Universitetet i Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study programme/specialisation: Master of Mathematics and Physics	Spring semester, 2019 Open/~~Confidential~~
Writer: Isra Khawar	…Isra Khawar… (signature of writer)

Faculty Supervisor:
Jan Terje Kvaløy
External Supervisor(s):
Hartwig Kørner

Thesis title:

Overall and relative Survival for Cancer Patients

Credits (ECTS): 60

Keywords: Overall survival , Kaplan-Meier estimate , Cox's proportional hazard model , Incurable cancer patients , Net survival , Relative survival , Excess hazard ratio, Univariate and Multivariate analysis , Relsurv package in R	Number of pages ……57…… Stavanger, 14/2019 date/year

Title page for Master's Thesis
Faculty of Science and Technology

University of Stavanger

Overall and Relative Survival of Cancer Patients

Isra Khawar

Department of Mathematics and Natural Science

University of Stavanger

Submission Date: June 2019

Supervisor: Jan Terje Kvaløy

Preface

I would like to thank my supervisor Jan Terje Kvaløy for his continued guidance and invaluable feedback throughout the year. I would also like to acknowledge my co-supervisor professor Hartwig Kørner. I am very thankful to him for letting me analyze data on non-curable colorectal cancer patients, and for very helpful feedback during the work.

Abstract

In this thesis, basic concepts of survival analysis such as censoring, truncation and survival functions are described. Measures of survival (i.e overall survival, net survival and relative survival ratio) and regression models such as Cox regression for overall hazard ratios and excess hazard regression model for excess hazard ratios are discussed. Cox regression model estimates the overall risk(hazard) whereas excess mortality provided by relative survival estimates the risk due to cancer. Kaplan-Meier curves are used to estimate the survival curve, to estimate regression coefficient, partial likelihood estimate is used. The main focus is to study the comparison between overall survival and relative survival ratio and apply this on non-curable colon and rectum data, derived from a research project on patients who received non-curative treatment due to incurable disease or other reasons preventing curative surgery. The data is obtained from Cancer Registry of Norway, Norweigan patient registry and population data from Statistics Norway between 2008 and 2015. The results provided by comparison show how much change the risk of death is, if death only because of cancer is considered and when other causes of death are involved. Regression analysis is done in two ways, 1. univariate analysis in which each covariate affect the analysis individually and 2. multivariate analysis in which all covariates togther affect the analysis. The software R is used for analysis and to plot survival curves and other graphs used.

Contents

1 Introduction to Survival Analysis

Analyzing time-to-event (survival times)data is called survival analysis. The time to event data shows the time span from well defined time origin til the well defined end point of interest (event). The terms survival analysis and survival data are in generally used more often than time-to-event analysis and time-to-event data but term "Time to event" is more clear and precise to use. The time origin and end point must be well defined. For instance, in study of a particular type of cancer, the time point of diagnosis of that type of cancer is chosen to be time origin and the death due to that particular cancer would be the end point. Or a study might follow people from birth (time origin) until the occurrence of a disease(end point). This is how the time length can be measured. The time to event data is usually collected prospectively in time such as data is collected for clinical experiment or data from potential cohort study. Sometimes data can also be collected retrospectively through accessing medical records or by interviewing patients who have that certain disease.

Time to death is the event of interest in most of the medical studies. But in cancer the time between a response to treatment and reappearance or disease -free time is another essential measure. Also the event and duration of observation is important to express. For example time interval between confirmed response and first relapse of cancer. The time to event data can include survival time, response to a given treatment, patient's attributes associated to response, survival and disease growth.

A particular problem linked to time to event analysis come to light from the fact that not all the individuals have experienced event so eventually survival times will not be known for a part of the study group. For example the individuals could have different events such as in the above example where the event of interest is death due to cancer but the patient died due to accident or they may drop out of a study. The other feasibility is that the study might finish at a certain point of time and individuals have not had their event yet and thus their event time will not have been noticed. This is known as censoring. These incomplete observations needs to be handled in a proper way. This is why 'special 'techniques are needed in time-to-event analysis. Additionally time to event data are skewed and seldom normally distributed,

therefore simple techniques established on normal distribution cannot be used accurately.

Observed survival and relative survival are two analysis which I am going to describe later in my thesis where I will be using data consisting of colon and rectum patients taken from the Norwegian registry named as the Norwegian register for cancer of the colon and rectum. My major focus will be on non curative colon and cancer patients and their survival rate after applying both the mentioned analysis.
In observed survival, the risk of death is not only considered cancer but other causes like death due to heart attack is also included but in relative survival we only take into account death due to cancer. Other causes of death does not affect the survival of cancer patients.
The introduction is based on [1, 2, 3]

2 Basic Concepts in Survival Analysis

2.1 Censoring

A main source for this subsection is [4]. Apart from survival analysis censoring may arise in other applications, whereby not all survival data hold censored observations. However, this is one such topic that unites a lot of applications to survival analysis because censored survival data are so common and censoring needs special treatment. Censoring has many forms and there are different causes of occurance of censoring . The primary difference is in between left censoring and right censoring.

2.1.1 Right Censoring

In survival data T is the time from start of observation until an event happens and some cases become right censored as observation breaks off before the event arise. Accordingly, if T is said to be the event as person's age at death(in years), the event is right censored at age 50 if you may only know that T>50. This concept is also not confined to event times only. The income is right censored at $75,000, if the only thing you know is that a person's income is more then $75,000 per year.

Example. Figure 2.1 shows data from a study in which all the persons go through heart surgery at time 0 and followed up to 3 years. The horizontal axis shows time in years after surgery and horizontal lines tagged A to E represents different person. The vertical line at 3 is the point at which we stop following the patients. An X specify that death occured at that point in time. Deaths occured at point 3 or before time 3 is observed and hence are uncensored but on the other hand, deaths occuring after time point 3 are not observed thus are

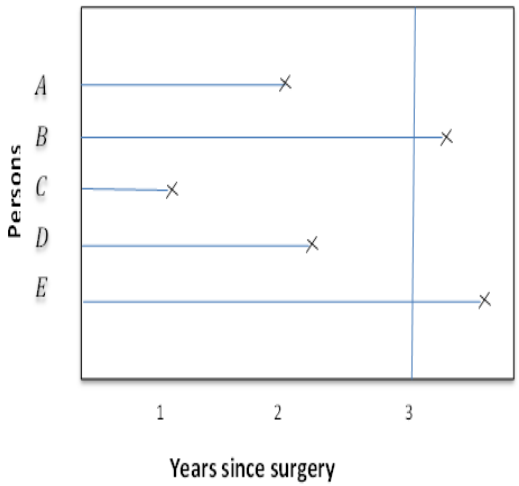


Figure 2.1: Image showing right censoring

censored at time 3. Consequently, A,C and D are uncensored, while B and E are right-censored.

2.1.2 Left Censoring

Left censoring occur when we only know that T is less than some value. This concept is not only applicable for event time but any kind of variables. For survival data left censoring most probably occur when some of the individuals may have already experienced the event when observing a sample at a time is just started.

Example. In the study of menarche(the beginning of menstruation) if you start observing girls of age 12 and you get to know some of the girls have already started menstruating so the age of menarche is called to be left censored at age 12 except if you can get informauion on the starting date for those girls.

2.1.3 Interval Censoring

Interval censoring is more common then left censoring. Both left censoring and right censoring together makes interval censoring. When you only know about variable T is $a<T<b$ for some values of a and b then T is interval censored. Interval censoring arise in survival data when the observations are made at specific time points and retroactive information on the exact timing of event cannot be achieved.

Example. For HIV infection, sample of people is followed.The time of infection between 2 an 3 would be interval censoring if a person who is not infected at

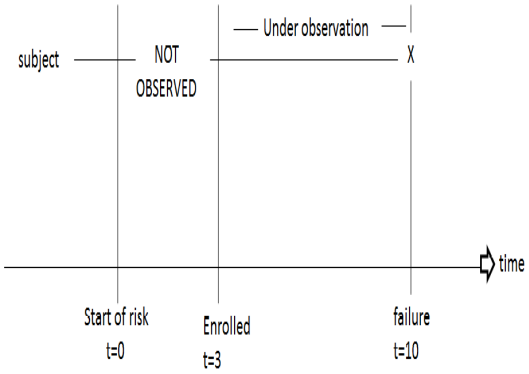


Figure 2.2: Image representation left truncation

the end of year 2 is then found to be infected at the end of year 3.

2.2 Truncation

For subsection (2.2), sources are [5, 6, 7].
Another factor which affects the survival data by giving rise to incomplete observations is truncation. Interval over which the subject was not observed but is not failed as well, is known as truncation. The statistical difficualty is if the subject had failed, he or she have never been observed. In truncated survival time data, survival times are excluded systematically from one's sample. The following are three types of truncation from which left truncation is most common.

2.2.1 Left Truncation

The period of ignorance in left truncation widen from on or before the beginning of study(at t=0) to sometime after time t=0. The Figure 2.2 explains the left truncation. The subject is not observed for some time after the start time but come under observation. Later if they have not had the event. This is why left trucation arise as we confront a subject who enrolled sometimes after the onset of risk.
This subject is only added to the study if he or she has not failed earlier,before the threshold. For example only those individual who survive the initial stage of myocardial infarction and reach the hospital will be included in the study. If an individual has been admitted to the hospital and is added to the study where the time t=0 is the time of infarction. For different patient it may happen at different times but those patients will never be entered into study if they die before reaching to the hospital."Delayed entry" is sometimes used for left truncated data.

2.2.2 Interval Truncation

Interval truncation is just an adoption of left truncation where an individual enters in the study at time zero but disappear for some time and report back to the study generating a gap in between observation.This is what the issue is that individual could have died when he or she disappear and can never report back.

2.2.3 Right Truncation

In this case only those individuals are added to the study who have experienced the exit event by some specific date but there is a point after which the subject who hasn't experienced exit event is not observed anymore and consequently long survival times are excluded systematically .

2.3 Functions of Survival Times

This subsection is based on references[1, 3, 8, 9, 10].
Before analysing the survival data, some related functions needs to be described such as survival function, density function, hazard function and cumulative hazard function from which survival and hazard functions are of particular interest[8]. In traditionally established statistical models, density and cumulative distributions are used but due to the incomplete observations in survival data(censored and truncated data) these standard functions are not appropriate. So survival and hazard functions are considered more suitable.

2.3.1 Survival Function

Survival function is defined to be the probability of surviving beyond a specified time t. Survival function is denoted by S(t) where $0 < t < \infty$.The formula is given in (2.1).

$$S(t) = P(T \geq t) = 1 - F(t),\ t > 0 \tag{2.1}$$

where T is the random variable under study(time to event) t is a fixed numberand F(t) is the cumulative distribution function of T[9]. S(t)= 1 at t=0 and S(t) = 0 at t=∞.The graph of the survival functionS(t) is called the survival curve which begins at S(t)=1 and as t increases to ∞, S(t) decreases to 0. The survival curve can be estimated by the Kaplan-Meier method (and will be disscussed later). See Figure 2.3 for an example of survival curve.

2.3.2 Density Function

The probability density function $f(t)$ is defined as the rate of event every unit time[1]. We can calculate the density function by taking the derivative of the survival function, which is as follows:

$$\frac{d}{dt}S(t) = \frac{d}{dt}(1 - F(t))$$

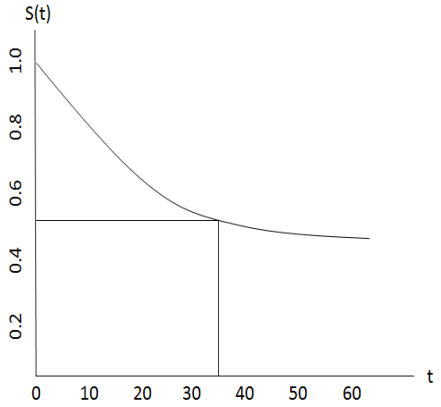


Figure 2.3: An example of a survival curve

and from the definition of distribution function we get:

$$\frac{d}{dt}S(t) = -f(t)$$

$$f(t) = -\frac{d}{dt}S(t) \tag{2.2}$$

The equation (2.2) shows the relation of probability density function with survival function.

Probability density function, also known as unconditional failure rate[10] is intuitively defined as:

$$P(t \leq T < t + \Delta t) \approx \Delta t f(t)$$

Equation (2.3) is the traditional matematical definition of probability density function as a limit.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, t > 0 \tag{2.3}$$

The definition described by the formula in equation (2.3) is well illustrated by Figure 2.4 , which shows that the probability of an observation lies in interval $(t, t + \Delta t)$ is fairly approximated by the area of rectangle with sides of length Δt and $f(t)$ [9]

2.3.3 Hazard Function

To understand survival analysis, hazard function is an important concept which we can say is a kind of density function f(t). The difference is that hazard function is conditional while density function is an unconditional probability.

