indicatorA=as.numeric(TA<CA) #event indicator for A
  indicatorB=as.numeric(TB<CB)#event indicator for B
#estimation of lambda for A(under the independent model)(opposite from the coef)
  modelAT=survreg(Surv(min_timeA,indicatorA)~1,dist='exponential')
#estimation of lambda for B(under the independent model)(opposite from the coef)
  modelBT=survreg(Surv(min_timeB,indicatorB)~1,dist='exponential')
  biaslambdaA[i]=as.numeric(modelAT$coefficients)+lambda #bias lambda for A
  biaslambdaB[i]=as.numeric(modelBT$coefficients)+2*lambda #bias lambda for B
#estimation of mu for A(under the independent model)(opposite from the coef)
  modelAC=survreg(Surv(min_timeA,1-indicatorA)~1,dist='exponential')
#estimation of mu for B(under the independent model)(opposite from the coef)
  modelBC=survreg(Surv(min_timeB,1-indicatorB)~1,dist='exponential')
  biasmuA[i]=as.numeric(modelAC$coefficients)+newtonsim(delta,lambda,k) #bias mu for A
  biasmuB[i]=as.numeric(modelBC$coefficients)+2*newtonsim(delta,2*lambda,k) #bias mu for B
  treatg=c(rep(0,100),rep(1,100)) #treatment variable when we put the 2 groups together
#estimation under the independent model of treament(opposite of coef)
  modelg=survreg(Surv(c(min_timeA,min_timeB),c(indicatorA,indicatorB))~treatg,dist="exponential")


  biastreat[i]=2*lambda+modelg$coefficients[2]-biaslambdaA[i]
  }
  meantreat=mean(biastreat)
  meanA=mean(biaslambdaA)
  meanB=mean(biaslambdaB)
  meanA2=mean(biasmuA)
  meanB2=mean(biasmuB)
  list(biaslambdaA=biaslambdaA,biaslambdaB=biaslambdaB,biasmuA=biasmuA,biasmuB=biasmuB,
       biastreat=biastreat,meanA=meanA,meanB=meanB,meanA2=meanA2,meanB2=meanB2,meantreat=meantreat)}
  #upperAmu
  #upperAlambda
  #upperBmu
  #upperBlambda
  #lowerAmu
  #lowerBmu
  #lowerAlambda
  #lowerBlambda
```

Bibliography

[1] Modelling Survival Data in Medical Research third edition , David Collett (2015).

[2] Statistical Models and Methods for Lifetime Data Second Edition , JERALD F. LAWLESS(2003).

[3] Applied Survival Analysis Using R , Dirk F. Moore(2016).

[4] APPLIED Survival analysis, Regression Modeling of Time-to-Event Data Second Edition DAVID W.HOSMER , STANLEY LEMESHOW ,SUSANNE MAY.

[5] Applications of a Parametric Model for Informative Censoring ,Fotios Siannis (2004).

[6] Application of inverse probability weights in survival analysis ,Guoqiao Wang,Inmaculada Aban (2015).

[7] Sensitivity analysis for informative censoring in parametric survival models ,FOTIOS SIANNIS,JOHN COPAS,GUOBING LU (2005).

[8] Informative Cencoring in relative survival Anamarija Rebolj Kodre ,Maja Pohar Perme (2013).

[9] Fitting Parametric Survival Models With Time- Dependent Covariates TROND PETERSEN (1986) .

[10] Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals, Stanley Xu,Colleen Ross,Marsha A. Raebel,Susan Shetterly,Christopher Blanchette, David Smith (2009).

[11] Weighting Regressions by Propensity Scores, David A. Freedman,Richard A. Berk (2008).

[12] On Estimation in Relative Survival,Maja Pohar Perme et al.,(2012).

[13] The relative survival rate :A statistical methodology ,Fred Ederer et al ,(1961).

[14] Instructions to ibm 650 programmers in processing survival computations, methodological note no. 10, end results evaluation section. Technical report, National Cancer Institute, Bethesda MD Ederer F, Heise H. (1959).

[15] Simple Parametric and Nonparametric Models for Excess and Relative Mortality,Per Kragh Andersen and Michael Væth ,(1989).

[16] The Relative Survival Rate: A Statistical Methodology, pp. 101–121. Bethesda, Maryland, U.S.: National Cancer Institute Monograph 6.Ederer, F., Axtell, L. M., and Cutler, S. J. (1961).

[17] Martingale Based Residuals for survival models by Terry Therneau, Patricia Grambsch, and Thomas Fleming Technical Report Series 40 (April 1988) .

[18] The estimation and modeling of cause-specific cumulative incidence functions using time-dependent weights,Paul C. Lambert,The Stata Journal (2017).

[19] Estimating and modeling relative survival,Paul W. Dickman,Enzo Coviello,The Stata Journal (2015).

[20] Estimating net survival: the importance of allowing for informative censoring,Coraline Danieli et al,Statistics in medicine(2012).

[21] Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics 1982.

[22] Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Statistics in Medicine 2007; 26(10):2214–2228.