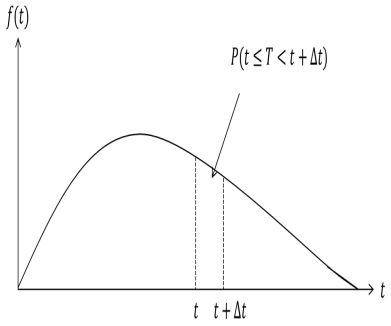


Figure 2.4: Graph of probability density function

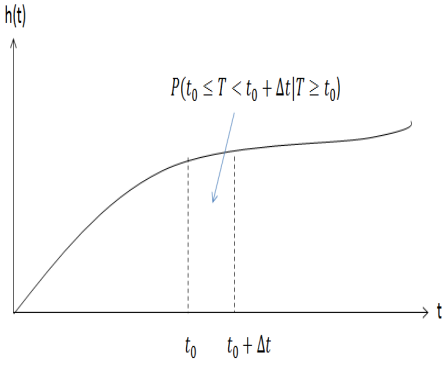


Figure 2.5: Hazard Functon

Hazard function also known as instantaneous failure rate is defined as the probability that the event lies in an interval $(t, t+\Delta t)$, given that it has not happened prior to t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, t > 0 \tag{2.4}$$

$$h(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t)$$

Equation (2.4) explains that the probability of a person who dies in a short interval$(t, t + \Delta t)$ where the individual has already survived the time t. [3]
The graphical interpretation of hazard function can be seen in Figure 2.5

2.3.4 Cumulative Hazard Function

By taking integral of hazard function we get cumulative hazard function which is comparatively easier to estimate non parametric models than hazard and

density functions.That is why it is considered to be an important function[9]. The formula is given in equation (2.5)

$$H(t) = \int_0^t h(x)dx, t \geq 0 \tag{2.5}$$

2.4 Parametric vs Semi-parametric vs Non parametric

Subsection (2.4) has references[1, 12, 14].
In survival analysis parametric and non-parametric approaches are used to estimate the quantities describing survival data so it is important and necessary to describe these approaches before moving ahead.

2.4.1 Parametric Approach

In parametric approach we assume to have a distribution with particular type of parametric form for example normal distribution, weibull distribution etc. We make assumptions on functional form that are used in distribution we assume and maximum likelihood procedure can be used to estimate the parameters. The most common assumption we made for parametric model is that data follow some specific probability distribution

2.4.2 Non Parametric Approach

This method of estimation does not assume any specific distribution. In the distribution of survival times setting non parametric method is quite simple and useful for example to abridge the survival data and to make simple comparisons but for the complex condition, it is difficult for these methods to deal with such situation[1][12]. Non parametric methods are generally used more to analyse the survival data as it is less restricted then the parametric method. We make few assumptions about the observed data. I am going to use the most common method for non parametric estimation of the survival function which is Kaplan-Meier estimator.

2.4.3 Semi-Parametric Approach

Semi-parametric method consists of models with both parametric and non parametric elements. It also focuses on effects of the covariates. The most well known example of semi parametric model is Cox proportional hazard model. (which we will disscuss later) [14]
In this thesis I will explain some well known examples of non parametric and semi parametric models.

2.5 Kaplan-Meier Estimate(KM)

The subsection (2.5) has references [2, 10, 11, 13]. Kaplan-Meier includes computing probability of survival within a small interval of time. It is also known

as "Product Limit estimate"(PL). As we know about the censored and truncation factors which give rise to incomplete observations and we cannot eliminate them as each individual, as long as they are event-free, contribute information to the calculation and also we do not want to make our sample size smaller by excluding those individuals. Also excluding the censored cases will lead to biased estimator. Kaplan-Meier is considered the simplest way of estimating probabilities of survival for both censored and uncensored survival times.

We calculate the probability of survival at distinct times by dividing the number of subjects survived to the number of subjects at risk. Where those subjects who are censored are not considered as "at risk" therefore are not added to the denominator n. Mathematically we can express the estimated survival probability at a certain time point as: $1 - \frac{d}{n}$ where,

d =no of subjects died / no of events

n =no of subjects live at the start of the day

The total probability of survival or cumulative probability in the period of follow up is obtained by multiplying all the probabilities of survival at all specific times within specified interval.

To make it more understandable, lets say the distinct event times are $t_1 < t_2 < t_3 < t_4 < t_5 < \ldots < t_j$ where j patients have the events within the follow up period and at time t_1 the probability is p_1. At time t_2 the probability is p_2 after the patients have survived time t_1, and at t_j the probability is p_j after surviving time t_{j-1}.

The probability of surviving beyond time t_j is estimated as:

$$\hat{S}(t_j) = \hat{P}(T > t_j) = p_1.p_2.p_3...p_{t_j}$$

The Kaplan- Meier estimate could then be found by:

$$\hat{S}(t_k) = \prod_{t_k < t} S(t_{k-1})(1 - \frac{d_k}{n_k}) \qquad 1 < k < j \tag{2.6}$$

In equation (2.6) $S(t_{k-1})$ is the probability of survival computed at time t_{k-1} , $d_k$is the number of subjects died at t_k and $n_k$is the number of subjects alive just beforet_k, where S(0)=1

For example to find the probability of survival of a patient two days after kidney transplant could be found as the number of patients survived the day one$S(t_{k-1})$ multiplied by the probability of patient survived the second day given that patient has survived first day$(1 - \frac{d_k}{n_k})$. The second one is the conditional probability that means for the patients/subjects to remian in the study they should have survived the first period of time.

Example. Figure 2.6 illustrates the survival function drawn by taking a hypothetical data of group of patients entered in clinical trial receiving anti-retroviral therapy for HIV infection. The data shows the time of event i.e death, occured among the patients that is:

$6, 12, 21, 27, 32, 39, 43, 43, 46^*, 89, 115^*, 139^*, 181^*, 211^*$

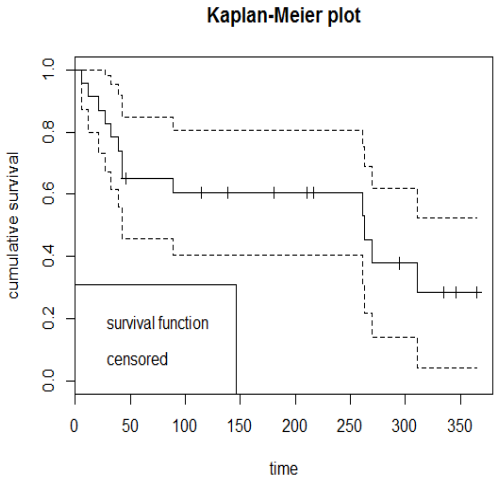


Figure 2.6: Plot of Kaplan-Meier estimates group of patients receiving ARV therapy

$, 217^*, 261, 263, 270, 295^*, 311, 335^*, 346^*, 365^*$(* means right censored observation). From Figure 2.6 we can see the estimated probability is the step function that remain unchanged even if there is a censored observation in between. The X-axis (horizontal lines) show the time past after entry into studies and the Y-axis (vertical lines) shows the estimated survival probabilities. The time t when the cumulative probability is 0.5 i.e S(t) = 0.5 is called median survival time which according to this example is t=263. We can use different statistical programs to plot Kaplan-Meier curve such as SPSS, R, Sigma plot etc. Here in our example we have used R to plot the curve.

2.6 Comparison of Kaplan-Meier Estimates

The citations for the following subsection (2.6) are [8, 11, 13].
The Kaplan-Meier curves can be compared to see the difference between them. For example we can check if a particular treatment, lets say A given to patients is less or more effective then the new treatment B given to other group of patients. The survival patterns in the survival curves such as horizontal and vertical gaps can be compared. The gap in horizontal direction means one from the two groups took longer time to experience the event (death) and the gap in vertical direction means that one group had survived more then the other group. That means both the directions are two sides of same result so we don't need to check both the directions at the same time. In clinical trials comparison of survival curves are particularly taken into account. The difference must be statistically significant otherwise both the estimates are considered same.

The method we are going to use to compare the survival curves is "log-rank test" which is the most common method. In each group this method calculates the chi square$\left(X^2\right)$ of each event time and sums the result. And the final chi-square is obtained by adding all the results from each group to compare the complete curves.

Log Rank Test

In this method we compare the curves of two different groups of patients and test whether the difference between their survival times are statistically different or not using statistical hypothesis test by testing a null hypothesis. Null hypothesis states that there is no difference between the curves regarding survival. We calculate the log rank test statistics as follows:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{2.7}$$

where
0_1 =Total number of observed events(patients died) in group 1
0_2 =Total number of observed events in group 2
E_1 =Total number of expected event(death) in group 1
E_2 =Total number of expected event (death) in group 2
The total number of expected events in any of the group are the sum of all the expected events calculated at different times (at the time of each event) and the expected number of events at the time of each event in a group is computed by multiplying the risk of event at that time with all the patients alive at the start of an event in that group (i.e lets say the total number of patients are 46, 23 in each group and at day 6 the risk of event is calculated as 1/46=0.0217395 where all the patients are alive at the start of the day and 1 died,hence in group 2 the expected number of event at day 6 would be 23x 0.0217395= 0.5). Once we get the sum of all expcted events in group 2 (E_2) we can get (E_1) by subtracting $E_2$from $O_1 + O_2$. Lets take an example.

Example. Following the previous example for Kaplan-Meier plot and name it as group1(ART therapy), lets take another hypothetical data for the patients entered in clinical trail for receiving a new Ayurvedic therapy for HIV infection:

$9, 13, 27, 38, 45^*, 49, 49, 79^*, 93, 118^*, 118^*, 126, 159^*, 211^*,$

$218, 229^*, 263^*, 298^*, 301, 333, 346^*, 353^*, 362^*$($*$ right censored observations) and name it as group 2 (Ayurvedic therapy) . For these two groups of patients, Figure 2.7 illustrates the difference between the survival curves of these two groups. The Figure 2.7 is constructed by using a package in statistical program R for the given data in the examples.
It can be seen from the Figure 2.7 that there is no big difference between the two curves. But to check the significant difference accuratly, we calculate test statistics (which is computed by using the formula in equation (2.7)) and compare it with the critical value (the value from chi-suare table) for one degree of freedom. If the test statistical value is less then the critical value, we accept the

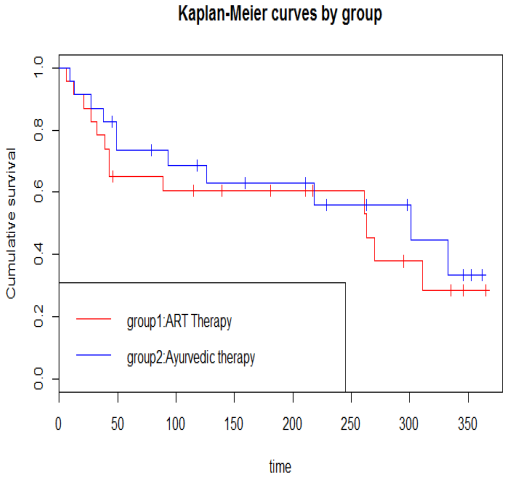


Figure 2.7: Plots of Kaplan-Meier estimates of two different groups of patients

	N	observed	expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
group 1	23	13	11.5	0.184	0.357
group 2	23	11	12.5	0.170	0.357
chisq= 0.4 on 1 degree of freedom , p= 0.5					

Table 1: Result of log rank test in statistical programming R

null hypothesis. The second method (which we have used) to draw the significance is using the statistical package in R for log rank test where the chisquare value is used to calculate p value which is then compared to the significant level ($P = 0.5$ in our case). In our example, the Table 1 shows the result of formula applied in R for the log rank test. According to the result we get chisqr =0.4 with p value= 0.5,we can see that the p value is the same as the significant level which means there is no significant difference between group 1 and group 2 (we accept the null hypothesis). The overall result of both the therapies are same regarding the survival.

2.7 Cox Proportional Hazard Model

The following subsection (2.7) is based on the references [1, 9, 10, 16, 17]. The Cox model is a semi parametric model. No matter if there is censored data or time-to-events are descrete or continuous, Cox model is widely used in survival data analysis.

2.7.1 Proportional Hazards Assumption

The one important property or we can say the prime assumption of Cox model is the proportional hazards, defined as the two hazard functions $h_1(t)$ and $h_0(t)$ from two independent distributions are proportional if:

$$h_1(t) = \psi^x h_0(t),\ \forall t > 0,\ x = 0, 1,\ \psi > 0 \tag{2.8}$$

where ψ is the positive proportionality constant that does not depend on t and $h_0(t)$is a baseline hazard. The proportional hazards would not be used for all the cases. For example if we take two groups, women and men and let the hazard function be age-specific mortality for these groups. Since it is widely known that in all ages men have larger mortality then women, hence we can plausibly assume the proportional hazards for this case which would mean that relative advantage for women in all ages is equaly large then men. This assumption must always be carefully examined and this could be done by using Schoenfeld residuals. and will descibe later in the thesis.

2.7.2 Cox's Proportional Hazard Model

Sometimes it is interesting to know if a person's attributes are associated with the occurrence of a certain disease. For example in public health research, it is checked whether the characteristics like exalted cholesterol level, cigarette smoking or having a history of heart disease are associated to the expansion of cardiovascular disease. These characteristics/attributes are called covariates or risk factors. The effect of such factors on time to event can be modelled by Cox model. On the other hand, hazard is the probability of experiencing the event given that patients have survived certain period of time.
The Cox model is a regression model for time-to-event data assuming that the covariates will affect the survival times. It enables to test the difference between survival times of different groups of patients allowing other factors(i.e covariate) to be taken into account. The proportional hazard assumption is the base for Cox's regression model. Using $\beta = log(\psi)$ if we rewrite equation(2.8), we can estimate hazard function as:

$$h_x(t) = h_0(t)e^{\beta x}, t > 0,\ x = 0, 1;\ -\infty < \beta < \infty \tag{2.9}$$

which is the form of Cox model of two groups, where t is the survival time,$h_x(t)$is the hazard at time t, β is the parameter to be estimated, $h_0(t)$ is the baseline hazard (hazard when all the covariates are equal to zero) and x is the covariate (also called explanatory variables). And because of these two term β and $h_0(t)$, the Cox model is called a semi parametric model as $h_0(t)$ is non parametric and β is parametric part.
Parameter β can be interpreted as the hazard function is multiplied by e^{β}everytime when covaritae x increases one unit. For example to represent two groups say A and B if covariate x takes the value 0 and 1, we say group B has a risk of e^{β}times than group A.

$$e^{\beta x} = \frac{h_1(t)}{h_0(t)}, \quad \forall t \geq 0 \tag{2.10}$$

The equation(2.10) is called hazard ratio or hazard rate (the risk of failure).
The general proportional hazard model for set of p covariates $\boldsymbol{x}_i = (x_{1i}, x_{2i}, ..., x_{pi})^T$, take the following form:

$$h_i(t, \boldsymbol{x}_i) = h_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{x}_i} \; t > 0 \tag{2.11}$$

where regression coefficient $\boldsymbol{\beta} = (\beta_{1i}, \beta_{2i}, ..., \beta_{pi})^T, \; i = 1, 2, ...d$, baseline hazard $h_0(t)$ is the hazard with all the covariates equal to zero $(x_{1i}, x_{2i}, ..., x_{pi} = 0)$. If we have two patients with the same score on all covariates except covariate m then:

$$e^{\beta_m} = \frac{h_1(t)}{h_0(t)}, \quad \forall t \geq 0$$

The effect of covariate x_m could be read as if x_m increases 1 unit, the hazard is multiplied by e^{β_m}.

2.7.3 Estimation

In equation (2.11) two components need to be estimated. First and most important, the regression coefficient $\boldsymbol{\beta}$ and the baseline hazard $h_0(t)$.

In Cox's proportional hazard model, the unknown parameter $\boldsymbol{\beta}$ can be estimated by partial likelihood.[9]

Partial Likelihood The standard likelihood function cannot be used as we do not have any knowledge about baseline hazard $h_0(t)$, it does not have any specific form(unspecified), also we do not model the censoring distribution and is therefore removed out of the formula by Cox. That is why Cox model likelihood function is called "partial likelihood Function". It clearly studies probabilities of failed subjects. Rregression parameter $\boldsymbol{\beta}$ for Cox model is obtained by maximizing the partial likelihood and to do so first we find out the equation for partial likelihhod.
Assume that $t_{(i)} = t_{(1)}, t_{(2)}, ..., t_{(d)}$ be the true failure times with one failure at each time and $R(t_{(i)})$ is the risk set consisting of the subjects under observation i.e have not been censored or have not failed by time $t_i, \; i = 1, 2, ..., d$.
Then the full likelihood is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{k} P(No.i \; dies | One \; event \; occur, R_i)$$

$$L_i(\boldsymbol{\beta}) = \frac{h_i(t_i)}{\sum_{l \epsilon R_i} h_l(t_i)}$$

Using equation(2.9)

$$L_i(\boldsymbol{\beta}) = \frac{h_0(t_{(i)})e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_i}}{\sum_{l\epsilon R_i} h_0(t_{(i)})e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}$$

from the denominator and numerator, the baseline hazards cancle out, hence we get the final form of partial likelihood:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_i}}{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}} \tag{2.12}$$

Now, maximum partial likelihood estimate of $\boldsymbol{\beta}$ can be calculated as follows. The log partial likelihood is given by

$$l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta})) = log[\prod_{i=1}^{k} \frac{e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_i}}{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}]$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{k} [\boldsymbol{\beta}^{\mathbf{T}} x_i - log\{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}\}] \tag{2.13}$$

The *Scorefunction* $U(\boldsymbol{\beta})$ is defined as the first derivative of log likelihood function, given by

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}(l(\boldsymbol{\beta})) = \boldsymbol{x}_i - \frac{\sum_{l\epsilon R_i} \boldsymbol{x}_l . e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}\boldsymbol{x}_l}} \tag{2.14}$$

We get estimator $\hat{\boldsymbol{\beta}}$of parameter $\boldsymbol{\beta}$, by setting the score function (equation(2.14)) equal to zero.(why prof has removed subscript l from x on numerator?)

$$\hat{\boldsymbol{\beta}} = \boldsymbol{x}_i - \frac{\sum_{l\epsilon R_i} \boldsymbol{x}_i . e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}} = 0 \tag{2.15}$$

By taking the negative of the derivative of score function (or second derivative of log likelihood) we can find the partial likelihood observed *Informationmatrix* $I(\beta)$.

$$I(\boldsymbol{\beta}) = -[\frac{\partial}{\partial \boldsymbol{\beta}}(U(\boldsymbol{\beta}))] = -\frac{\partial}{\partial \boldsymbol{\beta}}[\boldsymbol{x}_i - \frac{\sum_{l\epsilon R_i} \boldsymbol{x}_i . e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}]$$

$$I(\boldsymbol{\beta}) = -[\frac{\sum_{l\epsilon R_i} \boldsymbol{x}_i \boldsymbol{x}_i^{'} . e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}}{\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}} - \frac{[\sum_{l\epsilon R_i} \boldsymbol{x}_i . e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}][\sum_{l\epsilon R_i} \boldsymbol{x}_i^{'} . e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l}]}{(\sum_{l\epsilon R_i} e^{\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_l})^2}] \tag{2.16}$$

Equation(2.16) also known as minus the Hessian Matrix is used to produce the standard errors for the regression coeffficients.(from wikipedia)
After we obtain maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$. then asymptotically,

$$\hat{\boldsymbol{\beta}} \sim N(\beta_0, I^{-1}(\hat{\boldsymbol{\beta}}))$$

where$I^{-1}(\hat{\boldsymbol{\beta}})$ is the inverse of information matrix at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\beta_0$is a true value. This approximate distribution is used to construct confidence interval and test the hypothesis $H_0 : \beta = \beta_0$

For example $\hat{\boldsymbol{\beta}} \pm z_{\frac{\alpha}{2}}[J^{-1}(\hat{\boldsymbol{\beta}})]^{\frac{1}{2}}$ is a $(1-\alpha)$ CI (confidense interval) of $\boldsymbol{\beta}$. (reference : NC State university, Dr. Daowen Zhang's lecture notes, chapter 6) In my thesis I asume only one event occur at one event time. I am not working with tied events but if it occur there are special ways to deal with it.

Base-line Hazard The baseline hazard could be estimated by using cumulative hazard function.

Let d_j be the number of events and $R_j$is the risk set at t_j. The estimator is as follows:

$$\hat{H}_0(t) = \sum_{j:t_j \leq t} \frac{d_j}{\sum_{l \epsilon R_j} e^{\hat{\boldsymbol{\beta}} \boldsymbol{x}_l}} \tag{2.17}$$

and if $\hat{\boldsymbol{\beta}} = 0$,equation (2.17) is shortened to:

$$\hat{H}_0(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}$$

where $n_j$is the size at R_j[9].
In R, to perform a Cox regression, *coxph* function in the package "survival" is used and in the summary of this given function, $exp(coef)$ gives the proportionality constant $(\psi)$and also we can plot a diagram to show the proportionality constant.

2.7.4 Schoenfeld Residuals

For the proportional hazard regression model, Schoenfeld recommended a chi squared goodness of fit statistic which exploited the residuals of the shape "Expected- Observed". Schoenfeld residuals is define as:

$$r_k(\hat{\boldsymbol{\beta}}) = X_{(k)} - \bar{x}(\boldsymbol{\beta}, t_k)\, k = 1, ..., d$$

where d is the total number of events, X_k is the subject with event k at event time t_k. And $\bar{x}(\beta, t_k)$is the weighted average of X. I will not go in detail of how can Schoenfeld residuals is solved manually but we will check how R tests it. In R, function *cox.zph*() from *survival* package is used to test the proportionality assumption for each covariate based on set of scaled Schoenfeld residuals versus suitable transformation of time. If the result shows higher chi-square, means the assumption is violated. We can also plot the graph of Schoenfeld residuals returned by *cox.zph*() by simply using the *plot* function. *Cox.zph*() provide a smoothing spline showed by solid line (horizontal line) in a graph with covering ± 2 standard error around the fit. A systematic deviation from the horizontal line showes non proportionality assumption.

The corresponding cumulative hazard function $H_1(t)$ and $H_0(t)$ can also hold if equation (2.8) holds following: [16, 17].

$$H_1(t) = \psi H_0(t),\ \forall t \geq 0$$

Plots for the smooth Schoenfeld residuals for all the covariates discussed in section 3 are given in appendex B.

3 Introduction and First Analysis of Data

In this section I will introduce my data by describing the background of my data and how we get it.

3.1 Back ground:

The data I am working on consists of data on patients suffering from colorectal cancer and therefore it is important to get an idea about what colorectal cancer is.

3.1.1 Colorectal Cancer (CRC):

This is a type of cancer that develops in the colon or rectum (parts of large intestine) and is therefore called colorectal cancer. This cancer may spread to the other parts of the body like lungs, liver etc which is called metastasic(stage IV) stage and is considered incurable. This cancer has four stages of disease where the first three (I-III) stages are curable by surgical resection of the tumor, sometimes combined with chemotherapy and/or radiation. For stage IV some of the treatment options are: removal of primary tumor, oncological treatment, multimodel treatment or no tumor related treatment.

3.1.2 Data:

The data I am working on are derived from a research project on patients who received non-curative treatment for CRC due to incurable disease or other reasons preventing curative surgery. The project has been approved by the regional Ethics committee (REK Sør-Øst 2016/409), and parts of the data have been made available for statistical evaluation within the current master project. Data are obtained from the following two Norwegian registries from the year 2008 to 2014:

1. Cancer Registry of Norway (Kreftregisteret)/Norwegian Colorectal Cancer Registry (Norsk kvalitetsregister for kreft i tykk- og endetarm).

2. NPR: Norwegian Patient Registry

The complete data set consists of N=30404 observations with all patients diagnosed with CRC during the study period. Patients with non-metastatic CRC

are considered curable (stage I to III), III), and those with metastatic disease as incurable (stage IV). Some patients are either unfit for surgery, or do not wish surgery, and will not receive curative treatment but need palliative care as those with stage IV disease. My focus in the thesis is the survival of those patients with stage four (IV) and those patients who were unable to receive curative treatment. The number of observations treated non curatively are N=10663; 35.1% of the entire study population. The time scale used in data is "days since diagnosis to death".
The data consists of 55 variables from which the following variables, considered to possibly be important for the survival, are studied:

1. Treatment category

- 1:no resection
- 2:no treatment
- 3:primary resection of tumor
- 4:oncological treatment only
- 5:curative attempt– resection of metastases without resection of primary tumor
- 6:resection of primary tumor and metastases
- 7:primary resection + oncology

2. Age category

- category1: age <66
- category2: age between 66-79
- category3: 80+

3. Stage category

- 1. Stage 4: Incurable
- 2. Stage 5: Unknown

4. Gender

5. Tumor location(Colon and rectum)

6. Metastasis status

- M0: No metastasis
- M+: Metastasis

7. Site of metastasis

- 0: No metastasis
- 1: Liver
- 2: Lung
- 3: Liver+Lung
- 4: Multiple sites
- 5: unknown

8. Resection of metastasis

- .00: No resection of metastasis
- 1.00: Resection of metastasis

9. Tumor location category

- 1: Right Colon
- 2: Left Colon nonsigmoid
- 3: Sigmoid
- 4: Rectosigmoid
- 5: Rectum
- 6: Unspecified

10. Chemotherapy

- 0: No chemo
- 1: Chemo

11. Radiation

- 0: No radiation
- 1: Radiation

12. ASA category

- 1: Category 1-2
- 2: Category 3-4
- 3: Unknown

13. Charleson Comorbidity Index (CCI)

- Group0: 0-1 comorb
- Group1: 2+ comorb

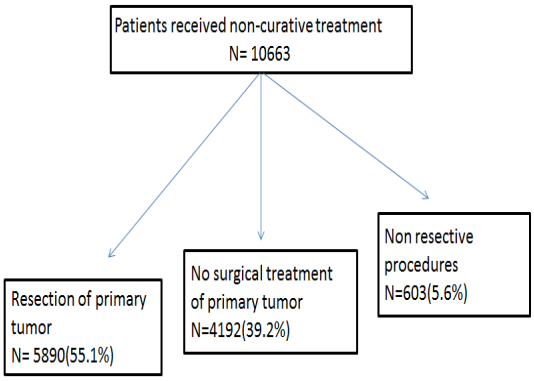


Figure 3.1: Treatment options for Non curative CRC with number of patients receieving the treatment

3.1.3 Treatment Options for Given Data:

For my data (N=10663) the incurable treatment options are:

- Non surgical treatment
- Non-resective surgical treatment
- Resection of primary tumor

The number of patients treated non curatively using the three options above are shown by the Figure 3.1. These treatment options are then further divided into 7 subgroups. The treatment category basically contain 2 subgroups "M0: no metastases" and "M+: metastases" and then M+ is further divided into six subgroups (categorical variables). The treatment category is as follows:

1: M0, no resection

2: M+, no treatment

3: M+, primary resection

4: M+, oncology only

5: M+, curative attempt

6: M+, primary and metastases resection

7: M+, primary resection + oncology

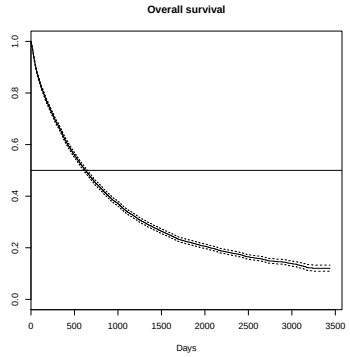


Figure 3.2: Overall survival for the population in non curable CRC

3.2 Statistical Analysis:

Kaplan-Meier and Cox regression model(explained in section 1) are used to fit the model, to get overall survival, plotting the survival curve and how covariates affect the survival. The significant level is taken as a p value <0.05. I will describe the results by p value, hazard ratio and confidence interval from the summary of Cox model. The criterian for hazard ratio is if HR >1 means high hazard of death and if HR$<$ 1 means hazard is low and survival is better.

3.2.1 Overall Survival:

After diagnosis, the probability of survival of patients after certain time point t is called overall survival. Overall survival is associated with the overall hazard rate λ_O such that:

$$S_O(t) = exp\left[-\int_0^t \lambda_O(u)du\right] \tag{3.1}$$

We can see in the Figure 3.2, the survival curve is gradually decreasing with the passage of time. At the end of the study almost 86% of patients had died that means overall survival is only 14%. If we look at the median survival time, we got to know that 50% of the patients would die until the 625th day. This shows that the overall survival is not good which is not surprising since this includes incurable patients.

3.2.2 Univariate Cox Analysis

- Treatment category

From the Figure 3.3 we can see that category 6 (where the patients receive "primary and metastas resection" treatment) shows good prognosis and category 2 (where patients do not receive any treatment) has worst prognosis. After analysing the data, results show that every category is significant and different

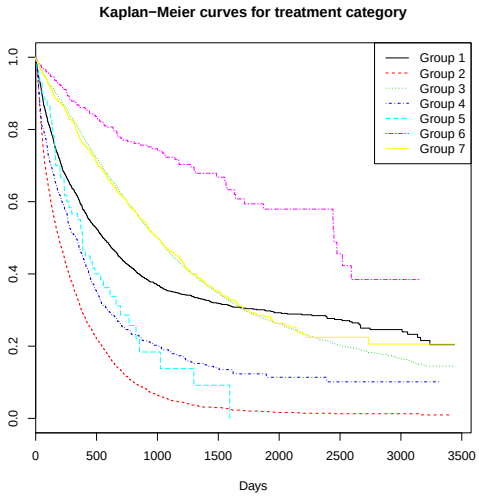


Figure 3.3: Rplot for 7 subgroups of treatment category

from reference category as the p value as shown in the Table 2 is less than the specified p value(0.05). Also the hazard ratio(HR) which tells the effect of each covariate on survival is shown in the Table 2 The HR for category 3,6 and 7 reduces the the risk of death by a factor of 0.78, 0.34 and 0.75 respectively compared to the reference category whereas for category 2, 4, 5 and 7, the hazard is high so they don't have good prognosis.

- Age category

The categorical covariate "age" consists of 3 categories where patiens are divided by age such as

category1: age <66

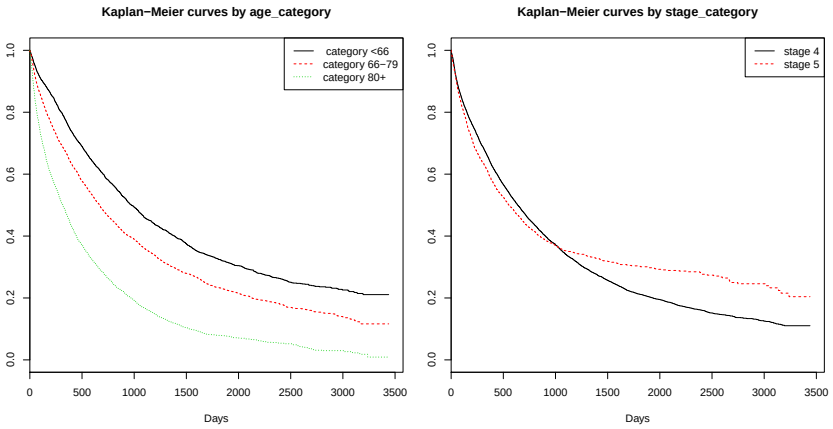
category2: age between 66-79

category3: 80+

The hazard ratio, p value and CI for the category 2 and 3 relative to 1 is given in the Table 2 and we can clearly see, both the age groups (category 2 and 3) as compare to group 1 has lower prognosis. The hazard is high by factors 1.35 and 2.43 (more than 1).This can also be seen clearly in Figure 3.4 (a).
Also the p value is far less then 0.05 which means the covariate "Age" is quite significant.

- Stages category

Stage category consists of 2 following stages:



(a) (b)

Figure 3.4: Kplan-Meier curves for (a): age, (b): stages of disease

1. Stage 4: Incurable

2. Stage 5: Unknown

The stage 5 is found to have slightly better survival than stage 4. The risk of death in stage 5 is 0.7, lower than stage 4, see Table 2. Figure 3.4 (b) shows the survival in the beginning was litte lower than 4 but after about 1000 days the curve went up and sustained above stage 4. Due to the crossing curves the p-value will be inaccurate. This covariate violates the proportionality assumption as you can see the Figure, the curves cross each other and thus are not constant over time.

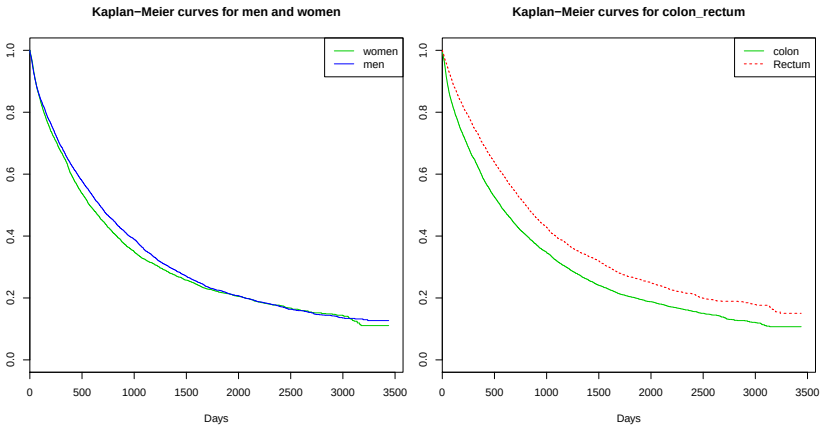
- Gender

Normaly in cancer studies it is seen that survival of female is better than male but for our data the survival for men is slightly better than the women. HR and p value is given in Table 2 and Figure 3.5 (c) shows the Kaplan curve.

- Tumor location(Colon_Rectum)

This category shows the presence of cancer in colon or rectum. More than 70 percent of patients have rectum cancer in our data. As it is clear from Kaplan-Meier Figure 3.5 (d) that patients with rectum cancer has little better survival than patients with colon cancer. Rectum cancer patients have (0.23) lower risk of death than colon cancer patients. P value is quite high that means this variable is very significant.

- Metastasis yes or no



(c) (d)

Figure 3.5: Kaplan-Meier curve for (c): Gender and (d): colon_Rectum(place of disease)

Patients with metastasis have higher hazard then the non metastasis patients. See the Table 2. In Figure 3.6 (e) before approx. 1000th day the patients with metastasis has good prognosis which means people can survive more than those of without metastasis. This is one of an example of non proportional model.

- Site of metastasis

This variable has following six groups:

0: No metastasis

1: Liver

2: Lung

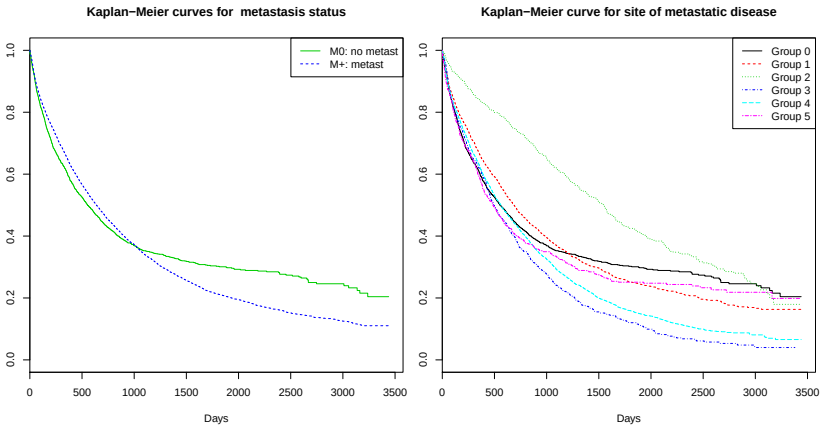
3: Liver+Lung

4: Others or multiple locations

5: unknown

According to the Cox summary, group 1 (liver) and group 5 (unknown) appears to be not significantly different relative to group 0 (reference group) as the p values are more than 0.05 and confidence intervals contain 1. Group 1 and 2 relative to group 0 (no metastasis) has hazard ratio of 0.96 and 0.57 respectively whereas rest of the groups have higher death risk. Group 2 with patients having metastasis in lungs has better survival than the rest of the groups as can be seen by Figure 3.6 (f) and group 3 has worst prognosis.

- Resection of metastasis (yes or no)



(e) (f)

Figure 3.6: Kaplan-Meier for (e): "status of metastasis"(either presense of metastatic disease or not) and (f): "site of metastatic disease"

This Variable has two groups of patients.

0: No resection of metastasis

1: Resection of metastasis

It is very clear from the plot 3.7 (g) that patients with resection of metastases have quite better survival than the other group. The resection of metastasis reduces the risk of death by factor 0.44. See Table 2

- Tumor location category

Following are the 6 locations of tumor:

1: Right Colon

2: Left Colon nonsigmoid

3: Sigmoid

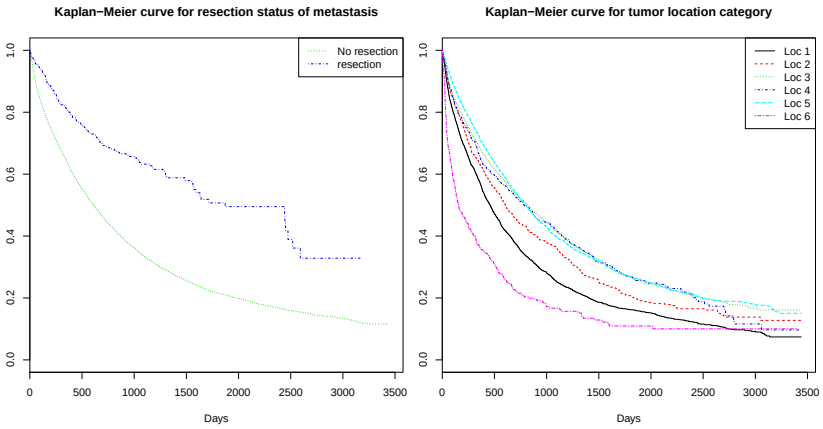
4: Rectosigmoid

5: Rectum

6: Unspecified

The survival of all locations is better relative to loction 1 except 6 which is unspecified and has worst prognosis. See Figure 3.7 (h)

- Chemotherapy



(g) (h)

Figure 3.7: Survival curves for covariates (a):"resection of metastasis" and (b): "Location of tumor"

This covariate consist of 2 groups of patients either receiving chemotherapy for cancer or not.

Group0: No chemotherapy

Group1: Chemotherapy

There is no huge difference in survival between two groups, see Figure 3.8 (i) but group 2 with patients who have gone through chemotherapy has better survival(HR= 0.92) than those without chemotherapy. See Table 2

- Radiation

Another therapy than chemo for cancer treatment is radiation. This covariate consists of following two groups

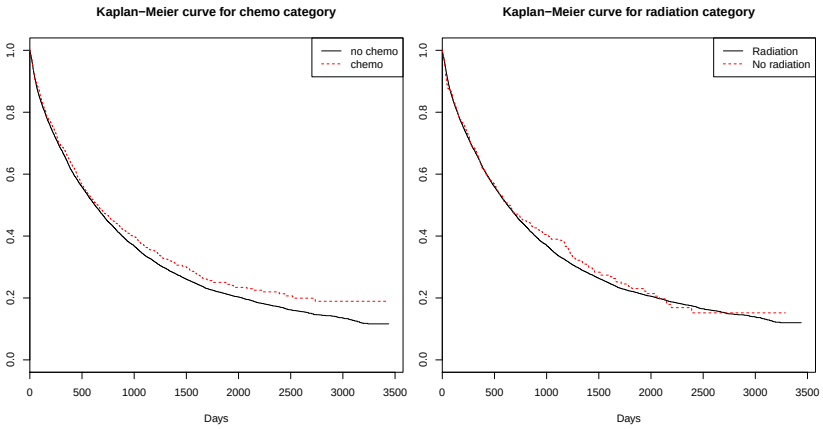
Group0: No radiation

Group1: Radiation

From the Cox summary given in Table 2, this covariate seems to be insignificant as p value (0.5) is more then the specified p value (0.05). Even though HR is less than 1, both groups are not significantly different from each other as 1 is included in the 95% confidence interval. See Figure 3.8 (j) .

- ASA

The American Society of Anesthesiologists (ASA) score also called as ASA-PS(physical status) score use to evaluate the physical status of all surgical patients. It has following five different classification:



(i) (j)

Figure 3.8: Kaplan-Meier curves for(i): "chemo" and (j): "radiation"

ASAI: A normal healthy patient.

ASA II: A patient with mild systemic disease.

ASAIII: A patient with a severe systemic disease that limits activity but is not incapacitating. .

ASAIV A patient with a severe systemic disease that is a persistent threat to life.

ASAV: A dying patient not expected to survive. [15]

In this covariate, following are the values given to the ASA categories

1: Category 1-2

2: Category 3-4

3: Unknown

Relative to group 1 which is "category (1-2)", both the other groups have a very high hazard rate which means "category 1-2" has quite better survival than the other groups as can be seen in the plot 3.9 (k).
Also group 2 and 3 have non proportional curves. See Table 2 for Cox summary.

- Charleson Comorbidity Index

Group0: 0-1 comorb

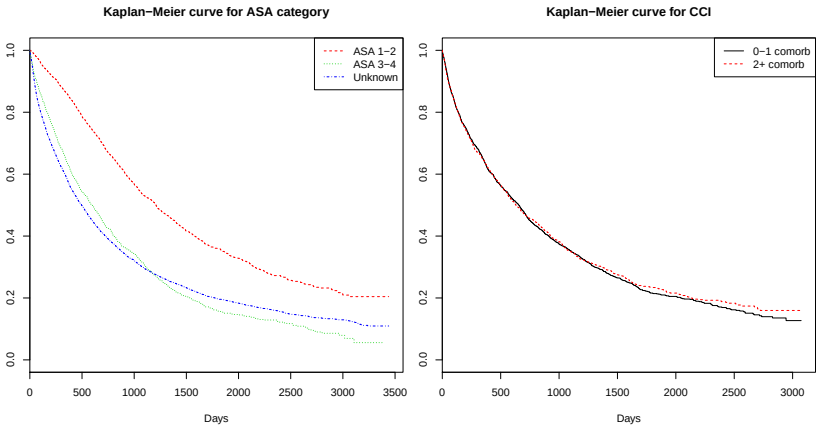
Group1: 2+ comorb

From the plot 3.9 (l), there seems to be no difference between the curves except after 1500 days where patients with "2+ comorbs" have little better survival than "0-1 comorb". HR is only 2% lower than the group1 (reference group) but p value is 0.6, higher than 0.05 so this covariate found to be not significant for the analysis.

Variables	HR	P	CI
Treatment category		<0.0001	
category 1	ref		
2	2.97	<0.0001	(2.77, 3.19)
3	0.78	<0.0001	(0.73, 0.83)
4	1.82	<0.0001	(1.64, 2.02)
5	1.78	<0.0001	(1.33, 2.39)
6	0.34	<0.0001	(0.27, 0.42)
7	0.75	<0.0001	(0.67, 0.84)
Age category		<0.0001	
<66	ref		
66-79	1.35	<0.0001	(1.28, 1.42)
80+	2.43	<0.0001	(2.30, 2.58)
Stages category			
stage 4	ref		
stage unknown	0.93	0.035	(0.88, 0.99)
Gender		0.01	
female	ref		
male	0.94	0.01	(0.90, 0.98)
Tumor location		<0.0001	
Colon	ref		
Rectum	0.77	<0.0001	(0.74, 0.81)
Metastasis status		0.03	
No metastasis	ref		
metastasis	1.06	0.035	(1.00, 1.03)
Site of metastasis		<0.0001	
group 0	ref		
1	0.96	0.39	(0.90, 1.04)
2	0.57	<0.0001	(0.51, 0.64)
3	1.38	<0.0001	(1.25, 1.52)
4	1.23	<0.0001	(1.15, 1.31)
5	1.09	0.08	(0.98, 1.21)

Variables	HR	P	CI
Resection of metastasis		<0.0001	
0.no resection	ref		
1.resection	0.44	<0.0001	(0.37, 0.52)
Tumor location category		<0.0001	
1	ref		
2	0.80	<0.0001	(0.73, 0.88)
3	0.67	<0.0001	(0.63, 0.72)
4	0.70	<0.0001	(0.63, 0.78)
5	0.66	<0.0001	(0.63, 0.70)
6	1.54	<0.0001	(1.35, 1.75)
Chemotherapy		0.03	
No chemotherapy	ref		
chemotherapy	0.92	0.02	(0.85, 0.99)
Radiation		0.5	
No radiation	ref		
Radiation	0.96	0.49	(0.85, 1.07)
ASA category		<0.0001	
1	ref		
2	1.81	<0.0001	(1.67, 1.96)
3	1.83	<0.0001	(1.72, 1.94)
CCI		0.6	
2+ comorb	ref		
0-1 comorb	0.98	0.62	(0.91, 1.05)

Table 2: characteristics of patients with cholorectal cancer from 2008 to 2014(Univariate Analysis)



(k) (l)

Figure 3.9: Kaplan-Meier plots for (k): ASA scores and (l): CCI

3.2.3 Multivariate Analysis:

Now I am going to analyse how all selected covariates together affect the survival. I will add all the covariates selected for univariate Cox analysis except radiation and CCI for not being significant (as $p>0.05$). Also I will not include ASA category as both ASA and CCI has 67% and 60% missing values respectively. After performing Multivariate Cox regression Analysis on 10 covariates in R, we get the results mentioned in Table(3)

Variables	HR	P value	CI
i.Treatmeant category			
1	ref		
2	2.65	$<$0.0001	(2.35, 3.00)
3	0.68	$<$0.0001	(0.60, 0.76)
4	2.16	$<$0.0001	(1.80, 2.59)
5	$1.31 \cdot 10^{5}$	0.95	$(4.27 \cdot 10^{-182}, 4.04 \cdot 10^{191})$
6	$2.42 \cdot 10^{4}$	0.96	$(7.88 \cdot 10^{-183}, 7.46 \cdot 10^{190})$
7	0.79	0.01	(0.65, 0.96)
ii.Age			
$<$66	ref		
66-79	1.38	$<$0.0001	(1.30, 1.45)
80+	2.47	$<$0.0001	(2.33, 2.62)
iii.Stages Category			
Stage 4	ref		
Unknown	NA	NA	NA
iv. Gender			
Female	ref		
Male	1.05	0.02	(1.00, 1.09)
v. Colon-Rectum			
Colon	ref		
Rectum	0.63	$<$0.0001	(0.60, 0.67)
vi. M-status			
No metastasis	ref		
metastasis	NA	NA	NA
vii. Site of metastasis			
0	ref		
1	1.25	$<$0.001	(1.12, 1.39)
2	0.80	0.001	(0.70, 0.92)
3	1.50	$<$0.0001	(1.33, 1.70)
4	1.36	$<$0.0001	(1.22, 1.51)
5	NA	NA	NA
viii. Resection of Met			
No resection	ref		
Resection	$1.32 \cdot 10^{-5}$	0.95	$(4.30 \cdot 10^{-192}, 4.07 \cdot 10^{181})$
ix. Location of tumor			
1	ref		
2	0.89	0.019	(0.81, 0.98)
3	0.67	$<$0.0001	(0.63, 0.71)
4	0.69	$<$0.0001	(0.62, 0.77)
5	NA	NA	NA
6	0.85	0.01	(0.75, 0.97)
x. Chemotherapy			
No chemotherapy	ref		
Chemotherapy	0.81	0.005	(0.69, 0.94)

Table 3: Results for Multivariate Cox regression Analysis of 11 covariates

As we can see through the Table there are some covariates and some subgroup of covariate have written "NA" instead of some values which means missing values. So we have to find a way to remove this "NA" values. And to do so if we really go through into the information given for covariates, we will see that some covariates have overlapping information. For example covariate "stage category" contains the same information about metastasis as covariate "Metastasis status" i.e in "stage category" category 4 represents the presence of metastasis and in "Metastasis status" group 1 shows the same. Similary covariate "Metastasis loation", "Resection of metastasis" and "Tumor location category" also carry the same information about metastasis. These 5 are the covariates causing the overlap information and the prevention is only removing unnecessary covariates. Removing one by one the covariates, I get my best option "Tumor location category" as my final covariate. This one is selected as it does not only contain required information about metastasis but some other information which cannot be deleted. So basically all the information in rest deleted covariates are covered by mentioned selected covariate.
After removing the not needed covarites and performing Multivariate Cox Analysis We get the result given in Table(4)

3.2.4 Final Results

Likelihood-ratio-test = 3631 on 15 df, p=<2e-16

The p value of the final model is quite low(<0.0001) which shows the model is quite significant. All the covariates in Table(4) are significant.

- Category 7 in treatment category is not significant as p value is greater than 0.05 but altogether the covariate is significant with category 6 "primary and metastaes resection" having the good prognosis with better survival of 0.62. And category 2,4 and 5 found to have a worst prognosis.
- The covariate age is related to the poor prognosis as increased age, increased the risk of death. The 95 % confidence interval includes 1 means both the age groups (66-79) and (80+) are not significantly different from reference group(<66).
- Being male or female are not significantly different from each other. Sex with HR=1.05 indicates increased risk of death.
- The overall tumor location category seems quite significant for the analysis. Every category has a good relationship with decreased risk of death. Location 5 which is "Rectum" has comparatively good survival than the other locations. And location 2 that is "Left Colon nonsigmoid" has poor survival than other locations.
- Chemo category with HR= 0.77 is associated with good prognosis. Chemotherapy reduces the risk of death by facor 0.77

Variables	HR	P value	CI
Treatment category			
1	ref		
2	3.38	<0.0001	(3.14, 3.64)
3	0.83	<0.0001	(0.77, 0.89)
4	2.52	<0.0001	(2.12, 3.00)
5	2.23	<0.0001	(1.65, 3.02)
6	0.38	<0.0001	(0.30, 0.47)
7	0.99	0.90	(0.84, 1.16)
Age category			
<66	ref		
66-79	1.35	<0.0001	(1.28, 1.43)
80+	2.41	<0.0001	(2.27, 2.55)
Gender			
Female	ref		
Male	1.05	0.02	(1.00, 1.10)
Tumor location category			
1	ref		
2	0.89	0.01	(0.80, 0.97)
3	0.65	<0.0001	(0.61, 0.69)
4	0.67	<0.0001	(0.60, 0.74)
5	0.61	<0.0001	(0.57, 0.64)
6	0.81	0.001	(0.70, 0.92)
chemotherapy			
No chemotherapy	ref		
Chemotherapy	0.77	0.0009	(0.66, 0.89)

Table 4: Final Multivariate Cox Analysis results after removing overlapping information

4 Further Measures of Survival Analysis

4.1 Relative Survival

One is often interested in estimating survival/mortality, based on cause specific data (i.e cancer) but most of the registries do not provide information about the specific cause of death as the information about cause of death collected is either unreliable or unavailable because of misclassification error or intrinsic uncertainty. As a result, interpretation and comparisons between countries and time periods is not easy due to the unassurity that whether the change in survival among the group of patients is due to change in risk of death from cancer or other causes of death. In such situations, the relative survival analysis provide information about survival/mortality of disease without knowing the cause of death.

The idea of relative survival is constructed to provide the probability of survival with the disease of interest when the cause of death is not known or not required and relative survival is defined as ratio of overall survival of patients dying of all causes of death to expected survival of comparable group with same demographical structure i.e age, gender and birth distribution as the patient groups (study cohort). Expected survival is generally estimated from the life table and is the total survival in a normal population life Table. Relative survival captures mortality both directly and indirectly related to the disease (for example, death due to treatment complications, suicide triggered by disease, etc.).
Assuming that cause of disease of interest and all other causes are independent. Let T_d indicate the time to death related to the disease, T_p indicate the time to death assuming the risk of death from all other causes, $T_o = minimum(T_d, T_p)$ be the observed time to death, all calculated with an accepTable reference point such as date of diagnosis. The overall survival probability to time t would then be:

$$P(T_o \geq t) = P(T_d \geq t)P(T_p \geq t) \tag{4.1}$$

$$P(T_d \geq t) = \frac{P(T_o \geq t)}{P(T_p \geq t)}$$

where $P(T_d \geq t)$ is the probability of survival to time t with disease of interest wihout the effect of other causes of death and termed as relative survival oftenly denoted as $S_R(t)$,$S_O(t) = P(T_o \geq t)$ is the observed probability of survival and can be estimated from the data and $P(T_p \geq t)$denoted as $S_P(t)$ is the expected or population survival that have all other causes of death except the disease of

interest. It can be estimated from population mortality Tables. For patients of size N, expected survival equals $S_P(t) = \frac{1}{N}\sum_i S_{p_i}(t)$. [18] hence,

$$S_R(t) = \frac{S_O(t)}{S_P(t)} \tag{4.2}$$

The ratio describes in equation (4.2) how our patients' survival compares to that of the general population. A measure of excess mortality is provided by relative survival, adepted by the cohort without the information about cause of death. The idea of relative survival inspired Researchers as it is related to the idea of "cure". In estimating cancer prognsis the researchers were concerned to know if and when the overall survival for both cancer patients and general population could be on same level. This level is called the cure point so when the excess deaths related to cancer became zero no patients died due to cancer. In the next subsection (4.2) I will explain what excess mortality is?

4.2 Excess Hazard

One needs precise cause of death data to directly estimate the cause specific mortality, but when the data is given for all causes of deaths so one can instead estimate excess mortality(excess hazard) also known as disease specific mortality by:

$$Total\,Mortality = Population\,Mortality + Excess\,Mortality$$

where population mortality is considered for a normal population with same age, gender and birth profile as the patient group.

A hazard function at time t for every individual diagnosed with disease i.e cancer is modelled as the sum of population hazard and the excess hazard due to the disease, following:

$$h_o(t) = h_p(t) + h_e(t) \tag{4.3}$$

where

$h_0(t)$ is the observed hazard for every individual in given data

$h_p(t)$ is the hazard every patient takes because of his age, sex and cohort year

$h_e(t)$ is the excess hazard specific for the disease

Now, if we integrate equation (4.3) and apply exponential we get:

$$e^{-\int h_o(t)dt} = e^{-\int (h_p(t)+h_e(t))dt}$$

$$e^{-\int h_o(t)dt} = e^{-\int (h_p(t)dt} + e^{-\int (h_e(t)dt} \tag{4.4}$$

by using cumulative survival function equation $S(t) = e^{-\int h(t)dt}$in equation (4.4)we get the expression:

$$S_O(t) = S_P(t) \times S_R(t) \tag{4.5}$$

This equation (4.5) gives the same form as in eqaution (4.2).The equation (4.3) is known as additive model, especially subjugated in cancer research. Also known as relative survival model as it can also be written as equation (4.5).
We can make a regression model for the excess hazard of the form:

$$h_e(t) = h_{e_0}(t)e^{\boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}}$$

In relative survival, the additive model is the preferred regression model to estimate the excess hazard. It provides better fit to data than other models, in general such models are considered reasonable for population based epidemiological studies. Population hazard does not depend on clinical covariates for example tumor specific covariate in cancer such as stage or histology but presume to depend on subset of covariates such as (typically) age, sex and birth distribution.
There are several methods to fit the additive model from which 3 are: (i) Hakulinen–Tenkanen additive survival method, (ii) The glm model with the Poisson error structure, (iii) The Esteve additive survival model.
For my data I am going to use the default method in R which is "The Esteve additive survival model". [20]

4.3 Net Survival

Net survival is another measure of estimating cause specific mortality referred to as the probability that a patient is still alive where the feasible cause of death is only the disease of interest (i.e cancer). Net survival is evaluated on survival scale when solely the disease specific mortality for each individual $h_{e_i}(t)$ is considered and an assumption is made that when the other causes of death are removed, the excess hazard will remain unaltered. The reason to use this measure is to get the measure that does no depend on the probability of dying due to other causes therefore is used when interest is in comparing populations with different mortality.
An alternative interpretation of net survival is known as the marginal relative survival ratio where a new unbaised estimator is described by Perme et al in which net survival is estimated as weighted average. [21] Relative survival ratio is the ratio of averages, likewise net survival can be written as average of ratios. Net survival for a cohort, is estimated as the weighted average of individual-specific net survival. These weights actually are the inverse of individual-specific expected survival probabilities of each individual. As a consequence of weights the number of people and number of deaths observed are increased in order to account for number of people and deaths not observed by reason of mortality caused by competing risks. To this end, we define the individual relative survival ratio through time and over individual as

$$S_{N_i}(t) = exp\{-\int_0^t h_{e_i}(u)du\} = \frac{exp\{-\int_0^t h_{o_i}(u)du\}}{exp\{-\int_0^t h_{p_i}(u)du\}}$$

$$S_{N_i} = \frac{S_{Oi}(t)}{S_{Pi}(t)}$$

So the net survival can be written as average of ratios of cohort of size N:

$$S_N(t) = \frac{1}{N}\sum_{i=1}^{N} S_{N_i} = \frac{1}{N}\sum_{i=1}^{N} \frac{S_{Oi}(t)}{S_{Pi}(t)} \tag{4.6}$$

in contrast to the relative survival ratio which is the ratio of averages:

$$S_R(t) = \frac{\frac{1}{N}\sum_{i=1}^{N} S_{Oi}(t)}{\frac{1}{N}\sum_{i=1}^{N} S_{Pi}(t)}$$

Net survival, contrary to relative survival is more suitable when comparing two cohorts with different population survival as it is not affected by population mortality hazard.

4.4 The Relsurv Package in R

To estimate the net survival and relative survival, function "*rs.surv*"and to fit the additive model the function "*rsadd*" in package "*relsurv*"in R is used.

-rs.surve:

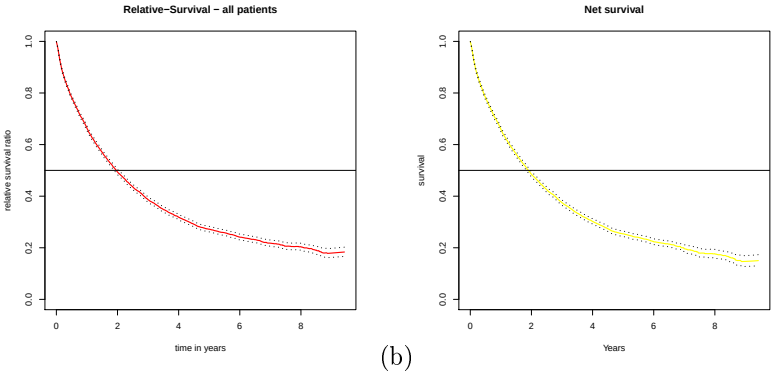
To calculate an estimate of relative survival or net survival, different methods are given as an argument and user can choose among them. The methods applied are the pohar-perma method, ederer 1, ederer 11 and hakulinen. Pohar perma is the default method and estimates the net seurvival with cumulative hazard. Other methods estimate the relative survival ratio.

-rsadd:

The function rsadd is used to fit the additive model to the data using the different method of estimation through *method*argument. The methods are (as descibed in section 4.2): Hakulinen–Tenkanen defined by "*glm.bin*", glm model with poisson error and Esteve method defined by "*glm.poi*"and "*max.lik*". The Esteve(or maximum likelihood method) is the default method.

-Usage The basic syntax for both the functions are same:

rs.surv(*formula*, *data*, *ratetable*)
rsadd(*formula*, *data*, *ratetable*)
where the syntax of the argument *formula*is same as we used for function *coxph* and *survfit* in survival package which is:
formula = Surv(time, status) ~ 1
formula = Surv(time, status) ~ x for rsadd where x is a covariate(or sum of covariates)
The object Surve holds the follow up time and status for censoring. The value 1 to the right of ~ sign represent the entire cohort but the curves for a subgroup



(a) (b)

Figure 5.1: (a)Relative Survival, (b) Net survival

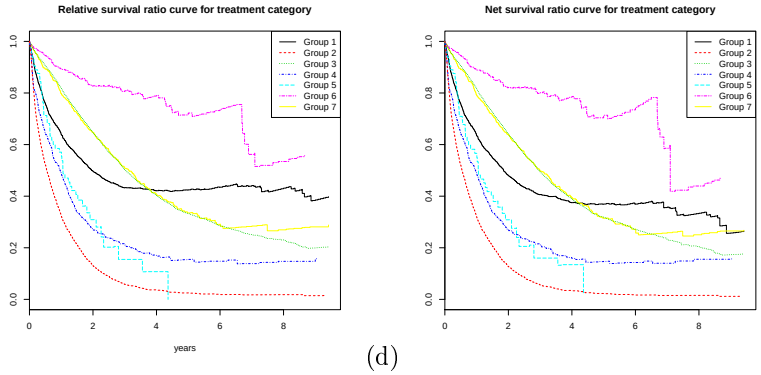
of certain covariates can also be estimated by writing that covariate or sum of covariates to the right of ~ sign (like in rsadd formula)

The argument *data*contains the data of observed cohort (patients) and the population data (mortality Table) should be stated in*ratetable*.The covariates in population data should be organised in the same manner as in the observed cohort. And the population data can be organised as a *ratetable* object. The factor 365.24 is used whenever the transformation between day and year is required. for example age is always expressed in days so if the age is in years it must be multiplied by factor 365.24.
Section 4 is based on sites [18, 19, 20, 21, 22]

5 Data Analysis

5.1 Relative Survival and Net Survival- All Patients

For my data the disease of interest is the non curative CRC. We see the relative survival reaches 50% at almost 2 years(716 days). After 8 years, relative survival ratio seems 20% according to the Figure 5.1 (a) which implies that survival of our observed cohort is at 20% of the survival of their population counterparts. The relative survival ratio is higher than the 8 year observed survival which is 14 %. See observed survival Figure 3.2 but the difference is not large, and means mostly patients in observed cohort died because of the disease.
At approximately 2 years(695 days) net survival reaches 50 %, see Figure 5.1(b). After 8 years, net survival of our cohort is approximaley 17 % which means in these years the number of patients who had died of non curative colon and rectum cancer would be 83 % in a hypothetical world where the cancer was the only thing the patients could die from. The net survival, as well is not larger than the overall survival. Net survival and relative survival ratio estimates are very close to each other as you can see both have almost same 2 years of survival and the difference at the end of the study is also not large.



(c) (d)

Figure 5.2: (c)Relative survival, (d) net survival for treatment category

5.2 Relative and Net Survival - Treatment Category

Relative survival ratio of all the categorise is higher than the overall survival, for example 6 year the relative survival ratio for group 6 (primary and metastas resection) is roughly estimated at 0.73 which means after 6 years, survival of patients in group 6 is at 73% of the survival of their counterparts. In Cox model, overall survival for group 6 was roughly estimated at 60%, See Figure 5.2 (c). Net survival at year 6 is estimated almost 0.73 for group 6. See Figure 5.2 (d). That means 27% of patients would die due to cancer in the first 6 year. Net surviavl and relative survival ratio estimates are almost equal. Relative and net survival for group 2 (no treatment) at year 6 is estimated at 0.02 which means the prognosis for group 2 is worst.

5.3 Excess Hazard Regression Model

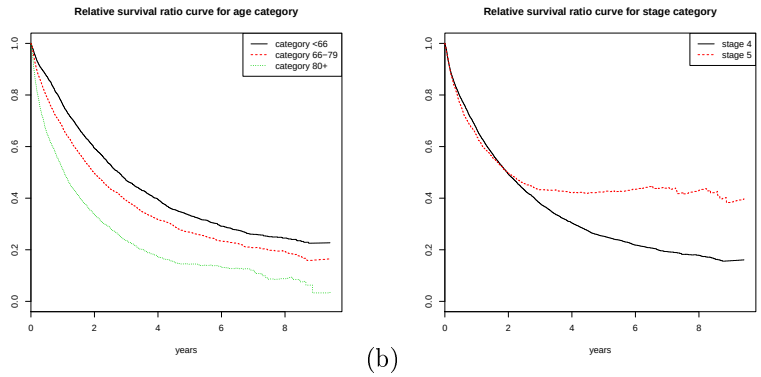
In this subsection I will study the impact of all the covariates described in section 3 on the relative survival and the excess hazard. I am not applying net survival separately since both net and relative survivals are almost same and due to the link between excess hazard and relative survival experienced in subsection 4.2.

5.3.1 Univariate Analysis

I am going to compare my results obtained from Cox regression model for observed survival with excess hazard regression for the cancer patients survival. The column HR(ex) contain excess hazard ratios (mortality due to disease only i.e colon-rectum cancer) (see subsection 4.2)

- Treatment category

Excess hazard ratio for category 6 which is “primary and metastas resection” treatment is 0.25 compared to hazard ratio of 0.34 in Cox model. The risk of



(a) (b)

Figure 5.3: Relative survival curve for (a): age catgory, (b): stage category

death for category 2 (i.e no treatment) has increased from factor 2.97 to 3.76. see Figure 5.2 and Table 5.

- Age category

Prognosis for both the age group ("66-79" and "80+") relative to reference group "<66" is still worse but the relative survival is higher than overall survival of Cox model as patients died from other factors are not included now. See Figure 5.3 (a).

- Stage category

The relative survival ratio of stage 5 is 0.19 compared to Cox's observd survival of 0.7. Also p value has changed and is more significant but considered inacurate because both the curves in Figure 5.3 (b) are crossing each other.

- Gender

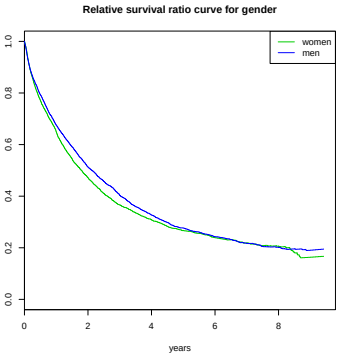
This covariate is now more significant and the excess hazard is 0.92 compared to Cox HR of 0. 94. Male patients have lower risk of death than female patients. see Table 5.

- Tumor location

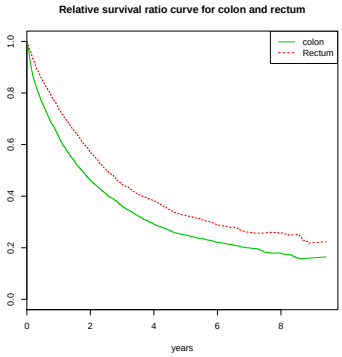
Relative to reference group (colon location) the rectum cancer patients has lower risk of death with excess hazard ratio of 0.74 which was 0.77 in Cox model. see Tables 2 and 5.

- Metastasis status

The excess hazard for patients with metastasis has increased by factor 1.22, compared to patients without metastasis. While in Cox modeling the hazard ratio was 1.06. So the patients with metastasis still has poor prognosis relative to patients with no metastasis. See Figure 5.5 (e).

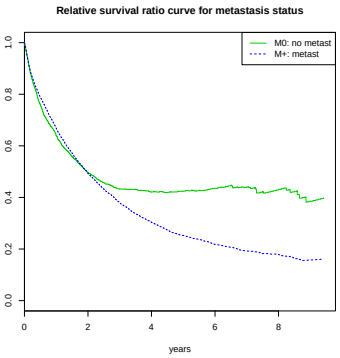


(c)

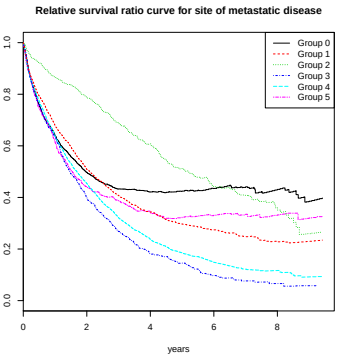


(d)

Figure 5.4: Relative survival curve for (c): gender, (d): colon-rectum



(e)



(f)

Figure 5.5: Relative survival curve for (e): Metastasis status, (f): Site of metastasis

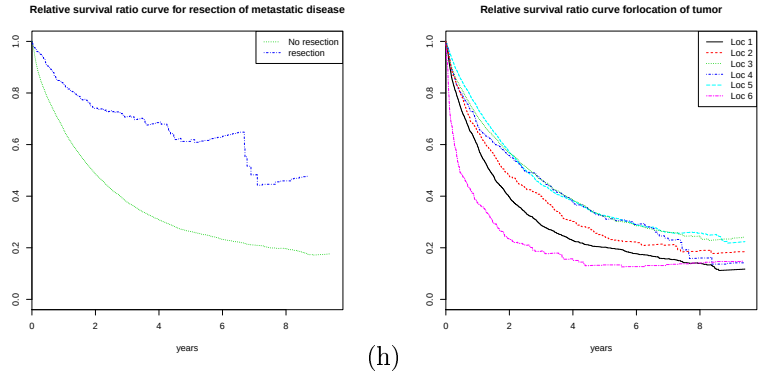
- Site of metastasis

Group 1 and 5 are now significantly different relevant to group 0 (reference group) compared to Cox model. All the groups have increased hazard except group 2 . Relative to reference group, presence of metastasis in lungs (group 2) has lower hazard ratio of 0.55 and in Cox HR was 0.53. Group 3 has worst prognosis, see Figure 5.5 (f) and Table 5.

- Resection of metastasis

The excess hazard ratio for resection of metastasis is 0.34 compared to HR of 0.44 in Cox model. The risk of death for the patients with resection of metastasis is lower relative to patients with no resection of metstasis as the difference between the curves can be seen from the Figure 5.6 (g).

- Location of tumor category



(g) (h)

Figure 5.6: Relative survival curve for (g): Resection of metastasis, (h): Location of tumor

Relative to reference location, all the locations except 6 has better survival. The excess hazard ratio of locations 3, 4 and 5 is lower while for location 6 is higher than hazard ratio of Cox model. Location 2 has same hazard ratio as in Cox model. Figure 5.6 (h).

- Chemotherapy

The excess hazard ratio is same as in Cox hazard ratio i.e chemotherapy has 8% better survival compared to no chemotherapy but the covariate is now significant (0.05). See Figure 5.7 (i), Table 5 and Table 2 for comparison.

- Radiation therapy

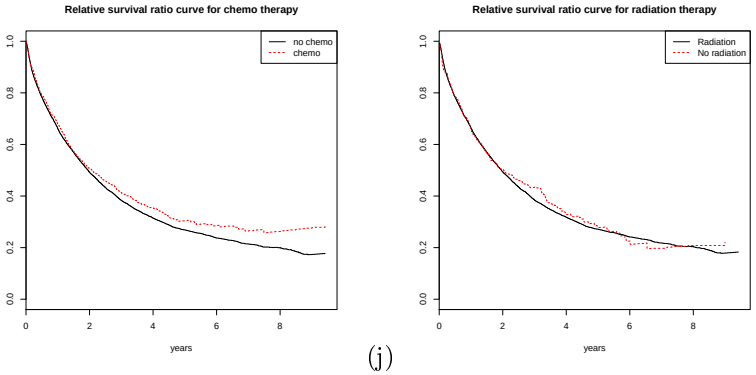
The covariate is still insignificant as p value is less than the specified p value(0.05). Both the groups are not significantly different from each other and can be clearly seen from the Figure 5.7(j).

- ASA category

Both the groups 2 and 3 still have poor prognosis relative to group 1. Group 2 "Catgory 2-3" and group 3 "category unknown" have non-proportional curves, see Figure 5.8 (k). Compared to "observed survival" in Cox regression, excess hazard ratio is a little lower in group 2 but higher in group 3. See Table 5

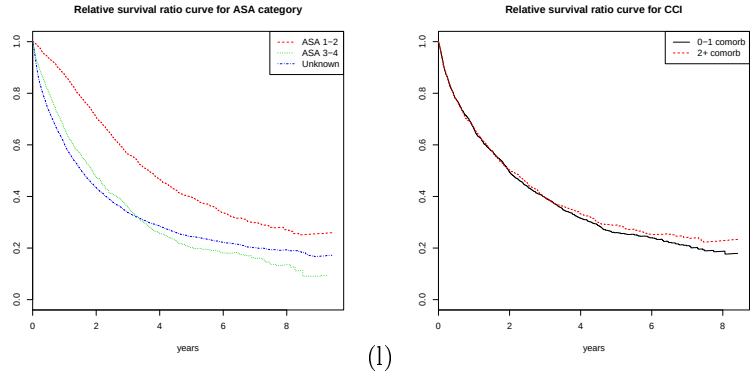
- Charleson Comorbidity Index

In Table 5, p value is not significant (0.38$>$ 0.05). And excess hazard for 2+ comorb is 0.96 compared to Cox's observed hazard of 0.98. See Figure 5.8 (l).



(i) (j)

Figure 5.7: Relative survival curve for (i): Chemo therapy, (j): Radiation therapy



(k) (l)

Figure 5.8: Relative survival curve for (k): ASA category, (l): CCI

Variables	HR(ex)	P	CI
Treatment category		<0.0001	
category 1	ref		
2	3.76	<0.0001	(3.45, 4.10)
3	0.86	0.0006	(0.80, 0.94)
4	2.20	<0.0001	(1.96, 2.48)
5	2.14	<0.0001	(1.56, 2.94)
6	0.25	<0.0001	(0.17, 0.35)
7	0.84	0.01	(0.73, 0.96)
Age category		<0.0001	
<66	ref		
66-79	1.27	<0.0001	(1.20, 1.34)
80+	2.07	<0.0001	(1.94, 2.21)
Stages category		<0.0001	
stage 4	ref		
stage unknown	0.81	<0.0001	(0.75, 0.87)
Gender		0.002	
female	ref		
male	0.92	0.002	(0.88, 0.97)
Tumor location		<0.0001	
Colon	ref		
Rectum	0.74	<0.0001	(0.70, 0.79)
Metas_status		<0.0001	
No Metastasis	ref		
metastasis	1.22	<0.0001	(1.13, 1.32)
Site of metastasis		<0.0001	
group 0	ref		
1	1.11	0.01	(1.01, 1.21)
2	0.55	<0.0001	(0.47, 0.63)
3	1.64	<0.0001	(1.46, 1.83)
4	1.44	<0.0001	(1.33, 1.57)
5	1.20	0.002	(1.06, 1.35)

Variables	HR(ex)	P	CI
Resection of Met		<0.0001	
0.no resection	ref		
1.resection	0.34	<0.0001	(0.27, 0.43)
Tumor location category		<0.0001	
1	ref		
2	0.80	<0.0001	(0.71, 0.89)
3	0.64	<0.0001	(0.60, 0.69)
4	0.67	<0.0001	(0.59, 0.76)
5	0.63	<0.0001	(0.59, 0.67)
6	1.66	<0.0001	(1.45, 1.90)
Chemotherapy		0.05	
No chemotherapy	ref		
chemotherapy	0.92	0.05	(0.85, 1.00)
Radiation		0.62	
No radiation	ref		
Radiation	0.96	0.62	(0.85, 1.10)
ASA category		<0.0001	
1	ref		
2	1.79	<0.0001	(1.63, 1.96)
3	1.87	<0.0001	(1.75, 2.00)
CCI		0.38	
2+ comorb	ref		
0-1 comorb	0.96	0.38	(0.89, 1.04)

Table 5: Univariate excess hazard analysis of patients with cholorectal cancer

Variables	HR(excess)	P value	CI
Treatment category			
1	ref		
2	3.94	<0.0001	(3.62, 4.28)
3	0.86	0.0005	(0.80, 0.94)
4	2.94	<0.0001	2.43, 3.56)
5	2.55	<0.0001	(1.84, 3.52)
6	0.28	<0.0001	(0.20, 0.40)
7	1.05	0.57	(0.88, 1.26)
Age category			
<66	ref		
66-79	1.27	<0.0001	(1.20, 1.35)
80+	2.04	<0.0001	(1.91, 2.17)
Gender			
Female	ref		
Male	1.02	0.30	(0.97, 1.07)
Tumor location category			
1	ref		
2	0.89	0.02	0.79, 0.98)
3	0.63	<0.0001	(0.58, 0.67)
4	0.65	<0.0001	(0.57, 0.72)
5	0.57	<0.0001	(0.53, 0.60)
6	0.80	0.001	(0.69, 0.91)
Chemotherapy			
No chemotherapy	ref		
Chemotherapy	0.75	0.001	(0.63, 0.89)

Table 6: Final multivariate excess analysis

5.4 Multivariate Analysis

5.4.1 Final Results

The relative survival ratio of combination of covariates given in Table 6 is very significant as the overall p value is <0.0001, which is quite less than 0.05. After comparing the final results with multivariates results of Cox model, we get:

- Treatment covariate as a whole is significant but catgory 7 is insignificant as p value is 0.57 which is greater than 0.05. Compared to category 1, the excess hazard ratio for categories 2, 3, 4, 5 and 7 are 3.94, 0.86, 2.94, 2.55 and 1.05 respectively. The same hazard ratios from the Cox model are lower, and are 3.38, 0.83, 2.52, 2.23 and 0.99. However, for category 6 the excess hazard ratio is lower (0.28) than the Cox hazard ratio (0.38). So with category 6 "primary and metastasis resection" treatment, patients survive longer, whereas category 2 "no treatment" has worst prognosis so it is better to be in the group that receives any treatment. See Tables 6 and 4.

- The excess hazard ratio for both the categories of age is higher than the reference category but is lower than hazard ratio for Cox model, that is because in the excess hazard model, other factors of death than disease are not considered. This shows the increasing age increases the risk of death due to cancer. See Table 6.

- In gender covariate, both the excess hazard and hazard ratio for Cox model are not very different (HRexcess=1.05 ,HRoverall=1.03) which shows being male is not significantly different to being female. See Tables 4 and 6 .

- The overall tumor location covariate is significant. The relative survival ratios for all the locations are little higher compared to overall survival from which location 5 "rectum" has better prognosis and location 2 "Left Colon nonsigmoid" has worse prognosis. See Table 6.

- The excess hazard ratio for chemotherapy is 0.75 compared to hazard ratio in Cox model of 0.73, see Tables 4 and 6 for comparison. So patients who recieved chemotherapy relative to those who do not, tend to live longer.

6 Summary

The aim of this thesis was to provide the introduction to various measures of survival (i.e the overall survival, net survival and relative survival) and regression models such as Cox model and excess hazard model. The thesis statistically evaluates a part of data taken from a research project on patients who received non-curative treatment for CRC due to incurable disease or other reasons preventing curative surgery. To analyse the survival of non curative patients, Kaplan-Meier and Cox regression model were used to fit the model, to get overall survival, plot the survival curve and study how covariates affect the survival. Partial likelihood estimation was discussed to estimate the regression coefficient $\boldsymbol{\beta}$. On the other hand, relative survival ratio and excess hazard regression model were used to estimate the cause specific mortality, to plot the relative survival curve and study how covariate affect the survival when other causes of deaths do not affect. In excess hazard, population mortality is considered for a normal population with same age, sex and birth profile as the patient group. In both regression models, all thirteen covariates were analysed through univariate analysis method to check each covariate's individual effect and multivariate analysis to check overall effect of all covariats. In multivariate analysis, all those insignifcant or overlapping covariates were thrown out of the group and as a result we got the following significant ones:

- Treatment category
- Age category
- Gender
- Tumor location category
- Chemotherapy

Specifically the focus was to compare the overall survival of cohort with the survival of their counterparts in general population to check either the patients died mostly from the disease or from other causes and what was the risk of death in each category of overall survival compared to general population. The results of Cox and excess hazard regression for all mentioned significant covarites were compared by comparing overall hazard ratio related to overall survival and excess hazard ratio related to relative survival. The results from the two methods are found to be not highly different from each other which indicates the patients with non curative disease are mostly died due to cancer.

References

[1] Kartsonaki, C. (2016). Survival analysis. Diagnostic Histopathology, 22(7), 263-270..

[2] Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. British journal of cancer, 89(2), 232.

[3] Lee, Elisa T., and John Wang. Statistical methods for survival data analysis. Vol. 476. John Wiley & Sons, 2003.

[4] Allison, Paul D. Survival analysis using SAS: a practical guide. Sas Institute, 2010.

[5] Cleves, M., Gould, W., Gould, W. W., Gutierrez, R., & Marchenko, Y. (2008). An introduction to survival analysis using Stata. Stata press

[6] Jenkins, S. P. (2005). Survival analysis. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK, 42, 54-56

[7] Aalen, O., Borgan, O., & Gjessing, H. (2008). Survival and event history analysis: a process point of view. Springer Science & Business Media.

[8] Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. Critical care, 8(5), 389.

[9] Broström, G. (2012). Event history analysis with R. CRC Press.

[10] Lee, E. T., & Go, O. T. (1997). Survival analysis in public health research. Annual review of public health, 18(1), 105-134.

[11] Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research, 1(4), 274.

[12] Hoskin, T. (2012). Parametric and nonparametric: Demystifying the terms. In Mayo Clinic (pp. 1-5).

[13] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology—Head and Neck Surgery, 143(3), 331-336.

[14] Powell, J. (2008). The New Palgrave Dictionary of Economics Online.

[15] Cullen, D. J., Apolone, G., Greenfield, S., Guadagnoli, E., & Cleary, P. (1994). ASA physical status and age predict morbidity after three surgical procedures. Annals of surgery, 220(1), 3.

[16] Fox, J. (2002). Cox proportional-hazards regression for survival data. An R and S-PLUS companion to applied regression, 2002.

[17] Xue, Y., & Schifano, E. D. (2017). Diagnostics for the Cox model. Communications for Statistical Applications and Methods, 24(6), 583-604.

[18] Reeves, G. K., Beral, V., Bull, D., & Quinn, M. (1999). Estimating relative survival among people registered with cancer in England and Wales. British journal of cancer, 79(1), 18.

[19] Mariotto, A. B., Noone, A. M., Howlader, N., Cho, H., Keel, G. E., Garshell, J., ... & Schwartz, L. M. (2014). Cancer survival: an overview of measures, uses, and interpretation. Journal of the National Cancer Institute Monographs, 2014(49), 145-186.

[20] Pohar, M., & Stare, J. (2006). Relative survival analysis in R. Computer methods and programs in biomedicine, 81(3), 272-278.

[21] Bajpai, Ram & Chaturvedi, Himanshu & Pandey, Arvind. (2014). Relative Survival: A Useful Tool in Population Based Health Studies. American Journal of Mathematics and Statistics 2014, 4(1): 38-45. 4. 38-45. 10.5923/j.ajms.20140401.06.

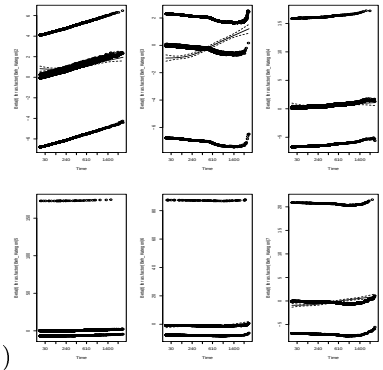
[22] Perme, M. P., & Pavlic, K. (2018). Nonparametric Relative Survival Analysis with the R Package relsurv. Journal of Statistical Software, 87(1), 1-27.

A. Norweigan Names for Covariates used in my data

1. Beh_Kategori (Treatment category)
2. Age_kat(Age catgory)
3. Stadium_kat(Stage category)
4. KJOENN(Gender)
5. Colon_Rectum(Tumor location)
6. M_status(Metastasis status)
7. Met_lok(Site of metastasis)
8. Met_kir_appr(Resection of metastasis)
9. Tu_Lok_Kat(Tumor location category)
10. Kjemo(Chemo therapy)
11. Radiatio(Radiation therapy)
12. ASA_kat(ASA category)
13. CCI_kat(Charleson comorbidity index)

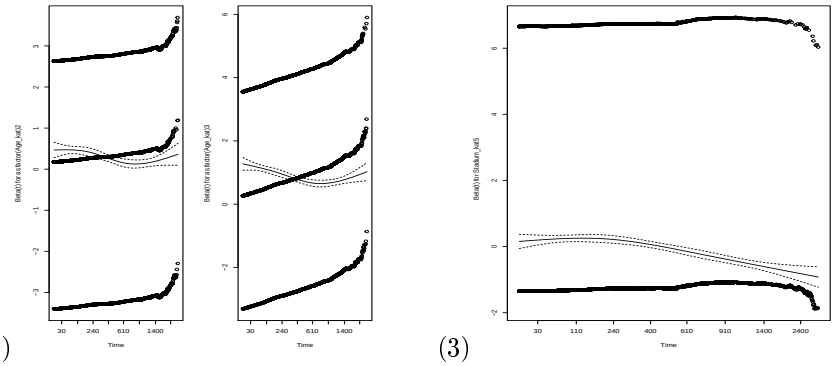
B. The graph of Schoenfeld residuals -All covariates

The scheoenfeld residuals are used to examined the proportionality assumption. In subection (2.7.4) it is mentioned that by using the simple *plot* function, the graph of Schoenfeld residuals returned by $cox.zph()$can be plotted. Following are the Schoenfeld graphs for each mentioned covariate.



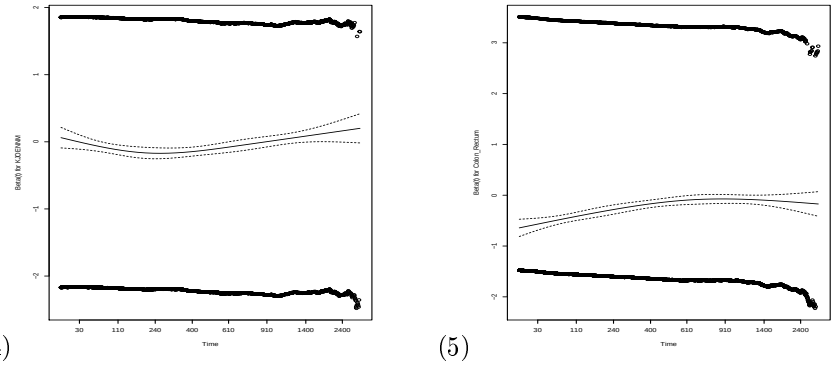
(1)

Figure .1: Smooth Schoenfeld residuals for (1) treatment category



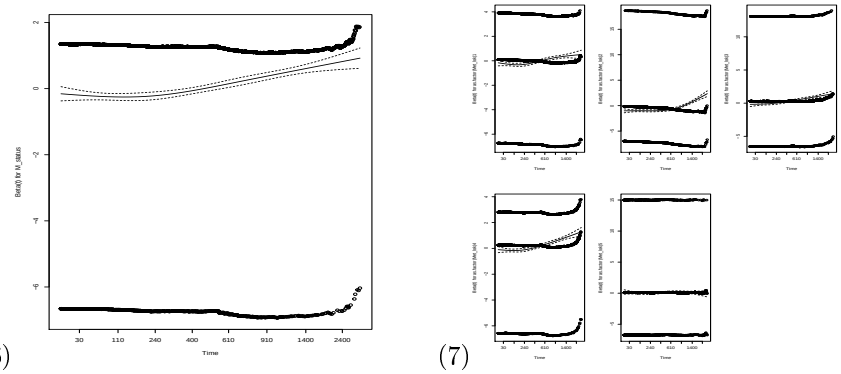
(2) (3)

Figure .2: Smooth Schoenfeld residuals for (2): age category and (3): stage category



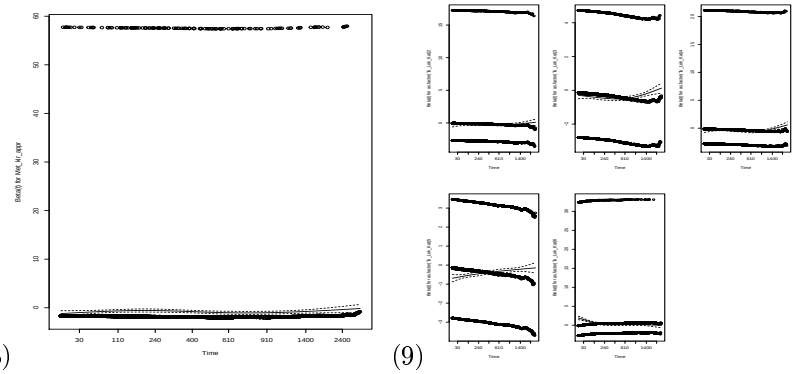
(4) (5)

Figure .3: Smooth Schoenfeld residuals for (4) gender and (5) tumor location



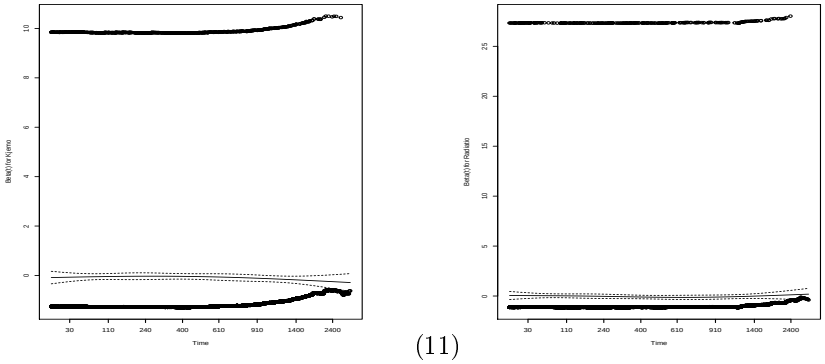
(6) (7)

Figure .4: Smooth Schoenfeld residuals for (6) metastasis status and (7) site of metastasis



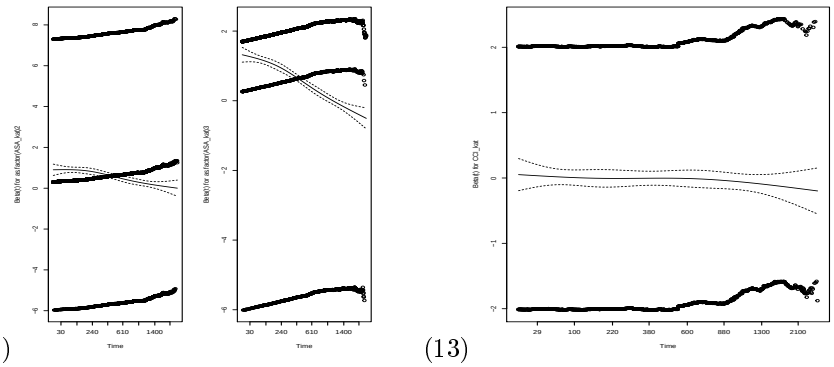
(8) (9)

Figure .5: Smooth Schoenfeld residuals for (8) resection of metastasis and (9) tumor location category



(10) (11)

Figure .6: Smooth Schoenfeld residuals for (10) chemotherapy and (11) radiation



(12) (13)

Figure .7: Smooth Schoenfeld residuals for (12) ASA category and (13) CCI